2006 European School of High-Energy Physics

Aronsborg, Sweden
18 June–1 July 2006

Proceedings
Editor: R. Fleischer
Abstract

The European School of High-Energy Physics is intended to give young experimental and phenomenological 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, Monte Carlo generators, relativistic heavy-ion physics, the flavour dynamics and CP violation in the Standard Model, cosmology, and high-energy neutrino astronomy with IceCube.
European School of High-Energy Physics

Aronsborg, Sweden
18 June - 1 July, 2006

Organizing Committee
L. Brand - Göttingen Univ.
P. Carlson - KTH, Stockholm
F. Erecinska - Lund Univ.
A. M. Ghezzi - CERN
I. Hoffmalm - Uppsala Univ.
R. Frischemakher - CERN
E. Lidin - CERN & Univ. Bergen
S. Ochsenkühle - JINR
J. Ramsman - Uppsala Univ.
T. Mikrut - CERN School Administrator
Y. Ternovskaya - JINR School Secretary

Discussion Leaders
J. Ramsman - Uppsala Univ.
L. Lindqvist - Lund Univ.
A. Providni - JINR Moscow
A. Arbuzov - JINR
S. Pacetti - CERN
G. Corcella - Rome Univ.

International Advisors
A. Sissakian - JINR
A. Skirinsky - BINP, Novosibirsk
N. E. Tsybin - IHEP, Prokine

Enquiries and Correspondence
Danielle Merrit
CERN Schools of Physics
CERN/DSTY
Ch-1211 Geneva 23, Switzerland
Tel: +41 22 767 9161
Fax: +41 22 767 5660
E-mail: Physics.School@cern.ch

Tatyana Donskova
International Department
Joint Institute for Nuclear Research
141980 Dubna, Moscow Region, Russia
Tel: +709521 43448
Fax: +709521 43599
E-mail: phs@jinr.ru

Applications are invited from young experimentalists and phenomenologists in
High-Energy Physics

Up-to-date information and application forms at
http://www.cern.ch/PhysicSchool/

Closing date for applications: 28 February, 2006
Preface

Ninety-eight students from twenty-five different countries attended the fourteenth in the new series of the European School of High-Energy Physics which took place in Aronsborg, Sweden, from 18 June to 1 July, 2006. The School was hosted in Aronsborg Konferenshotell, a nice four-star hotel located by Lake Mälaren, the largest lake in Sweden. According to the tradition of the School, the students shared twin rooms, mixing nationalities and in particular Eastern participants with Western ones.

Tord Ekelöf from the University of Uppsala was the local director of the School, and was assisted efficiently by Johan Rathsman, also from Uppsala University. Paola Errola, Lund University, helped in the initial phase of setting up the school. Six students from Uppsala, Henrik Pettersson, Oscar Stahl, Bjarte Mohn, Elias Coniavitis, Martin Flechl and Sophie Ohlsson, assisted with a multitude of tasks such as photocopying and transport of PCs and participants.

Our thanks go to the local team for their help and assistance. Their efforts contributed in major ways to the success of the School.

Our thanks are also 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 their personal contribution in answering questions and explaining points of theory.

We are very grateful to Danielle Métal and Tatyana Donskova for their untiring efforts in the lengthy preparations for and the day-to-day care of the School. Their efficient teamwork and continuous care of the students and their needs were highly appreciated.

Our special thanks also go to the hotel management and staff who were always ready to assist the School participants in a most friendly manner.

The participants spent their free time biking or walking in the forests bordering Lake Mälaren, swimming in the lake, or in sports activities using the multitude of facilities at the hotel. The hotel's own disco was a popular place late at night even though it was often competing in popularity with a wide-screen TV display set up to watch the World Cup soccer matches.

Excursions were organized to Uppsala and the Botanical Garden and to Stockholm, and on 23 June the School took part in the traditional Midsummer Night festivities, first by raising the 'May Pole', followed by folk dancing at lunch time, and in the evening by a special Midsummer Night 'Smörgåsbord' dinner accompanied by a Swedish schnaps tasting together with traditional Swedish songs led by the two excellent performers, Leiv Lönnblad and Tord Ekelöf. A local musician, Roger Andersson, had been invited to play folk music on a special instrument called a nyckelharpa.

Thanks go to the Swedish Research Council and the Swedish Academy of Sciences and to INTAS, who sponsored the School, and to the Natural Science Faculty of Uppsala University and the Municipality of Balsta, who offered the welcome drink.

However, the success of the 2006 School was to a large extent due to the students themselves. Their posters were of excellent quality both technically and in content, and throughout the School they participated actively during the lectures, in the discussion sessions, and with genuine interest in the different activities and excursions.

Egil Lillestøl
on behalf of the Organizing Committee
### People in the photograph

<table>
<thead>
<tr>
<th>1. Balev</th>
<th>Spasimír</th>
<th>59. Wilke</th>
<th>Lotte</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Conti</td>
<td>Glaucoma</td>
<td>60. Zwahlen</td>
<td>Nicolas</td>
</tr>
<tr>
<td>3. Spousta</td>
<td>Mattin</td>
<td>61. Uhlig</td>
<td>Selma</td>
</tr>
<tr>
<td>4. Ellis</td>
<td>Nick</td>
<td>62. Aharrouche</td>
<td>Mohamed</td>
</tr>
<tr>
<td>5. Sadykov</td>
<td>Renat</td>
<td>63. Diaz Valdès</td>
<td>Rafael</td>
</tr>
<tr>
<td>6. Mariáni</td>
<td>Camillo</td>
<td>64. Stahl</td>
<td>Thorsten</td>
</tr>
<tr>
<td>8. Kolesnikov</td>
<td>Vladimir</td>
<td>66. Hofmann</td>
<td>Franziska</td>
</tr>
<tr>
<td>9. Nowak</td>
<td>Krzysztof</td>
<td>67. Vala</td>
<td>Martin</td>
</tr>
<tr>
<td>10. Giacomini</td>
<td>Gabriele</td>
<td>68. Banie</td>
<td>Danilo</td>
</tr>
<tr>
<td>11. Lyubushkin</td>
<td>Vladimír</td>
<td>69. Lönndblad</td>
<td>Leif</td>
</tr>
<tr>
<td>12. Di Ruzza</td>
<td>Benedetto</td>
<td>70. Da Silva Costa</td>
<td>Carlos Filipe</td>
</tr>
<tr>
<td>13. Vendiotti</td>
<td>Stefano</td>
<td>71. Lavesson</td>
<td>Nils</td>
</tr>
<tr>
<td>14. Sozzi</td>
<td>Federica</td>
<td>72. Perfetto</td>
<td>Francesco</td>
</tr>
<tr>
<td>15. Czyzula</td>
<td>Zofia</td>
<td>73. Fiechl</td>
<td>Martin</td>
</tr>
<tr>
<td>16. Mastrandrea</td>
<td>Paolo</td>
<td>74. Cervi</td>
<td>Alan</td>
</tr>
<tr>
<td>17. Runyantsev</td>
<td>Leonid</td>
<td>75. Unidentified</td>
<td>Georgy</td>
</tr>
<tr>
<td>18. McLenray</td>
<td>Larry</td>
<td>76. Starov</td>
<td>Ioannis</td>
</tr>
<tr>
<td>19. Resende</td>
<td>Bernardo</td>
<td>77. Katsmanos</td>
<td>Paolo</td>
</tr>
<tr>
<td>20. Kolosou</td>
<td>Iro</td>
<td>78. Bellan</td>
<td>Henrik</td>
</tr>
<tr>
<td>23. Descamps</td>
<td>Julien</td>
<td>81. Dedes</td>
<td>Andreas Werner</td>
</tr>
<tr>
<td>24. Toms</td>
<td>Konstantin</td>
<td>82. Jung</td>
<td>Johan</td>
</tr>
<tr>
<td>25. Kondrateva</td>
<td>Natalia</td>
<td>83. Rathsman</td>
<td>Giovanni</td>
</tr>
<tr>
<td>26. Linn</td>
<td>Alexander</td>
<td>84. Stragusa</td>
<td>Anastasios</td>
</tr>
<tr>
<td>27. Donskova</td>
<td>Tatyana</td>
<td>85. Papageorgiou</td>
<td>Giuseppe</td>
</tr>
<tr>
<td>28. Bighi</td>
<td>Ikaros</td>
<td>86. Cerati</td>
<td>Matthias</td>
</tr>
<tr>
<td>29. Kocskar-Gorzo</td>
<td>Katherine</td>
<td>87. Kirsch</td>
<td>Stefan</td>
</tr>
<tr>
<td>30. Ashimova</td>
<td>Akmanl</td>
<td>88. Koenig</td>
<td>Egil</td>
</tr>
<tr>
<td>31. Wren</td>
<td>Adam</td>
<td>89. Lilleshul</td>
<td>Carlo</td>
</tr>
<tr>
<td>32. Borel</td>
<td>Jérome</td>
<td>90. Bombonati</td>
<td>Ingo</td>
</tr>
<tr>
<td>33. Kaussen</td>
<td>Gordon</td>
<td>91. Torchiani</td>
<td>Jan-Christopher</td>
</tr>
<tr>
<td>34. Strasser</td>
<td>Barbara</td>
<td>92. Winter</td>
<td>Nathalie</td>
</tr>
<tr>
<td>35. Bangert</td>
<td>Andrea</td>
<td>93. Chon-Sen</td>
<td>Umberto</td>
</tr>
<tr>
<td>36. Métrial</td>
<td>Danielle</td>
<td>94. De Sanctis</td>
<td>Robert</td>
</tr>
<tr>
<td>37. Timlin</td>
<td>Claire</td>
<td>95. Feischar</td>
<td>Elena</td>
</tr>
<tr>
<td>38. Mendes Sanaiva</td>
<td>Joao Gentil</td>
<td>96. Bruna</td>
<td>Andrey</td>
</tr>
<tr>
<td>39. Flossdorf</td>
<td>Alexander</td>
<td>97. Arbusov</td>
<td>Rupert</td>
</tr>
<tr>
<td>40. Ferreira da Gama Velho Arruda</td>
<td>Maria Luisa</td>
<td>98. Leutner</td>
<td>Oscar</td>
</tr>
<tr>
<td>41. Gadfort</td>
<td>Thomas</td>
<td>99. Stål</td>
<td>Christian</td>
</tr>
<tr>
<td>42. Poghosyan</td>
<td>Tavevik</td>
<td>100. Helebrant</td>
<td>Eric</td>
</tr>
<tr>
<td>43. Bruna</td>
<td>Grzegorz</td>
<td>101. Conte</td>
<td>Volker</td>
</tr>
<tr>
<td>44. Houbin</td>
<td>Pieter</td>
<td>102. Michels</td>
<td>Heléna</td>
</tr>
<tr>
<td>45. Asryan</td>
<td>Angel</td>
<td>103. Riceud</td>
<td>Alexei</td>
</tr>
<tr>
<td>46. Mills</td>
<td>Corinne</td>
<td>104. Pivovarov</td>
<td>Catalin</td>
</tr>
<tr>
<td>47. Jeffery</td>
<td>Ben</td>
<td>105. Ristea</td>
<td>Konstantinos</td>
</tr>
<tr>
<td>48. Parsamyan</td>
<td>Bakur</td>
<td>106. Bachas</td>
<td>Elias</td>
</tr>
<tr>
<td>49. Richter</td>
<td>Svenja</td>
<td>107. Conaviatis</td>
<td>Mark</td>
</tr>
<tr>
<td>50. Széke</td>
<td>Andrzej</td>
<td>108. Owen</td>
<td>Gennaro</td>
</tr>
<tr>
<td>51. Spreitzer</td>
<td>Teresa</td>
<td>109. Williams</td>
<td>Kai</td>
</tr>
<tr>
<td>52. Mack</td>
<td>Philipp</td>
<td>110. Corcella</td>
<td>Silvje</td>
</tr>
<tr>
<td>53. Bloch</td>
<td>Christoph</td>
<td>111. Rosphe</td>
<td>Henning</td>
</tr>
<tr>
<td>54. Ludwig</td>
<td>Inga</td>
<td>112. Puscoli</td>
<td>Daniel</td>
</tr>
<tr>
<td>55. Eliasson</td>
<td>Peder</td>
<td>113. Gaat</td>
<td>Philip</td>
</tr>
<tr>
<td>56. Gouleit</td>
<td>Marc</td>
<td>114. Walker</td>
<td>Tord</td>
</tr>
<tr>
<td>57. Fiore</td>
<td>Salvatore</td>
<td>115. von Doetinchem</td>
<td>Philip</td>
</tr>
</tbody>
</table>
Contents

Preface
   E. Lillestøl ........................................................................................................ v

Photograph of participants ............................................................................. vi

The Standard Model of electroweak interactions
   A. Pich ........................................................................................................... 1

Monte Carlo generators
   T. Sjöstrand .................................................................................................. 51

Relativistic heavy-ion physics: three lectures
   L. McLerran .................................................................................................. 75

Flavour dynamics and CP violation in the Standard Model: A crucial past and an essential future
   I. I. Bigi ......................................................................................................... 115

Cosmology
   V. A. Rubakov ............................................................................................. 197

High-energy neutrino astronomy with IceCube
   O. Bømer ....................................................................................................... 253

List of Organizing Committee Members ....................................................... 265
List of Lecturers ............................................................................................. 265
List of Discussion Leaders ............................................................................ 265
List of Students .............................................................................................. 266
List of Posters ................................................................................................. 267
The Standard Model of electroweak interactions

A. Pich
IFIC, University of València – CSIC,
València, Spain

Abstract
Gauge invariance is a powerful tool to determine the dynamics of the electroweak and strong forces. The particle content, structure and symmetries of the Standard Model Lagrangian are discussed. Special emphasis is given to the many phenomenological tests which have established this theoretical framework as the Standard Theory of electroweak interactions.

1 Introduction
The Standard Model (SM) is a gauge theory, based on the symmetry group $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$, which describes strong, weak, and electromagnetic interactions, via the exchange of the corresponding spin-1 gauge fields: eight massless gluons and one massless photon, respectively, for the strong and electromagnetic interactions, and three massive bosons, $W^\pm$ and $Z$, for the weak interaction. The fermionic matter content is given by the known leptons and quarks, which are organized in a three-fold family structure:

\[
\begin{bmatrix}
\nu_e & u \\
\nu_e & d
\end{bmatrix}, \quad
\begin{bmatrix}
\nu_\mu & c \\
\mu^- & s
\end{bmatrix}, \quad
\begin{bmatrix}
\nu_\tau & t \\
\tau^- & b
\end{bmatrix},
\]

where (each quark appears in three different colours)

\[
\begin{bmatrix}
\nu_l & q_u \\
l^- & q_d
\end{bmatrix} \equiv \left( \begin{bmatrix} \nu_l \\ l^- \end{bmatrix} \right)_L, \quad \left( \begin{bmatrix} q_u \\ q_d \end{bmatrix} \right)_L, \quad l_R, \quad q_{uR}, \quad q_{dR},
\]

plus the corresponding antiparticles. Thus, the left-handed fields are $SU(2)_L$ doublets, while their righthanded partners transform as $SU(2)_L$ singlets. The three fermionic families in Eq. (1) appear to have identical properties (gauge interactions); they differ only by their mass and their flavour quantum number.

The gauge symmetry is broken by the vacuum, which triggers the Spontaneous Symmetry Breaking (SSB) of the electroweak group to the electromagnetic subgroup:

\[
SU(3)_C \otimes SU(2)_L \otimes U(1)_Y \xrightarrow{SSB} SU(3)_C \otimes U(1)_{\text{QED}}.
\]

The SSB mechanism generates the masses of the weak gauge bosons, and gives rise to the appearance of a physical scalar particle in the model, the so-called Higgs. The fermion masses and mixings are also generated through the SSB.

The SM constitutes one of the most successful achievements in modern physics. It provides a very elegant theoretical framework, which is able to describe the known experimental facts in particle physics with high precision. These lectures [1] provide an introduction to the electroweak sector of the SM, i.e., the $SU(2)_L \otimes U(1)_Y$ part [2–5]. The strong $SU(3)_C$ piece is discussed in more detail in Ref. [6]. The power of the gauge principle is shown in Section 2, where the simpler Lagrangians of quantum electrodynamics and quantum chromodynamics are derived. The electroweak theoretical framework is presented in Sections 3 and 4, which discuss, respectively, the gauge structure and the SSB mechanism. Section 5 summarizes the present phenomenological status and shows the main precision tests performed at the $Z$ peak. The flavour structure is discussed in Section 6, where knowledge of the quark mixing angles is briefly reviewed and the importance of $CP$ violation tests is emphasized. Finally, a few comments on open questions, to be investigated at future facilities, are given in the summary.
Some useful but more technical information has been collected in several appendices: a minimal amount of quantum field theory concepts are given in Appendix A; Appendix B summarizes the most important algebraic properties of $SU(N)$ matrices; and a short discussion on gauge anomalies is presented in Appendix C.

2 Gauge invariance

2.1 Quantum electrodynamics

Let us consider the Lagrangian describing a free Dirac fermion:

$$\mathcal{L}_0 = i \bar{\psi}(x) \gamma^\mu \partial_\mu \psi(x) - m \bar{\psi}(x) \psi(x).$$

(4)

$\mathcal{L}_0$ is invariant under global $U(1)$ transformations

$$\psi(x) \xrightarrow{U(1)} \psi'(x) \equiv \exp \{ i Q \theta \} \psi(x),$$

(5)

where $Q\theta$ is an arbitrary real constant. The phase of $\psi(x)$ is then a pure convention-dependent quantity without physical meaning. However, the free Lagrangian is no longer invariant if one allows the phase transformation to depend on the space–time coordinate, i.e., under local phase redefinitions $\theta = \theta(x)$, because

$$\partial_\mu \psi(x) \xrightarrow{U(1)} \exp \{ i Q \theta \} \left( \partial_\mu + i Q \partial_\mu \theta \right) \psi(x).$$

(6)

Thus, once a given phase convention has been adopted at the reference point $x_0$, the same convention must be taken at all space–time points. This looks very unnatural.

The ‘gauge principle’ is the requirement that the $U(1)$ phase invariance should hold locally. This is only possible if one adds an extra piece to the Lagrangian, transforming in such a way as to cancel the $\partial_\mu \theta$ term in Eq. (6). The needed modification is completely fixed by the transformation (6): one introduces a new spin-1 (since $\partial_\mu \theta$ has a Lorentz index) field $A_\mu(x)$, transforming as

$$A_\mu(x) \xrightarrow{U(1)} A'_\mu(x) \equiv A_\mu(x) - \frac{1}{e} \partial_\mu \theta,$$

(7)

and defines the covariant derivative

$$D_\mu \psi(x) \equiv \left[ \partial_\mu + ieQA_\mu(x) \right] \psi(x),$$

(8)

which has the required property of transforming like the field itself:

$$D_\mu \psi(x) \xrightarrow{U(1)} (D_\mu \psi)'(x) \equiv \exp \{ i Q \theta \} D_\mu \psi(x).$$

(9)

The Lagrangian

$$\mathcal{L} \equiv i \bar{\psi}(x) \gamma^\mu D_\mu \psi(x) - m \bar{\psi}(x) \psi(x) = \mathcal{L}_0 - eQ A_\mu(x) \bar{\psi}(x) \gamma^\mu \psi(x)$$

(10)

is then invariant under local $U(1)$ transformations.

The gauge principle has generated an interaction between the Dirac spinor and the gauge field $A_\mu$, which is nothing else than the familiar vertex of Quantum Electrodynamics (QED). Note that the corresponding electromagnetic charge $Q$ is completely arbitrary. If one wants $A_\mu$ to be a true propagating field, one needs to add a gauge-invariant kinetic term

$$\mathcal{L}_{\text{Kin}} \equiv -\frac{1}{4} F_{\mu\nu} A_\mu A_\nu,$$

(11)

where $F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu$ is the usual electromagnetic field strength. A possible mass term for the gauge field, $\mathcal{L}_m = \frac{1}{2} m^2 A_\mu A_\mu$, is forbidden because it would violate gauge invariance; therefore, the photon field is predicted to be massless. Experimentally, we know that $m_\gamma < 6 \cdot 10^{-17}$ eV [7].
The total Lagrangian in Eqs. (10) and (11) gives rise to the well-known Maxwell equations:
\[ \partial_{\mu}F^{\mu\nu} = J^{\nu} = eQ\overline{\psi}\gamma^{\nu}\psi, \]  
where \( J^{\nu} \) is the fermion electromagnetic current. From a simple gauge-symmetry requirement, we have deduced the right QED Lagrangian, which leads to a very successful quantum field theory.

### 2.1.1 Lepton anomalous magnetic moments

![Feynman diagrams](image)

**Fig. 1:** Feynman diagrams contributing to the lepton anomalous magnetic moment

The most stringent QED test comes from the high-precision measurements of the \( e [8] \) and \( \mu [9] \) anomalous magnetic moments \( a_{\ell} = (g^{\ell}_0 - 2)/2 \), where \( g^{\ell}_0 = g^{\ell}_1 (e/2m_{\ell}) S^{\ell}_0 \):

\[ a_e = (159.652 \pm 0.76) \times 10^{-12}, \quad a_\mu = (11659.208 \pm 6.3) \times 10^{-10}. \]  

To a measurable level, \( a_e \) arises entirely from virtual electrons and photons; these contributions are fully known to \( O(\alpha^4) \) and some \( O(\alpha^5) \) corrections have been already computed [10–14]. The impressive agreement achieved between theory and experiment has promoted QED to the level of the best theory ever built to describe Nature. The theoretical error is dominated by the uncertainty in the input value of the QED coupling \( \alpha \equiv e^2/(4\pi) \). Turning things around, \( a_e \) provides the most accurate determination of the fine structure constant [15]:

\[ \alpha^{-1} = 137.035 \, 999 \, 710 \pm 0.000 \, 000 \, 096. \]  

The anomalous magnetic moment of the muon is sensitive to small corrections from virtual heavier states; compared to \( a_e \), they scale with the mass ratio \( m_\mu/m_e^2 \). Electroweak effects from virtual \( W^\pm \) and \( Z \) bosons amount to a contribution of \( (15.4 \pm 0.2) \times 10^{-10} \) [10, 11], which is larger than the present experimental precision. Thus \( a_\mu \) allows one to test the entire SM. The main theoretical uncertainty comes from strong interactions. Since quarks have electric charge, virtual quark–antiquark pairs induce *hadronic vacuum polarization* corrections to the photon propagator [Fig. 1(c)]. Owing to the non-perturbative character of the strong interaction at low energies, the light-quark contribution cannot be reliably calculated at present. This effect can be extracted from the measurement of the cross-section \( \sigma(e^+e^- \rightarrow \text{hadrons}) \) and from the invariant-mass distribution of the final hadrons in \( \tau \) decays, which unfortunately provide slightly different results [16–18]:

\[ a^{\text{th}}_\mu = \begin{cases} 
(11659 \, 180.2 \pm 5.6) \times 10^{-10} & (e^+e^- \text{ data}), \\
(11659 \, 199.7 \pm 6.3) \times 10^{-10} & (\tau \text{ data}). 
\end{cases} \]  

The quoted uncertainties include also the smaller *light-by-light scattering* contributions [Fig. 1(d)] [19]. The difference between the SM prediction and the experimental value (13) corresponds to \( 3.3 \sigma (e^+e^-) \) or \( 0.9 \sigma (\tau) \). New precise \( e^+e^- \) and \( \tau \) data sets are needed to settle the true value of \( a^{\text{th}}_\mu \).
2.2 Quantum chromodynamics

2.2.1 Quarks and colour

The large number of known mesonic and baryonic states clearly signals the existence of a deeper level of elementary constituents of matter: quarks. Assuming that mesons are \( M \equiv q\bar{q} \) states, while baryons \( B \equiv qqq \), one can nicely classify the entire hadronic spectrum. However, in order to satisfy the Fermi–Dirac statistics one needs to assume the existence of a new quantum number, colour, such that each species of quark may have \( N_C = 3 \) different colours: \( q^\alpha, \alpha = 1, 2, 3 \) (red, green, blue). Baryons and mesons are then described by the colour-singlet combinations

\[
B = \frac{1}{\sqrt{6}} \epsilon^{\alpha\beta\gamma} |q_\alpha q_\beta q_\gamma\rangle, \quad M = \frac{1}{\sqrt{3}} \delta^{\alpha\beta} |q_\alpha \bar{q}_\beta\rangle.
\]

In order to avoid the existence of non-observed extra states with non-zero colour, one needs to further postulate that all asymptotic states are colourless, i.e., singlets under rotations in colour space. This assumption is known as the confinement hypothesis, because it implies the non-observability of free quarks: since quarks carry colour they are confined within colour-singlet bound states.

A direct test of the colour quantum number can be obtained from the ratio

\[
R_{e^+e^-} \equiv \frac{\sigma(e^+e^- \rightarrow \text{hadrons})}{\sigma(e^+e^- \rightarrow \mu^+\mu^-)}.
\]

The hadronic production occurs through \( e^+e^- \rightarrow \gamma^*, Z^* \rightarrow q\bar{q} \rightarrow \text{hadrons} \) (Fig. 2). Since quarks are assumed to be confined, the probability to hadronize is just one; therefore, summing over all possible quarks in the final state, we can estimate the inclusive cross-section into hadrons. The electroweak production factors which are common with the \( e^+e^- \rightarrow \gamma^*, Z^* \rightarrow \mu^+\mu^- \) process cancel in the ratio (17). At energies well below the \( Z \) peak, the cross-section is dominated by the \( \gamma \)-exchange amplitude; the ratio \( R_{e^+e^-} \) is then given by the sum of the quark electric charges squared:

\[
R_{e^+e^-} \approx N_C \sum_{f=1}^{N_f} Q_f^2 = \begin{cases} 
\frac{2}{3} N_C = 2, & (N_f = 3 : u, d, s) \\
\frac{4}{3} N_C = 4, & (N_f = 4 : u, d, s, c) \\
\frac{5}{3} N_C = 5, & (N_f = 5 : u, d, s, c, b)
\end{cases}
\]

The measured ratio is shown in Fig. 3. Although the simple formula (18) cannot explain the complicated structure around the different quark thresholds, it gives the right average value of the cross-section (away from thresholds), provided that \( N_C \) is taken to be three. The agreement is better at larger energies. Notice that strong interactions have not been taken into account; only the confinement hypothesis has been used.

Electromagnetic interactions are associated with the fermion electric charges, while the quark flavours (up, down, strange, charm, bottom, top) are related to electroweak phenomena. The strong forces are flavour conserving and flavour independent. On the other side, the carriers of the electroweak interaction (\( \gamma, Z, W^\pm \)) do not couple to the quark colour. Thus it seems natural to take colour as the charge associated with the strong forces and try to build a quantum field theory based on it [20, 21].
2.2.2 Non-Abelian gauge symmetry

Let us denote $q_{f}^{\alpha}$ a quark field of colour $\alpha$ and flavour $f$. To simplify the equations, let us adopt a vector notation in colour space: $q_{f}^{T} \equiv (q_{f}^{1}, q_{f}^{2}, q_{f}^{3})$. The free Lagrangian

$$L_{0} = \sum_{f} \bar{q}_{f} (i\gamma^{\mu} \partial_{\mu} - m_{f}) q_{f}$$

is invariant under arbitrary global $SU(3)_{C}$ transformations in colour space,

$$q_{f}^{\alpha} \rightarrow (q_{f}^{\alpha})' = U_{\alpha \beta}^{\dagger} q_{f}^{\beta}, \quad U U^{\dagger} = U^{\dagger} U = 1, \quad \det U = 1.$$  \hspace{1cm} (20)

The $SU(3)_{C}$ matrices can be written in the form

$$U = \exp \left\{ i \frac{\lambda^{a}}{2} \theta_{a} \right\},$$ \hspace{1cm} (21)

where $\frac{1}{2} \lambda^{a}$ ($a = 1, 2, \ldots, 8$) denote the generators of the fundamental representation of the $SU(3)_{C}$ algebra, and $\theta_{a}$ are arbitrary parameters. The matrices $\lambda^{a}$ are traceless and satisfy the commutation relations

$$\left[ \frac{\lambda^{a}}{2}, \frac{\lambda^{b}}{2} \right] = i f^{abc} \frac{\lambda^{c}}{2},$$ \hspace{1cm} (22)

with $f^{abc}$ the $SU(3)_{C}$ structure constants, which are real and totally antisymmetric. Some useful properties of $SU(3)$ matrices are collected in Appendix B.

As in the QED case, we can now require the Lagrangian to be also invariant under local $SU(3)_{C}$ transformations, $\theta_{a} = \theta_{a}(x)$. To satisfy this requirement, we need to change the quark derivatives by covariant objects. Since we have now eight independent gauge parameters, eight different gauge bosons $G_{a}^{\mu}(x)$, the so-called gluons, are needed:

$$D^{\mu} q_{f} \equiv \left[ \partial^{\mu} + ig_{s} \frac{\lambda^{a}}{2} G_{a}^{\mu}(x) \right] q_{f} = [\partial^{\mu} + ig_{s} G^{\mu}(x)] q_{f}. $$ \hspace{1cm} (23)

Notice that we have introduced the compact matrix notation

$$[G^{\mu}(x)]_{\alpha \beta} \equiv \left( \frac{\lambda^{a}}{2} \right)_{\alpha \beta} G_{a}^{\mu}(x).$$ \hspace{1cm} (24)
Fig. 4: Interaction vertices of the QCD Lagrangian

We want \( D^\mu q_f \) to transform in exactly the same way as the colour-vector \( q_f \); this fixes the transformation properties of the gauge fields:

\[
D^\mu \longrightarrow (D^\mu)' = U D^\mu U^\dagger, \quad G^\mu \longrightarrow (G^\mu)' = U G^\mu U^\dagger + \frac{i}{g_s} (\partial^\mu U) U^\dagger.
\]  
(25)

Under an infinitesimal \( SU(3)_C \) transformation,

\[
q_f^a \longrightarrow (q_f^a)' = q_f^a + i \left( \frac{\lambda^a}{2} \right)_{\alpha\beta} \delta \theta_\alpha q_f^\beta,
\]

\[
G_a^\mu \longrightarrow (G_a^\mu)' = G_a^\mu - \frac{1}{g_s} \partial^\mu (\delta \theta_\alpha) - f^{\alpha\beta\gamma} \delta \theta_\beta G_\gamma^\mu.
\]
(26)

The gauge transformation of the gluon fields is more complicated than the one obtained in QED for the photon. The non-commutativity of the \( SU(3)_C \) matrices gives rise to an additional term involving the gluon fields themselves. For constant \( \delta \theta_\alpha \), the transformation rule for the gauge fields is expressed in terms of the structure constants \( f^{abc} \); thus, the gluon fields belong to the adjoint representation of the colour group (see Appendix B). Note also that there is a unique \( SU(3)_C \) coupling \( g_s \). In QED it was possible to assign arbitrary electromagnetic charges to the different fermions. Since the commutation relation (22) is non-linear, this freedom does not exist for \( SU(3)_C \).

To build a gauge-invariant kinetic term for the gluon fields, we introduce the corresponding field strengths:

\[
G^{\mu\nu}(x) = \frac{-i}{g_s} [D^\mu, D^\nu] = \partial^\mu G^\nu - \partial^\nu G^\mu + ig_s [G^\mu, G^\nu] \equiv \frac{\lambda^a}{2} G_a^{\mu\nu}(x),
\]

\[
G_a^{\mu\nu}(x) = \partial^\mu G_a^\nu - \partial^\nu G_a^\mu - g_s f^{abc} G_b^\mu G_c^\nu.
\]
(27)

Under a gauge transformation,

\[
G^{\mu\nu} \longrightarrow (G^{\mu\nu})' = U G^{\mu\nu} U^\dagger,
\]

(28)

and the colour trace \( \text{Tr}(G^{\mu\nu} G_{\mu\nu}) = \frac{1}{2} G_a^{\mu\nu} G_{a\mu\nu} \) remains invariant.

Taking the proper normalization for the gluon kinetic term, we finally have the \( SU(3)_C \) invariant Lagrangian of Quantum Chromodynamics (QCD):

\[
\mathcal{L}_{\text{QCD}} \equiv -\frac{1}{4} G_a^{\mu\nu} G_a^{\mu\nu} + \sum_f \bar{q}_f (i\gamma^\mu D_\mu - m_f) q_f.
\]
(29)

It is worth while to decompose the Lagrangian into its different pieces:

\[
\mathcal{L}_{\text{QCD}} = -\frac{1}{4} (\partial^\mu G_a^\nu - \partial^\nu G_a^\mu) (\partial_\mu G_a^\nu - \partial_\nu G_a^\mu) + \sum_f \bar{q}_f^a (i\gamma^\mu \partial_\mu - m_f) q_f^a
\]
The first line contains the correct kinetic terms for the different fields, which give rise to the corresponding propagators. The colour interaction between quarks and gluons is given by the second line; it involves the $SU(3)_C$ matrices $\lambda^a$. Finally, owing to the non-Abelian character of the colour group, the $G^\mu^\nu_a G_a^\mu^\nu$ term generates the cubic and quartic gluon self-interactions shown in the last line; the strength of these interactions (Fig. 4) is given by the same coupling $g_s$ which appears in the fermionic piece of the Lagrangian.

In spite of the rich physics contained in it, the Lagrangian (29) looks very simple because of its colour symmetry properties. All interactions are given in terms of a single universal coupling $g_s$, which is called the strong coupling constant. The existence of self-interactions among the gauge fields is a new feature that was not present in QED; it seems then reasonable to expect that these gauge self-interactions could explain properties like asymptotic freedom (strong interactions become weaker at short distances) and confinement (the strong forces increase at large distances), which do not appear in QED [6].

Without any detailed calculation, one can already extract qualitative physical consequences from $\mathcal{L}_{QCD}$. Quarks can emit gluons. At lowest order in $g_s$, the dominant process will be the emission of a single gauge boson; thus the hadronic decay of the $Z$ should result in some $Z \to q\bar{q}G$ events, in addition to the dominant $Z \to q\bar{q}$ decays. Figure 5 clearly shows that 3-jet events, with the required kinematics, indeed appear in the LEP data. Similar events show up in $e^+e^-$ annihilation into hadrons, away from the $Z$ peak. The ratio between 3-jet and 2-jet events provides a simple estimate of the strength of the strong interaction at LEP energies ($s = M_Z^2$): $\alpha_s \equiv g_s^2/(4\pi) \sim 0.12$.

3 Electroweak unification

3.1 Experimental facts

Low-energy experiments have provided a large amount of information about the dynamics underlying flavour-changing processes. The detailed analysis of the energy and angular distributions in $\beta$ decays, such as $\mu^- \to e^-\bar{\nu}_e\nu_\mu$ or $n \to p e^-\bar{\nu}_e$, made clear that only the left-handed (right-handed) fermion (antifermion) chiralities participate in those weak transitions; moreover, the strength of the interaction appears to be universal. This is further corroborated through the study of other processes like $\pi^- \to e^-\bar{\nu}_e$ or $\pi^- \to \mu^-\bar{\nu}_\mu$, which show that neutrinos have left-handed chiralities while anti-neutrinos are right-handed.
From neutrino scattering data, we learned the existence of different neutrino types ($\nu_e \neq \nu_\mu$) and that there are separately conserved lepton quantum numbers which distinguish neutrinos from antineutrinos; thus we observe the transitions $\bar{\nu}_e p \to e^+ n$, $\nu_e n \to e^- p$, $\bar{\nu}_\mu p \to \mu^+ n$ or $\nu_\mu n \to \mu^- p$, but we do not see processes like $\nu_e p \not\to e^+ n$, $\bar{\nu}_e n \not\to e^- p$, $\nu_\mu p \not\to \mu^+ n$ or $\bar{\nu}_\mu n \not\to \mu^- p$.

Together with theoretical considerations related to unitarity (a proper high-energy behaviour) and the absence of flavour-changing neutral-current transitions ($\mu^- \not\to e^- e^- e^+$), the low-energy information was good enough to determine the structure of the modern electroweak theory [23]. The intermediate vector bosons $W^\pm$ and $Z$ were theoretically introduced and their masses correctly estimated, before their experimental discovery. Nowadays, we have accumulated huge numbers of $W^\pm$ and $Z$ decay events, which bring much direct experimental evidence of their dynamical properties.

3.1.1 Charged currents

The interaction of quarks and leptons with the $W^\pm$ bosons (Fig. 6) exhibits the following features:

- Only left-handed fermions and right-handed antifermions couple to the $W^\pm$. Therefore, there is a 100% breaking of parity $P$ (left $\leftrightarrow$ right) and charge conjugation $C$ (particle $\leftrightarrow$ antiparticle). However, the combined transformation $CP$ is still a good symmetry.

- The $W^\pm$ bosons couple to the fermionic doublets in Eq. (1), where the electric charges of the two fermion partners differ in one unit. The decay channels of the $W^-$ are then:

$$W^- \to e^- \bar{\nu}_e, \mu^- \bar{\nu}_\mu, \tau^- \bar{\nu}_\tau, d' \bar{u}', s' \bar{c}' . \quad (31)$$

Owing to the very high mass of the top quark [24], $m_t = 171$ GeV $> M_W = 80.4$ GeV, its on-shell production through $W^- \to b' \bar{t}'$ is kinematically forbidden.

- All fermion doublets couple to the $W^\pm$ bosons with the same universal strength.

- The doublet partners of the up, charm, and top quarks appear to be mixtures of the three quarks with charge $-\frac{1}{3}$:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V \begin{pmatrix} d \\ s \\ b \end{pmatrix} , \quad V V^\dagger = V^\dagger V = 1 . \quad (32)$$

Thus, the weak eigenstates $d', s', b'$ are different than the mass eigenstates $d, s, b$. They are related through the $3 \times 3$ unitary matrix $V$, which characterizes flavour-mixing phenomena.

- The experimental evidence of neutrino oscillations shows that $\nu_e$, $\nu_\mu$ and $\nu_\tau$ are also mixtures of mass eigenstates. However, the neutrino masses are tiny: $|m_{\nu_3}^2 - m_{\nu_2}^2| \sim 2.5 \times 10^{-3}$ eV$^2$, $m_{\nu_2}^2 - m_{\nu_1}^2 \sim 8 \times 10^{-5}$ eV$^2$ [7].
3.1.2 Neutral currents

The neutral carriers of the electromagnetic and weak interactions have fermionic couplings (Fig. 7) with the following properties:

- All interacting vertices are flavour conserving. Both the $\gamma$ and the $Z$ couple to a fermion and its own antifermion, i.e., $\gamma f \bar{f}$ and $Z f \bar{f}$. Transitions of the type $\mu \not\rightarrow e \gamma$ or $Z \not\rightarrow e^\pm \mu^\mp$ have never been observed.
- The interactions depend on the fermion electric charge $Q_f$. Fermions with the same $Q_f$ have exactly the same universal couplings. Neutrinos do not have electromagnetic interactions ($Q_\nu = 0$), but they have a non-zero coupling to the $Z$ boson.
- Photons have the same interaction for both fermion chiralities, but the $Z$ couplings are different for left-handed and right-handed fermions. The neutrino coupling to the $Z$ involves only left-handed chiralities.
- There are three different light neutrino species.

3.2 The $SU(2)_L \otimes U(1)_Y$ theory

Using gauge invariance, we have been able to determine the right QED and QCD Lagrangians. To describe weak interactions, we need a more elaborated structure, with several fermionic flavours and different properties for left- and right-handed fields; moreover, the left-handed fermions should appear in doublets, and we would like to have massive gauge bosons $W^\pm$ and $Z$ in addition to the photon. The simplest group with doublet representations is $SU(2)$. We want to include also the electromagnetic interactions; thus we need an additional $U(1)$ group. The obvious symmetry group to consider is then

$$G \equiv SU(2)_L \otimes U(1)_Y,$$

where $L$ refers to left-handed fields. We do not specify, for the moment, the meaning of the subindex $Y$ since, as we shall see, the naive identification with electromagnetism does not work.

For simplicity, let us consider a single family of quarks, and introduce the notation

$$\psi_1(x) = \begin{pmatrix} u \\ d \end{pmatrix}_L, \quad \psi_2(x) = u_R, \quad \psi_3(x) = d_R. \quad (34)$$

Our discussion will also be valid for the lepton sector, with the identification

$$\psi_1(x) = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \quad \psi_2(x) = \nu_{eR}, \quad \psi_3(x) = e^- R. \quad (35)$$

As in the QED and QCD cases, let us consider the free Lagrangian

$$\mathcal{L}_0 = i \bar{u}(x) \gamma^\mu \partial_\mu u(x) + i \bar{d}(x) \gamma^\mu \partial_\mu d(x) = \sum_{j=1}^3 i \bar{\psi}_j(x) \gamma^\mu \partial_\mu \psi_j(x). \quad (36)$$
$\mathcal{L}_0$ is invariant under global $G$ transformations in flavour space:

$$
\begin{align*}
\psi_1(x) & \xrightarrow{G} \psi'_1(x) = \exp\{iy_1\beta\} U_L \psi_1(x), \\
\psi_2(x) & \xrightarrow{G} \psi'_2(x) = \exp\{iy_2\beta\} \psi_2(x), \\
\psi_3(x) & \xrightarrow{G} \psi'_3(x) = \exp\{iy_3\beta\} \psi_3(x),
\end{align*}
$$

(37)

where the $SU(2)_L$ transformation

$$
U_L \equiv \exp\left\{i \frac{\sigma^i}{2} \alpha^i \right\} \quad (i = 1, 2, 3)
$$

(38)

acts only on the doublet field $\psi_1$. The parameters $y_i$ are called hypercharges, since the $U(1)_Y$ phase transformation is analogous to the QED one. The matrix transformation $U_L$ is non-Abelian as in QCD. Notice that we have not included a mass term in Eq. (36) because it would mix the left- and right-handed fields [see Eq. (A.17)], therefore spoiling our symmetry considerations.

We can now require the Lagrangian to be also invariant under local $SU(2)_L \otimes U(1)_Y$ gauge transformations, i.e., with $\alpha^i = \alpha^i(x)$ and $\beta = \beta(x)$. In order to satisfy this symmetry requirement, we need to change the fermion derivatives by covariant objects. Since we have now four gauge parameters, $\alpha^i(x)$ and $\beta(x)$, four different gauge bosons are needed:

$$
\begin{align*}
D_\mu \psi_1(x) & \equiv \left[\partial_\mu + ig \tilde{W}_\mu(x) + ig' y_1 B_\mu(x)\right] \psi_1(x), \\
D_\mu \psi_2(x) & \equiv \left[\partial_\mu + ig' y_2 B_\mu(x)\right] \psi_2(x), \\
D_\mu \psi_3(x) & \equiv \left[\partial_\mu + ig' y_3 B_\mu(x)\right] \psi_3(x),
\end{align*}
$$

(39)

where

$$
\tilde{W}_\mu(x) \equiv \frac{\sigma^i}{2} W^{i\mu}(x)
$$

(40)

denotes a $SU(2)_L$ matrix field. Thus we have the correct number of gauge fields to describe the $W^\pm$, $Z$ and $\gamma$.

We want $D_\mu \psi_j(x)$ to transform in exactly the same way as the $\psi_j(x)$ fields; this fixes the transformation properties of the gauge fields:

$$
\begin{align*}
B_\mu(x) & \xrightarrow{G} B'_\mu(x) \equiv B_\mu(x) - \frac{1}{g'} \partial_\mu \beta(x), \\
\tilde{W}_\mu & \xrightarrow{G} \tilde{W}'_\mu \equiv U_L(x) \tilde{W}_\mu U_L^\dagger(x) + \frac{i}{g} \partial_\mu U_L(x) U_L^\dagger(x),
\end{align*}
$$

(41)

(42)

where $U_L(x) \equiv \exp\left\{i \frac{g'}{2} \alpha^i(x)\right\}$. The transformation of $B_\mu$ is identical to the one obtained in QED for the photon, while the $SU(2)_L$ $W^i_\mu$ fields transform in a way analogous to the gluon fields of QCD. Note that the $\psi_j$ couplings to $B_\mu$ are completely free as in QED, i.e., the hypercharges $y_j$ can be arbitrary parameters. Since the $SU(2)_L$ commutation relation is non-linear, this freedom does not exist for the $W^i_\mu$: there is only a unique $SU(2)_L$ coupling $g$.

The Lagrangian

$$
\mathcal{L} = \sum_{j=1}^{3} i \bar{\psi}_j(x) \gamma^\mu D_\mu \psi_j(x)
$$

(43)

is invariant under local $G$ transformations. In order to build the gauge-invariant kinetic term for the gauge fields, we introduce the corresponding field strengths:

$$
B_{\mu\nu} \equiv \partial_\mu B_\nu - \partial_\nu B_\mu,
$$

(44)

10
\[ \tilde{W}_{\mu\nu} \equiv -\frac{i}{g} \left[ \left( \partial_\mu + ig \tilde{W}_\mu \right), \left( \partial_\nu + ig \tilde{W}_\nu \right) \right] = \partial_\mu \tilde{W}_\nu - \partial_\nu \tilde{W}_\mu + ig \left[ \tilde{W}_\mu, \tilde{W}_\nu \right], \]

\[ \tilde{W}_{\mu\nu} \equiv \frac{\sigma^i}{2} \tilde{W}^i_{\mu\nu}, \quad \tilde{W}^i_{\mu\nu} = \partial_\mu \tilde{W}^i_\nu - \partial_\nu \tilde{W}^i_\mu - g \epsilon^{ijk} \tilde{W}^j_\mu \tilde{W}^k_\nu. \]

\( B_{\mu\nu} \) remains invariant under \( G \) transformations, while \( \tilde{W}_{\mu\nu} \) transforms covariantly:

\[ B_{\mu\nu} \xrightarrow{G} B_{\mu\nu}, \quad \tilde{W}_{\mu\nu} \xrightarrow{G} U_L \tilde{W}_{\mu\nu} U_L^\dagger. \]

Therefore, the properly normalized kinetic Lagrangian is given by

\[ \mathcal{L}_{\text{kin}} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{2} \text{Tr} \left[ \tilde{W}_{\mu\nu} \tilde{W}^{\mu\nu} \right] = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} \tilde{W}^i_{\mu\nu} \tilde{W}^{i\mu\nu}. \]

Since the field strengths \( \tilde{W}^i_{\mu\nu} \) contain a quadratic piece, the Lagrangian \( \mathcal{L}_{\text{kin}} \) gives rise to cubic and quartic self-interactions among the gauge fields. The strength of these interactions is given by the same \( SU(2)_L \) coupling \( g \) which appears in the fermionic piece of the Lagrangian.

The gauge symmetry forbids the writing of a mass term for the gauge bosons. Fermionic masses are also not possible, because they would communicate the left- and right-handed fields, which have different transformation properties, and therefore would produce an explicit breaking of the gauge symmetry. Thus, the \( SU(2)_L \otimes U(1)_Y \) Lagrangian in Eqs. (43) and (48) contains only massless fields.

### 3.3 Charged-current interaction

![Fig. 8: Charged-current interaction vertices](image)

The Lagrangian (43) contains interactions of the fermion fields with the gauge bosons,

\[ \mathcal{L} \rightarrow -g \bar{\psi}_1 \gamma^\mu \tilde{W}_\mu \psi_1 - g' B_\mu \sum_{j=1}^3 y_j \bar{\psi}_j \gamma^\mu \psi_j. \]

The term containing the \( SU(2)_L \) matrix

\[ \tilde{W}_\mu = \frac{\sigma^i}{2} \tilde{W}^i_\mu = \frac{1}{2} \begin{pmatrix} W^3_\mu & \sqrt{2} W^3_\mu \\ -\sqrt{2} W^3_\mu & -W^3_\mu \end{pmatrix} \]

gives rise to charged-current interactions with the boson field \( W^i_\mu \equiv (W^1_\mu + i W^2_\mu)/\sqrt{2} \) and its complex-conjugate \( W^i_\mu \equiv (W^1_\mu - i W^2_\mu)/\sqrt{2} \) (Fig. 8). For a single family of quarks and leptons,

\[ \mathcal{L}_{\text{CC}} = -\frac{g}{\sqrt{2}} \left\{ W^i_\mu \left[ \bar{u} \gamma^\mu (1 - \gamma_5) d + \bar{e} \gamma^\mu (1 - \gamma_5) e \right] + \text{h.c.} \right\}. \]

The universality of the quark and lepton interactions is now a direct consequence of the assumed gauge symmetry. Note, however, that Eq. (51) cannot describe the observed dynamics, because the gauge bosons are massless and, therefore, give rise to long-range forces.
3.4 Neutral-current interaction

Equation (49) contains also interactions with the neutral gauge fields $W^3_\mu$ and $B_\mu$. We would like to identify these bosons with the $Z$ and the $\gamma$. However, since the photon has the same interaction with both fermion chiralities, the singlet gauge boson $B_\mu$ cannot be equal to the electromagnetic field. That would require $y_1 = y_2 = y_3$ and $g'y_j = e Q_j$, which cannot be simultaneously true.

Since both fields are neutral, we can try with an arbitrary combination of them:

$$\begin{pmatrix} W^3_\mu \\ B_\mu \end{pmatrix} \equiv \begin{pmatrix} \cos \theta_W & \sin \theta_W \\ -\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix}.$$  \hspace{1cm} (52)

The physical $Z$ boson has a mass different from zero, which is forbidden by the local gauge symmetry. We will see in the next section how it is possible to generate non-zero boson masses, through the SSB mechanism. For the moment, we just assume that something breaks the symmetry, generating the $Z$ mass, and that the neutral mass eigenstates are a mixture of the triplet and singlet $SU(2)_L$ fields. In terms of the fields $Z$ and $\gamma$, the neutral-current Lagrangian is given by

$$\mathcal{L}_{NC} = -\sum_j \overline{\psi}_j \gamma^\mu \left\{ A_\mu \left[ g \frac{\sigma_3}{2} \sin \theta_W + g'y_j \cos \theta_W \right] + Z_\mu \left[ g \frac{\sigma_3}{2} \cos \theta_W - g'y_j \sin \theta_W \right] \right\} \psi_j.$$  \hspace{1cm} (53)

In order to get QED from the $A_\mu$ piece, one needs to impose the conditions:

$$g \sin \theta_W = g' \cos \theta_W = e, \hspace{1cm} Y = Q - T_3,$$  \hspace{1cm} (54)

where $T_3 \equiv \sigma_3/2$ and $Q$ denotes the electromagnetic charge operator

$$Q_1 \equiv \begin{pmatrix} Q_{u/\nu} & 0 \\ 0 & Q_{d/e} \end{pmatrix}, \hspace{1cm} Q_2 = Q_{u/\nu}, \hspace{1cm} Q_3 = Q_{d/e}.$$  \hspace{1cm} (55)

The first equality relates the $SU(2)_L$ and $U(1)_Y$ couplings to the electromagnetic coupling, providing the wanted unification of the electroweak interactions. The second identity fixes the fermion hypercharges in terms of their electric charge and weak isospin quantum numbers:

Quarks: $y_1 = Q_u - \frac{1}{2} = Q_d + \frac{1}{2} = \frac{1}{6}$, \hspace{0.5cm} $y_2 = Q_u = \frac{2}{3}$, \hspace{0.5cm} $y_3 = Q_d = -\frac{1}{3}$,

Leptons: $y_1 = Q_\nu - \frac{1}{2} = Q_e + \frac{1}{2} = -\frac{1}{2}$, \hspace{0.5cm} $y_2 = Q_\nu = 0$, \hspace{0.5cm} $y_3 = Q_e = -1$.

A hypothetical right-handed neutrino would have both electric charge and weak hypercharge equal to zero. Since it would not couple either to the $W^\pm$ bosons, such a particle would not have any kind of interaction (sterile neutrino). For aesthetic reasons, we shall not consider right-handed neutrinos any longer.

Using the relations (54), the neutral-current Lagrangian can be written as

$$\mathcal{L}_{NC} = \mathcal{L}_{QED} + \mathcal{L}_{NC}^Z,$$  \hspace{1cm} (56)
Table 1: Neutral-current couplings

<table>
<thead>
<tr>
<th></th>
<th>u</th>
<th>d</th>
<th>νe</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 v_f</td>
<td>1 - (\frac{8}{3}) sin^2 (\theta_W)</td>
<td>-1 + (\frac{4}{3}) sin^2 (\theta_W)</td>
<td>1</td>
<td>-1 + 4 sin^2 (\theta_W)</td>
</tr>
<tr>
<td>2 a_f</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

where

\[
\mathcal{L}_{\text{QED}} = -e A_\mu \sum_j \bar{\psi}_j \gamma^ \mu Q_j \psi_j \equiv -e A_\mu J^\mu_{\text{em}}
\]

is the usual QED Lagrangian and

\[
\mathcal{L}_{\text{NC}}^Z = -\frac{e}{2 \sin \theta_W \cos \theta_W} J^\mu_{\text{em}} Z_\mu
\]

contains the interaction of the \(Z\) boson with the neutral fermionic current

\[
J^\mu_Z \equiv \sum_j \bar{\psi}_j \gamma^ \mu (\sigma_3 - 2 \sin^2 \theta_W Q_j) \psi_j = J^\mu_3 - 2 \sin^2 \theta_W J^\mu_{\text{em}}.
\]

In terms of the more usual fermion fields, \(\mathcal{L}_{\text{NC}}^Z\) has the form (Fig. 9)

\[
\mathcal{L}_{\text{NC}}^Z = -\frac{e}{2 \sin \theta_W \cos \theta_W} Z_\mu \sum_f \bar{f} \gamma^ \mu (v_f - a_f \gamma_5) f,
\]

where \(a_f = T^f_3\) and \(v_f = T^f_3 \left(1 - 4 |Q_f| \sin^2 \theta_W\right)\). Table 1 shows the neutral-current couplings of the different fermions.

3.5 Gauge self-interactions

![Gauge boson self-interaction vertices](image)

Fig. 10: Gauge boson self-interaction vertices

In addition to the usual kinetic terms, the Lagrangian (48) generates cubic and quartic self-interactions among the gauge bosons (Fig. 10):

\[
\mathcal{L}_3 = i e \cot \theta_W \left\{ \left( \partial^\mu W^\nu - \partial^\nu W^\mu \right) W^\dagger_\mu Z_\nu - \left( \partial^\mu W^{\nu\dagger} - \partial^\nu W^{\mu\dagger} \right) W_\mu Z^\dagger_\nu + W_\mu W^\dagger_\nu \left( \partial^\mu Z^{\nu\dagger} - \partial^\nu Z^{\mu\dagger} \right) \right\}
\]

\[
+ i e \left\{ \left( \partial^\mu W^\nu - \partial^\nu W^\mu \right) W^\dagger_\mu A_\nu - \left( \partial^\mu W^{\nu\dagger} - \partial^\nu W^{\mu\dagger} \right) W_\mu A^\dagger_\nu + W_\mu W^\dagger_\nu \left( \partial^\mu A^{\nu\dagger} - \partial^\nu A^{\mu\dagger} \right) \right\};
\]

\[
\mathcal{L}_4 = -\frac{e^2}{2 \sin^2 \theta_W} \left\{ \left( W^\dagger_\mu W^\mu \right)^2 - W^\dagger_\mu W^{\mu\dagger} W_\nu W^\nu \right\} - e^2 \cot^2 \theta_W \left\{ W^\dagger_\mu W^\mu Z_\nu Z^\dagger_\nu - W^\dagger_\mu Z^{\mu\dagger} W_\nu Z^\nu \right\}
\]
\[-e^2 \cot \theta_W \left\{ 2W^\mu W^\nu Z^\mu A^\nu - W^\mu_\mu Z^\mu W^\nu A^\nu - W^\mu_\mu A^\mu W^\nu Z^\nu \right\} \]
\[-e^2 \left\{ W^\mu_\mu W^\nu A^\nu - W^\mu_\mu A^\mu W^\nu A^\nu \right\} .\]

Notice that at least a pair of charged \( W \) bosons are always present. The \( SU(2)_L \) algebra does not generate any neutral vertex with only photons and \( Z \) bosons.

### 4 Spontaneous symmetry breaking

So far, we have been able to derive charged- and neutral-current interactions of the type needed to describe weak decays; we have nicely incorporated QED into the same theoretical framework and, moreover, we have got additional self-interactions of the gauge bosons, which are generated by the non-Abelian structure of the \( SU(2)_L \) group. Gauge symmetry also guarantees that we have a well-defined renormalizable Lagrangian. However, this Lagrangian has very little to do with reality. Our gauge bosons are massless particles; while this is fine for the photon field, the physical \( W^\pm \) and \( Z \) bosons should be quite heavy objects.

In order to generate masses, we need to break the gauge symmetry in some way; however, we also need a fully symmetric Lagrangian to preserve renormalizability. This dilemma may be solved by the possibility of getting non-symmetric results from an invariant Lagrangian.

Let us consider a Lagrangian, which

1. is invariant under a group \( G \) of transformations;
2. has a degenerate set of states with minimal energy, which transform under \( G \) as the members of a given multiplet.

If one of those states is arbitrarily selected as the ground state of the system, the symmetry is said to be spontaneously broken.

A well-known physical example is provided by a ferromagnet: although the Hamiltonian is invariant under rotations, the ground state has the spins aligned into some arbitrary direction; moreover, any higher-energy state, built from the ground state by a finite number of excitations, would share this anisotropy. In a Quantum Field Theory, the ground state is the vacuum; thus the SSB mechanism will appear when there is a symmetric Lagrangian, but a non-symmetric vacuum.
The horse in Fig. 11 illustrates in a very simple way the phenomenon of SSB. Although the left and right carrots are identical, Nicolás must take a decision if he wants to get food. What is important is not whether he goes left or right, which are equivalent options, but that the symmetry gets broken. In two dimensions (discrete left–right symmetry), after eating the first carrot Nicolás would need to make an effort to climb the hill in order to reach the carrot on the other side; however, in three dimensions (continuous rotation symmetry) there is a marvelous flat circular valley along which Nicolás can move from one carrot to the next without any effort.

The existence of flat directions connecting the degenerate states of minimal energy is a general property of the SSB of continuous symmetries. In a Quantum Field Theory it implies the existence of massless degrees of freedom.

4.1 Goldstone theorem

Let us consider a complex scalar field \( \phi(x) \), with Lagrangian

\[
\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi), \quad V(\phi) = \mu^2 \phi^\dagger \phi + h \left( \phi^\dagger \phi \right)^2.
\]

(62)

\( \mathcal{L} \) is invariant under global phase transformations of the scalar field

\[
\phi(x) \rightarrow \phi'(x) \equiv \exp \{ i \theta \} \phi(x).
\]

(63)

In order to have a ground state the potential should be bounded from below, i.e., \( h > 0 \). For the quadratic piece there are two possibilities, shown in Fig. 12:

1. \( \mu^2 > 0 \): The potential has only the trivial minimum \( \phi = 0 \). It describes a massive scalar particle with mass \( \mu \) and quartic coupling \( h \).

2. \( \mu^2 < 0 \): The minimum is obtained for those field configurations satisfying

\[
|\phi_0| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}} > 0, \quad V(\phi_0) = -\frac{h}{4} v^4.
\]

(64)

Owing to the \( U(1) \) phase-invariance of the Lagrangian, there is an infinite number of degenerate states of minimum energy, \( \phi_0(x) = \frac{v}{\sqrt{2}} \exp \{ i \theta \} \). By choosing a particular solution, \( \theta = 0 \) for example, as the ground state, the symmetry gets spontaneously broken. If we parametrize the excitations over the ground state as

\[
\phi(x) \equiv \frac{1}{\sqrt{2}} \left[ v + \varphi_1(x) + i \varphi_2(x) \right],
\]

(65)
where \( \varphi_1 \) and \( \varphi_2 \) are real fields, the potential takes the form

\[
V(\phi) = V(\phi_0) - \mu^2 \varphi_1^2 + hv \varphi_1 (\varphi_1^2 + \varphi_2^2) + \frac{h}{4} (\varphi_1^2 + \varphi_2^2)^2.
\]

(66)

Thus \( \varphi_1 \) describes a massive state of mass \( m_{\varphi_1}^2 = -2\mu^2 \), while \( \varphi_2 \) is massless.

The first possibility \( (\mu^2 > 0) \) is just the usual situation with a single ground state. The other case, with SSB, is more interesting. The appearance of a massless particle when \( \mu^2 < 0 \) is easy to understand: the field \( \varphi_2 \) describes excitations around a flat direction in the potential, i.e., into states with the same energy as the chosen ground state. Since those excitations do not cost any energy, they obviously correspond to a massless state.

The fact that there are massless excitations associated with the SSB mechanism is a completely general result, known as the Goldstone theorem [25]: if a Lagrangian is invariant under a continuous symmetry group \( G \), but the vacuum is only invariant under a subgroup \( H \subset G \), then there must exist as many massless spin-0 particles (Goldstone bosons) as broken generators (i.e., generators of \( G \) which do not belong to \( H \)).

### 4.2 The Higgs–Kibble mechanism

At first sight, the Goldstone theorem has very little to do with our mass problem; in fact, it makes it worse since we want massive states and not massless ones. However, something very interesting happens when there is a local gauge symmetry [26, 27].

Let us consider [3] an \( SU(2)_L \) doublet of complex scalar fields

\[
\phi(x) \equiv \begin{pmatrix} \phi^{(+)}(x) \\ \phi^{(0)}(x) \end{pmatrix}.
\]

(67)

The gauged scalar Lagrangian of the Goldstone model in Eq. (62),

\[
\mathcal{L}_S = (D_\mu \phi) \dagger D^\mu \phi - \mu^2 \varphi_1 \phi - h (\varphi_1^2) \varphi_1 \phi \left( \varphi_1^2 + \varphi_2^2 \right), \quad (h > 0, \mu^2 < 0),
\]

(68)

\[
D_\mu \phi = \left[ \partial_\mu + ig \tilde{W}^\mu + ig' y_\phi B_\mu \right] \phi, \quad y_\phi = Q_\phi - T_3 = \frac{1}{2},
\]

(69)

is invariant under local \( SU(2)_L \otimes U(1)_Y \) transformations. The value of the scalar hypercharge is fixed by the requirement of having the correct couplings between \( \phi(x) \) and \( A^\mu(x) \); i.e., the photon does not couple to \( \phi^{(0)} \), and \( \phi^{(+)} \) has the right electric charge.

The potential is very similar to the one considered before. There is an infinite set of degenerate states with minimum energy, satisfying

\[
|0| \langle \phi^{(0)} |0 \rangle | = \sqrt{-\mu^2 / 2h} \equiv \frac{v}{\sqrt{2}}.
\]

(70)

Note that we have made explicit the association of the classical ground state with the quantum vacuum. Since the electric charge is a conserved quantity, only the neutral scalar field can acquire a vacuum expectation value. Once we choose a particular ground state, the \( SU(2)_L \otimes U(1)_Y \) symmetry gets spontaneously broken to the electromagnetic subgroup \( U(1)_{\text{QED}} \), which by construction still remains a true symmetry of the vacuum. According to the Goldstone theorem three massless states should then appear.

Now, let us parametrize the scalar doublet in the general form

\[
\phi(x) = \exp \left\{ i \frac{\sigma_i}{2} \theta^i(x) \right\} \frac{1}{\sqrt{2}} \left( \begin{array}{c} 1 \\ 0 \end{array} \right), \quad v + H(x),
\]

(71)
with four real fields $\theta^i(x)$ and $H(x)$. The crucial point is that the local $SU(2)_L$ invariance of the Lagrangian allows us to rotate away any dependence on $\theta^i(x)$. These three fields are precisely the would-be massless Goldstone bosons associated with the SSB mechanism.

The covariant derivative (69) couples the scalar multiplet to the $SU(2)_L \otimes U(1)_Y$ gauge bosons. If one takes the physical (unitary) gauge $\theta^i(x) = 0$, the kinetic piece of the scalar Lagrangian (68) takes the form:

$$\left( D_\mu \phi \right)^\dagger D^\mu \phi \xrightarrow{\theta^i = 0} \frac{1}{2} \partial_\mu H^\mu H + (v + H)^2 \left\{ \frac{g^2}{4} W^\mu_1 W^\mu_1 + \frac{g^2}{8 \cos^2 \theta_W} Z^\mu_2 Z^\mu_2 \right\}. \quad (72)$$

The vacuum expectation value of the neutral scalar has generated a quadratic term for the $W^\pm$ and the $Z$, i.e., those gauge bosons have acquired masses:

$$M_Z \cos \theta_W = M_W = \frac{1}{2} v g. \quad (73)$$

Therefore, we have found a clever way of giving masses to the intermediate carriers of the weak force. We just add $L_S$ to our $SU(2)_L \otimes U(1)_Y$ model. The total Lagrangian is invariant under gauge transformations, which guarantees the renormalizability of the associated Quantum Field Theory [28]. However, SSB occurs. The three broken generators give rise to three massless Goldstone bosons which, owing to the underlying local gauge symmetry, can be eliminated from the Lagrangian. Going to the unitary gauge, we discover that the $W^\pm$ and the $Z$ (but not the $\gamma$, because $U(1)_{\text{QED}}$ is an unbroken symmetry) have acquired masses, which are moreover related as indicated in Eq. (73). Notice that Eq. (52) has now the meaning of writing the gauge fields in terms of the physical boson fields with definite mass.

It is instructive to count the number of degrees of freedom (d.o.f.). Before the SSB mechanism, the Lagrangian contains massless $W^\pm$ and $Z$ bosons, i.e., $3 \times 2 = 6$ d.o.f., due to the two possible polarizations of a massless spin-1 field, and four real scalar fields. After SSB, the three Goldstone modes are ‘eaten’ by the weak gauge bosons, which become massive and, therefore, acquire one additional longitudinal polarization. We have then $3 \times 3 = 9$ d.o.f. in the gauge sector, plus the remaining scalar particle $H$, which is called the Higgs boson. The total number of d.o.f. remains of course the same.

### 4.3 Predictions

We have now all the needed ingredients to describe the electroweak interaction within a well-defined Quantum Field Theory. Our theoretical framework implies the existence of massive intermediate gauge bosons, $W^\pm$ and $Z$. Moreover, the Higgs–Kibble mechanism has produced a precise prediction\(^1\) for the $W^\pm$ and $Z$ masses, relating them to the vacuum expectation value of the scalar field through Eq. (73). Thus $M_Z$ is predicted to be bigger than $M_W$ in agreement with the measured masses [29,30]:

$$M_Z = 91.1875 \pm 0.0021 \text{ GeV}, \quad M_W = 80.398 \pm 0.025 \text{ GeV}. \quad (74)$$

From these experimental numbers, one obtains the electroweak mixing angle

$$\sin^2 \theta_W = 1 - \frac{M_W^2}{M_Z^2} = 0.223. \quad (75)$$

We can easily get an independent estimate of $\sin^2 \theta_W$ from the decay $\mu^- \rightarrow e^- \nu_e \bar{\nu}_\mu$. The momentum transfer $q^2 = (p_\mu - p_{e^-})^2 = (p_e + p_{\nu_e})^2 \lesssim m_\mu^2$ is much smaller than $M_W^2$. Therefore the

\(^1\)Note, however, that the relation $M_Z \cos \theta_W = M_W$ has a more general validity. It is a direct consequence of the symmetry properties of $L_S$ and does not depend on its detailed dynamics.
W propagator in Fig. 6 shrinks to a point and can be well approximated through a local four-fermion interaction, i.e.,

$$\frac{g^2}{M_W^2 - q^2} \approx \frac{g^2}{M_W^2} = \frac{4\pi\alpha}{\sin^2\theta_W M_W^2} \equiv 4\sqrt{2} G_F . \quad (76)$$

The measured muon lifetime, $\tau_\mu = (2.197019 \pm 0.000021) \cdot 10^{-6} \text{ s}$ [31], provides a very precise determination of the Fermi coupling constant $G_F$:

$$\frac{1}{\tau_\mu} = \Gamma_\mu = \frac{G_F^2 m_\mu^5}{192 \pi^3} f(m_e^2/m_\mu^2) \left( 1 + \delta_{\text{RC}} \right), \quad f(x) \equiv 1 - 8x + 8x^3 - x^4 - 12x^2 \log x . \quad (77)$$

Taking into account the radiative corrections $\delta_{\text{RC}}$, which are known to $O(\alpha^2)$ [32, 33], one gets [31]:

$$G_F = (1.166371 \pm 0.000006) \cdot 10^{-5} \text{ GeV}^{-2} . \quad (78)$$

The measured values of $\alpha^{-1} = 137.035999710 (96)$, $M_W$ and $G_F$ imply

$$\sin^2 \theta_W = 0.215 , \quad (79)$$

in very good agreement with Eq. (75). We shall see later that the small difference between these two numbers can be understood in terms of higher-order quantum corrections. The Fermi coupling also gives a direct determination of the electroweak scale, i.e., the scalar vacuum expectation value:

$$v = \left( \sqrt{2} G_F \right)^{-1/2} = 246 \text{ GeV} . \quad (80)$$

### 4.4 The Higgs boson

The scalar Lagrangian in Eq. (68) has introduced a new scalar particle into the model: the Higgs $H$. In terms of the physical fields (unitary gauge), $\mathcal{L}_S$ takes the form

$$\mathcal{L}_S = \frac{1}{4} h v^4 + \mathcal{L}_H + \mathcal{L}_{HG^2} , \quad (81)$$

where

$$\mathcal{L}_H = \frac{1}{2} \partial_\mu H \partial^\mu H - \frac{1}{2} M_H^2 H^2 - \frac{M_H^2}{2v} H^3 - \frac{M_H^2}{8v^2} H^4 , \quad (82)$$

Fig. 13: Higgs couplings to the gauge bosons
The Standard Model of Electroweak Interactions

\[ \mathcal{L}_{HC^2} = M_W^2 W^\mu W^\nu \left\{ 1 + \frac{2}{v} H + \frac{H^2}{v^2} \right\} + \frac{1}{2} M_Z^2 Z^\mu Z^\nu \left\{ 1 + \frac{2}{v} H + \frac{H^2}{v^2} \right\} \]  

(83)

and the Higgs mass is given by

\[ M_H = \sqrt{-2\mu^2} = \sqrt{2h v}. \]  

(84)

The Higgs interactions (Fig. 13) have a very characteristic form: they are always proportional to the mass (squared) of the coupled boson. All Higgs couplings are determined by \( M_H, M_W, M_Z, \) and the vacuum expectation value \( v. \)

So far the experimental searches for the Higgs have only provided a lower bound on its mass, corresponding to the exclusion of the kinematical range accessible at LEP and the Tevatron [7]:

\[ M_H > 114.4 \text{ GeV} \]  

(95% C.L.)

(85)

4.5 Fermion masses

\[ \frac{m_f}{v} \]

Fig. 14: Fermionic coupling of the Higgs boson

A fermionic mass term \( \mathcal{L}_m = -m \bar{\psi} \psi = -m \left( \bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L \right) \) is not allowed, because it breaks the gauge symmetry. However, since we have introduced an additional scalar doublet into the model, we can write the following gauge-invariant fermion-scalar coupling:

\[ \mathcal{L}_Y = -c_1 \left( \bar{\psi}_L \right)_L \left( \phi^{(+)} \right)_L \psi_R - c_2 \left( \bar{\psi}_L \right)_L \left( \phi^{(0)*} \right)_L \psi_R - c_3 \left( \bar{\psi}_L \right)_L \left( \phi^{(+)} \right)_L e + \text{h.c.,} \]  

(86)

where the second term involves the \( \mathcal{C} \)-conjugate scalar field \( \phi^c \equiv i \sigma_2 \phi^*. \) In the unitary gauge (after SSB), this Yukawa-type Lagrangian takes the simpler form

\[ \mathcal{L}_{Y'} = -\frac{1}{\sqrt{2}} (v + H) \left\{ c_1 \bar{d} d + c_2 \bar{u} u + c_3 \bar{e} e \right\}. \]  

(87)

Therefore, the SSB mechanism generates also fermion masses:

\[ m_d = c_1 \frac{v}{\sqrt{2}}, \quad m_u = c_2 \frac{v}{\sqrt{2}}, \quad m_e = c_3 \frac{v}{\sqrt{2}}. \]  

(88)

Since we do not know the parameters \( c_i, \) the values of the fermion masses are arbitrary. Note, however, that all Yukawa couplings are fixed in terms of the masses (Fig. 14):

\[ \mathcal{L}_Y = -\left( 1 + \frac{H}{v} \right) \left\{ m_d \bar{d} d + m_u \bar{u} u + m_e \bar{e} e \right\}. \]  

(89)

5 Electroweak phenomenology

In the gauge and scalar sectors, the SM Lagrangian contains only four parameters: \( g, g', \mu^2, \) and \( h. \) One could trade them for \( \alpha, \theta_W, M_W, \) and \( M_H. \) Alternatively, we can choose as free parameters:

\[ G_F = (1.166 371 \pm 0.000 006) \cdot 10^{-5} \text{ GeV}^{-2} \]  

[31],

19
\[ \alpha^{-1} = 137.035999710 \pm 0.000000096 \quad [15], \]
\[ M_Z = (91.1875 \pm 0.0021) \text{ GeV} \quad [29, 30] \]

and the Higgs mass \( M_H \). This has the advantage of using the three most precise experimental determinations to fix the interaction. The relations
\[ \sin^2 \theta_W = 1 - \frac{M_Z^2}{M_W^2}, \quad M_W^2 \sin^2 \theta_W = \frac{\alpha}{\sqrt{2} G_F} \]
determine then \( \sin^2 \theta_W = 0.212 \) and \( M_W = 80.94 \text{ GeV} \). The predicted \( M_W \) is in good agreement with the measured value in Eq. (74).

At tree level (Fig. 15), the decay widths of the weak gauge bosons can easily be computed. The \( W \) partial widths,
\[ \Gamma(W^- \to \bar{\nu}_i l^-) = \frac{G_F M_W^3}{6 \pi \sqrt{2}}, \quad \Gamma(W^- \to \bar{u}_i d_j) = N_C |V_{ij}|^2 \frac{G_F M_W^3}{6 \pi \sqrt{2}}, \]
are equal for all leptonic decay modes (up to small kinematical mass corrections). The quark modes involve also the colour quantum number \( N_C = 3 \) and the mixing factor \( V_{ij} \) relating weak and mass eigenstates, \( d'_i = V_{ij} d_j \). The \( Z \) partial widths are different for each decay mode, since its couplings depend on the fermion charge:
\[ \Gamma(Z \to f\bar{f}) = N_f \frac{G_F M_W^3}{6 \pi \sqrt{2}} (|v_f|^2 + |a_f|^2), \]
where \( N_l = 1 \) and \( N_q = N_C \). Summing over all possible final fermion pairs, one predicts the total widths \( \Gamma_W = 2.09 \text{ GeV} \) and \( \Gamma_Z = 2.48 \text{ GeV} \), in excellent agreement with the experimental values \( \Gamma_W = (2.147 \pm 0.060) \text{ GeV} \) and \( \Gamma_Z = (2.4952 \pm 0.0023) \text{ GeV} \) [29, 30]. The universality of the \( W \) couplings implies
\[ \text{Br}(W^- \to \bar{\nu}_i l^-) = \frac{1}{3 + 2N_C} = 11.1\%, \]
where we have taken into account that the decay into the top quark is kinematically forbidden. Similarly, the leptonic decay widths of the \( Z \) are predicted to be \( \Gamma_l \equiv \Gamma(Z \to l^+ l^-) = 84.85 \text{ MeV} \). As shown in Table 2, these predictions are in good agreement with the measured leptonic widths, confirming the universality of the \( W \) and \( Z \) leptonic couplings. There is, however, an excess of the branching ratio \( W \to \tau \bar{\nu}_\tau \) with respect to \( W \to e \bar{\nu}_e \) and \( W \to \mu \bar{\nu}_\mu \), which represents a 2.8 \( \sigma \) effect [29, 30]. The universality of the leptonic \( W \) couplings can also be tested indirectly, through weak decays mediated by charged-current interactions. Comparing the measured decay widths of leptonic or semileptonic decays which differ only by the lepton flavour, one can test experimentally that the \( W \) interaction is indeed the same, i.e., that \( g_e = g_\mu = g_\tau \equiv g \). As shown in Table 3, the present data verify the universality of the leptonic charged-current couplings to the 0.2% level.
Table 2: Measured values of $\text{Br}(W^- \to l^- \bar{l}^-)$ and $\Gamma(Z \to l^+ l^-)$ [29, 30]. The average of the three leptonic modes is shown in the last column (for a massless charged lepton $l$).

<table>
<thead>
<tr>
<th></th>
<th>$e$</th>
<th>$\mu$</th>
<th>$\tau$</th>
<th>$l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Br}(W^- \to \bar{\nu} l^-)$ (%)</td>
<td>$10.65 \pm 0.17$</td>
<td>$10.59 \pm 0.15$</td>
<td>$11.44 \pm 0.22$</td>
<td>$10.84 \pm 0.09$</td>
</tr>
<tr>
<td>$\Gamma(Z \to l^+ l^-)$ (MeV)</td>
<td>$83.92 \pm 0.12$</td>
<td>$83.99 \pm 0.18$</td>
<td>$84.08 \pm 0.22$</td>
<td>$83.985 \pm 0.086$</td>
</tr>
</tbody>
</table>

Table 3: Experimental determinations of the ratios $g_l/g\nu$ [18, 34]

<table>
<thead>
<tr>
<th></th>
<th>$g_\tau/g_\mu$</th>
<th>$g_\mu/g_\nu$</th>
<th>$g_l/g_\nu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma_{\tau \to \nu_\tau \bar{\nu}<em>\tau}/\Gamma</em>{\mu \to \nu_\mu \bar{\nu}_\mu}$</td>
<td>$1.0004 \pm 0.0022$</td>
<td>$0.996 \pm 0.005$</td>
<td>$0.979 \pm 0.017$</td>
</tr>
<tr>
<td>$\Gamma_{\tau \to \nu_\tau \bar{\nu}<em>\tau}/\Gamma</em>{\tau \to e \bar{e}_e}$</td>
<td>$1.0000 \pm 0.0020$</td>
<td>$1.0017 \pm 0.0015$</td>
<td>$1.012 \pm 0.009$</td>
</tr>
<tr>
<td>$\Gamma_{W \to \mu \bar{\nu}<em>\mu}/\Gamma</em>{W \to e \bar{e}_e}$</td>
<td>$0.997 \pm 0.010$</td>
<td>$</td>
<td>g_\tau/g_\mu</td>
</tr>
</tbody>
</table>

Another interesting quantity is the $Z$ decay width into invisible modes,

$$\frac{\Gamma_{\text{inv}}}{\Gamma_l} = \frac{N_\nu \Gamma(Z \to \bar{\nu} \nu)}{\Gamma_l} = \frac{2 N_\nu}{(1 - 4 \sin^2 \theta_W)^2 + 1}, \quad (95)$$

which is usually normalized to the charged leptonic width. The comparison with the measured value, $\Gamma_{\text{inv}}/\Gamma_l = 5.942 \pm 0.016$, provides very strong experimental evidence for the existence of three different light neutrinos.

5.1 Fermion-pair production at the $Z$ peak

![Fig. 16: Tree-level contributions to $e^+e^- \to \bar{f} f$ and kinematical configuration in the centre-of-mass system](image)

Additional information can be obtained from the study of the process $e^+e^- \to \gamma, Z \to \bar{f} f$ (Fig. 16). For unpolarized $e^+$ and $e^-$ beams, the differential cross-section can be written, at lowest order, as

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{8\pi} N_f \left\{ A (1 + \cos^2 \theta) + B \cos \theta - h_f \left[ C (1 + \cos^2 \theta) + D \cos \theta \right] \right\}, \quad (96)$$

where $h_f = \pm 1$ denotes the sign of the helicity of the produced fermion $f$, and $\theta$ is the scattering angle between $e^-$ and $f$ in the centre-of-mass system. Here,

$$A = 1 + 2 v_e v_f \text{Re} \chi + \left( v_e^2 + a_e^2 \right) \left( v_f^2 + a_f^2 \right) |\chi|^2,$$
\[ B = 4 a_e a_f \text{Re}(\chi) + 8 v_e a_v a_f |\chi|^2, \]
\[ C = 2 v_e a_f \text{Re}(\chi) + 2 (v_e^2 + a_e^2) v_f a_f |\chi|^2, \]
\[ D = 4 a_e v_f \text{Re}(\chi) + 4 v_e a_e (v_f^2 + a_f^2) |\chi|^2, \]
(97)

and \( \chi \) contains the \( Z \) propagator
\[ \chi = \frac{G_F M_Z^2}{\sqrt{2} \pi \alpha} \frac{s}{s - M_Z^2 + i s \Gamma_Z / M_Z}. \]
(98)

The coefficients \( A, B, C \) and \( D \) can be experimentally determined by measuring the total cross-section, the forward–backward asymmetry, the polarization asymmetry, and the forward–backward polarization asymmetry, respectively:
\[ \sigma(s) = \frac{4 \pi \alpha^2}{3s} N_f A, \quad A_{FB}(s) \equiv \frac{N_F - N_B}{N_F + N_B} = \frac{3B}{8 A}, \]
\[ A_{FB,Pol}(s) = \frac{\sigma^{(h_f=+1)} - \sigma^{(h_f=-1)}}{\sigma^{(h_f=+1)} + \sigma^{(h_f=-1)}} = -\frac{C}{A}, \]
\[ A_{FB,Pol} = \frac{N_F^{(h_f=+1)} - N_B^{(h_f=-1)}}{N_F^{(h_f=+1)} + N_B^{(h_f=-1)}}, \quad A_{FB,Pol}(M_Z^2) = \frac{3D}{8 A}. \]
(99)

Here, \( N_F \) and \( N_B \) denote the number of \( f \)'s emerging in the forward and backward hemispheres, respectively, with respect to the electron direction. The measurement of the final fermion polarization can be done for \( f = \tau \) by measuring the distribution of the final \( \tau \) decay products.

For \( s = M_Z^2 \), the real part of the \( Z \) propagator vanishes and the photon-exchange terms can be neglected in comparison with the \( Z \)-exchange contributions \( (\Gamma_Z^2 / M_Z^2) \ll 1 \). Equations (99) become then,
\[ \sigma^{0,f} \equiv \sigma(M_Z^2) = \frac{12 \pi}{M_Z^2} \frac{\Gamma_e \Gamma_f}{\Gamma_Z^2}, \quad A_{FB}^{0,f} \equiv A_{FB}(M_Z^2) = \frac{3}{4} P_e P_f, \]
\[ A_{Pol}^{0,f} \equiv A_{Pol}(M_Z^2) = P_f, \quad A_{FB,Pol}^{0,f} \equiv A_{FB,Pol}(M_Z^2) = \frac{3}{4} P_e, \]
(100)

where \( \Gamma_f \) is the \( Z \) partial decay width into the \( \bar{f}f \) final state, and
\[ P_f \equiv -A_f = \frac{-2 v_f a_f}{v_f^2 + a_f^2} \]
(101)

is the average longitudinal polarization of the fermion \( f \), which only depends on the ratio of the vector and axial-vector couplings.

With polarized \( e^+e^- \) beams, which have been available at SLC, one can also study the left–right asymmetry between the cross-sections for initial left- and right-handed electrons, and the corresponding forward–backward left–right asymmetry:
\[ A_{LR}^{0,f} \equiv A_{LR}(M_Z^2) = \frac{\sigma_L(M_Z^2) - \sigma_R(M_Z^2)}{\sigma_L(M_Z^2) + \sigma_R(M_Z^2)} = -P_e, \quad A_{FB,LR}^{0,f} \equiv A_{FB,LR}(M_Z^2) = \frac{3}{4} P_f. \]
(102)

At the \( Z \) peak, \( A_{LR}^{0,f} \) measures the average initial lepton polarization, \( P_e \), without any need for final particle identification, while \( A_{FB,LR}^{0,f} \) provides a direct determination of the final fermion polarization.

\( P_f \) is a very sensitive function of \( \sin^2 \theta_W \). Small higher-order corrections can produce large variations on the predicted lepton polarization because \( |v_f| = \frac{1}{2} |1 - 4 \sin^2 \theta_W| \ll 1 \). Therefore \( P_f \) provides an interesting window to search for electroweak quantum effects.
Fig. 17: The photon vacuum polarization (left) generates a charge screening effect, making \( \alpha(s) \) smaller at larger distances.

5.2 QED and QCD corrections

Before trying to analyse the relevance of higher-order electroweak contributions, it is instructive to consider the numerical impact of the well-known QED and QCD corrections. The photon propagator gets vacuum polarization corrections, induced by virtual fermion–antifermion pairs. This kind of QED loop correction can be taken into account through a redefinition of the QED coupling, which depends on the energy scale. The resulting QED running coupling \( \alpha(s) \) decreases at large distances. This can be intuitively understood as the charge screening generated by the virtual fermion pairs (Fig. 17). The physical QED vacuum behaves as a polarized dielectric medium. The huge difference between the electron and \( Z \) mass scales makes this quantum correction relevant at LEP energies [15, 29, 30]:

\[
\alpha(m_e^2)^{-1} = 137.035\,999\,710 \pm 96 \quad > \quad \alpha(M_W^2)^{-1} = 128.93 \pm 0.05 .
\]

The running effect generates an important change in Eq. (91). Since \( G_F \) is measured at low energies, while \( M_W \) is a high-energy parameter, the relation between both quantities is modified by vacuum-polarization contributions. Changing \( \alpha \) by \( \alpha(M_W^2) \), one gets the corrected predictions:

\[
\sin^2 \theta_W = 0.231 , \quad M_W = 79.96 \text{ GeV} .
\]

The experimental value of \( M_W \) is in the range between the two results obtained with either \( \alpha \) or \( \alpha(M_W^2) \), showing its sensitivity to quantum corrections. The effect is more spectacular in the leptonic asymmetries at the \( Z \) peak. The small variation of \( \sin^2 \theta_W \) from 0.212 to 0.231 induces a large shift on the vector \( Z \) coupling to charged leptons from \( \nu_l = -0.076 \) to \( -0.038 \), changing the predicted average lepton polarization \( P_l \) by a factor of two.

So far, we have treated quarks and leptons on an equal footing. However, quarks are strongly interacting particles. The gluonic corrections to the decays \( Z \to \bar{q}q \) and \( W^- \to \bar{u}d \) can be directly incorporated into the formulae given before by taking an ‘effective’ number of colours:

\[
N_C \quad \implies \quad N_C \left\{ 1 + \frac{\alpha_s}{\pi} + \ldots \right\} \approx 3.115 ,
\]

where we have used the value of \( \alpha_s \) at \( s = M_Z^2 \), \( \alpha_s(M_Z^2) = 0.119 \pm 0.002 \) [7, 35].

Note that the strong coupling also ‘runs’. However, the gluon self-interactions generate an anti-screening effect, through gluon-loop corrections to the gluon propagator, which spread out the QCD charge [6]. Since this correction is larger than the screening of the colour charge induced by virtual quark–antiquark pairs, the net result is that the strong coupling decreases at short distances. Thus QCD has the required property of asymptotic freedom: quarks behave as free particles when \( Q^2 \to \infty \) [36, 37].

QCD corrections increase the probability of the \( Z \) and the \( W^\pm \) to decay into hadronic modes. Therefore, their leptonic branching fractions become smaller. The effect can be easily estimated from Eq. (94). The probability of the decay \( W^- \to \bar{\nu}_e e^- \) gets reduced from 11.1% to 10.8%, improving the agreement with the measured value in Table 2.

23
5.3 Higher-order electroweak corrections

Quantum corrections offer the possibility to be sensitive to heavy particles, which cannot be kinematically accessed, through their virtual loop effects. In QED and QCD the vacuum polarization contribution of a heavy fermion pair is suppressed by inverse powers of the fermion mass. At low energies, the information on the heavy fermions is then lost. This ‘decoupling’ of the heavy fields happens in theories with only vector couplings and an exact gauge symmetry [38], where the effects generated by the heavy particles can always be reabsorbed into a redefinition of the low-energy parameters.

The SM involves, however, a broken chiral gauge symmetry. This has the very interesting implication of avoiding the decoupling theorem [38]. The vacuum polarization contributions induced by a heavy top generate corrections to the $W^{\pm}$ and $Z$ propagators (Fig. 18), which increase quadratically with the top mass [39]. Therefore, a heavy top does not decouple. For instance, with $m_t = 171$ GeV, the leading quadratic correction to the second relation in Eq. (91) amounts to a sizeable 3% effect. The quadratic mass contribution originates in the strong breaking of weak isospin generated by the top and bottom quark masses, i.e., the effect is actually proportional to $m_t^2 - m_b^2$.

Owing to an accidental $SU(2)_C$ symmetry of the scalar sector (the so-called custodial symmetry), the virtual production of Higgs particles does not generate any quadratic dependence on the Higgs mass at one loop [39]. The dependence on $M_H$ is only logarithmic. The numerical size of the corresponding correction in Eq. (91) varies from a 0.1% to a 1% effect for $M_H$ in the range from 100 to 1000 GeV.

Higher-order corrections to the different electroweak couplings are non-universal and usually smaller than the self-energy contributions. There is one interesting exception, the $Zb\bar{b}$ vertex (Fig. 19), which is sensitive to the top quark mass [40]. The $Zf\bar{f}$ vertex gets one-loop corrections where a virtual $W^{\pm}$ is exchanged between the two fermionic legs. Since the $W^{\pm}$ coupling changes the fermion flavour, the decays $Z \to \bar{d}d, \bar{s}s, \bar{b}b$ get contributions with a top quark in the internal fermionic lines, i.e., $Z \to \bar{t}t \to \bar{d}_id_j$. Notice that this mechanism can also induce the flavour-changing neutral-current decays $Z \to \bar{d}_id_j$ with $i \neq j$. These amplitudes are suppressed by the small CKM mixing factors $|V_{tj}V_{ti}^*|^2$. However, for the $Z \to b\bar{b}$ vertex, there is no suppression because $|V_{tb}| \approx 1$.

The explicit calculation [40–43] shows the presence of hard $m_t^2$ corrections to the $Z \to b\bar{b}$ vertex. This effect can be easily understood [40] in non-unitary gauges where the unphysical charged scalar $\phi(\pm)$ is present. The fermionic couplings of the charged scalar are proportional to the fermion masses; therefore the exchange of a virtual $\phi(\pm)$ gives rise to a $m_t^2$ factor. In the unitary gauge, the charged scalar has been ‘eaten’ by the $W^{\pm}$ field; thus the effect comes now from the exchange of a longitudinal $W^{\pm}$, with terms proportional to $q^{\mu}q^{\nu}$ in the propagator that generate fermion masses. Since the $W^{\pm}$...
couples only to left-handed fermions, the induced correction is the same for the vector and axial-vector $Z\bar{b}b$ couplings and, for $m_t = 171$ GeV, amounts to a 1.6% reduction of the $Z \rightarrow \bar{b}b$ decay width [40]. The ‘non-decoupling’ present in the $Z\bar{b}b$ vertex is quite different from the one happening in the boson self-energies. The vertex correction is not dependent on the Higgs mass. Moreover, while any kind of new heavy particle coupling to the gauge bosons would contribute to the $W$ and $Z$ self-energies, the possible new physics contributions to the $Z\bar{b}b$ vertex are much more restricted and, in any case, different. Therefore, the independent experimental measurement of the two effects is very valuable in order to disentangle possible new physics contributions from the SM corrections. In addition, since the ‘non-decoupling’ vertex effect is related to $W_L$ exchange, it is sensitive to the SSB mechanism.

5.4 Standard Model electroweak fit

The leptonic asymmetry measurements from LEP and SLD can all be combined to determine the ratios $v_l/a_l$ of the vector and axial-vector couplings of the three charged leptons, or equivalently the effective electroweak mixing angle

$$\sin^2 \theta^\text{lept}_{\text{eff}} = \frac{1}{4} \left( 1 - \frac{v_l}{a_l} \right).$$

The sum $(v_l^2 + a_l^2)$ is derived from the leptonic decay widths of the $Z$, i.e., from Eq. (93) corrected with a multiplicative factor $(1 + \frac{2}{3} \frac{\alpha}{\pi})$ to account for final-state QED corrections. The signs of $v_l$ and $a_l$ are fixed by requiring $a_e < 0$.

The resulting 68% probability contours are shown in Fig. 20, which provides strong evidence of the electroweak radiative corrections. The good agreement with the SM predictions, obtained for low values of the Higgs mass, is lost if only the QED vacuum polarization contribution is taken into account, as indicated by the point with an arrow. Notice that the uncertainty induced by the input value of $\alpha(M_Z^2)^{-1} = 128.93 \pm 0.05$ is sizeable. The measured couplings of the three charged leptons confirm lepton universality in the neutral-current sector. The solid contour combines the three measurements assuming universality.

![Fig. 20: Combined LEP and SLD measurements of $\sin^2 \theta^\text{lept}_{\text{eff}}$ and $\Gamma_l$ (left) and the corresponding effective vector and axial-vector couplings $v_l$ and $a_l$ (right). The shaded region shows the SM prediction. The arrows point in the direction of increasing values of $m_t$ and $M_H$. The point shows the predicted values if, among the electroweak radiative corrections, only the photon vacuum polarization is included. Its arrow indicates the variation induced by the uncertainty in $\alpha(M_Z^2)$ [29, 30].](image)
Fig. 21: Measurements of $A_l$, $A_b$ (SLD) and $A_{FB}^{0,b}$. The arrows pointing to the left (right) show the variations of the SM prediction with $M_H = 300 \pm 700$ GeV ($m_t = 172.7 \pm 2.9$ GeV). The small arrow oriented to the left shows the additional uncertainty from $\alpha(M_Z^2)$ [29,30].

The neutrino couplings can also be determined from the invisible $Z$ decay width, by assuming three identical neutrino generations with left-handed couplings, and fixing the sign from neutrino scattering data. Alternatively, one can use the SM prediction for $\Gamma_{inv}$ to get a determination of the number of light neutrino flavours [29,30]:

$$N_\nu = 2.9840 \pm 0.0082 \, .$$

(107)

Figure 21 shows the measured values of $A_l$ and $A_b$ together with the joint constraint obtained from $A_{FB}^{0,b}$ (diagonal band). The direct measurement of $A_b$ at SLD agrees well with the SM prediction; however, a much lower value is obtained from the ratio $\frac{2}{3} A_{FB}^{0,b}/A_l$. This is the most significant discrepancy observed in the $Z$-pole data. Heavy quarks ($\frac{4}{3} A_{FB}^{0,b}/A_b$) seem to prefer a high value of the Higgs mass, while leptons ($A_l$) favour a light Higgs. The combined analysis prefers low values of $M_H$, because of the influence of $A_l$.

The strong sensitivity of the ratio $R_b \equiv \Gamma(Z \to bb)/\Gamma(Z \to \text{hadrons})$ to the top quark mass is shown in Fig. 22. Owing to the $|V_{td}|^2$ suppression, such a dependence is not present in the analogous ratio $R_d$. Combined with all other electroweak precision measurements at the $Z$ peak, $R_b$ provides a determination of $m_t$ in good agreement with the direct and most precise measurement at the Tevatron. This is shown in Fig. 23, which compares the information on $M_W$ and $m_t$ obtained at LEP1 and SLD, with the direct measurements performed at LEP2 and the Tevatron. A similar comparison for $m_t$ and $M_H$ is also shown. The lower bound on $M_H$ obtained from direct searches excludes a large portion of the 68% C.L. allowed domain from precision measurements.

Taking all direct and indirect data into account, one obtains the best constraints on $M_H$. The global electroweak fit results in the $\Delta \chi^2 = \chi^2 - \chi^2_{\text{min}}$ curve shown in Fig. 24. The lower limit on $M_H$ obtained from direct searches is close to the point of minimum $\chi^2$. At 95% C.L., one gets [29,30]

$$114.4 \, \text{GeV} < M_H < 144 \, \text{GeV} \, .$$

(108)

The fit provides also a very accurate value of the strong coupling constant, $\alpha_s(M_Z^2) = 0.1186 \pm 0.0027$, in very good agreement with the world average value $\alpha_s(M_Z^2) = 0.119 \pm 0.002$ [7,35]. The largest
Fig. 23: Comparison (left) of the direct measurements of $M_W$ and $m_t$ (LEP2 and Tevatron data) with the indirect determination through electroweak radiative corrections (LEP1 and SLD). Also shown in the SM relationship for the masses as function of $M_H$. The figure on the right makes the analogous comparison for $m_t$ and $M_H$ [29, 30].

Fig. 24: $\Delta \chi^2 = \chi^2 - \chi^2_{\text{min}}$ versus $M_H$, from the global fit to the electroweak data. The vertical band indicates the 95% exclusion limit from direct searches [29, 30].

Fig. 25: Comparison between the measurements included in the combined analysis of the SM and the results from the global electroweak fit [29, 30]

discrepancy between theory and experiment occurs for $A_{FB}^{0,b}$, with the fitted value being nearly 3 $\sigma$ larger than the measurement. As shown in Fig. 25, a good agreement is obtained for all other observables.

5.5 Gauge self-interactions

At tree level, the $W$-pair production process $e^+e^- \rightarrow W^+W^-$ involves three different contributions (Fig. 26), corresponding to the exchange of $\nu_e$, $\gamma$ and $Z$. The cross-section measured at LEP2 agrees very well with the SM predictions. As shown in Fig. 27, the $\nu_e$-exchange contribution alone would lead to an unphysical growing of the cross-section at large energies and, therefore, would imply a violation of unitarity. Adding the $\gamma$-exchange contribution softens this behaviour, but a clear disagreement with the data persists. The $Z$-exchange mechanism, which involves the $ZWW$ vertex, appears to be crucial in
order to explain the data.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig26}
\caption{Feynman diagrams contributing to $e^+e^- \rightarrow W^+W^-$ and $e^+e^- \rightarrow ZZ$}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig27}
\caption{Measured energy dependence of $\sigma(e^+e^- \rightarrow W^+W^-)$ (left) and $\sigma(e^+e^- \rightarrow ZZ)$ (right). The three curves shown for the $W$-pair production cross-section correspond to only the $\nu_e$-exchange contribution (upper curve), $\nu_e$ exchange plus photon exchange (middle curve), and all contributions including also the $ZWW$ vertex (lower curve). Only the $e$-exchange mechanism contributes to $Z$-pair production [29, 30].}
\end{figure}

Since the $Z$ is electrically neutral, it does not interact with the photon. Moreover, the SM does not include any local $ZZZ$ vertex. Therefore, the $e^+e^- \rightarrow ZZ$ cross-section only involves the contribution from $e$ exchange. The agreement of the SM predictions with the experimental measurements in both production channels, $W^+W^-$ and $ZZ$, provides a test of the gauge self-interactions. There is a clear signal of the presence of a $ZWW$ vertex, with the predicted strength, and no evidence for any $\gamma ZZ$ or $ZZZ$ interactions. The gauge structure of the $SU(2)_L \otimes U(1)_Y$ theory is nicely confirmed by the data.

### 5.6 Higgs decays

The couplings of the Higgs boson are always proportional to some mass scale. The $Hf\bar{f}$ interaction grows linearly with the fermion mass, while the $HWW$ and $HZZ$ vertices are proportional to $M_W^2$ and $M_Z^2$, respectively. Therefore, the most probable decay mode of the Higgs will be the one into the heaviest possible final state. This is clearly illustrated in Fig. 28. The $H \rightarrow b\bar{b}$ decay channel is by far the dominant one below the $W^+W^-$ production threshold. When $M_H$ is large enough to allow the production of a pair of gauge bosons, $H \rightarrow W^+W^-$ and $H \rightarrow ZZ$ become dominant. For $M_H > 2m_t$, the $H \rightarrow t\bar{t}$ decay width is also sizeable, although smaller than the $WW$ and $ZZ$ ones because of the different dependence of the corresponding Higgs coupling with the mass scale (linear instead of quadratic).

The total decay width of the Higgs grows with increasing values of $M_H$. The effect is very strong
above the $W^+W^-$ production threshold. A heavy Higgs becomes then very broad. At $M_H \sim 600$ GeV, the width is around 100 GeV; while for $M_H \sim 1$ TeV, $\Gamma_H$ is already of the same size as the Higgs mass itself.

The design of the LHC detectors has taken into account all these very characteristic properties in order to optimize the future search for the Higgs boson.

6 Flavour dynamics

We have learned experimentally that there are six different quark flavours $u, d, s, c, b, t$, three different charged leptons $e, \mu, \tau$ and their corresponding neutrinos $\nu_e, \nu_\mu, \nu_\tau$. We can nicely include all these particles into the SM framework, by organizing them into three families of quarks and leptons, as indicated in Eqs. (1) and (2). Thus, we have three nearly identical copies of the same $SU(2)_L \otimes U(1)_Y$ structure, with masses as the only difference.

Let us consider the general case of $N_G$ generations of fermions, and denote $\nu'_j, l'_j, u'_j, d'_j$ the members of the weak family $j$ ($j = 1, \ldots, N_G$), with definite transformation properties under the gauge group. Owing to the fermion replication, a large variety of fermion-scalar couplings are allowed by the gauge symmetry. The most general Yukawa Lagrangian has the form

$$\mathcal{L}_Y = - \sum_{jk} \left\{ \left( \bar{u}'_j, \bar{d}'_j \right)_L \begin{pmatrix} c^{(d)}_{jk} & \phi^{(+)}(0) \\ \phi^{(0)} & \phi^{(-)}(0) \end{pmatrix} d'_{kR} + c^{(u)}_{jk} \left( \begin{pmatrix} \phi^{(0)*} \\ -\phi^{(-)} \end{pmatrix} \right)^T u'_{kR} \right\},$$

where $c^{(d)}_{jk}, c^{(u)}_{jk}$ and $c^{(l)}_{jk}$ are arbitrary coupling constants.

After SSB, the Yukawa Lagrangian can be written as

$$\mathcal{L}_Y = - \left( 1 + \frac{H}{v} \right) \left\{ \bar{\nu}'_j \mathbf{M}'_d \nu'_{kR} + \bar{u}'_j \mathbf{M}'_u u'_{kR} + \bar{l}'_j \mathbf{M}'_l l'_{kR} + \text{h.c.} \right\}.$$

Here, $d', u'$ and $l'$ denote vectors in the $N_G$-dimensional flavour space, and the corresponding mass matrices are given by

$$(\mathbf{M}'_d)_{ij} \equiv c^{(d)}_{ij} \frac{v}{\sqrt{2}}, \quad (\mathbf{M}'_u)_{ij} \equiv c^{(u)}_{ij} \frac{v}{\sqrt{2}}, \quad (\mathbf{M}'_l)_{ij} \equiv c^{(l)}_{ij} \frac{v}{\sqrt{2}}.$$

The diagonalization of these mass matrices determines the mass eigenstates $d_j, u_j$ and $l_j$, which are linear combinations of the corresponding weak eigenstates $d'_j, u'_j$ and $l'_j$, respectively.
The matrix $M'_d$ can be decomposed as $M'_d = H_d U_d = S'_d M_d S_d U_d$, where $H_d \equiv \sqrt{M'_d M'_d}$ is an Hermitian positive-definite matrix, while $U_d$ is unitary. $H_d$ can be diagonalized by a unitary matrix $S_d$; the resulting matrix $M_d$ is diagonal, Hermitian and positive definite. Similarly, one has $M'_u = H_u U_u = S'_u M_u S_u U_u$ and $M'_l = H_l U_l = S'_l M_l S_l U_l$. In terms of the diagonal mass matrices

$$M_d = \text{diag}(m_d, m_s, m_b, \ldots), \quad M_u = \text{diag}(m_u, m_c, m_t, \ldots), \quad M_l = \text{diag}(m_e, m_\mu, m_\tau, \ldots),$$

the Yukawa Lagrangian takes the simpler form

$$\mathcal{L}_Y = - \left( 1 + \frac{H}{v} \right) \left\{ \bar{d} M_d d + \bar{u} M_u u + \bar{l} M_l l \right\},$$

where the mass eigenstates are defined by

$$d_L \equiv S_d d'_L, \quad u_L \equiv S_u u'_L, \quad l_L \equiv S_l l'_L,$$
$$d_R \equiv S_d u'_d d', \quad u_R \equiv S_u u'_R, \quad l_R \equiv S_l u'_l l'_R.$$

Note that the Higgs couplings are proportional to the corresponding fermions masses.

Since, $\bar{T'_L} f'_L = \bar{T}_L f_L$ and $\bar{T'_R} f'_R = \bar{T}_R f_R$ ($f = d, u, l$), the form of the neutral-current part of the $SU(2)_L \otimes U(1)_Y$ Lagrangian does not change when expressed in terms of mass eigenstates. Therefore, there are no flavour-changing neutral currents in the SM (GIM mechanism [5]). This is a consequence of treating all equal-charge fermions on the same footing.

However, $\bar{\pi}_L d'_L = \bar{\pi}_L S_u S'_d d_L \equiv \bar{\pi}_L V d_L$. In general, $S_u \neq S_d$; thus, if one writes the weak eigenstates in terms of mass eigenstates, a $N_G \times N_G$ unitary mixing matrix $V$, called the Cabibbo–Kobayashi–Maskawa (CKM) matrix [45,46], appears in the quark charged-current sector:

$$\mathcal{L}_{\text{CC}} = -\frac{g}{2\sqrt{2}} \left\{ W^\mu_\mu \left[ \sum_{ij} \bar{u}_i \gamma^\mu (1 - \gamma_5) V_{ij} d_j + \sum_l \bar{l}_l \gamma^\mu (1 - \gamma_5) l \right] + \text{h.c.} \right\}. \quad (115)$$

The matrix $V$ couples any ‘up-type’ quark with all ‘down-type’ quarks (Fig. 29).

If neutrinos are assumed to be massless, we can always redefine the neutrino flavours in such a way as to eliminate the analogous mixing in the lepton sector: $\bar{\nu}'_L l'_L = \bar{\nu}_L S'_l S_l l_L \equiv \bar{\nu}_L V l_L$. Thus we have lepton-flavour conservation in the minimal SM without right-handed neutrinos. If sterile $\nu_R$ fields are included in the model, one would have an additional Yukawa term in Eq. (109), giving rise to a neutrino mass matrix $(M'_\nu)_{ij} \equiv c_{\nu_{ij}} v/\sqrt{2}$. Thus the model could accommodate non-zero neutrino masses and lepton-flavour violation through a lepton mixing matrix $V_L$ analogous to the one present

---

2The condition $\det M'_f \neq 0$ ($f = d, u, l$) guarantees that the decomposition $M'_f = H_f U_f$ is unique: $U_f \equiv H^{-1}_f M'_f$. The matrices $S_f$ are completely determined (up to phases) only if all diagonal elements of $M_f$ are different. If there is some degeneracy, the arbitrariness of $S_f$ reflects the freedom to define the physical fields. If $\det M_f = 0$, the matrices $U_f$ and $S_f$ are not uniquely determined, unless their unitarity is explicitly imposed.
in the quark sector. Note, however, that the total lepton number \( L \equiv L_e + L_\mu + L_\tau \) would still be conserved. We know experimentally that neutrino masses are tiny and there are strong bounds on lepton-flavour violating decays: \( \text{Br}(\mu^\pm \rightarrow e^\pm e^-) < 1.0 \cdot 10^{-12} \) [47], \( \text{Br}(\mu^\pm \rightarrow e^\pm \gamma) < 1.2 \cdot 10^{-11} \) [48], \( \text{Br}(\tau^\pm \rightarrow \mu^\pm \gamma) < 4.5 \cdot 10^{-8} \) [49, 50] ... However, we do have a clear evidence of neutrino oscillation phenomena.

The fermion masses and the quark mixing matrix \( V \) are all determined by the Yukawa couplings in Eq. (109). However, the coefficients \( c_{ij}^{(f)} \) are not known; therefore we have a bunch of arbitrary parameters. A general \( N_G \times N_G \) unitary matrix is characterized by \( N_G^2 \) real parameters: \( N_G(N_G - 1)/2 \) moduli and \( N_G(N_G + 1)/2 \) phases. In the case of \( V \), many of these parameters are irrelevant, because we can always choose arbitrary quark phases. Under the phase redefinitions \( u_i \rightarrow e^{i\phi_i} u_i \) and \( d_j \rightarrow e^{i\phi_j} d_j \), the mixing matrix changes as \( V_{ij} \rightarrow V_{ij} e^{i(\theta_{ij} - \phi_i)} \); thus, \( 2N_G - 1 \) phases are unobservable. The number of physical free parameters in the quark-mixing matrix then gets reduced to \( (N_G - 1)^2 \): \( N_G(N_G - 1)/2 \) moduli and \( (N_G - 1)(N_G - 2)/2 \) phases.

In the simpler case of two generations, \( V \) is determined by a single parameter. One then recovers the Cabibbo rotation matrix [45]

\[
V = \begin{pmatrix}
\cos \theta_C & \sin \theta_C \\
-\sin \theta_C & \cos \theta_C
\end{pmatrix}.
\]

With \( N_G = 3 \), the CKM matrix is described by three angles and one phase. Different (but equivalent) representations can be found in the literature. The Particle data Group [7] advocates the use of the following one as the ‘standard’ CKM parametrization:

\[
V = \begin{pmatrix}
c_{12} c_{13} & s_{12} c_{13} & s_{13} e^{-i\delta_{13}} \\
-s_{12} c_{23} - c_{12} s_{23} s_{13} e^{i\delta_{13}} & c_{12} c_{23} - s_{12} s_{23} s_{13} e^{i\delta_{13}} & -c_{12} s_{23} - s_{12} c_{23} s_{13} e^{i\delta_{13}} \\
s_{12} s_{23} - c_{12} c_{23} s_{13} e^{i\delta_{13}} & -c_{12} s_{23} - s_{12} c_{23} s_{13} e^{i\delta_{13}} & c_{23} c_{13}
\end{pmatrix}.
\]

Here \( c_{ij} \equiv \cos \theta_{ij} \) and \( s_{ij} \equiv \sin \theta_{ij} \), with \( i \) and \( j \) being ‘generation’ labels \( (i, j = 1, 2, 3) \). The real angles \( \theta_{12}, \theta_{23} \) and \( \theta_{13} \) can all be made to lie in the first quadrant, by an appropriate redefinition of quark field phases; then, \( c_{ij} \geq 0 \), \( s_{ij} \geq 0 \) and \( 0 \leq \delta_{13} \leq 2\pi \).

Notice that \( \delta_{13} \) is the only complex phase in the SM Lagrangian. Therefore it is the only possible source of \( CP \)-violation phenomena. In fact, it was for this reason that the third generation was assumed to exist [46], before the discovery of the \( b \) and the \( \tau \). With two generations, the SM could not explain the observed \( CP \) violation in the \( K \) system.

### 6.1 Quark mixing

Fig. 30: Determinations of \( V_{ij} \) are done in semileptonic quark decays (left) where a single quark current is present. Hadronic decay modes (right) involve two different quark currents and are more affected by QCD effects (gluons can couple everywhere).

Our knowledge of the charged-current parameters is unfortunately not so good as in the neutral-current case. In order to measure the CKM matrix elements, one needs to study hadronic weak decays

\[31\]
of the type \( H \rightarrow H' l^- \bar{\nu}_l \) or \( H \rightarrow H' l^+ \nu_l \), which are associated with the corresponding quark transitions \( d_j \rightarrow u_i l^- \bar{\nu}_l \) and \( u_i \rightarrow d_j l^+ \nu_l \) (Fig. 30). Since quarks are confined within hadrons, the decay amplitude

\[
T[H \rightarrow H' l^- \bar{\nu}_l] \approx \frac{G_F}{\sqrt{2}} V_{ij} \langle H' | \bar{u}_i \gamma^\mu (1 - \gamma^5) d_j | H \rangle \left( \bar{l}_\mu (1 - \gamma^5) \nu_l \right)
\]

always involves an hadronic matrix element of the weak left current. The evaluation of this matrix element is a non-perturbative QCD problem, which introduces unavoidable theoretical uncertainties.

One usually looks for a semileptonic transition where the matrix element can be fixed at some kinematical point by a symmetry principle. This has the virtue of reducing the theoretical uncertainties to the level of symmetry-breaking corrections and kinematical extrapolations. The standard example is a \( 0^- \rightarrow 0^- \) decay such as \( K \rightarrow \pi l \nu \), \( D \rightarrow K l \nu \) or \( B \rightarrow D l \nu \). Only the vector current can contribute in this case:

\[
\langle P'(k') | \bar{u}_i \gamma^\mu d_j | P(k) \rangle = C_{PP'} \left\{ (k + k')^\mu f_+(t) + (k - k')^\mu f_-(t) \right\} .
\]

Here, \( C_{PP'} \) is a Clebsch–Gordan factor and \( t = (k - k')^2 \equiv q^2 \). The unknown strong dynamics is fully contained in the form factors \( f_\pm(t) \). In the limit of equal quark masses, \( m_u - m_d = 0 \), the divergence of the vector current is zero; thus \( q_\mu (\bar{u}_i \gamma^\mu d_j) = 0 \), which implies \( f_-(t) = 0 \) and, moreover, \( f_+(0) = 1 \) to all orders in the strong coupling because the associated flavour charge is a conserved quantity\(^3\). Therefore, one only needs to estimate the corrections induced by the quark mass differences.

Since \( q^2 \left| \bar{\nu}_l (1 - \gamma_5) \nu_l \right| \sim m_l \), the contribution of \( f_-(t) \) is kinematically suppressed in the electron and muon modes. The decay width can then be written as

\[
\Gamma(P \rightarrow P' l \nu) = \frac{G_F^2 M_P^3}{192 \pi^3} |V_{ij}|^2 C_{PP'}^2 |f_+(0)|^2 \mathcal{I} \left( 1 + \delta_{RC} \right),
\]

where \( \delta_{RC} \) is an electroweak radiative correction factor and \( \mathcal{I} \) denotes a phase-space integral, which in the \( m_l = 0 \) limit takes the form

\[
\mathcal{I} \approx \int_0^{(M_P - M_{P'})^2} \frac{dt}{M_P^8} \lambda^{3/2}(t, M_P^2, M_{P'}^2) \left| \frac{f_+(t)}{f_+(0)} \right|^2 .
\]

The usual procedure to determine \( |V_{ij}| \) involves three steps:

1. Measure the shape of the \( t \) distribution. This fixes \( |f_+(t)/f_+(0)| \) and therefore determines \( \mathcal{I} \).
2. Measure the total decay width \( \Gamma \). Since \( G_F \) is already known from \( \mu \) decay, one gets then an experimental value for the product \( |f_+(0)| |V_{ij}| \).
3. Get a theoretical prediction for \( f_+(0) \).

It is important to realize that theoretical input is always needed. Thus, the accuracy of the \( |V_{ij}| \) determination is limited by our ability to calculate the relevant hadronic input.

The conservation of the vector and axial-vector QCD currents in the massless quark limit allows for accurate determinations of the light-quark mixings \( |V_{ud}| \) and \( |V_{us}| \). The present values are shown in Table 4, which takes into account the recent changes in the \( K \rightarrow \pi e^+ \nu_e \) data \([7, 34]\) and the new \( |V_{us}| \) determinations from Cabibbo suppressed tau decays \([52]\) and from the ratio of decay amplitudes \( \Gamma(K^+ \rightarrow \mu^+ \bar{\nu}_\mu)/\Gamma(\pi^+ \rightarrow \mu^+ \bar{\nu}_\mu) \) \([53–55]\). Since \( |V_{ub}|^2 \) is tiny, these two light quark entries provide a sensible test of the unitarity of the CKM matrix:

\[
|V_{ud}|^2 + |V_{us}|^2 + |V_{ub}|^2 = 0.9980 \pm 0.0012 .
\]

\(^3\)This is completely analogous to the electromagnetic charge conservation in QED. The conservation of the electromagnetic current implies that the proton electromagnetic form factor does not get any QED or QCD correction at \( q^2 = 0 \) and, therefore, \( Q(p) = 2 Q(u) + Q(d) = |Q(e)| \). A detailed proof can be found in Ref. [51].
Table 4: Direct determinations of the CKM matrix elements $V_{ij}$. For $|V_{tb}|$, 95% C.L. limits are given

<table>
<thead>
<tr>
<th>CKM entry</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>V_{ud}</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>0.9746 ± 0.0019</td>
<td>$n \rightarrow p e^- \bar{\nu}_e$ [7]</td>
</tr>
<tr>
<td></td>
<td>0.9728 ± 0.0030</td>
<td>$\pi^+ \rightarrow \pi^0 e^+ \nu_e$ [62]</td>
</tr>
<tr>
<td></td>
<td>0.97378 ± 0.00027</td>
<td>average</td>
</tr>
<tr>
<td>$</td>
<td>V_{us}</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>0.2220 ± 0.0033</td>
<td>$\tau$ decays [52]</td>
</tr>
<tr>
<td></td>
<td>$0.2226^{+0.0026}_{-0.0014}$</td>
<td>$K^+/\pi^+ \rightarrow \mu^+ \nu_\mu$, $V_{ud}$ [7, 53–55]</td>
</tr>
<tr>
<td></td>
<td>0.226 ± 0.005</td>
<td>Hyperon decays [64–66]</td>
</tr>
<tr>
<td></td>
<td>0.2230 ± 0.0015</td>
<td>average</td>
</tr>
<tr>
<td>$</td>
<td>V_{cd}</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>0.230 ± 0.011</td>
<td>$\nu d \rightarrow c X$ [7]</td>
</tr>
<tr>
<td></td>
<td>0.227 ± 0.010</td>
<td>average</td>
</tr>
<tr>
<td>$</td>
<td>V_{cs}</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$0.94^{+0.35}_{-0.29}$</td>
<td>$W^+ \rightarrow c s$ [7]</td>
</tr>
<tr>
<td></td>
<td>0.974 ± 0.013</td>
<td>$W^+ \rightarrow$ had., $V_{uj}$, $V_{cd}$, $V_{cb}$ [29, 30]</td>
</tr>
<tr>
<td>$</td>
<td>V_{cb}</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>0.0417 ± 0.0007</td>
<td>$b \rightarrow c l \bar{\nu}_l$ [7, 67]</td>
</tr>
<tr>
<td></td>
<td>0.0413 ± 0.0006</td>
<td>average</td>
</tr>
<tr>
<td>$</td>
<td>V_{ub}</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>0.0045 ± 0.0003</td>
<td>$b \rightarrow u l \bar{\nu}_l$ [7, 67]</td>
</tr>
<tr>
<td></td>
<td>0.0044 ± 0.0003</td>
<td>average</td>
</tr>
<tr>
<td>$</td>
<td>V_{tb}</td>
<td>/\sqrt{\sum_q</td>
</tr>
<tr>
<td>$</td>
<td>V_{tb}</td>
<td>$</td>
</tr>
</tbody>
</table>

It is important to notice that at the quoted level of uncertainty radiative corrections play a crucial role.

In the limit of very heavy quark masses, QCD has additional symmetries [56–59] which can be used to make rather precise determinations of $|V_{cb}|$, either from exclusive decays such as $B \rightarrow D^* l \bar{\nu}_l$ [60, 61] or from the inclusive analysis of $b \rightarrow c l \bar{\nu}_l$ transitions. The control of theoretical uncertainties is much more difficult for $|V_{ub}|$, $|V_{cd}|$ and $|V_{cs}|$, because the symmetry arguments associated with the light and heavy quark limits get corrected by sizeable symmetry-breaking effects.

The most precise determination of $|V_{cd}|$ is based on neutrino and antineutrino interactions. The difference of the ratio of double-muon to single-muon production by neutrino and antineutrino beams is proportional to the charm cross-section off-valence $d$ quarks and, therefore, to $|V_{cd}|$. A direct determination of $|V_{cs}|$ can also be obtained from charm-tagged $W$ decays at LEP2. Moreover, the ratio of the
total hadronic decay width of the $W$ to the leptonic one provides the sum \cite{29, 30}

$$\sum_{i=\text{u,c}} \sum_{j=\text{d,s,b}} |V_{ij}|^2 = 1.999 \pm 0.025 .$$ \hspace{1cm} (123)

Although much less precise than Eq. (122), this result tests unitarity at the 1.25% level. From Eq. (123) one can also obtain a tighter determination of $|V_{us}|$, using the experimental knowledge on the other CKM matrix elements, i.e., $|V_{ud}|^2 + |V_{us}|^2 + |V_{ub}|^2 + |V_{cd}|^2 + |V_{cb}|^2 = 1.0512 \pm 0.0058$. This gives the most accurate and final value of $|V_{us}|$ quoted in Table 4.

The measured entries of the CKM matrix show a hierarchical pattern, with the diagonal elements being very close to one, the ones connecting the two first generations having a size

$$\lambda \approx |V_{us}| = 0.2230 \pm 0.0015 ,$$ \hspace{1cm} (124)

deleting to all orders in $\lambda$ \cite{71},

$$V = \begin{bmatrix} 1 - \frac{\lambda^2}{2} & \frac{\lambda}{2} & A\lambda^3 (\rho - i\eta) \\ -\lambda & 1 - \frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3 (1 - \rho - i\eta) & -A\lambda^2 & 1 \end{bmatrix} + O(\lambda^4) ,$$ \hspace{1cm} (125)

where

$$A \approx \frac{|V_{cb}|}{\lambda^2} = 0.831 \pm 0.014 , \quad \sqrt{\rho^2 + \eta^2} \approx \left| \frac{V_{ub}}{AV_{cb}} \right| = 0.478 \pm 0.033 .$$ \hspace{1cm} (126)

Defining to all orders in $\lambda$ \cite{72} $s_{12} \equiv \lambda$, $s_{23} \equiv A\lambda^2$ and $s_{13} e^{-i\delta_{13}} \equiv A\lambda^3 (\rho - i\eta)$, Eq. (125) just corresponds to a Taylor expansion of Eq. (117) in powers of $\lambda$.

### 6.2 CP violation

While parity and charge conjugation are violated by the weak interactions in a maximal way, the product of the two discrete transformations is still a good symmetry (left-handed fermions $\leftrightarrow$ right-handed antifermions). In fact, $CP$ appears to be a symmetry of nearly all observed phenomena. However, a slight violation of the $CP$ symmetry at the level of 0.2% is observed in the neutral kaon system and more sizeable signals of $CP$ violation have been recently established at the B factories. Moreover, the huge matter–antimatter asymmetry present in our Universe is a clear manifestation of $CP$ violation and its important role in the primordial baryogenesis.

The $CPT$ theorem guarantees that the product of the three discrete transformations is an exact symmetry of any local and Lorentz-invariant quantum field theory preserving micro-causality. Therefore, a violation of $CP$ requires a corresponding violation of time reversal. Since $T$ is an antunitary transformation, this requires the presence of relative complex phases between different interfering amplitudes.

The electroweak SM Lagrangian only contains a single complex phase $\delta_{13}$ ($\eta$). This is the sole possible source of $CP$ violation and, therefore, the SM predictions for $CP$-violating phenomena are quite constrained. The CKM mechanism requires several necessary conditions in order to generate an observable $CP$-violation effect. With only two fermion generations, the quark mixing mechanism cannot give rise to $CP$ violation; therefore, for $CP$ violation to occur in a particular process, all three generations are required to play an active role. In the kaon system, for instance, $CP$-violation effects can only appear
at the one-loop level, where the top quark is present. In addition, all CKM matrix elements must be non-zero and the quarks of a given charge must be non-degenerate in mass. If any of these conditions were not satisfied, the CKM phase could be rotated away by a redefinition of the quark fields. $\mathcal{CP}$-violation effects are then necessarily proportional to the product of all CKM angles, and should vanish in the limit where any two (equal-charge) quark masses are taken to be equal. All these necessary conditions can be summarized in a very elegant way as a single requirement on the original quark mass matrices $M_u$ and $M_d$ [73]:

$$\mathcal{CP} \text{ violation } \iff \det \left\{ \begin{array}{c} M_u' \ M_u'^\dagger, \\ M_d' \ M_d'^\dagger \end{array} \right\} \neq 0. \quad (127)$$

Without performing any detailed calculation, one can make the following general statements on the implications of the CKM mechanism of $\mathcal{CP}$ violation:

- Owing to unitarity, for any choice of $i, j, k, l$ (between 1 and 3),

$$\text{Im} \left[ \begin{array}{c} V_{ij} \ V_{ik}^* \ V_{lk} \ V_{lj} \end{array} \right] = \mathcal{J} \sum_{m,n=1}^{3} \epsilon_{ilm} \epsilon_{jkn}, \quad (128)$$

$$\mathcal{J} = c_{12} c_{23}^2 s_{12} s_{23} s_{13} \sin \delta_{13} \approx A^2 \lambda^6 \eta < 10^{-4}. \quad (129)$$

Any $\mathcal{CP}$-violation observable involves the product $\mathcal{J}$ [73]. Thus, violations of the $\mathcal{CP}$ symmetry are necessarily small.

- In order to have sizeable $\mathcal{CP}$-violating asymmetries $A \equiv (\Gamma - \Gamma)/(\Gamma + \Gamma)$, one should look for very suppressed decays, where the decay widths already involve small CKM matrix elements.

- In the SM, $\mathcal{CP}$ violation is a low-energy phenomenon, in the sense that any effect should disappear when the quark mass difference $m_c - m_u$ becomes negligible.

- $B$ decays are the optimal place for $\mathcal{CP}$-violation signals to show up. They involve small CKM matrix elements and are the lowest-mass processes where the three quark generations play a direct (tree-level) role.

The SM mechanism of $\mathcal{CP}$ violation is based on the unitarity of the CKM matrix. Testing the constraints implied by unitarity is then a way to test the source of $\mathcal{CP}$ violation. The unitarity tests in Eqs. (122) and (123) involve only the moduli of the CKM parameters, while $\mathcal{CP}$ violation has to do with their phases. More interesting are the off-diagonal unitarity conditions:

$$V_{ud}^* V_{us} + V_{cd}^* V_{cs} + V_{td}^* V_{ts} = 0, \quad (130)$$

$$V_{ub}^* V_{ub} + V_{cb}^* V_{cb} + V_{tb}^* V_{tb} = 0, \quad (131)$$

$$V_{ub}^* V_{ud} + V_{cb}^* V_{cd} + V_{tb}^* V_{td} = 0. \quad (132)$$

These relations can be visualized by triangles in a complex plane which, owing to Eq. (128), have the same area $|\mathcal{J}|/2$. In the absence of $\mathcal{CP}$ violation, these triangles would degenerate into segments along the real axis.

In the first two triangles, one side is much shorter than the other two (the Cabibbo suppression factors of the three sides are $\lambda$, $\lambda$ and $\lambda^5$ in the first triangle, and $\lambda^4$, $\lambda^2$ and $\lambda^2$ in the second one). This is why $\mathcal{CP}$ effects are so small for $K$ mesons (first triangle), and why certain asymmetries in $B_s$ decays are predicted to be tiny (second triangle). The third triangle looks more interesting, since the three sides have a similar size of about $\lambda^2$. They are small, which means that the relevant $b$-decay branching ratios are small, but once enough $B$ mesons have been produced, the $\mathcal{CP}$-violation asymmetries are sizeable. The present experimental constraints on this triangle are shown in Fig. 31, where it has been scaled by dividing its sides by $V_{cb}^* V_{cd}$. This aligns one side of the triangle along the real axis and makes its length equal to 1; the coordinates of the three vertices are then $(0,0)$, $(1,0)$ and $(\bar{\rho}, \bar{\eta}) \equiv (1 - \lambda^2/2)(\rho, \eta)$.
One side of the unitarity triangle has already been determined in Eq. (126) from the ratio $|V_{ub}/V_{cb}|$. The other side can be obtained from the measured mixing between the $B_d^0$ and $\bar{B}_d^0$ mesons (Fig. 32), $\Delta M_d = 0.507 \pm 0.004$ ps$^{-1}$ [67], which fixes $|V_{tb}|$. Additional information has been provided by the recent observation of $B_d^0-\bar{B}_d^0$ oscillations at CDF, implying $\Delta M_s = 17.77 \pm 0.12$ ps$^{-1}$ [74]. From the experimental ratio $\Delta M_d/\Delta M_s = 0.0286 \pm 0.0003$, one obtains $|V_{td}|/|V_{ts}|$. A more direct constraint on the parameter $\eta$ is given by the observed $CP$ violation in $K^0 \rightarrow 2\pi$ decays. The measured value of $|\varepsilon_K| = (2.232 \pm 0.007) \times 10^{-3}$ [7] determines the parabolic region shown in Fig. 31.

$B^0$ decays into $CP$ self-conjugate final states provide independent ways to determine the angles of the unitarity triangle [75, 76]. The $B^0$ (or $\bar{B}^0$) can decay directly to the given final state $f$, or do it after the meson has been changed to its antiparticle via the mixing process. $CP$-violating effects can then result from the interference of these two contributions. The time-dependent $CP$-violating rate asymmetries contain direct information on the CKM parameters. The gold-plated decay mode is $B^0_d \rightarrow J/\psi K_S$, which gives a clean measurement of $\beta \equiv -\arg(V_{td}^* V_{tb}/V_{td} V_{tb}^*)$, without strong-interaction uncertainties. Including the information obtained from other $b \rightarrow c\bar{c}s$ decays, one gets [67]:

$$\sin 2\beta = 0.68 \pm 0.03.$$  

(133)

Many additional tests of the CKM matrix from different $B$ decay modes are being pursued at the $B$ factories. Determinations of the other two angles of the unitarity triangle, $\alpha \equiv -\arg(V_{td}^* V_{ub}/V_{td} V_{ub}^*)$ and $\gamma \equiv -\arg(V_{td}^* V_{ub}/V_{td} V_{ub}^*)$, have already been obtained [67, 78], and are included in the global fit shown in Fig. 31 [77, 79]. Complementary and very valuable information could be also obtained from the kaon decay modes $K^+ \rightarrow \pi^+\nu\bar{\nu}$, $K_L \rightarrow \pi^0\nu\bar{\nu}$ and $K_L \rightarrow \pi^0 e^+e^-$ [80].
The Standard Model of Electroweak Interactions

![Graph showing measured fluxes of $^8$B solar neutrinos of $\nu_\mu$ or $\nu_\tau$ type ($\phi_{\mu,\tau}$) versus the flux of $\nu_e$ ($\phi_e$) [83]](image)

### 6.3 Lepton mixing

The so-called ‘solar neutrino problem’ has been a long-standing question since the very first chlorine experiment at the Homestake mine [81]. The flux of solar $\nu_e$ neutrinos reaching the Earth has been measured by several experiments to be significantly below the standard solar model prediction [82]. More recently, the Sudbury Neutrino Observatory has provided strong evidence that neutrinos do change flavour as they propagate from the core of the Sun [83], independently of solar model flux predictions. SNO is able to detect neutrinos through three different reactions: the charged-current process $\nu_e d \rightarrow e^- p p$ which is only sensitive to $\nu_e$, the neutral current transition $\nu_\mu d \rightarrow \nu_\tau p n$ which has equal probability for all active neutrino flavours, and the elastic scattering $\nu_x e^- \rightarrow \nu_x e^-$ which is also sensitive to $\nu_\mu$ and $\nu_\tau$, although the corresponding cross-section is a factor 6.48 smaller than the $\nu_e$ one. The measured neutrino fluxes, shown in Fig. 33, demonstrate the existence of a non-$\nu_e$ component in the solar neutrino flux at the 5.3 $\sigma$ level. The SNO results are in good agreement with the Super-Kamiokande solar measurements [84] and have been further reinforced with the more recent KamLAND data, showing that $\bar{\nu}_e$ from nuclear reactors disappear over distances of about 180 km [85].

Another evidence of oscillations has been obtained from atmospheric neutrinos. The known discrepancy between the experimental observations and the predicted ratio of muon to electron neutrinos has become much stronger with the high precision and large statistics of Super-Kamiokande [86]. The atmospheric anomaly appears to originate in a reduction of the $\nu_\mu$ flux, and the data strongly favours the $\nu_\mu \rightarrow \nu_\tau$ hypothesis. This result has been confirmed by K2K [87] and MINOS [88], observing the disappearance of accelerator $\nu_\mu$'s at distances of 250 and 735 km, respectively. Super-Kamiokande has recently reported statistical evidence of $\nu_\tau$ appearance at the 2.4 $\sigma$ level [86]. The direct detection of the produced $\nu_\tau$ is the main goal of the ongoing CERN to Gran Sasso neutrino programme.

Thus, we now have clear experimental evidence that neutrinos are massive particles and there is mixing in the lepton sector. Figures 34 and 35 show the present information on neutrino oscillations, from solar, atmospheric, accelerator, and reactor neutrino data. A global analysis, combining the full set of data, leads to the following preferred ranges for the oscillation parameters [7]:

\[
\Delta m^2_{21} = (8.0 \pm 0.4) \cdot 10^{-5} \text{ eV}^2, \quad 1.9 \cdot 10^{-3} < |\Delta m^2_{32}| / \text{eV}^2 < 3.0 \cdot 10^{-3}, \quad (134)
\]

\[
\sin^2(2\theta_{12}) = 0.86 \pm 0.03, \quad \sin^2(2\theta_{23}) > 0.92, \quad \sin^2(2\theta_{13}) < 0.19, \quad (135)
\]

where $\Delta m^2_{ij} \equiv m_i^2 - m_j^2$ are the mass squared differences between the neutrino mass eigenstates $\nu_{i,j}$ and $\theta_{ij}$ the corresponding mixing angles in the standard three-flavour parametrization [7]. The ranges indicate 90% C.L. bounds. In the limit $\theta_{13} = 0$, solar and atmospheric neutrino oscillations decouple.
because $\Delta m^2_{\odot} \ll \Delta m^2_{\text{atm}}$. Thus, $\Delta m^2_{21}$, $\theta_{12}$ and $\theta_{13}$ are constrained by solar data, while atmospheric experiments constrain $\Delta m^2_{32}$, $\theta_{23}$ and $\theta_{13}$. The angle $\theta_{13}$ is strongly constrained by the CHOOZ reactor experiment [89]. New planned reactor experiments T2K and NO$\nu$A are expected to achieve sensitivities around $\sin^2(2\theta_{13}) \sim 0.01$.

\[ \Delta L = -\frac{c_{ij}}{\Lambda} \bar{L}_i \tilde{\phi}^T \phi^j L^c_j + \text{h.c.,} \]

where $\nu^c_{iR} \equiv C \nu^T_{iR}$ denotes the charge-conjugated field. The Majorana mass matrix $M_{ij}$ could have an arbitrary size, because it is not related to the ordinary Higgs mechanism. Since both fields $\nu^c_{iR}$ and $\nu^c_{iR}$ absorb $\nu$ and create $\bar{\nu}$, the Majorana mass term mixes neutrinos and anti-neutrinos, violating lepton number by two units. Clearly, new physics is called for.

Adopting a more general effective field theory language, without any assumption about the existence of right-handed neutrinos or any other new particles, one can write the most general $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ invariant Lagrangian in terms of the known low-energy fields (left-handed neutrinos only). The SM is the unique answer with dimension four. The first contributions from new physics appear through dimension-5 operators, and have also a unique form which violates lepton number by two units [90]:

\[ \Delta \mathcal{L} = -\frac{c_{ij}}{\Lambda} \bar{L}_i \tilde{\phi}^T \phi^j L^c_j + \text{h.c.,} \]

where $L_i$ denotes the $i$-flavoured $SU(2)_L$ lepton doublet, $\tilde{\phi} \equiv i \tau^2 \phi^*$ and $L^c_i \equiv C L^T_i$. Similar operators with quark fields are forbidden, owing to their different hypercharges, while higher-dimension operators would be suppressed by higher powers of the new-physics scale $\Lambda$. After SSB, $\langle \phi^{(0)} \rangle = v/\sqrt{2}$, $\Delta \mathcal{L}$ generates a Majorana mass term for the left-handed neutrinos, with $M_{ij} = c_{ij} v^2 / \Lambda$. Thus, Majorana neutrino masses should be expected on general symmetry grounds. Taking $m_\nu \gtrsim 0.05 \text{ eV}$, as suggested

\[ m_\nu \sim m^2 / \Lambda \]

4This relation generalizes the well-known see-saw mechanism ($m_\nu \sim m^2 / \Lambda$) [91, 92].

Fig. 34: Allowed regions for $2\nu$ oscillations for the combination of solar ($\nu_e$) and KamLAND ($\bar{\nu}_e$) data, assuming $\mathcal{CPT}$ symmetry [83]

Fig. 35: MINOS allowed regions for $\nu_{\mu}$ disappearance oscillations, compared with K2K and Super-Kamiokande results [88]
Table 5: Best published limits (90% C.L.) on lepton-flavour-violating decays [7, 49, 50]

<table>
<thead>
<tr>
<th>Decay</th>
<th>Limit (90% C.L.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Br}(\mu^- \to e^- \gamma)$</td>
<td>$&lt; 1.2 \cdot 10^{-11}$</td>
</tr>
<tr>
<td>$\text{Br}(\mu^- \to e^- 2\gamma)$</td>
<td>$&lt; 7.2 \cdot 10^{-11}$</td>
</tr>
<tr>
<td>$\text{Br}(\mu^- \to e^- e^+)$</td>
<td>$&lt; 1.0 \cdot 10^{-12}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to \mu^- \gamma)$</td>
<td>$&lt; 4.5 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to e^- \gamma)$</td>
<td>$&lt; 1.1 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to e^- \mu^+)$</td>
<td>$&lt; 1.1 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to e^- K^0_S)$</td>
<td>$&lt; 5.6 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to \mu^- K^0_S)$</td>
<td>$&lt; 4.9 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to \mu^- e^+ \pi^-)$</td>
<td>$&lt; 0.7 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to \Lambda\pi^-)$</td>
<td>$&lt; 7.2 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to e^- \pi^0)$</td>
<td>$&lt; 1.4 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to e^+ \pi^-)$</td>
<td>$&lt; 1.2 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to \mu^- \pi^0)$</td>
<td>$&lt; 1.1 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\text{Br}(\tau^- \to \mu^- e^+ \pi^-)$</td>
<td>$&lt; 1.3 \cdot 10^{-7}$</td>
</tr>
</tbody>
</table>

by atmospheric neutrino data, one gets $\Lambda/c_{ij} \lesssim 10^{15}$ GeV, amazingly close to the expected scale of Grand Unification.

With non-zero neutrino masses, the leptonic charged-current interactions involve a flavour mixing matrix $V_L$. The data on neutrino oscillations imply that all elements of $V_L$ are large, except for $(V_L)_{e3} < 0.18$; therefore the mixing among leptons appears to be very different from the one in the quark sector. The number of relevant phases characterizing the matrix $V_L$ depends on the Dirac or Majorana nature of neutrinos, because if one rotates a Majorana neutrino by a phase, this phase will appear in its mass term which will no longer be real. With only three Majorana (Dirac) neutrinos, the $3 \times 3$ matrix $V_L$ involves six (four) independent parameters: three mixing angles and three (one) phases.

The smallness of neutrino masses implies a strong suppression of neutrinoless lepton-flavour-violating processes, which can be avoided in models with other sources of lepton-flavour violation, not related to $m_{\nu}$.

Table 5 shows the best published limits on lepton-flavour-violating decays. The B Factories are pushing the experimental limits on neutrinoless $\tau$ decays beyond the $10^{-7}$ level, increasing in a drastic way the sensitivity to new physics scales. Future experiments could further push some limits to the $10^{-9}$ level, allowing one to explore interesting and totally unknown phenomena. Complementary information will be provided by the MEG experiment, which will search for $\mu^+ \to e^+ \gamma$ events with a sensitivity of $10^{-13}$ [93]. There are also ongoing projects at J-PARC aiming to study $\mu \to e$ conversions in muonic atoms, at the $10^{-18}$ level.

At present, we still ignore whether neutrinos are Dirac or Majorana fermions. Another important question to be addressed in the future concerns the possibility of leptonic CP violation and its relevance for explaining the baryon asymmetry of our Universe through leptogenesis.

7 Summary

The SM provides a beautiful theoretical framework which is able to accommodate all our present knowledge on electroweak and strong interactions. It is able to explain any single experimental fact and, in some cases, it has successfully passed very precise tests at the 0.1% to 1% level. In spite of this impressive phenomenological success, the SM leaves too many unanswered questions to be considered as a complete description of the fundamental forces. We do not understand yet why fermions are replicated in three (and only three) nearly identical copies. Why the pattern of masses and mixings is what it is. Are the masses the only difference among the three families? What is the origin of the SM flavour structure? Which dynamics is responsible for the observed CP violation?

In the gauge and scalar sectors, the SM Lagrangian contains only four parameters: $g, g', \mu^2$, and $h$. We can trade them by $\alpha, M_Z, G_F,$ and $M_H^2$; this has the advantage of using the three most precise experimental determinations to fix the interaction. In any case, one describes a lot of physics with only four inputs. In the fermionic flavour sector, however, the situation is very different. With $N_G = 3$, we have 13 additional free parameters in the minimal SM: 9 fermion masses, 3 quark mixing angles and 1 phase. Taking into account non-zero neutrino masses, we have three more mass parameters plus the leptonic mixings: three angles and one phase (three phases) for Dirac (or Majorana) neutrinos.
Clearly, this is not very satisfactory. The source of this proliferation of parameters is the set of unknown Yukawa couplings in Eq. (109). The origin of masses and mixings, together with the reason for the existing family replication, constitute at present the main open problem in electroweak physics. The problem of fermion mass generation is deeply related to the mechanism responsible for the electroweak SSB. Thus, the origin of these parameters lies in the most obscure part of the SM Lagrangian: the scalar sector. The dynamics of flavour appears to be terra incognita which deserves a careful investigation.

The SM incorporates a mechanism to generate $C\mathcal{P}$ violation, through the single phase naturally occurring in the CKM matrix. Although the present laboratory experiments are well described, this mechanism is unable to explain the matter–antimatter asymmetry of our Universe. A fundamental explanation of the origin of $C\mathcal{P}$-violating phenomena is still lacking.

The first hints of new physics beyond the SM have emerged recently, with convincing evidence of neutrino oscillations showing that $\nu_e \rightarrow \nu_{\mu, \tau}$ and $\nu_\mu \rightarrow \nu_\tau$ transitions do occur. The existence of lepton-flavour violation opens a very interesting window to unknown phenomena.

The Higgs particle is the main missing block of the SM framework. The successful tests of the SM quantum corrections with precision electroweak data confirm the assumed pattern of SSB, but do not prove the validity of the minimal Higgs mechanism embedded in the SM. The present experimental bounds (108) put the Higgs hunting within the reach of the new generation of detectors. The LHC should find out whether such scalar field indeed exists, either confirming the SM Higgs mechanism or discovering completely new phenomena.

Many interesting experimental signals are expected to be seen in the near future. New experiments will probe the SM to a much deeper level of sensitivity and will explore the frontier of its possible extensions. Large surprises may well be expected, probably establishing the existence of new physics beyond the SM and offering clues to the problems of mass generation, fermion mixing, and family replication.

Acknowledgements

I want to thank the organizers for the charming atmosphere of this school and all the students for their many interesting questions and comments. This work has been supported by the EU MRTN-CT-2006-035482 (FLAVIAnet), MEC (Spain, FPA2004-00996), and Generalitat Valenciana (GVACOMP2007-156).
Appendices

A Basic inputs from quantum field theory

A.1 Wave equations

The classical Hamiltonian of a non-relativistic free particle is given by $H = \frac{\vec{p}^2}{2m}$. In quantum mechanics, energy and momentum correspond to operators acting on the particle wave function. The substitutions $H = i\hbar \frac{\partial}{\partial t}$ and $\vec{p} = -i\hbar \vec{\nabla}$ lead then to the Schrödinger equation:

$$i\hbar \frac{\partial}{\partial t} \psi(\vec{x}, t) = \frac{\hbar^2}{2m} \nabla^2 \psi(\vec{x}, t). \quad \text{(A.1)}$$

We can write the energy and momentum operators in a relativistic covariant way as $H = \frac{\vec{p}^2}{2m}$, where we have adopted the usual natural units convention $\hbar = c = 1$. The relation $E^2 = \vec{p}^2 + m^2$ determines the Klein–Gordon equation for a relativistic free particle:

$$(\Box + m^2) \phi(x) = 0, \quad \Box \equiv \partial^\mu \partial_\mu = \frac{\partial^2}{\partial t^2} - \nabla^2. \quad \text{(A.2)}$$

The Klein–Gordon equation is quadratic on the time derivative because relativity puts the space and time coordinates on an equal footing. Let us investigate whether an equation linear in derivatives could exist. Relativistic covariance and dimensional analysis restrict its possible form to

$$(i \gamma^\mu \partial_\mu - m) \psi(x) = 0. \quad \text{(A.3)}$$

Since the r.h.s. is identically zero, we can fix the coefficient of the mass term to be $-1$; this just determines the normalization of the four coefficients $\gamma^\mu$. Notice that $\gamma^\mu$ should transform as a Lorentz four-vector. The solutions of Eq. (A.3) should also satisfy the Klein–Gordon relation of Eq. (A.2). Applying an appropriate differential operator to Eq. (A.3), one can easily obtain the required quadratic equation:

$$- (i \gamma^\nu \partial_\nu + m) (i \gamma^\mu \partial_\mu - m) \psi(x) = 0 \equiv (\Box + m^2) \psi(x). \quad \text{(A.4)}$$

Terms linear in derivatives cancel identically, while the term with two derivatives reproduces the operator $\Box \equiv \partial^\mu \partial_\mu$ provided the coefficients $\gamma^\mu$ satisfy the algebraic relation

$$\left\{ \gamma^\mu, \gamma^\nu \right\} \equiv \gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2 \delta^{\mu\nu}, \quad \text{(A.5)}$$

which defines the so-called Dirac algebra. Eq. (A.3) is known as the Dirac equation.

Obviously the components of the four-vector $\gamma^\mu$ cannot simply be numbers. The three $2 \times 2$ Pauli matrices satisfy $\left\{ \sigma^i, \sigma^j \right\} = 2 \delta^{ij}$, which is very close to the relation (A.5). The lowest-dimensional solution to the Dirac algebra is obtained with $D = 4$ matrices. An explicit representation is given by:

$$\gamma^0 = \left( \begin{array}{cc} I_2 & 0 \\ 0 & -I_2 \end{array} \right), \quad \gamma^i = \left( \begin{array}{cc} 0 & \sigma^i \\ -\sigma^i & 0 \end{array} \right). \quad \text{(A.6)}$$

Thus, the wave function $\psi(x)$ is a column vector with four components in the Dirac space. The presence of the Pauli matrices strongly suggests that it contains two components of spin $\frac{1}{2}$. A proper physical analysis of its solutions shows that the Dirac equation describes simultaneously a fermion of spin $\frac{1}{2}$ and its own antiparticle [94].

It is useful to define the following combinations of gamma matrices:

$$\sigma^{\mu\nu} \equiv \frac{i}{2} \left[ \gamma^\mu, \gamma^\nu \right], \quad \gamma^5 \equiv \gamma^0 \gamma^1 \gamma^2 \gamma^3 = -\frac{i}{4!} \epsilon_{\mu\nu\rho\sigma} \gamma^\mu \gamma^\nu \gamma^\rho \gamma^\sigma. \quad \text{(A.7)}$$
In the explicit representation (A.6),
\[ \sigma_{ij} = \epsilon^{ijk} \begin{pmatrix} \sigma^k & 0 \\ 0 & \sigma^k \end{pmatrix}, \quad \sigma^{0i} = i \begin{pmatrix} 0 & \sigma^i \\ \sigma^i & 0 \end{pmatrix}, \quad \gamma_5 = \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix}. \] (A.8)

The matrix \( \sigma_{ij} \) is then related to the spin operator. Some important properties are
\[ \gamma^0 \gamma^\mu \gamma^0 = \gamma^\mu \dagger, \quad \gamma^0 \gamma_5 \gamma^0 = -\gamma_5 = -\gamma_5 \dagger, \quad \{ \gamma_5, \gamma^\mu \} = 0, \quad (\gamma_5)^2 = I_4. \] (A.9)

Specially relevant for weak interactions are the chirality projectors \( (P_L + P_R = 1) \)
\[ P_L \equiv \frac{1 - \gamma_5}{2}, \quad P_R \equiv \frac{1 + \gamma_5}{2}, \quad P^2_R = P_R, \quad P^2_L = P_L, \quad P_L P_R = P_R P_L = 0, \] (A.10)
which allow us to decompose the Dirac spinor in its left-handed and right-handed chirality parts:
\[ \psi(x) = [P_L + P_R] \psi(x) \equiv \psi_L(x) + \psi_R(x). \] (A.11)

In the massless limit, the chiralities correspond to the fermion helicities.

A.2 Lagrangian formalism

The Lagrangian formulation of a physical system provides a compact dynamical description and makes it easier to discuss the underlying symmetries. Like in classical mechanics, the dynamics is encoded in the action
\[ S = \int d^4x \ L[\phi_i(x), \partial \mu \phi_i(x)]. \] (A.12)

The integral over the four space–time coordinates preserves relativistic invariance. The Lagrangian density \( L \) is a Lorentz-invariant functional of the fields \( \phi_i(x) \) and their derivatives. The space integral \( L = \int d^3x \ L \) would correspond to the usual non-relativistic Lagrangian.

The principle of stationary action requires the variation \( \delta S \) of the action to be zero under small fluctuations \( \delta \phi_i \) of the fields. Assuming that the variations \( \delta \phi_i \) are differentiable and vanish outside some bounded region of space–time (which allows an integration by parts), the condition \( \delta S = 0 \) determines the Euler–Lagrange equations of motion for the fields:
\[ \frac{\partial L}{\partial \phi_i} - \partial \mu \left( \frac{\partial L}{\partial (\partial \mu \phi_i)} \right) = 0. \] (A.13)

One can easily find appropriate Lagrangians to generate the Klein–Gordon and Dirac equations. They should be quadratic on the fields and Lorentz invariant, which determines their possible form up to irrelevant total derivatives. The Lagrangian
\[ L = \partial^\mu \phi^* \partial_\mu \phi - m^2 \phi^* \phi \] (A.14)
describes a complex scalar field without interactions. Both the field \( \phi(x) \) and its complex conjugate \( \phi^*(x) \) satisfy the Klein–Gordon equation; thus, \( \phi(x) \) describes a particle of mass \( m \) without spin and its antiparticle. Particles which are their own antiparticles (i.e., with no internal charges) have only one degree of freedom and are described through a real scalar field. The appropriate Klein–Gordon Lagrangian is then
\[ L = \frac{1}{2} \partial^\mu \phi \partial_\mu \phi - \frac{1}{2} m^2 \phi^2. \] (A.15)

The Dirac equation can be derived from the Lagrangian density
\[ L = \bar{\psi} (i \gamma^\mu \partial_\mu - m) \psi. \] (A.16)
The adjoint spinor $\bar{\psi}(x) = \psi^\dagger(x) \gamma^0$ closes the Dirac indices. The matrix $\gamma^0$ is included to guarantee the proper behaviour under Lorentz transformations: $\bar{\psi}\psi$ is a Lorentz scalar, while $\bar{\psi} \gamma^\mu \psi$ transforms as a four-vector [94]. Therefore, $\mathcal{L}$ is Lorentz invariant as it should be.

Using the decomposition (A.11) of the Dirac field in its two chiral components, the fermionic Lagrangian adopts the form:

$$\mathcal{L} = \bar{\psi}_L i \gamma^\mu \partial_\mu \psi_L + \bar{\psi}_R i \gamma^\mu \partial_\mu \psi_R - m \left( \bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L \right).$$

(A.17)

Thus, the two chiralities decouple if the fermion is massless.

### A.3 Symmetries and conservation laws

Let us assume that the Lagrangian of a physical system is invariant under some set of continuous transformations

$$\phi_i(x) \rightarrow \phi'_i(x) = \phi_i(x) + \epsilon \delta \phi_i(x) + O(\epsilon^2),$$

(A.18)

i.e., $\mathcal{L}[\phi_i(x), \partial_\mu \phi_i(x)] = \mathcal{L}[\phi'_i(x), \partial_\mu \phi'_i(x)]$. One finds then that

$$\delta \mathcal{L} = 0 = \sum_i \left\{ \left[ \frac{\partial \mathcal{L}}{\partial \phi_i} - \partial^\mu \left( \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi_i)} \right) \right] \delta \phi_i + \partial^\mu \left[ \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi_i)} \delta \phi_i \right] \right\}. \tag{A.19}$$

If the fields satisfy the Euler–Lagrange equations of motion (A.13), the first term is identically zero; therefore the system has a conserved current:

$$J_\mu \equiv \sum_i \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi_i)} \delta \phi_i, \quad \partial_\nu J_\mu = 0. \tag{A.20}$$

This allows us to define a conserved charge

$$Q \equiv \int d^3 x \ J^0. \tag{A.21}$$

The condition $\partial_\mu J_\mu = 0$ guarantees that $dQ = 0$, i.e., that $Q$ is a constant of motion.

This result, known as Noether’s theorem, can be easily extended to general transformations involving also the space–time coordinates. For every continuous symmetry transformation which leaves the Lagrangian invariant, there is a corresponding divergenceless Noether’s current and, therefore, a conserved charge. The selection rules observed in Nature, where there exist several conserved quantities (energy, momentum, angular momentum, electric charge, etc.), correspond to dynamical symmetries of the Lagrangian.

### A.4 Classical electrodynamics

The well-known Maxwell equations,

$$\nabla \cdot \vec{B} = 0, \quad \nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0, \tag{A.22}$$

$$\nabla \cdot \vec{E} = \rho, \quad \nabla \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{J}, \tag{A.23}$$

summarize a large amount of experimental and theoretical work and provide a unified description of the electric and magnetic forces. The first two equations in (A.22) are easily solved, writing the electromagnetic fields in terms of potentials:

$$\vec{E} = -\nabla V - \frac{\partial \vec{A}}{\partial t}, \quad \vec{B} = \nabla \times \vec{A}. \tag{A.24}$$
It is very useful to rewrite these equations in a Lorentz covariant notation. The charge density $\rho$ and the electromagnetic current $\vec{J}$ transform as a four-vector $A^\mu \equiv \left( \rho, \vec{J} \right)$. The same is true for the potentials which combine into $A^\mu \equiv \left( V, \vec{A} \right)$. The relations (A.24) between the potentials and the fields then take a very simple form, which defines the field strength tensor:

$$F^{\mu\nu} \equiv \partial^\mu A^\nu - \partial^\nu A^\mu = \begin{pmatrix} 0 & -E_1 & -E_2 & -E_3 \\ E_1 & 0 & -B_3 & B_2 \\ E_2 & B_3 & 0 & -B_1 \\ E_3 & -B_2 & B_1 & 0 \end{pmatrix}, \quad \tilde{F}^{\mu\nu} \equiv \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma}. \quad (A.25)$$

In terms of the tensor $F^{\mu\nu}$, the covariant form of the Maxwell equations turns out to be very transparent:

$$\partial_\mu F^{\mu\nu} = 0, \quad \partial_\mu F^{\mu\nu} = J^\nu. \quad (A.26)$$

The electromagnetic dynamics is clearly a relativistic phenomenon, but Lorentz invariance was not very explicit in the original formulation of Eqs. (A.22) and (A.23). Once a covariant formulation is adopted, the equations become much simpler. The conservation of the electromagnetic current appears now as a natural compatibility condition:

$$\partial_\nu J^\nu = \partial_\nu \partial_\mu F^{\mu\nu} = 0. \quad (A.27)$$

In terms of potentials, $\partial_\mu \tilde{F}^{\mu\nu}$ is identically zero while $\partial_\mu F^{\mu\nu} = J^\nu$ adopts the form

$$\Box A^\nu - \partial^\nu \left( \partial_\mu A^\mu \right) = J^\nu. \quad (A.28)$$

The same dynamics can be described by many different electromagnetic four-potentials, which give the same field strength tensor $F^{\mu\nu}$. Thus, the Maxwell equations are invariant under gauge transformations:

$$A^\mu \longrightarrow A'^\mu = A^\mu + \partial^\mu \Lambda. \quad (A.29)$$

Taking the Lorentz gauge $\partial_\mu A^\mu = 0$, Eq. (A.28) simplifies to

$$\Box A^\nu = J^\nu. \quad (A.30)$$

In the absence of an external current, i.e., with $J^\mu = 0$, the four components of $A^\mu$ satisfy then a Klein–Gordon equation with $m = 0$. The photon is therefore a massless particle.

The Lorentz condition $\partial_\mu A^\mu = 0$ still allows for a residual gauge invariance under transformations of the type (A.29), with the restriction $\Box \Lambda = 0$. Thus, we can impose a second constraint on the electromagnetic field $A^\mu$, without changing $F^{\mu\nu}$. Since $A^\mu$ contains four fields ($\mu = 0, 1, 2, 3$) and there are two arbitrary constraints, the number of physical degrees of freedom is just two. Therefore, the photon has two different physical polarizations.

**B SU(N) algebra**

$SU(N)$ is the group of $N \times N$ unitary matrices, $UU^\dagger = U^\dagger U = 1$, with $\det U = 1$. Any $SU(N)$ matrix can be written in the form

$$U = \exp \left\{ i T^a \theta_a \right\}, \quad a = 1, 2, \ldots, N^2 - 1, \quad (B.1)$$

with $T^a = \lambda^a / 2$ Hermitian, traceless matrices. Their commutation relations

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

define the $SU(N)$ algebra. The $N \times N$ matrices $\lambda^a / 2$ generate the fundamental representation of the $SU(N)$ algebra. The basis of generators $\lambda^a / 2$ can be chosen so that the structure constants $f^{abc}$ are real and totally antisymmetric.
The Standard Model of Electroweak Interactions

For \( N = 2 \), \( \lambda^a \) are the usual Pauli matrices,

\[
\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},
\]

which satisfy the commutation relation

\[
[\sigma_i, \sigma_j] = 2i \epsilon_{ijk} \sigma_k.
\]

Other useful properties are: \( \{\sigma_i, \sigma_j\} = 2 \delta_{ij} \) and \( \text{Tr}(\sigma_i \sigma_j) = 2 \delta_{ij} \).

For \( N = 3 \), the fundamental representation corresponds to the eight Gell-Mann matrices:

\[
\lambda^1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda^2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda^4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},
\]

\[
\lambda^5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \lambda^6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \lambda^7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda^8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.
\]

They satisfy the anticommutation relation

\[
\left\{ \lambda^a, \lambda^b \right\} = \frac{4}{N} \delta^{ab} I_N + 2 d^{abc} \lambda^c,
\]

where \( I_N \) denotes the \( N \)-dimensional unit matrix and the constants \( d^{abc} \) are totally symmetric in the three indices.

For \( SU(3) \), the only non-zero (up to permutations) \( f^{abc} \) and \( d^{abc} \) constants are

\[
\frac{1}{2} f^{123} = f^{147} = -f^{156} = f^{246} = f^{257} = f^{345} = -f^{367} = \frac{1}{\sqrt{3}} f^{458} = \frac{1}{\sqrt{3}} f^{678} = \frac{1}{2}, \tag{B.7}
\]

\[
d^{146} = d^{157} = d^{247} = d^{256} = d^{344} = d^{355} = -d^{366} = -d^{377} = \frac{1}{2}, \tag{B.8}
\]

\[
d^{118} = d^{228} = d^{338} = -2 d^{148} = -2 d^{558} = -2 d^{668} = -2 d^{778} = -d^{888} = \frac{1}{2}.
\]

The adjoint representation of the \( SU(N) \) group is given by the \( (N^2 - 1) \times (N^2 - 1) \) matrices \( (T_A)^a_{bc} \equiv -i f^{abc} \), which satisfy the commutation relations (B.2). The following equalities

\[
\text{Tr} \left( \lambda^a \lambda^b \right) = 4 T_F \delta_{ab}, \quad \text{Tr} \left( \lambda^a \lambda^b \right) = 4 C_F \delta_{ab},
\]

\[
\text{Tr} (T_A T_B) = f^{acd} f^{bde} = C_A \delta_{ab}, \quad C_A = N,
\]

define the \( SU(N) \) invariants \( T_F, C_F \) and \( C_A \). Other useful properties are

\[
(\lambda^a)^\alpha_\beta (\lambda^b)^\gamma_\delta = 2 \delta_{\alpha \beta \gamma \delta} - \frac{2}{N} \delta_{\alpha \beta} \delta_{\gamma \delta}, \quad \text{Tr} \left( \lambda^a \lambda^b \lambda^c \right) = 2 (d^{abc} + i f^{abc}), \tag{B.10}
\]

\[
\sum_b d^{abb} = 0, \quad d^{abc} d^{deb} = \left( N - \frac{4}{N} \right) \delta_{ae}, \quad d^{fde} f^{bce} = 0.
\]
C Anomalies

Our theoretical framework is based on the local gauge symmetry. However, so far we have discussed only the symmetries of the classical Lagrangian. It happens sometimes that a symmetry of $\mathcal{L}$ gets broken by quantum effects, i.e., it is not a symmetry of the quantized theory; one says then that there is an ‘anomaly’. Anomalies appear in those symmetries involving both axial ($\overline{\psi}\gamma^{\mu}\gamma_5\psi$) and vector ($\overline{\psi}\gamma^{\mu}\psi$) currents, and reflect the impossibility of regularizing the quantum theory (the divergent loops) in a way which preserves the chiral (left/right) symmetries.

A priori there is nothing wrong with having an anomaly. In fact, sometimes they are even welcome. A good example is provided by the decay $\pi^0 \rightarrow \gamma \gamma$. Fortunately, there is an anomaly generated by a triangular quark loop (Fig. C.1) which couples the axial current at the quantum level:

\[
\langle A^3_\mu \rangle = \frac{\alpha}{4\pi} \epsilon^{\alpha\beta\sigma\rho} F_{\alpha\beta} F_{\sigma\rho} + \mathcal{O}(m_u + m_d). \tag{C.1}
\]

Since the $\pi^0$ couples to $A^3_\mu$, $\langle A^3_\mu | \pi^0 \rangle = 2i f_\pi p_\mu$, the $\pi^0 \rightarrow \gamma \gamma$ decay does finally occur, with a predicted rate

\[
\Gamma(\pi^0 \rightarrow \gamma \gamma) = \left(\frac{N_C}{3}\right)^2 \frac{\alpha^2 m_\pi^2}{64\pi^3 f_\pi^2} = 7.73 \text{ eV}, \tag{C.2}
\]

where $N_C = 3$ denotes the number of quark colours and the so-called pion decay constant, $f_\pi = 92.4 \text{ MeV}$, is known from the $\pi^- \rightarrow \mu^- \nu_\mu$ decay rate (assuming isospin symmetry). The agreement with the measured value, $\Gamma = 7.7 \pm 0.6 \text{ eV}$ [7], is excellent.

Anomalies are, however, very dangerous in the case of local gauge symmetries, because they destroy the renormalizability of the Quantum Field Theory. Since the $SU(2)_L \otimes U(1)_Y$ model is chiral (i.e., it distinguishes left from right), anomalies are clearly present. The gauge bosons couple to vector and axial-vector currents; we can then draw triangular diagrams with three arbitrary gauge bosons ($W^\pm$, $Z$, $\gamma$) in the external legs. Any such diagram involving one axial and two vector currents generates a breaking of the gauge symmetry. Thus, our nice model looks meaningless at the quantum level.

We have still one way out. What matters is not the value of a single Feynman diagram, but the sum of all possible contributions. The anomaly generated by the sum of all triangular diagrams connecting the three gauge bosons $G_a$, $G_b$ and $G_c$ is proportional to

\[
\mathcal{A} = \text{Tr} \left( \{ T^a, T^b \} T^c \right)_L - \text{Tr} \left( \{ T^a, T^b \} T^c \right)_R, \tag{C.3}
\]

where the traces sum over all possible left- and right-handed fermions, respectively, running along the internal lines of the triangle. The matrices $T^a$ are the generators associated with the corresponding gauge bosons; in our case, $T^a = \sigma_a/2$, $Y$.

In order to preserve the gauge symmetry, one needs a cancellation of all anomalous contributions, i.e., $\mathcal{A} = 0$. Since $\text{Tr}(\sigma_k) = 0$, we have an automatic cancellation in two combinations of generators: $\text{Tr}(\{\sigma_i, \sigma_j\} \sigma_k) = 2 \delta^{ij} \text{Tr}(\sigma_k) = 0$ and $\text{Tr}(\{Y, Y\} \sigma_k) \propto \text{Tr}(\sigma_k) = 0$. However, the other two
combinations $\text{Tr}(\{\sigma_i, \sigma_j\} Y)$ and $\text{Tr}(Y^3)$ turn out to be proportional to $\text{Tr}(Q)$, i.e., to the sum of fermion electric charges:

$$\sum_i Q_i = Q_e + Q_\nu + N_C (Q_u + Q_d) = -1 + \frac{1}{3} N_C = 0.$$  \hfill (C.4)

Equation (C.4) conveys a very important message: the gauge symmetry of the $SU(2)_L \otimes U(1)_Y$ model does not have any quantum anomaly, provided that $N_C = 3$. Fortunately, this is precisely the right number of colours to understand strong interactions. Thus, at the quantum level, the electroweak model seems to know something about QCD. The complete SM gauge theory based on the group $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ is free of anomalies and, therefore, renormalizable. The anomaly cancellation involves one complete generation of leptons and quarks: $\nu, e, u, d$. The SM would not make any sense with only leptons or quarks.

References


THE STANDARD MODEL OF ELECTROWEAK INTERACTIONS

Monte Carlo generators

T. Sjöstrand
CERN, Geneva, Switzerland, and
Lund University, Sweden

Abstract
The structure of events in high-energy collisions is complex and not predictable from first principles. Event generators allow the problem to be subdivided into more manageable pieces, some of which can be described from first principles, while others need to be based on appropriate models with parameters tuned to data. In these lectures we provide an overview, discuss how matrix elements are used, introduce the machinery for initial- and final-state parton showers, explain how matrix elements and parton showers can be combined for optimal accuracy, introduce the concept of multiple parton–parton interactions, comment briefly on the hadronization issue, and provide an outlook for the future.

1 Introduction
Given the current landscape in experimental high-energy physics, these lectures are focused on applications of event generators for hadron colliders like the Tevatron and the Large Hadron Collider (LHC). Much of the material would also be relevant for $e^+e^-$ machines like the Large Electron–Positron collider (LEP) and the International Linear Collider (ILC) or $e^\pm p$ machines like the Hadron–Electron Ring Accelerator (HERA), but with some differences not discussed here. Heavy-ion physics is not addressed at all since it involves rather different aspects, specifically the potential formation of a quark–gluon plasma. Furthermore, within the field of high-energy $pp/\bar{p}\bar{p}$ collisions, the emphasis will be on the common aspects of QCD physics that occur in all collisions, rather than on those aspects that are specific to a particular physics topic, such as $B$ production or supersymmetry. Heavy ions and other physics topics are instead covered by other lectures at this school.

Section 2 contains a first overview of the physics picture and the generator landscape. Then Section 3 describes the usage of matrix elements, Section 4 the important topics of initial- and final-state showers, and Section 5 shows how showers can be matched to different hard processes. The issue of multiple interactions and their role in minimum-bias and underlying-event physics is introduced in Section 6, followed by some brief comments on hadronization in Section 7. The article concludes with an outlook on the ongoing generator-development work in Section 8.

Slides for these and other similar lectures [1] are complementary to this writeup in style and contents, including many (colour) illustrations absent here. Other useful resources include the Les Houches Guidebook to Monte Carlo Generators for Hadron Collider Physics [2] and a recent review on QCD physics at the Tevatron and LHC [3].

2 Overview
In real life, machines produce events that are stored by the data acquisition system of a detector. In the virtual reality, event generators like HERWIG [4] and PYTHIA [5] play the role of machines like the Tevatron and the LHC, and detector simulation programs like GEANT 4 [6] play the role of detectors like ATLAS or CMS. The real and virtual worlds can share the same event reconstruction framework and subsequent physics analysis. It is by understanding how an original physics input is distorted step by step in the better controlled virtual world that an understanding can be gained of what may be happening in the real world. For approximate studies the detector simulation and reconstruction steps can be short-cut, so that generators can be used directly in the physics studies.
A number of physics analyses would not be feasible without generators. Specifically, a proper understanding of the (potential) signal and background processes is important to separate the two. The key aspect of generators here is that they provide a detailed description of the final state so that, ideally, any experimental observable or combination of observables can be predicted and compared with data. Thereby generators can be used at various stages of an experiment: when optimizing the detector and its trigger design to the intended physics programme, when estimating the feasibility of a specific physics study, when devising analysis strategies, when evaluating acceptance corrections, and so on.

However, it should always be kept in mind that generators are not perfect. They suffer from having to describe a broad range of physics, some of which is known from first principles, while other parts are modelled in different frameworks. (In the latter case, a generator actually acts as a vehicle of ideology, where ideas are disseminated in prepackaged form from theorists to experimentalists.) Given the limited resources, different authors may also have invested more or less time on specific physics topics, and therefore these may be more or less well modelled. It always pays to shop around, and to compare several approaches before drawing too definite conclusions. Blind usage of a generator is not encouraged: then you are the slave rather than the master.

Why then Monte Carlo event generators? Basically because Einstein was wrong: God does throw dice! In quantum mechanics, calculations provide the probability for different outcomes of a measurement. Event by event, it is impossible to know beforehand what will happen: anything that is at all allowed could be next. It is only when averaging over large event samples that the expected probability distributions emerge—provided we did the right calculation to high enough accuracy. In generators, (pseudo)random numbers are used to make choices intended to reproduce the quantum mechanical probabilities for different outcomes at various stages of the process.

The build-up of the structure in an event occurs in several steps, and can be summarized as follows:

- Initially two hadrons are coming in on a collision course. Each hadron can be viewed as a bag of partons—quarks and gluons.
- A collision between two partons, one from each side, gives the hard process of interest, be it for physics within or beyond the Standard Model: \( ug \rightarrow ug, u\overline{d} \rightarrow W^+, gg \rightarrow h^0 \), etc. (Actually, the bulk of the cross-section results in rather mundane events, with at most rather soft jets, or events of a simple elastic or diffractive character that are not easily described as partonic processes. Such events usually are filtered away at an early stage, however.)
- When short-lived ‘resonances’ are produced in the hard process, such as the top, \( W^\pm \) or \( Z^0 \), their decay has to be viewed as part of this process itself, since spin correlations, for example, are transferred from the production to the decay stages.
- A collision implies accelerated colour (and often electromagnetic) charges, and thereby bremsstrahlung can occur. Emissions that can be associated with the two incoming colliding partons are called Initial-State Radiation (ISR). As we shall see, such emissions can be modelled by so-called space-like parton showers.
- Emissions that can be associated with outgoing partons are instead called Final-State Radiation (FSR), and can be approximated to be time-like parton showers. Often the distinction between a hard process and ISR and FSR is ambiguous, as we shall see.
- So far we extracted only one parton from each incoming hadron to undergo a hard collision. But the hadron is made up of a multitude of further partons, and so further parton pairs may collide within one single hadron–hadron collision — multiple interactions (MI). (Not to be confused with pile-up events, when several hadron pairs collide during a bunch–bunch crossing, but with obvious analogies.)
- Each of these further collisions may also be associated with its ISR and FSR.
- The colliding partons take a fraction of the energy of the incoming hadrons, but much of the energy
remains in the beam remnants, which continue to travel essentially in the original directions. These remnants also carry colours that compensate the colour taken away by the colliding partons.

– At short time-scales, when partons are close to each other, the principle of asymptotic freedom tells us that we can think of each parton as freely moving along its trajectory. However, as time goes by and the partons recede from each other, confinement forces become significant. The structure and time evolution of these force fields cannot be described from first principles within any calculational technique currently at our disposal, so models have to be introduced. One common approach is to assume that a separate confinement field is stretched between each colour and its matching anticolour, with each gluon considered as a simple sum of a colour and an anticolour, and all colours distinguishable from each other (the $N_C \to \infty$ limit).

– Such fields can break up by the production of new quark–antiquark pairs that screen the endpoint colours, and where a quark from one break (or from an endpoint) can combine with an antiquark from an adjacent break to produce a primary hadron. This process is called hadronization.

– Many of those primary hadrons are unstable and decay further at various time-scales. Some are sufficiently long-lived that their decays are visible in a detector, or are (almost) stable. Thereby we have reached scales where the event-generator description has to be matched to a detector-simulation framework.

– It is only at this stage that experimental information can be obtained and used to reconstruct what may have happened at the core of the process, e.g., whether a Higgs particle was produced or not.

The Monte Carlo method allows these steps to be considered sequentially, and within each step to define a set of rules that can be used iteratively to construct a more and more complex state, maybe ending with hundreds of particles moving out in different directions. Since each particle contains of the order of ten degrees of freedom (flavour, mass, momentum, production vertex, lifetime, etc.) we realize that thousands of choices are involved for a typical event. The aim is to have a sufficiently realistic description of these choices such that both the average behaviour and the fluctuations around this average are well described.

Schematically, the cross-section for a range of final states is provided by

$$\sigma_{\text{final state}} = \sigma_{\text{hard process}} \mathcal{P}_{\text{tot,hard process} \to \text{final state}},$$

properly integrated over the relevant phase-space regions and summed over possible ‘paths’ (of showering, hadronization, etc.) that lead from a hard process to the final state. That is, the dimensional quantities are associated with the hard process; subsequent steps are handled in a probabilistic approach.

The spectrum of event generators is very broad, from general-purpose ones to more specialized ones. HERWIG and PYTHIA are the two most commonly used among the former ones, with ISAJET [7] and SHERPA [8] as the other two main programs in this category. Among more specialized programs, many deal with the matrix elements for some specific set of processes, a few with topics such as parton showers or particle decays, but there are, for example, no freestanding programs that handle hadronization. In the end, many of the specialized programs are therefore used as ‘plug-ins’ for the general-purpose ones.

### 3 Matrix elements and their usage

The Feynman rules can be derived from the Lagrangian of a theory, and from them matrix elements can be calculated. Combined with phase space it allows the calculation of cross-sections. As a simple example consider the scattering of quarks in QCD, say $u(1) \bar{d}(2) \to u(3) \bar{d}(4)$, a process similar to Rutherford scattering but with gluon exchange instead of photon exchange. The Mandelstam variables are defined as $\hat{s} = (p_1 + p_2)^2$, $\hat{t} = (p_1 - p_3)^2$ and $\hat{u} = (p_1 - p_4)^2$. In the centre-of-mass frame of the collision $\hat{s}$ is the squared total energy and $\hat{t}, \hat{u} = -\hat{s}(1 \mp \cos \hat{\theta})/2$ where $\hat{\theta}$ is the scattering angle. The
The differential cross-section is then

$$\frac{d\sigma}{dt} = \frac{\pi}{s^2} \cdot \frac{4}{9} \cdot \frac{\alpha_s^2}{t^2} \cdot \frac{s^2 + \hat{u}^2}{t^2},$$

which diverges roughly like $dp_{\perp}^2/p_{\perp}^4$ for transverse momentum $p_{\perp} \to 0$. We shall come back to this issue when discussing multiple interactions; for now suffice it to say that some lower cut-off $p_{\perp min}$ needs to be introduced. Similar cross-sections, differing mainly by colour factors, are obtained for $qg \to qg$ and $gg \to gg$. A few further QCD graphs, like $gg \to q\bar{q}$, are less singular and give smaller contributions.

These cross-sections then have to be convoluted with the flux of the incoming partons $i$ and $j$ in the two incoming hadrons $A$ and $B$:

$$\sigma = \sum_{i,j} \iiint dx_1 dx_2 \frac{d\hat{\sigma}_{ij}}{dt} f_i^{(A)}(x_1, Q^2) f_j^{(B)}(x_2, Q^2).$$

The parton density functions (PDFs) of gluons and sea quarks are strongly peaked at small momentum fractions $x_1 \approx E_i/E_A, x_2 \approx E_j/E_B$. This further enhances the peaking of the cross-section at small $p_{\perp}$ values. Nevertheless, with high machine luminosity the jet cross-section can be studied to quite high values.

The cross-section of other processes can be suppressed by two main effects. First, for massive particles the $p_{\perp}$ spectrum is strongly dampened below the respective mass scale, and it is only above it that these precesses have a chance to stand up above the QCD background. Second, the processes may involve electroweak (or other small) couplings rather than strong ones.

In order to address the physics of interest, a large number of processes, both within the Standard Model and in various extensions of it, have to be available in generators. Indeed many can also be found in the general-purpose ones, but by far not enough. Further, processes are often available here only to lowest order, while experimental interest may be in higher orders, with more jets in the final state, either as a signal or as a potential background. So a wide spectrum of matrix-element-centred programs are available [9], some quite specialized and others more generic.

The way these programs can be combined with a general-purpose generator is illustrated in Fig. 1.

Fig. 1: Example of how different programs can be combined in the event-generation chain

The cross-section of other processes can be suppressed by two main effects. First, for massive particles the $p_{\perp}$ spectrum is strongly dampened below the respective mass scale, and it is only above it that these precesses have a chance to stand up above the QCD background. Second, the processes may involve electroweak (or other small) couplings rather than strong ones.

In order to address the physics of interest, a large number of processes, both within the Standard Model and in various extensions of it, have to be available in generators. Indeed many can also be found in the general-purpose ones, but by far not enough. Further, processes are often available here only to lowest order, while experimental interest may be in higher orders, with more jets in the final state, either as a signal or as a potential background. So a wide spectrum of matrix-element-centred programs are available [9], some quite specialized and others more generic.

The way these programs can be combined with a general-purpose generator is illustrated in Fig. 1. In the study of Supersymmetry (SUSY) it is customary to define a model in terms of a handful of parameters, e.g., specified at some large Grand Unification scale. It is then the task of a spectrum calculator
to turn this into a set of masses, mixings, and couplings for the physical states to be searched for. Separately, the matrix elements can be calculated with these properties as unknown parameters, and only when the two are combined is it possible to speak of physically relevant matrix-element expressions. These matrix elements now need to be combined with PDFs and sampled in phase space, preferably with some preweighting procedure so that regions of phase space with high cross-sections are sampled more frequently. The primarily produced SUSY particles typically are unstable and undergo sequential decays down to a lightest supersymmetric particle (LSP), again with branching ratios and angular distributions that should be properly modelled. The LSP would be neutral and escape undetected, while other decay products would be normal quarks and leptons.

It is at this stage that general-purpose programs take over. They describe the showering associated with the above process, the presence of additional interactions in the same hadron–hadron collision, the structure of beam remnants, and the hadronization and decays. They would still rely on the externally supplied PDFs, and potentially make use of programs dedicated to \( \tau \) and \( B \) decays, where spin information and form factors require special encoding. Even after the event has been handed on to the detector-simulation program, some parts of the generator may be used in the simulation of secondary interactions and decays.

Several standards have been developed to further this interoperability. The Les Houches Accord (LHA) for user processes [10] specifies how parton-level information about the hard process and sequential decays can be encoded and passed on to a general-purpose generator. Originally it was defined in terms of two FORTRAN common blocks, but more recently a standard Les Houches Event File format [11] offers a language-independent alternative approach. The Les Houches Accord Parton Density Functions (LHAPDF) library [12] makes different PDF sets available in a uniform framework. The SUSY Les Houches Accord (SLHA) [13] allows a standardized transfer of masses, mixings, couplings and branching ratios from spectrum calculators to other programs. Finally, the HepMC C++ event record [14] succeeds the HEPEVT FORTRAN one [15] as a standard way to transfer information from a generator on to the detector-simulation stage. One of the key building blocks for several of these standards is the PDG codes for all the most common particles [16], also in some scenarios for physics beyond the Standard Model.

The \( 2 \rightarrow 2 \) processes above are about the simplest one can imagine at a hadron collider. In reality one needs to go on to higher orders. In \( \mathcal{O}(\alpha_s^3) \) two new kinds of graphs enter. One kind is where one additional parton is present in the final state, i.e. \( 2 \rightarrow 3 \) processes. The cross-section for such processes is almost always divergent when one of the parton energies vanishes (soft singularities) or two partons become collinear (collinear singularities). The other kind is loop graphs, with an additional intermediate parton not present in the final state, i.e., a correction to the \( 2 \rightarrow 2 \) processes. Strictly speaking, at \( \mathcal{O}(\alpha_s^2) \) one picks up the interference between the lowest-order graph and the loop graph, and this interference has negative divergences that exactly cancel the positive ones above, with only finite terms surviving. For inclusive event properties such next-to-leading-order (NLO) calculations lead to an improved accuracy of predictions, but for more exclusive studies the mathematical cancellation of singularities has to be supplemented by more physical techniques, which is far from trivial.

The tricky part of the calculations is the virtual corrections. NLO is now state-of-the-art, with NNLO still in its infancy. If one is content with Born-level diagrams only, i.e., without any loops, it is possible to go to quite high orders, with up to something like eight partons in the final state. These partons have to be kept well separated to avoid the phase-space regions where the divergences become troublesome. In order to cover also regions where partons become soft/collinear we therefore next turn our attention to parton showers.
4 Parton showers

To iterate, the emission rate for a branching such as $q \rightarrow qg$ diverges when the gluon either becomes collinear with the quark or when the gluon energy vanishes. The QCD pattern is similar to that for $e \rightarrow e\gamma$ in QED, except with a larger coupling, and actually a coupling that also increases for smaller relative $p_\perp$ in a branching, thereby further enhancing the divergence. Furthermore the non-Abelian character of QCD leads to $g \rightarrow gg$ branchings with similar divergences, without any correspondence in QED. The third main branching, $g \rightarrow q\overline{q}$ with its $\gamma \rightarrow e^+e^-$ QED equivalence, does not have the soft divergence and is less important.

Now, if the rate for one emission of a gluon is big, then also the rate for two or more will be big, and thus the need for high orders and many loops in matrix-element-based descriptions. With showers we introduce two new concepts that make like easier:

1) an iterative structure that allows simple expressions for $q \rightarrow qg$, $g \rightarrow gg$ and $g \rightarrow q\overline{q}$ branchings to be combined to build up complex multiparton final states, and

2) a Sudakov factor that offers a physical way to handle the cancellation between real and virtual divergences. Neither of the simplifications is exact, but together they allow us to provide sensible approximate answers for the structure of emissions in soft and collinear regions of phase space.

4.1 The shower approach

The starting point is to ‘factorize’ a complex $2 \rightarrow n$ process, where $n$ represents a large number of partons in the final state, into a simple core process, maybe $2 \rightarrow 2$, convoluted with showers, Fig. 2. To begin with, in a simple $ud \rightarrow ud$ process the incoming and outgoing quarks must be on the mass shell, i.e., satisfy $p^2 = E^2 - p_\perp^2 = m^2_q \sim 0$, at long time-scales. By the uncertainty principle, however, the closer one comes to the hard interaction, i.e, the shorter the time-scales considered, the more off-shell the partons may be. (In the old-fashioned perturbative language, particles were always on mass-shell, and the uncertainty relation allowed energy not to be conserved temporarily. In the modern Feynman-graph language four-momentum is conserved at each vertex, but intermediate ‘propagator’ particles need not be on the mass shell. The final physics is the same in both languages.)

Thus the incoming quarks may radiate a succession of harder and harder gluons, while the outgoing ones radiate softer and softer gluons. One definition of hardness is how off-shell the quarks are, $|p^2| = |E^2 - p^2|$, but we shall encounter other variants later. In the initial-state radiation (ISR) part of the cascade these virtualities are space-like, $p^2 < 0$, hence the alternative name space-like showers.
Correspondingly the final-state radiation (FSR) is characterized by time-like virtualities, $p^2 > 0$, and hence also called time-like showers.

To see where this distinction comes from, consider the kinematics of an arbitrary branching $a \rightarrow bc$, with $a$ defined to be moving along the $+z$ axis. Then it is useful to introduce light-cone momenta $p_\pm = E \pm p_z$, so that the relation $p_+^2 = E^2 - p_y^2 - p_z^2 = m^2$ translates to $p_+ p_- = m^2 + p_x^2 = m^2 + p_\perp^2$. Now define the splitting of $p_+$ by $p_{+b} = z p_+$ and $p_{+c} = (1 - z) p_+$. Obviously $p_\perp = -p_\perp$ so $p_\perp^2 = p_{\perp b}^2 = p_{\perp c}^2$. It remains to ensure conservation of $p_-$ by $(m^2 + p_\perp^2)/p_+:

$$p_{-a} = p_{-b} + p_{-c} \iff \frac{m_a^2}{p_{+a}} = \frac{m_b^2 + p_{\perp b}^2}{zp_{+a}} + \frac{m_c^2 + p_{\perp c}^2}{(1 - z)p_{+a}} \iff m_a^2 = \frac{m_b^2}{z} + \frac{m_c^2}{1 - z} + \frac{p_{\perp}^2}{z(1 - z)}.$$ (4)

In an initial-state branching the incoming $a$ should be (essentially) massless, and if $c$ does not interact any further it should also be massless. This gives $m_b^2 = -(1 - z) p_\perp^2 < 0$, a virtuality that is acceptable if $b$ is on its way in to a hard scattering, i.e., is destined only to live for a short while. In ISR the $Q_i^2$ virtualities, such as $Q_1^2$ and $Q_2^2$ in Fig. 2, are usually defined as $-m_i^2$ to keep them positive definite. For a final-state branching, assume that $b$ and $c$ will not branch any further and thus are massless, while $a$ is an intermediate particle coming from the hard interaction. Then $m_a^2 = p_\perp^2/(z(1 - z)) > 0$, cf. $Q_3^2$ and $Q_4^2$ in Fig. 2, here with $Q_2^2 = +m_i^2$.

The cross-section for the whole $2 \rightarrow n$ graph is associated with the cross-section of the hard subprocess, with the approximation that the other $Q_i^2$ virtualities can be neglected in the matrix-element expression. In the limit that all the $Q_i^2 \ll Q^2$ this should be a good approximation. In other words, first the hard process can be picked without any reference to showers, and only thereafter are showers added with unit probability. But, of course, the showers do modify the event shape, so at the end of the day the cross-section is affected. For instance, the total transverse energy $E_{\perp \text{tot}}$ of an event is increased by ISR, so the cross-sections of events with a given $E_{\perp \text{tot}}$ are increased by the influx of events that started out with a lower $E_{\perp \text{tot}}$ in the hard process.

It is important that the hard-process scale $Q^2$ is picked to be the largest one, i.e., $Q^2 > Q_i^2$ in Fig. 2. If $Q_4^2 > Q_1^2$, for example, then the $ug \rightarrow ug$ subgraph ought to be chosen instead as the hard process, and the gluon of virtuality $Q_4^2$ ought to be part of the ISR off the incoming $d$. Without such a criterion one might double-count a given graph, or even count it once for every possible subgraph inside the complete $2 \rightarrow n$ graph. In addition, the approximation of neglecting virtualities in the hard-scattering matrix elements obviously becomes worse the more the incoming and outgoing partons are off-shell, another reason not to put a larger scale than necessary in the shower part.

4.2 Final-state radiation

Let us next turn to a more detailed presentation of the showering approach, and begin with the simpler final-state stage. This is most cleanly studied in the process $e^+ e^- \rightarrow \gamma^*/Z^0 \rightarrow q\bar{q}$. The first-order correction here corresponds to the emission of one additional gluon, by either of the two Feynman graphs

---

**Fig. 3:** The two Feynman graphs that contribute to $\gamma^*/Z^0(0) \rightarrow q(1)\bar{q}(2)g(3)$
in Fig. 3. Neglect quark masses and introduce energy fractions \( x_j = 2E_j/E_{\text{cm}} \) in the rest frame of the process. Then the cross-section is of the form

\[
\frac{d\sigma_{\text{ME}}}{\sigma_0} = \frac{\alpha_s}{2\pi} \frac{4}{3} \left( \frac{x_1^2 + x_2^2}{(1 - x_1)(1 - x_2)} \right) \, dx_1 \, dx_2 ,
\]

where \( \sigma_0 \) is the \( q\bar{q} \) cross-section, i.e., without the gluon emission.

Now study the kinematics in the limit \( x_2 \to 1 \). Since \( 1 - x_2 = m_{13}^2/E_{\text{cm}}^2 \) we see that this corresponds to the ‘collinear region’, where the separation between the quark and gluon vanishes. Equivalently, the virtuality \( Q^2 = Q_i^2 = m_{13}^2 \) of the intermediate quark propagator \( i \) in Fig. 3(a) vanishes. Although the full answer contains contributions from both graphs it is obvious that, in this region, the amplitude of the one in Fig. 3(a) dominates over the one in Fig. 3(b). We can therefore view the process as \( \gamma^*/Z^0 \to q\bar{q} \) followed by \( q \to qg \). Define the energy sharing in the latter branching by \( E_q = zE_i \) and \( E_g = (1 - z)E_i \).

The kinematics then are

\[
1 - x_2 = \frac{m_{13}^2}{E_{\text{cm}}^2} = \frac{Q_i^2}{E_{\text{cm}}^2} \implies dx_2 = \frac{dQ_i^2}{E_{\text{cm}}^2}
\]

\[
x_1 \approx z \implies dx_1 \approx dz
\]

\[
x_3 \approx 1 - z
\]

so that

\[
dP = \frac{d\sigma_{\text{ME}}}{\sigma_0} = \frac{\alpha_s}{2\pi} \left( \frac{dx_2 \cdot 4}{1 - x_2} \frac{x_2^2 + x_1^2}{3 (1 - x_1)} \right) \, dx_1 \approx \frac{\alpha_s}{2\pi} \frac{dQ_i^2}{Q_i^2} \left( \frac{4}{3} \frac{1 + z^2}{1 - z} \right) \, dz.
\]

Here \( dQ_i^2/Q_i^2 \) corresponds to the ‘collinear’ or ‘mass’ singularity and \( dz/(1 - z) = dE_g/E_g \) to the soft-gluon singularity.

The interesting aspect of Eq. (9) is that it is universal: whenever there is a massless quark in the final state, this equation provides the probability for the same final state except for the quark being replaced by an almost collinear \( qg \) pair (plus some other slight kinematics adjustments to conserve overall energy and momentum). That is reasonable: in a general process any number of distinct Feynman graphs may contribute and interfere in a nontrivial manner, but once we go to a collinear region only one specific graph will contribute, and that graph always has the same structure, in this case with an intermediate quark propagator. Corresponding rules can be derived for what happens when a gluon is replaced by a collinear \( gg \) or \( q\bar{q} \) pair. These rules are summarized by the DGLAP equations [17]

\[
dP_{a \to bc} = \frac{\alpha_s}{2\pi} \frac{dQ_i^2}{Q_i^2} \, P_{a \to bc}(z) \, dz.
\]

where

\[
P_{q \to qq} = 4 \frac{1 + z^2}{3 (1 - z)} ,
\]

\[
P_{g \to gg} = 3 \frac{(1 - z(1 - z))^2}{z^2} ,
\]

\[
P_{g \to q\bar{q}} = \frac{n_f}{2} \left( z^2 + (1 - z)^2 \right) \quad (n_f = \text{no. of quark flavours}).
\]

Furthermore, the rules can be combined to allow for the successive emission in several steps, e.g., where a \( q \to qg \) branching is followed by further branchings of the daughters. That way a whole shower develops, Fig. 4.

Such a picture should be reliable in cases where the emissions are strongly ordered, i.e., \( Q_1^2 \gg Q_2^2 \gg Q_3^2 \ldots \). Showers would not be useful if they could be applied only to strongly-ordered parton configurations, however. A further study of the \( \gamma^*/Z^0 \to q\bar{q}g \) example shows that the simple sum of the \( q \to qg \) and \( \gamma \to g\bar{q}g \) branchings reproduces the full matrix elements, with interference included, to better than a factor of 2 over the full phase space. This is one of the simpler cases, and of course one
should expect the accuracy to be worse for more complicated final states. Nevertheless, it is meaningful
to use the shower over the whole strictly-ordered, but not necessarily strongly-ordered, region $Q_2^1 >
Q_2^2 > Q_2^3 \ldots$ to obtain an approximate answer for multiparton topologies for which the complete matrix
elements would be too lengthy.

With the parton-shower approach, the big probability for one branching $q \to qg$ turns into a big
probability for several successive branchings. Nevertheless we did not tame the fact that probabilities
blow up in the soft and collinear regions. For sure, perturbation theory will cease to be meaningful at
$Q^2$ scales so small that $\alpha_s(Q^2)$ diverges; there confinement effects and hadronization phenomena take
over. Typically therefore some lower cut-off at around 1 GeV is used to regulate both soft and collinear
divergences: below such a scale no further branchings are simulated. Whatever perturbative effects may
remain are effectively pushed into the parameters of the nonperturbative framework. That way we avoid
the singularities, but we can still have ‘probabilities’ well above unity, which does not seem to make
sense.

This brings us to the second big concept of this section, the Sudakov (form) factor [18]. In the
context of particle physics it has a specific meaning related to the properties of the loop diagrams, but
more generally we can just see it as a consequence of the conservation of total probability

\[ \mathcal{P}(\text{nothing happens}) = 1 - \mathcal{P}(\text{something happens}) , \]

where the former is multiplicative in a time-evolution sense:

\[ \mathcal{P}_{\text{nothing}}(0 < t \leq T) = \mathcal{P}_{\text{nothing}}(0 < t \leq T_1) \mathcal{P}_{\text{nothing}}(T_1 < t \leq T) . \]

Now subdivide further, with $T_i = (i/n)T$, $0 \leq i \leq n$:

\[ \mathcal{P}_{\text{nothing}}(0 < t \leq T) = \lim_{n \to \infty} \prod_{i=0}^{n-1} \mathcal{P}_{\text{nothing}}(T_i < t \leq T_{i+1}) \]

\[ = \lim_{n \to \infty} \prod_{i=0}^{n-1} (1 - \mathcal{P}_{\text{something}}(T_i < t \leq T_{i+1})) \]

\[ = \exp \left( - \lim_{n \to \infty} \sum_{i=0}^{n-1} \mathcal{P}_{\text{something}}(T_i < t \leq T_{i+1}) \right) \]

\[ = \exp \left( - \int_0^T dt \mathcal{P}_{\text{something}}(t) \right) \]
That is, the probability for something to happen for the first time at time $T$ is the naive probability for this to happen, times the probability that this did not yet happen. As such it applies to a host of situations. Take the example of football (relevant at the time of the School). Assume that players are equally energetic and skillful from the first minute of the match to the last. Then the chance of scoring a goal is uniform in time, but the probability of scoring the first goal of the match is bigger at the beginning, because later on any goal could well be the second or third.

In physics a common example is that of radioactive decay. If the number of undecayed radioactive nuclei at time $t$ is $N(t)$, with initial number $N_0$ at time $t = 0$, then a naive ansatz would be $dN/dt = -cN_0$, where $c$ parametrizes the decay likelihood per unit of time. This equation has the solution $N(t) = N_0(1 - ct)$, which becomes negative for $t > 1/c$, because by then the probability for having had a decay exceeds unity. So what we did wrong was not to take into account that only an undecayed nucleus can decay, i.e., that the equation ought to have been $dN/dt = -cN(t)$ with the solution $N(t) = N_0 \exp(-ct)$. This is a nicely well-behaved expression, where the total probability for decays goes to unity only for $t \rightarrow \infty$. If $c$ had not been a constant but varied in time, $c = c(t)$, it is simple to show that the solution instead would have become

$$N(t) = N_0 \exp \left( - \int_0^t c(t') dt' \right) \quad \Rightarrow \quad \frac{dN}{dt} = -c(t)N_0 \exp \left( - \int_0^t c(t') dt' \right). \quad (17)$$

For a shower the relevant ‘time’ scale is something like $1/Q$, by the Heisenberg uncertainty principle. That is, instead of evolving to later and later times we evolve to smaller and smaller $Q^2$. Thereby the DGLAP Eq. (10) becomes

$$dP_{a \rightarrow bc} = \frac{\alpha_s}{2\pi} \frac{dQ^2}{Q^2} P_{a \rightarrow bc}(z) \, dz \, \exp \left( -\sum_{b,c} \int_{Q^2}^{Q_{\text{max}}^2} \frac{dQ'^2}{Q'^2} \int_{Q^2}^{Q_{\text{max}}^2} \frac{\alpha_s}{2\pi} P_{a \rightarrow bc}(z') \, dz' \right), \quad (18)$$

where the exponent (or simple variants thereof) is the Sudakov factor. As for the radioactive-decay example above, the inclusion of a Sudakov ensures that the total probability for a parton to branch never exceeds unity. Then you may have sequential radioactive decay chains, and you may have sequential parton branchings, but that is another story.

It is a bit deeper than that, however. Just as the standard branching expressions can be viewed as approximations to the complete matrix elements for real emission, the Sudakov is an approximation to the complete virtual corrections from loop graphs. The divergences in real and virtual emissions, so strange-looking in the matrix-element language, here naturally combine to provide a physical answer everywhere. What is not described in the shower, of course, is the non-universal finite parts of the real and virtual matrix elements.

The implementation of a cascade evolution now makes sense. Starting from a simple $q\bar{q}$ system the $q$ and $\bar{q}$ are individually evolved downwards from some initial $Q_{\text{max}}^2$ until they branch. At a branching the mother parton disappears and is replaced by two daughter partons, which in their turn are evolved downwards in $Q^2$ and may branch. Thereby the number of partons increases, until the lower cut-off scale is reached.

This does not mean that everything is uniquely specified. In particular, the choice of evolving in $Q^2 = m^2$ is by no means obvious. Any alternative variable $P^2 = f(z) Q^2$ would work equally well, since $dP^2/P^2 = dQ^2/Q^2$. Alternative evolution variables therefore include the transverse momentum, $\hat{p}_T^2 \approx z(1 - z)m^2$, and the energy-weighted emission angle $E^2\theta^2 \approx m^2/(z(1 - z))$.

Both these two alternative choices are favourable when the issue of coherence is introduced. Coherence means that emissions do not occur independently. For instance, consider $g_1 \rightarrow g_2 g_3$, followed
by an emission of a gluon either from 2 or 3. When this gluon is soft it cannot resolve the individual
colour charges of $g_2$ and $g_3$, but only the net charge of the two, which of course is the charge of $g_1$.
Thereby the multiplication of partons in a shower is reduced relative to naive expectations. As it turns
out, evolution in $p_\perp$ or angle automatically includes this reduction, while one in mass does not.

In the study of FSR, e.g., at LEP, three algorithms have been commonly used. The HERWIG
angular-ordered and PYTHIA mass-ordered ones are conventional parton showers as described above,
while the ARIADNE [19] $p_\perp$-ordered one is based on a picture of dipole emissions. That is, instead of
considering $a \rightarrow b c$ one studies $a b \rightarrow c d e$. One aspect of this is that, in addition to the branching parton,
ARIADNE also explicitly includes a ‘recoil parton’ needed for overall energy–momentum conservation.
Additionally emissions off $a$ and $b$ are combined in a well-defined manner.

All three approaches have advantages and disadvantages. As already mentioned, PYTHIA does
not inherently include coherence, but has to add that approximately by brute force. Both PYTHIA and
HERWIG break Lorentz invariance slightly. The HERWIG algorithm cannot cover the full phase space
with it emissions, but has to fill in some ‘dead zones’ using higher-order matrix elements. The ARIADNE
dipole picture does not include $g \rightarrow q\bar{q}$ branchings in a natural way. And so on.

When all is said and done, it turns out that all three algorithms do quite a decent job of describing
LEP data [20], but typically ARIADNE does best and HERWIG worst. Since ARIADNE uses PYTHIA for
hadronization the difference between those two is entirely due to the shower algorithms, while compar-
isons with HERWIG also are complicated by significant differences in hadronization.

4.3 Initial-state radiation

The structure of initial-state radiation (ISR) is more complicated than that of FSR, since the nontrivial
structure of the incoming hadrons enters the game. A proton is made up of three quarks, $uud$, plus the
gluons that bind them together. This picture is not static, however: gluons are continuously emitted and
absorbed by the quarks, and each gluon may in its turn temporarily split into two gluons or into a $q\bar{q}$
pair. Thus a proton is teeming with activity, and much of it in a nonperturbative region where we cannot
calculate. We are therefore forced to introduce the concept of a parton density $f_b(x, Q^2)$ as an empirical
distribution, describing the probability to find a parton of species $b$ in a hadron, with a fraction $x$ of the
hadron energy–momentum when the hadron is probed at a resolution scale $Q^2$.

While $f_b(x, Q^2)$ itself cannot be predicted, the change of $f_b$ with resolution scale can, once $Q^2$ is
large enough that perturbation theory should be applicable:

$$\frac{d f_b(x, Q^2)}{d (\ln Q^2)} = \sum_a \int_x^1 \frac{dz}{z} f_a(x', Q^2) \frac{\alpha_s}{2\pi} P_{a \rightarrow bc} \left( \frac{z}{x'} \right).$$

(19)

This is actually nothing but our familiar DGLAP equations. Before they were written in an exclusive
manner: given a parton $a$, what is the probability that it will branch to $b c$ during a change $d Q^2$? Here
the formulation is instead inclusive: given that the probability distributions $f_a(x, Q^2)$ of all partons $a$ are
known at a scale $Q^2$, how is the distribution of partons $b$ changed by the set of possible branchings $a \rightarrow b$
($+c$, here implicit)? The splitting kernels $P_{a \rightarrow bc}(z)$ are the same to leading order, but differ between ISR
and FSR in higher orders. In higher orders also the concept of $f_b(x, Q^2)$ as a positive definite probability is
lost, additional complications that we shall not consider any further here.

Even though Eqs. (10) and (19) are equivalent, the physics context is different. In FSR the outgoing
partons have been kicked to large time-like virtualities by the hard process and then cascade downwards
towards the mass shell. In ISR we rather start out with a simple proton at early times and then allow
more and more space-like virtualities as we get closer to the hard interaction. Not that big fluctuations
could not happen at early times — they do — but if they happen too early the uncertainty relation does
not allow them to live long enough to be of any interest to us. That is, the higher the virtuality, the later
the fluctuation has to occur.
Fig. 5: A cascade of successive branchings. The thick line represents the main chain of space-like partons leading in to the hard interaction (marked by a cross). The thin lines are partons that cannot be recombined, while dashed lines are further fluctuations that may (if space-like) or may not (if time-like) recombine. In this graph lines can represent both quarks and gluons.

So, when the hard scattering occurs, in some sense the initial-state cascade is already there, as a virtual fluctuation. Had no collision occurred the fluctuation would have collapsed back, but now one of the partons of the fluctuation is kicked out in a quite different direction and can no longer recombine with its sister parton from its last branching, nor with its aunt from the last-but-one branching. And so on for each preceding branching in the cascade that lead up to this particular parton. Post facto we therefore see that a chain of branchings with increasing $Q^2$ values built up an ISR shower, Fig. 5.

The obvious way to simulate this situation would be to pick partons in the two incoming hadrons from parton densities at some low $Q^2$ scale, and then use the exclusive formulation of Eq. (10) to construct a complete picture of partons available at higher $Q^2$ scales, event by event. The two sets of incoming partons could then be weighted by the cross-section for the process under study. A problem is that this may not be very efficient. We have to evolve for all possible fluctuations, but at best one particular parton will collide and most of the other fluctuations will collapse back. The cost may become prohibitive when the process of interest has a constrained phase space, like a light-mass Higgs which has to have the colliding partons matched up in a very narrow mass bin.

There are ways to speed up this ‘forwards evolution’ approach. However, the most common solution is instead to adopt a ‘backwards evolution’ point of view [21]. Here one starts at the hard interaction and then tries to reconstruct what happened ‘before’. To be more precise, the cross-section formula in Eq. (3) already includes the summation over all possible incoming shower histories by the usage of $Q^2$-dependent parton densities. Therefore what remains is to pick one exclusive shower history from the inclusive set that went into the $Q^2$ evolution. To do this, recast Eq. (19) as

$$dP_b = \frac{df_b}{f_b} = |d(\ln Q^2)| \sum_a \int dz \frac{x' f_a(x', t)}{x f_b(x, t)} \frac{\alpha_s}{2\pi} P_{a\to bc} \left( z = \frac{x}{x'} \right). \quad (20)$$

Then we have defined a conditional probability: if parton $b$ is present at scale $Q^2$, what is the probability that it will turn out to have come from a branching $a \to bc$ at some infinitesimally smaller scale? (Recall that the original Eq. (19) was defined for increasing virtuality.) Like for FSR this expression has to be modified by a Sudakov factor to preserve total probability, and this factor is again the exponent of the real-emission expression with a negative sign, integrated over $Q^2$ from an upper starting scale $Q^2_{\text{max}}$ down to the $Q^2$ of the hypothetical branching.

The approach is now clear. First a hard scattering is selected, making use of the $Q^2$-evolved parton densities. Then, with the hard process as upper maximum scale, a succession of ISR branchings are reconstructed at lower and lower $Q^2$ scales, going ‘backwards in time’ towards the early low-virtuality
initiators of the cascades. Again some cut-off needs to be introduced when the nonperturbative regime is reached.

Unfortunately the story does not end there. For FSR we discussed the need to take into account coherence effects and the possibility to use different variables. Such issues exist here as well, but also additional ones. For instance, evolution need not be strictly ordered in $Q^2$. [Equation (4) only gives you $Q^2_b > zQ^2_a$, not $Q^2_b > Q^2_a$.] Non-ordered chains in some cases can be important. Another issue is that there can be so many partons evolving inside a hadron that they become close-packed, which leads to additional recombinations. See the lectures of G. Ingelman for further details [22].

In summary, ISR and FSR share many aspects, but also differ. The DGLAP evolution with Sudakov factors allows a simple probabilistic framework, where an initial parton undergoes successive branchings. For ISR the branching (usually) is in terms of higher and higher space-like virtualities as the hard scattering approaches, while for FSR the branchings involve lower and lower time-like virtualities as the hard scattering recedes. In FSR both daughter partons appear on equal footing, in that both can be time-like and branch further. In ISR only the daughter parton on its way in to the hard scattering can be space-like virtual; its sister will become part of the final state and thus has to be on mass shell, or else be time-like and start an FSR shower of its own. And, last but not least, the whole FSR framework is considerably better understood than the ISR one.

5 Combining matrix elements and parton showers

As we have seen, both matrix elements (ME) and parton showers (PS) have advantages and disadvantages.

To recapitulate, ME allow a systematic expansion in powers of $\alpha_s$, and thereby offer a controlled approach towards higher precision. Calculations can be done with several (up to $\sim 8$) partons in the final state, so long as only Born-level results are asked for, and it is possible to tailor the phase-space cuts for these partons precisely to the experimental needs. Loop calculations are much more difficult, on the other hand, and the mathematically correct cancellation between real- and virtual-emission graphs in the soft/collinear regions is not physically sensible. Therefore ME cannot be used to explore the internal structure of a jet, and are difficult to match to hadronization models, which are supposed to take over in the very soft/collinear region.

PS, on the other hand, clearly are approximate and do not come with a guaranteed level of precision for well-separated jets. You cannot steer the probabilistic evolution of a shower too much, and therefore the efficiency for obtaining events in a specific region of phase space can be quite low. On the other hand, PS are universal, so for any new model you only need to provide the basic hard process and then PS will turn that into reasonably realistic multiparton topologies. The use of Sudakov factors ensures a physically sensible behaviour in the soft/collinear regions, and it is also here that the PS formalism is supposed to be most reliable. It is therefore possible to obtain a good picture of the internal structure of jets, and to provide a good match to hadronization models.

In a nutshell: ME are good for well-separated jets, PS for the structure inside jets. Clearly the two complement each other, and a marriage is highly desirable. To do this, without double-counting or gaps in the phase space coverage, is less trivial, and several alternative approaches have been developed. In the following we shall discuss three main options: merging, vetoed parton showers, and MC@NLO, roughly ordered in increasing complexity. Which of these to use may well depend on the task at hand.

5.1 Merging

The aspiration of merging is to cover the whole phase space with a smooth transition from ME to PS. The typical case would be a process where the lowest-order (LO) ME is known, as well as the next-to-
leading-order (NLO) real-emission one, say of an additional gluon. The shower should then reproduce

$$W_{\text{ME}} = \frac{1}{\sigma(\text{LO})} \frac{d\sigma(\text{LO} + g)}{d\text{(phasespace)}}$$

starting from a LO topology. If the shower populates phase space according to $W_{\text{PS}}$ this implies that a correction factor $W_{\text{ME}} / W_{\text{PS}}$ needs to be applied.

At first glance this does not appear to make sense: if all we do is get back $W_{\text{ME}}$, then what did we gain? However, the trick is to recall that the PS formula comes in two parts: the real-emission answer and a Sudakov factor that ensures total conservation of probability. What we have called $W_{\text{PS}}$ above should only be the real-emission part of the story. It is also this one that we know will agree with $W_{\text{ME}}$ in the soft and collinear regions. Actually, with some moderate amount of effort it is often possible to ensure that $W_{\text{ME}} / W_{\text{PS}}$ is of order unity over the whole phase space, and to adjust the showers in the hard region so that the ratio is always below unity, i.e., so that standard Monte Carlo rejection techniques can be used. What the Sudakov factor then does is introduce some ordering variable $Q^2$, so that the whole phase space is covered starting from ‘hard’ emissions and moving to ‘softer’ ones. At the end of the day this leads to a distribution over phase space like

$$W_{\text{actual}}(Q^2) = W_{\text{ME}}(Q^2) \exp \left( - \int_{Q^2}^{Q_{\text{max}}^2} W_{\text{ME}}(Q'^2) dQ'^2 \right).$$

That is, we have used the PS choice of evolution variable to provide an exponentiated version of the ME answer. As such it agrees with the ME answer in the hard region, where the Sudakov factor is close to unity, and with the PS in the soft/collinear regions, where $W_{\text{ME}} \approx W_{\text{PS}}$.

The method is especially convenient for resonance decays, such as $e^+ e^- \rightarrow \gamma^*/Z^0 \rightarrow q\bar{q}$ where it was first introduced [23]. In that case there is an added bonus: the full NLO answer, with virtual corrections included, is known to be $\sigma^{\text{NLO}} = \sigma^{\text{LO}} (1 + \alpha_s/\pi)$. So it is trivial to use the procedure above and rescale everything by $(1 + \alpha_s/\pi)$ to obtain a complete NLO answer. Note that the difference between using $\sigma(\text{LO})$ or $\sigma(\text{NLO})$ in the denominator of Eq. (21) only gives a difference to $O(\alpha_s^2)$, i.e., to NNLO.

In Pythia this approach is used for essentially all resonance decays in the Standard Model and minimal supersymmetric extensions thereof: $t \rightarrow b W^+, W^+ \rightarrow u\bar{d}$, $H \rightarrow b\bar{b}$, $\chi^0 \rightarrow q\bar{q}$, $\tilde{q} \rightarrow q\tilde{g}$, etc. [24]. It is also used in ISR to describe $q\bar{q} \rightarrow \gamma^*/Z^0 / W^\pm$ [25], but here the NLO corrections are more tricky, so the cross-section remains as provided by the LO number.

Merging is also used for several processes in Herwig, such as $\gamma^*/Z^0 \rightarrow q\bar{q}$, $t \rightarrow bW^+$ and $q\bar{q} \rightarrow \gamma^*/Z^0 / W^\pm$ [26]. A special problem here is that the angular-ordered algorithms, both for FSR and for ISR, leave some ‘dead zones’ of hard emissions that are kinematically forbidden for the shower to populate. It is therefore necessary to start directly from higher-order matrix elements in these regions. A consistent treatment still allows a smooth joining across the boundary.

### 5.2 Vetoed parton showers

In some sense vetoed parton showers are an extension of the merging approach above. The objective is still to combine the real-emission behaviour of ME with the emission-ordering-variable-dependent Sudakov factors of PS. While the merging approach only works for combining the LO and NLO expressions, however, the vetoed parton showers offer a generic approach for combining several different orders. Therefore it is likely to be a standard tool for many studies in the future.

To understand how the algorithm works, consider a lowest-order process such as $q\bar{q} \rightarrow W^\pm$. For each higher order one additional jet would be added to the final state, so long as only real-emission graphs are considered: in first order $q\bar{q} \rightarrow W^\pm g$, for example, in second order $q\bar{q} \rightarrow W^\pm gg$, for example, and so on. Call these (differential) cross-sections $\sigma_0$, $\sigma_1$, $\sigma_2$, . . . . It should then come as no surprise that
each $\sigma_i, i \geq 1$, contains soft and collinear divergences. We therefore need to impose some set of ME phase-space cuts, e.g., on invariant masses of parton pairs, or on parton energies and angular separation between them. When these cuts are varied, so that the mass or energy thresholds, for example, are lowered towards zero, all of these $\sigma_i, i \geq 1$, increase without bounds.

However, in the ME approach without virtual corrections there is no ‘detailed balance’, wherein the addition of cross-section to $\sigma_{i+1}$ is compensated by a depletion of $\sigma_i$. That is, if you have an event with $i$ jets at some resolution scale, and a lowering of the minimal jet energy reveals the presence of one additional jet, then you should reclassify the event from being $i$-jet to being $i + 1$-jet. Add one, subtract one, with no net change in $\sum_i \sigma_i$. So the trick is to use the Sudakovs of showers to ensure this detailed balance. Of course, in a complete description the cancellation between real and virtual corrections is not completely exact but leaves a finite net contribution, which is not predicted in this approach.

A few alternative algorithms exist along these lines. All share the three first steps as follows [27]:

1) Pick a hard process within the ME-cuts-allowed phase-space region, in proportions provided by the ME integrated over the respective allowed region, $\sigma_0 : \sigma_1 : \sigma_2 : \ldots$. Use for this purpose a $\alpha_s$ larger than the $\alpha_s$ values that will be used below.

2) Reconstruct an imagined shower history that describes how the event could have evolved from the lowest-order process to the actual final state. That provides an ordering of emissions by whatever shower-evolution variable is intended.

3) The ‘best-bet’ choice of $\alpha_s$ scale in showers is known to be the squared transverse momentum of the respective branching. Therefore a factor $W_\alpha = \prod_{\text{branchings}} \left( \alpha_s(\mathbf{p}_T^2)/\alpha_{s0} \right)$, provides the probability that the event should be retained.

Now the algorithms part ways. In the CKKW–L approach the subsequent steps are:

4) Evaluate Sudakov factors for all the ‘propagator’ lines in the shower history reconstructed in step 2, i.e. for intermediate partons that split into further partons, and also for the evolution of the final partons down to the ME cuts without any further emissions. This provides an acceptance weight $W_{\text{Sud}} = \prod_{\text{propagators}} \text{Sudakov}(Q^2_{\text{beg}}, Q^2_{\text{end}})$ where $Q^2_{\text{beg}}$ is the large scale where a parton is produced by a branching and $Q^2_{\text{end}}$ is either the scale at which the parton branches or the ME cuts, as may be the case.

4a) In the CKKW approach [28] the Sudakovs are evaluated by analytical formulae, which is fast.

4b) In the L approach [29] trial showers are used to evaluate Sudakovs, which is slower but allows a more precise modelling of kinematics and phase space than offered by the analytic expression.

5) Now the matrix-element configuration can be evolved further, to provide additional jets below the ME cuts used. In order to avoid double-counting of emissions, any branchings that might occur above the ME cuts must be vetoed.

The MLM approach [27] is rather different. Here the steps instead are:

4') Allow a complete parton shower to develop from the selected parton configuration.

5') Cluster these partons back into a set of jets, e.g., using a cone-jet algorithm, with the same jet-separation criteria as used when the original parton configuration was picked.

6') Try to match each jet to its nearest original parton.

7') Accept the event only if the number of clustered jets agrees with the number of original partons, and if each original parton is sensibly matched to its jet. This would not be the case if, for example, one parton gave rise to two jets, or two partons to one jet, or an original $b$ quark migrated outside of the clustered jet. The point of the MLM approach is that the probability of not generating any additional fatal jet activity during the shower evolution is provided by the Sudakovs used in step 4.
Fig. 6: MC@NLO applied to a $Z^0$ production. The region between the two curves is considered as ‘true’ $Z + 1$ jet events, with showers added to that. The rest — LO real, NLO virtual and NLO real in the shower approximation — are combined to cancel singularities and then showered as simple $Z$ events.

5.3 MC@NLO

MC@NLO [30] in some respects is the most ambitious approach: it aims to get not only real but also virtual contributions correctly included, so that cross-sections are accurate to NLO, and that NLO results are obtained for all observables when formally expanded in powers of $\alpha_s$. Thus hard emissions should again be generated according to ME, while soft and collinear ones should fall within the PS regime.

In simplified terms, the scheme works as follows:

1) Calculate the NLO ME corrections to an $n$-body process, including $n + 1$-body real corrections and $n$-body virtual ones.
2) Calculate analytically how a first branching in a shower starting from a $n$-body topology would populate $n + 1$-body phase space, excluding the Sudakov factor.
3) Subtract the shower expression from the $n + 1$ ME one to obtain the ‘true’ $n + 1$ events, and consider the rest as belonging to the $n$-body class. The PS and ME expressions agree in the soft and collinear limits, so the singularities there cancel, leaving finite cross-sections both for the $n$- and $n + 1$-body event classes.
4) Now add showers to both kinds of events.

A toy example, for the case of $Z^0$ production, is shown in Fig. 6. Several more complicated processes have been considered, such as $b\bar{b}$, $t\bar{t}$ and $W^+W^-$ production. A technical problem is that, although ME and PS converge in the collinear region, it is not guaranteed that ME is everywhere above PS. This is solved by having a small fraction of events with negative weights.

In summary, MC@NLO is superior in that it does provide the total cross-section for a process to NLO accuracy, and so is essential for a set of precision tests. The real-emission $n + 1$-body part is the same as used for merging, however, so for normalized event shapes the merging approach is as valid. (Any differences are of higher order.) Finally, if multijet topologies need to be studied, where several orders may contribute, vetoed showers are more appropriate. To each tool its task.
6 Multiple interactions

The cross-section for $2 \rightarrow 2$ QCD parton processes is dominated by $t$-channel gluon exchange, as we already mentioned, and thus diverges like $d\sigma^2_t / dp_t^2$ for $p_t \rightarrow 0$. Introduce a lower cut $p_{\perp \text{min}}$ and integrate the interaction cross-section above this, properly convoluted with parton densities. At LHC energies this $\sigma_{\text{int}}(p_{\perp \text{min}})$ reaches around 100 mb for $p_{\perp \text{min}} = 5$ GeV, and 1000 mb at around 2 GeV. Since each interaction gives two jets to lowest order, the jet cross-section is twice as big. This should be compared with an expected total cross-section of the order of 100 mb. (QCD is a confining theory and thus provides a finite total cross-section, unlike QED, where infinitely small scattering angles are allowed at infinitely large distances.) In addition, at least a third of the total cross-section is related to elastic scattering $pp \rightarrow pp$ and low-mass diffractive states $pp \rightarrow pX$ that could not contain jets.

So can it really make sense that $\sigma_{\text{int}}(p_{\perp \text{min}}) > \sigma_{\text{tot}}$? Yes, it can! The point is that each incoming hadron is a bunch of partons. You can have several (more or less) independent parton–parton interactions when these two bunches pass through each other. And the point is that an event with $n$ interactions above $p_{\perp \text{min}}$ counts once for the total cross-section but once for each interaction when the interaction rate is calculated. That is,

$$\sigma_{\text{tot}} = \sum_{n=0}^{\infty} \sigma_n \quad \text{while} \quad \sigma_{\text{int}} = \sum_{n=0}^{\infty} n \sigma_n,$$

where $\sigma_n$ is the cross-section for events with $n$ interactions. Thus $\sigma_{\text{int}} > \sigma_{\text{tot}}$ is equivalent to $\langle n \rangle > 1$, each event on the average contains more than one interaction. Furthermore, if interactions really do occur independently when two hadron pass by each other, then one would expect a Poisson distribution, $P_n = (\langle n \rangle)^n \exp(-\langle n \rangle)/n!$, so that several interactions could occur occasionally also when $\sigma_{\text{int}}(p_{\perp \text{min}}) < \sigma_{\text{tot}}$, e.g., for a larger $p_{\perp \text{min}}$ cut. Energy–momentum conservation ensures that interactions never are truly independent, and also other effects enter (see below), but the Poisson ansatz is still a useful starting point.

Multiple interactions (MI) can only be half the solution, however. The divergence for $p_{\perp \text{min}} \rightarrow 0$ would seem to imply an infinite average number of interactions. But what one should realize is that, in order to calculate the $d\sigma/dt$ matrix elements within standard perturbation theory, it has to be assumed that free quark and gluon states exist at negative and positive infinity. That is, the confinement of colour into hadrons of finite size has not been taken into account. So, obviously, perturbation theory has to be broken down by

$$p_{\perp \text{min}} \simeq \frac{\hbar}{r_p} \approx \frac{0.2 \text{ GeV} \cdot \text{fm}}{0.7 \text{ fm}} \approx 0.3 \text{ GeV} \simeq \Lambda_{\text{QCD}}.$$

The nature of the breakdown is also easy to understand: a small-$p_{\perp}$ gluon, to be exchanged between the two incoming hadrons, has a large transverse wavelength and thus almost the same phase across the extent of each hadron. The contributions from all the colour charges in a hadron thus add coherently, and that means that they add to zero since the hadron is a colour singlet.

What is then the typical scale of such colour screening effects, i.e., at what $p_{\perp}$ has the interaction rate dropped to approximately half of what it would have been if the quarks and gluons of a proton had all been free to interact fully independently? That ought to be related to the typical screening distance between a given colour and its opposite anticolour. When a proton contains many partons this characteristic screening distance can well be much smaller than the proton radius. Empirically we need to introduce a $p_{\perp \text{min}}$ scale of the order of 2 GeV to describe Tevatron data, i.e., of the order of 0.1 fm separation. It is not meaningful to take this number too seriously without a detailed model of the space–time structure of a hadron, however.

The 2 GeV number is very indirect and does not really tell exactly how the dampening occurs. One can use a simple recipe, with a step-function cut at this scale, or a physically more reasonable dampening
by a factor $p_\perp^4/(p_{\perp 0}^2 + p_\perp)^2$, plus a corresponding shift of the $\alpha_s$ argument,

$$\frac{d\hat{\sigma}}{dp_\perp^2} \propto \frac{\alpha_s^2(p_{\perp 0}^2)}{p_\perp^2} \frac{\alpha_s^2(p_{\perp 0}^2 + p_\perp^2)}{(p_{\perp 0}^2 + p_\perp^2)^2},$$

(25)

with $p_{\perp 0}$ a dampening scale that also lands at around 2 GeV. This translates into a typical number of 2–3 interactions per event at the Tevatron and 4–5 at the LHC. For events with jets or other hard processes the average number is likely to be higher.

### 6.1 Multiple-interaction models

The first studies of complete events based on perturbatively generated MI [31] started out from a minimally simple model:

1) Use a sharp $p_{\perp \text{min}}$ cut as the key tunable parameter.
2) Address only inelastic nondiffractive events, i.e., with $\sigma_{\text{nd}} \simeq (1/2 - 2/3)\sigma_{\text{tot}}$, so that the average number of interactions per such events is $\langle n \rangle = \sigma_{\text{int}}(p_{\perp \text{min}})/\sigma_{\text{nd}}$.
3) To a first approximation this gives a Poisson distribution in the number of interactions per event, with a fraction $P_0 = e^{-\langle n \rangle}$ of purely low-$p_\perp$ interactions.
4) The interactions are generated in an ordered sequence of decreasing $p_\perp$ values: $p_{\perp 1} > p_{\perp 2} > p_{\perp 3} > \ldots$. This is possible with the standard Sudakov kind of trick:

$$\frac{dP}{dp_{\perp 1}} = \frac{1}{\sigma_{\text{nd}} dp_{\perp}} \exp \left[ - \int_{p_{\perp}}^{p_{\perp 1}} \frac{1}{\sigma_{\text{nd}} dp_{\perp}} d\sigma \left( \frac{dP}{dp_{\perp}} \right) \right],$$

(26)

with a starting $p_{\perp 0} = E_{\text{cm}}/2$.
5) The ordering of emissions allows parton densities to be rescaled in $x$ after each interaction, so that energy–momentum is not violated. Thereby the actual distribution of the number of interactions becomes narrower than Poissonian, since the rate of a further interaction is reduced if previous ones already took away energy.
6) For technical reasons the model was simplified after the first interaction, so that there only $gg$ or $q\bar{q}$ outgoing pairs were allowed, and no showers were added to these additional $2 \rightarrow 2$ interactions.

This simple PYTHIA-based model is able to ‘explain’ a large set of experimental data.

Some additional features were then included in a more sophisticated variant.

1) Use the smooth turnoff of Eq. (25) all the way down to $p_\perp = 0$. Then an event has to contain at least one interaction to be an event at all, i.e., $p_{\perp 0}$ has to be selected sufficiently small that $\sigma_{\text{int}} > \sigma_{\text{nd}}$.
2) Hadrons are extended, and therefore partons are distributed in (transverse) coordinates. To allow a flexible parametrization and yet have an easy-to-work-with expression, a double Gaussian $\rho_{\text{matter}}(r) = N_1 \exp(-r^2/r_1^2) + N_2 \exp(-r^2/r_2^2)$ is used, where $N_2/N_1$ and $r_2/r_1$ are tunable parameters.
3) The matter overlap during a collision, calculated by

$$O(b) = \int d^3x \ d t \rho_{1,\text{matter}}^{\text{boosted}}(x, t) \rho_{2,\text{matter}}^{\text{boosted}}(x, t),$$

(27)

directly determines the average activity in events at different impact parameter $b$: $\langle n(b) \rangle \propto O(b)$. That is, central collisions tend to have more activity, peripheral less, but of course properly normalized so that the $b$-integrated interaction cross section agrees with standard perturbation theory (modulo the already-discussed dampening at small $p_\perp$). Thereby the $P_n$ distribution becomes broader than a Poisson one.
As before, several simplifications are necessary. This is the scenario that has been used in many of the experimental studies over the years.

More recently a number of improvements have been included [32].

1) The introduction of junction fragmentation, wherein the confinement field between the three quarks in a baryon is described as a Y-shaped topology, now allows the handling of topologies where several valence quarks are kicked out, thus allowing arbitrary flavours and showering in all interactions in an event.

2) Parton densities are rescaled not only for energy–momentum conservation, but also to take into account the number of remaining valence quarks, or that sea quarks have to occur in $q\overline{q}$ pairs.

3) The introduction of $p_\perp$-ordered showers allows the selection of new ISR branchings and new interactions to be interleaved in one common sequence of falling $p_\perp$ values. Thereby the competition between these two components, which both remove energy from the incoming beams, is modelled more realistically.

FSR is not yet interleaved, but also does not compete for beam energy. This scenario is not yet as well studied experimentally.

The traditional HERWIG soft underlying event (SUE) approach to this issue has its origin in the UA5 Monte Carlo. Here a number of clusters are distributed almost independently in rapidity and transverse momentum, but shifted so that energy–momentum is conserved, and the clusters then decay isotropically. The multiplicity distribution of clusters and their $y$ and $p_\perp$ spectra are tuned to give the observed inclusive hadron spectra. No jets are produced in this approach.

The JIMMY [33] program is an add-on to HERWIG. It replaces the SUE model with a MI-based one more similar to the PYTHIA ones above, e.g., with an impact-parameter-based picture for the multiple-interactions rate. Technical differences exist, for example, JIMMY interactions are not picked to be $p_\perp$-ordered and thus energy–momentum issues are handled differently.

The DPMJET/DTUJET/PHOJET family of programs [34] come from the ‘historical’ tradition of soft physics, wherein multiple $p_\perp \approx 0$ 'pomeron' exchanges fill a role somewhat similar to the hard MI above. Jet physics was originally not included, but later both hard and soft interactions have been allowed. One strong point is that this framework also allows diffractive events to be included as part of the same basic machinery.

### 6.2 Multiple-interaction studies

How do we know that MI exist? The key problem is that it is not possible to identify jets coming from $p_\perp \approx 2$ GeV partons. Therefore we either have to use indirect signals for the presence of interactions at this scale or we have to content ourselves with studying the small fraction of events where two interactions occur at visibly large $p_\perp$ values.

An example of the former is the total charged multiplicity distribution in high-energy $pp/\bar{p}p$ collisions. This distribution is very broad, and is even getting broader with increasing energy, measured in terms of the width over the average, $\sigma(n_{ch})/\langle n_{ch} \rangle$. By contrast, recall that for a Poisson distribution this quantity scales like $1/\sqrt{\langle n_{ch} \rangle}$ and thus is getting narrower. Simple models, with at most one interaction and with a fragmentation framework in agreement with LEP data, cannot explain this: they are way too narrow, and have the wrong energy behaviour. If MI are included, the additional variability in the number of interactions per event offers the missing piece [31]. The variable-impact parameter improves the description further.

Another related example is forward–backward correlations. Consider the charged multiplicity $n_f$ and $n_b$ in a forward and a backward rapidity bin, each of width one unit, separated by a central rapidity
gap of size \( \Delta y \). It is not unnatural that \( n_f \) and \( n_g \) are somewhat correlated in two-jet events, and for small \( \Delta y \) one may also be sensitive to the tails of jets. But the correlation coefficient, although falling with \( \Delta y \), still is appreciable even out to \( \Delta y = 5 \), and here again traditional one-interaction models come nowhere near. In a MI scenario each interaction provides additional particle production over a large rapidity range, and this additional number-of-MI variability leads to good agreement with data.

Direct evidence comes from the study of four-jet events. These can be caused by two separate interactions, but also by a single one where higher orders (call it ME or PS) have allowed two additional branchings in a basic two-jet topology. Fortunately the kinematics should be different. Assume the four jets are ordered in \( p_{\perp 1} > p_{\perp 2} > p_{\perp 3} > p_{\perp 4} \). If coming from two separate interactions the jets should pair up into two separately balancing sets, \( |p_{\perp 1} + p_{\perp 2}| \approx 0 \) and \( |p_{\perp 3} + p_{\perp 4}| \approx 0 \). If an azimuthal angle \( \varphi \) is introduced between the two jet axes this also should be flat if the interactions are uncorrelated. By contrast the higher-order graph offers no reason why the jets should occur in balanced pairs, and the \( \varphi \) distribution ought to be peaked at small values, corresponding to the familiar collinear singularity. The first to observe an MI signal this way was the AFS Collaboration [35] at the ISR (\( pp \) at 62 GeV), but with large uncertainties. A more convincing study was made by CDF [36], who obtained a clear signal in a sample with three jets plus a photon. In fact the deduced rate was almost a factor of three higher than naive expectations, but quite in agreement with the impact-parameter-dependent picture, wherein correlations of this kind are enhanced.

A topic that has been quite extensively studied in CDF is that of the jet pedestal [37], i.e., the increased activity seen in events with a jet, even away from the jet itself, and away from the recoiling jet that should be there. Some effects come from the showering activity, i.e., the presence of additional softer jets, but much of it rather finds its explanation in MI, as a kind of ‘trigger bias’ effect, as follows. (1) Central collisions tend to produce many interactions, peripheral ones few. (2) If an event has \( n \) interactions there are \( n \) chances that one of them is hard. Combine the two and one concludes that events with hard jets are biased towards central collisions and many additional interactions. The rise of the pedestal with trigger-jet energy saturates once \( \sigma_{\text{int}}(p_{\perp \text{min}} = p_{\perp \text{jet}}) \ll \sigma_{\text{nd}} \), however, because by then events are already maximally biased towards small impact parameter. And this is indeed what is observed in the data: a rapid rise of the pedestal up to \( p_{\perp \text{jet}} \approx 10 \) GeV, and then a slower increase that is mainly explained by showering contributions.

In more detailed studies of this kind of pedestal effects there are also some indications of a jet substructure in the pedestal, i.e., that the pedestal is indeed associated with the production of additional (soft) jet pairs.

In spite of many qualitative successes, and even some quantitative ones, one should not be led to believe that all is understood. Possibly the most troublesome issue is how colours are hooked up between all the outgoing partons that come from several different interactions. A first, already difficult, question is how colours are correlated between all the partons that are taken out from an incoming hadron. These colours are then mixed up by the respective scattering, in principle (approximately) calculable. But, finally, all the outgoing partons will radiate further and overlap with each other on the way out, and how much that may mess up colours is an open question.

A sensitive quantity is \( \langle p_{\perp} \rangle(n_{\text{ch}}) \), i.e., how the average transverse momentum of charged particles varies as a function of their multiplicity. If interactions are uncorrelated in colour this curve tends to be flat: each further interaction adds about as much \( p_{\perp} \) as \( n_{\text{ch}} \). If colours somehow would rearrange themselves, so that the confinement colour fields would not have to run criss-cross in the event, then the multiplicity would not rise as fast for each further interaction, and so a positive slope would result. The embarrassing part is that the CDF tunes tend to come up with values that are about 90\% on the way to being maximally rearranged [37], which is far more than one would have guessed. Obviously further modelling and tests are necessary here [38].

Another issue is whether the \( p_{\perp 0} \) regularization scale should be energy dependent. In olden days there was no need for this, but it became necessary when HERA data showed that parton densities rise
faster at small $x$ values than had commonly been assumed. This means that the partons become more close-packed and the colour screening increases faster with increasing collision energy. Therefore an energy-dependent $p_{\perp 0}$ is not unreasonable, but also cannot be predicted. Currently the default PYTHIA ansatz is $p_{\perp 0}(E_{cm}) = (2.0 \text{ GeV}) (E_{cm}/1.8 \text{ TeV})^{0.16}$, i.e., a predicted $p_{\perp 0} = 2.8 \text{ GeV}$ at 14 TeV. This gives a minimum-bias multiplicity of about 7 per unit of rapidity in the central region, and a pedestal under jet events of around 30 charged particles per unit [39]. However, these numbers are model and parameter dependent. PHOJET predicts about half as big a pedestal, and typical JIMMY tunes about 50% more, a priori leaving a big range of uncertainty to be resolved once LHC runs begin.

7 Hadronization

The physics mechanisms discussed so far are mainly being played out on the partonic level, while experimentalists observe hadrons. In between exists the very important hadronization phase, where all the outgoing partons end up confined inside hadrons of a typical 1 GeV mass scale. This phase cannot (so far?) be described from first principles, but has to involve some modelling. The main approaches in use today are string fragmentation and cluster fragmentation. These are described in the lectures of G. Ingelman [22], so this section will be very brief, with only a few comments.

Hadronization models start from some ideologically motivated principles, but then have to add ‘cookbook recipes’ with free parameters to arrive at a complete picture of all the nitty-gritty details. This should come as no surprise, given that there are hundereds of known hadron species to take into account, each with its mass, width, wavefunction, couplings, decay patterns and other properties that could influence the structure of the observable hadronic state, and with many of those properties being poorly or not at all known, either from theory (lattice QCD) or from experiment. In that sense, it is sometimes more surprising that models can work as well as they do than that they fail to describe everything.

The simpler initial state at an $e^+e^-$ collider, such as LEP, implies that this is the logical place to tune the hadronization framework to data [20], and thereafter those tunes can be applied to other studies. One such is the internal structure of jets in hadron collider, where the pattern in many respects is surprisingly well described.

On the other hand, at the HERA $e^\pm p$ collider it has been observed that the relative amount of strange-particle production is only $2/3$ of that at LEP, and of (anti)baryons only $1/2$. This has no simple explanation within the string fragmentation model, so it acts as a useful reminder that we still do not know as much as we should. Other examples could also be provided.

8 Summary and outlook

In these lectures we have followed the flow of generators roughly ‘inwards out’, i.e., from short-distance processes to long-distance ones. At the core lies the hard process, described by matrix elements. It is surrounded by initial- and final-state showers, that should be properly matched to the hard process. Multiple parton–parton interactions can occur, and the colour flow is tied up with the structure of beam remnants. At longer time-scales the partons turn into hadrons, many of which are unstable and decay further. This basic pattern is likely to remain in the future, but many aspects will change.

One such aspect, that stands a bit apart, is that of language. The traditional event generators, like PYTHIA and HERWIG, have been developed in FORTRAN — up until the end of the LEP era this was the main language in high-energy physics. But now the experimental community has largely switched to C++ and decisions have been taken, for example, at CERN to discontinue FORTRAN altogether. The older generators are still being used, hidden under C++ wrappers, but this can only be a temporary solution, for several reasons. One is that younger experimentalists often need to look into the code of generators and tailor some parts to specific needs of theirs, and if the code is in an unknown language this will not work. Another is that theory students who apply for non-academic positions are much better off if their resumés say ‘expert in object-oriented programming’ rather than ‘FORTRAN fan’.
A conversion program has thus begun on many fronts. SHERPA, as the youngest of the general-purpose generators, was conceived from the onset as a C++ package and thus is some steps ahead of the other programs in this respect. HERWIG++ [40] is a complete reimplementation of the HERWIG program, as is PYTHIA 8 of the current PYTHIA 6. Both conversions have taken longer than originally hoped, but progress is being made and first versions exist. THEPEG [41] is a generic toolkit for event generators, used by HERWIG++ and the upcoming new ARIADNE.

There are also other aspects where we have seen progress in recent years and can hope for more:

- Faster, better and more user-friendly general-purpose matrix-element generators with an improved sampling of phase space.
- New ready-made libraries of physics processes, in particular with full NLO corrections included.
- More precise parton showers.
- Better matching between matrix elements and parton showers.
- Improved models for minimum-bias physics and underlying events.
- Some upgrades of hadronization models and decay descriptions.

In general one would say that generators are getting better all the time, but at the same time the experimental demands are also getting higher, so it is a tight race. However, given that typical hadronic final states at the LHC will contain hundreds of particles and quite complex patterns buried in that, it is difficult to see that there are any alternatives.

At the same time as you need to use generators, you should remain critical and be on the lookout for bugs and bad modelling. Many years ago Bjorken worried about the passive attitude many experimentalists have towards the output of generators; they “carry the authority of data itself. They look like data and feel like data, and if one is not careful they are accepted as if they were data.” [42]. Do not fall into that trap!

References

    http://www.ippp.dur.ac.uk/montecarlo/BSM/.


[22] G. Ingelman, lectures given at this School (unpublished).


[34] M. Sandhoff and P. Skands, FERMILAB-CONF-05-518-T.


[37] J.D. Bjorken, in a talk given at the 75th anniversary celebration of the Max-Planck Institute of Physics, Munich, Germany, December 10th, 1992, as quoted in Beam Line, Winter 1992, Vol. 22, No. 4.
Relativistic heavy-ion physics: three lectures

L. McLerran
Brookhaven National Laboratory, Upton, NY 11973, USA and RIKEN BNL Research Center, Brookhaven National Laboratory, Upton, NY 11973, USA

Abstract

These lectures provide an introduction to the physics issues which are being studied in the collisions of ultrarelativistic heavy ions. The lectures are focused on the production of new states of matter. The quark–gluon plasma is thermal matter which once existed in the Big Bang. The colour glass condensate is a universal form of high energy density gluonic matter which is part of a hadron wavefunction and which controls the high-energy limit of strong interactions. The glasma is matter produced in the collisions of high-energy hadrons which evolves into a quark–gluon plasma. The glasma has interesting topological properties and may be responsible for the early thermalization seen at RHIC. I introduce the student to these topics, discuss results from experiments, and comment upon future opportunities.

1 Introduction

These lectures will introduce the student to the physics issues behind the study of new forms of matter, and the general issue of understanding the high-energy limit of QCD. The full programme of this study involves the collisions of protons on protons, deuterons on nuclei, nuclei on nuclei, and electrons on protons and nuclei. The reason for using nuclei is that one can achieve extraordinary energy densities of matter, and, because of the large size of nuclei relative to partons, more easily study effects associated with the bulk properties of matter. The highest energies are required, as this allows one to generate the highest energy densities, and as we shall see, at RHIC energies and higher, one can study novel effects associated with the high density of gluons in a hadron wavefunction.

Central to these experimental studies is the production of new forms of matter. This may be a Quark–Gluon Plasma (QGP) or a Colour Glass Condensate (CGC). The properties of these forms of matter are described below.

The outline of these lectures is

– New states of matter
In the first lecture, I describe the new forms of matter which may be produced in heavy-ion collisions. These are the quark–gluon plasma, the colour glass condensate, and the glasma.

– Space–time dynamics
This lecture describes the space–time dynamics of high-energy heavy-ion collisions. In this lecture, I illustrate how high energy density matter might be formed. I describe how the colour glass condensate may produce a glasma which evolves into the quark–gluon plasma, and eventually to a gas of ordinary hadrons.

– Experiment and theory
In the final lecture, I show how various experimental measurements might teach us about the properties of matter. Topics discussed are multiplicities and the colour glass condensate, low-transverse-momentum particles and the quark–gluon plasma, heavy vector meson production and confinement, the flavour dependence of the quark–gluon plasma, high-transverse-momentum particles and what they tell us about the CGC and the QGP, and identical particle correlations and what they tell us about the space–time evolution of the matter produced in collisions.
2 Lecture I: High-density matter

2.1 The goals of RHIC

The goal of nuclear physics has traditionally been to study matter at densities of the order of those in the atomic nucleus

\[ \epsilon \sim 0.15 \text{ GeV/fm}^3. \]  

High-energy nuclear physics has extended this study to energy densities several orders of magnitude higher. This extension includes the study of matter inside ordinary strongly interacting particles, such as the proton and the neutron, and producing new forms of matter at much higher energy densities in high-energy collisions of nuclei with nuclei, and various other probes.

There are at least three central issues of high-energy nuclear physics:

- **The properties of matter close to thermal equilibrium at energy densities greater than one or two orders that of nuclear matter**
  This matter is at such high densities that it is only simply described in terms of quarks and gluons and is generically referred to as the Quark–Gluon Plasma (QGP). The study of this matter may allow us to better understand the origin of the masses of ordinary particles such as nucleons, and of the confinement of quarks and gluons into hadrons. The QGP will be described below [1].

- **The study of the matter which controls high-energy strong interactions**
  This matter is believed to be universal (independent of the hadron), and exists over sizes large compared to the typical microphysics size scales important for high-energy strong interactions. (The microphysics size scale here is about 1 fm and the microphysics time scale is the time it takes light to fly 1 fm, \( t \sim 10^{-23} \text{ s.} \)) The matter appears in the wavefunctions of a hadron, and is called a Colour Glass Condensate (CGC) because it is composed of coloured particles, evolves on time scales long compared to microphysics time scales and therefore has properties similar to glasses, and a condensate since the phase-space density of gluons is very high. In collisions, this matter forms a glasma, which eventually evolves into a quark–gluon plasma, for hadrons of large enough size (for example, nuclei). The study of these forms of matter may allow us to better understand the typical features of strong interactions when they are truly strong, a problem which has eluded a basic understanding since strong interactions were first discovered. The CGC and the glasma will be described below [2].

- **The study of the structure of the proton, most notably spin**
  The structure of the proton and neutron is important as these particles form the ordinary matter from which we are composed. We would like to understand how valence quantum numbers such as baryon number, charge, and spin are distributed. RHIC has an active programme to study the spin of the proton [3].

Because I was asked to provide lectures on ultrarelativistic nuclear collisions, I shall discuss only the first two issues.

2.2 The quark–gluon plasma

This section describes the quark–gluon plasma, why it is important for astrophysics and cosmology, and why it provides a laboratory in which one can study the origin of mass and of confinement [1].

2.2.1 What is the quark–gluon plasma?

Matter at low energy densities is composed of electrons, protons, and neutrons. If we heat the system, we might produce thermal excitations which include light-mass strongly interacting particles such as the pion. Inside the protons, neutrons, and other strongly interacting particles are quarks and gluons. If we
As the energy density is decreased, the quark–gluon plasma condenses into a low-density gas of hadrons. Quarks are red, green, or blue and gluons are yellow.

At the earliest times in the Big Bang, temperatures are of order $T \sim 10^{19}$ GeV, quantum gravity is important, and despite the efforts of several generations of string theorists, we have little understanding. At somewhat lower temperatures, perhaps there is the grand unification of all the forces, except gravity. It might be possible that the baryon number of the universe is generated at this temperature scale. At much smaller temperatures, the interactions between the quarks and gluons become weak. This is a consequence of the asymptotic freedom of strong interactions: at short distances the strong interactions become weak.

The QGP surely existed during the Big Bang. In Fig. 2, the various stages of evolution in the Big Bang are shown.
lower temperatures, of order $T \sim 100$ GeV, electroweak symmetry breaking takes place. It is possible here that the baryon asymmetry of the universe might be produced. At temperatures of order $T \sim 1$ GeV, quarks and gluons become confined into hadrons. This is the temperature range appropriate for studies at RHIC and the LHC. At $T \sim 1$ MeV, the light elements are made. This temperature corresponds to an energy range which has been much studied, and is the realm of conventional nuclear physics. At temperatures of the order of an electronvolt, corresponding to the binding energies of electrons in atoms, the universe changes from an ionized gas to a lower-pressure gas of atoms, and structure begins to form.

The QGP is formed at energy densities of order $1$ GeV/fm$^3$. Matter at such energy densities probably exists inside the cores of neutron stars as shown in Fig. 3. Neutron stars are objects of about 10 km in radius and are composed of extremely high energy density matter. The typical energy density in the core is of the order of $1$ GeV/fm$^3$, and approaches zero at the surface. Unlike the matter in the Big Bang, this matter is cold and has temperature small compared to the Fermi energies of quarks. It is a cold, degenerate gas of quarks. At lower densities, this matter converts into a cold gas of nucleons.

Hot and dense matter with energy density of order $1$ GeV/fm$^3$ may have occurred in the supernova explosion which led to the neutron star’s formation. It may also occur in collisions of neutron stars and black holes, and may be the origin of the mysterious gamma-ray bursters. (Gamma-ray bursters are believed to be starlike objects which convert of the order of their entire mass into gamma rays.)

### 2.2.2 The quark–gluon plasma and ideal gases

At very high energy density, the coupling constant of QCD becomes weak. A gas of particles should to a good approximation become an ideal gas. Each species of particle contributes to the energy density of an ideal gas as

$$
\epsilon = \int \frac{d^3 p}{(2\pi)^3} \sum_i \frac{E_i}{e^{\beta E_i} \pm 1}
$$

where the $\pm$ is for bosons and the $+$ for fermions. The energy of each particle is $E_i$. At high temperatures, masses can be ignored, and the factor of $\pm 1$ in the denominator turns out to make a small difference. One finds therefore that

$$
\epsilon \sim \frac{\pi^2}{30} N T^4
$$
where $N$ is the number of particle degrees of freedom. At low temperatures when masses are important, only the lowest mass, strongly interacting particle degree of freedom contributes; the pion, and the energy density approaches zero as $\epsilon \sim e^{-m_\pi/T}$. For an ideal gas of pions, the number of pion degrees of freedom is three. For a QGP there are two helicities and eight colours for each gluon, and for each quark, three colours, two spins, and a quark–antiquark pair. The number of degrees of freedom is $N \sim 2 \times 8 + 4 \times 3 \times N_F$ where $N_F$ is the number of important quark flavours, which is about three if the temperature is below the charm quark mass so that $N \sim 50$.

There is about an order of magnitude change in the number of degrees of freedom between a hadron gas and a QGP. This is because the degrees of freedom of the QGP include colour. In the large $N_{\text{colour}}$ limit, the number of degrees of freedom of the plasma are proportional to $N_{\text{colour}}^2$, and in the confined phase is of order 1. In this limit, the energy density has an infinite discontinuity at the phase transition. There would be a limiting temperature for the hadronic world in the limit for which $N_{\text{colour}} \rightarrow \infty$, since at some temperature the energy density would go to infinity. This is the Hagedorn limiting temperature. (In the real world $N_{\text{colour}}$ is three, and there is a temperature at which the energy density changes by an order of magnitude in a narrow range.)

### 2.2.3 The quark–gluon plasma and fundamental physics issues

The nature of matter at high densities is an issue of fundamental interest. Such matter occurred during the Big Bang, and it is the ultimate and universal state of matter at very high energy densities.

A hypothetical phase diagram for QCD is shown in Fig. 4. The vertical axis is temperature, and the horizontal is a measure of the matter or baryon number density, the baryon number chemical potential [5]. The solid lines indicate a first-order phase transition, and the dashed line a rapid crossover. It is not known for sure whether or not the region marked crossover is or is not a true first-order phase transition. There are analytic arguments for the properties of matter at high density, but numerical computations are of insufficient resolution. At high temperature and fixed baryon number density, there are both analytic arguments and numerical computations of good quality. At high density and fixed temperature, one goes into a superconducting phase, perhaps multiple phases of superconducting quark matter. At high temperature and fixed baryon number density, the degrees of freedom are those of a quark–gluon plasma.

![The Evolving QCD Phase Transition](image)

**The Evolving QCD Phase Transition**

- $t \sim 1980$
  - Critical Temperature 150 - 200 MeV ($\mu_B = 0$)
  - Critical Density 1/2-2 Baryons/fm$^3$ ($T = 0$)

- $t \sim 1990$
  - Quark Gluon Plasma
  - Hadron Gas

- $t \sim 2000$
  - Quark Gluon Plasma
  - Hadron Gas
  - Color Superconductor

**Fig. 4:** A phase diagram for QCD collisions

I have shown this phase diagram as a function of time. What this means is that at various times people thought they knew what the phase diagram was. As time evolved, the picture changed. The latest
ideas are marked with the date 2000. The point of doing this is to illustrate that theoretical ideas in
the absence of experiment change with time. Physics is essentially an experimental science, and it is
very difficult to appreciate the richness which nature allows without knowing from experiment what is
possible.

Much of the information we have about QCD at finite energy density comes from lattice gauge
theory numerical simulation [5]. To see how lattice gauge theory works, recall that at finite temperature,
the grand canonical ensemble is given by

$$Z = \text{Tr } e^{-\beta H}. \quad (4)$$

This is similar to computing

$$Z = \langle e^{-itH} \rangle \quad (5)$$

where $-it = \beta$. That is we compute the expectation value of the time evolution operator for imaginary
time. This object has a path integral representation, which has been described to you in your elementary
field theory text books. Under the change of variables, the action becomes $iS = i \int dt L \rightarrow S =
- \int_0^\beta d\tau L$. Here $L$ is the Lagrangian.

The grand canonical ensemble has the representation

$$Z = \int [dA] e^{-S[A]} \quad (6)$$

for a system of pure gluons. The gluon fields satisfy periodic boundary conditions due to the trace in the
definition of the grand canonical ensemble. (Fermions may also be included, although the path integral
is more complicated, and the fermion fields are required to satisfy antiperiodic boundary conditions.)
Expectation values are computed as

$$\langle 0 \rangle = \frac{\text{Tr } O e^{-\beta H}}{\text{Tr } e^{-\beta H}}. \quad (7)$$

The way that lattice Monte Carlo simulates the grand canonical ensemble is by placing all of the
fields on a finite grid, so the path integral becomes finite dimensional. Then field configurations are
selectively sampled, as weighted by their action. This works because the factor of $e^{-\beta H}$ is positive and
real. (The method has essential complications for finite density systems, since there the action becomes
complex.)

Lattice gauge theory numerical studies, and analytic studies have taught us much about the properties
of these various phases of matter [5]. There have been detailed computations of the energy density
as a function of temperature. In Fig. 5 the energy density scaled by $T^4$ is plotted. This is essentially the
number of degrees of freedom of the system as a function of $T$. At a temperature of $T_c \sim 160–190$ MeV
the number of degrees of freedom changes very rapidly, possibly discontinuously. This is the location of
the transition from the hadron gas to the quark–gluon plasma.

In Fig. 6, the sound velocity is plotted as a function of temperature. The sound velocity increases
at high temperature asymptoting to its ideal gas value of $v_{\text{sound}}^2 \sim 1/3$. Near the phase transition, it
becomes very small. This is because the energy density jumps at the transition temperature, but the
pressure must be smooth and continuous. The sound velocity squared is $dP/d\epsilon$.

Lattice Monte Carlo simulation has also been used to study how the phase transition is related to
the confining force. In a theory with only gluons, the potential for sources of fundamental representation
colour charge grows linearly in the confined phase. (With dynamical fermions, the potential stops rising
at some distance when it is energetically favourable to produce quark–antiquark pairs which short out
the potential.)
We can understand how confinement might disappear at high temperature. At finite temperature, there is a symmetry of the pure gluon Yang–Mills system. Consider a Wilson line which propagates from \((0, \vec{x})\) to the point \((\beta, \vec{x})\). A Wilson line is a path-ordered phase,

\[
L(x) = P \exp i \int_0^\beta dt A^0(t, \vec{x}) .
\]  

One can show that the expectation value of this line gives the free energy of an isolated quark:

\[
e^{-\beta F} = \frac{1}{N_c} \text{Tr} (L(x))) .
\] 

Fig. 5: The energy density scaled by \(T^4\) as a function of temperature

Fig. 6: The sound velocity as a function of temperature
Now consider gauge transformations which maintain the periodic boundary conditions on the gauge fields (required by the trace in the definition of the grand canonical ensemble). The most general gauge transformation which does this is not periodic but solves

$$U(\beta, \vec{x}) = ZU(0, \vec{x}).$$

One can show that $[z, \tau^a] = 0$, and that $\nabla^i Z = 0$. $Z$ is an element of the gauge group so that $\det Z = 1$. These conditions require that

$$Z = e^{2\pi i j/N_c}.$$  

This symmetry under non-periodic gauge transformations is global, that is it does not depend upon the position in space. It may be broken. If it is realized, the free energy of a quark must be infinite since $L \rightarrow ZL$ under this transformation, and $\langle L \rangle = 0$. If the symmetry is broken, quarks can be free.

Lattice gauge computations have measured the quark–antiquark potential as a function of $T$, and at the deconfinement temperature, the potential changes from linear at infinity to constant. This is shown in Fig. 7.

![Fig. 7: The potential in pure gauge theory as a function of temperature](image)

In addition to confinement–deconfinement, there is an additional symmetry which might occur at high temperatures. In nature, the up and down quark masses are almost zero. This leads to a chiral symmetry, which is the rotation of fermion fields by $e^{i\gamma_5 \theta}$. This symmetry would require that either baryons are massless or occur in parity doublets. Neither arises in nature. The nucleon has a mass of about 1 GeV and has no opposite parity partner of almost equal mass. It is believed that this symmetry becomes broken, and as a consequence, the nucleon acquires mass, and that the pion becomes an almost massless Goldstone boson. It turns out that at the confinement–deconfinement phase transition, chiral symmetry is restored. This is seen in Fig. 8 where a quantity proportional to the nucleon mass is plotted as a function of $T$.

The chiral symmetry restoration phase transition can have interesting dynamical consequences. In the confined phase, the mass of a nucleon is of order $N_c \Lambda_{\text{QCD}}$, but in the deconfined phase is of order $T$. Therefore in the confined phase, the Boltzman weight $e^{-M/T}$ is very small. Imagine what happens...
as we go through the phase transition starting at a temperature above $T_c$. At first the system is entirely in QGP. As the system expands, a mixed phase of droplets of QGP and droplets of hadron gas forms. The nucleons like to stay in the QGP because their Boltzman weight is larger. As the system expands further, the droplets of QGP shrink, but most of the baryon number is concentrated in them. At the end of the mixed phase, one has made large-scale fluctuations in the baryon number. This scenario is shown in Fig. 9.

The confinement–deconfinement phase transition and the chiral symmetry restoration phase transition might be logically disconnected from one another. The confinement–deconfinement phase transition is related to a symmetry when the quark masses are infinite. The chiral transition is related to a symmetry when the quarks are massless. As a function of mass, one can follow the evolution of the phase transitions. At large and small masses there is a real phase transition marked by a discontinuity in physical quantities. At intermediate masses, there is probably a rapid transition, but not a real phase transition. It is believed that the real world has masses which make the transition closer to a cross-over than a phase transition, but the evidence from lattice Monte Carlo studies is very weak. In Fig. 10, the various possibilities are shown.

2.3 The colour glass condensate

This section describes the colour glass condensate, and why it is important for our understanding of basic properties of strong interactions [2], [6]. I argue that the colour glass condensate is a universal form of matter which controls the high-energy limit of all strong interaction processes and is the part of the hadron wavefunction important at such energies. Since the colour glass condensate is universal and controls the high-energy limit of all strong interactions, it is of fundamental importance.

2.3.1 What is the colour glass condensate?

The colour glass condensate is a new form of matter which controls the high-energy limit of strong interactions. It is universal and independent of the hadron which generated it. It should describe

- high-energy cross-sections
A very-high-energy hadron has contributions to its wavefunction from gluons, quarks, and anti-quarks with energies up to that of the hadron and all the way down to energies of the order of the scale of light-mass hadron masses, $E \sim 200 \text{ MeV}$. A convenient variable in which to think about these quark degrees of freedom is the typical energy of a constituent scaled by that of the hadron,

$$x = E_{\text{constituent}} / E_{\text{hadron}}.$$

Clearly the higher the energy of the hadron we consider, the lower the minimum $x$ of a constituent. Sometimes it is also useful to consider the rapidity of a constituent which is $y \sim \ln(1/x)$.

The density of small-$x$ partons is

$$\frac{dN}{dy} = xG(x,Q^2).$$

The scale $Q^2$ appears because the number of constituents one measures depends (weakly) upon the resolution scale of the probe with which one measures. (Resolution scales are measured in units of the inverse momentum of the probe, which is usually taken to be a virtual photon.) A plot of $xG(x,Q^2)$ for gluons at various $x$ and $Q^2$ measured at the HERA accelerator in protons [7] is shown in Fig. 11.

Note that the gluon density rises rapidly at small $x$ in Fig. 11. This is the so-called small-$x$ problem. It means that if we view the proton head-on at increasing energies, the low-momentum gluon density grows. This is shown in Fig. 12.

As the density of gluons per unit area per unit rapidity increases, the typical transverse separation of the gluons decreases. This means that the matter which controls high-energy strong interactions is very dense, and it means that the QCD interaction strength which is usually parametrized by the dimensionless
The glassy nature of the condensate arises because the fields associated with the condensate are generated by constituents of the proton at higher momentum. These higher momentum constituents have their time scales Lorentz time dilated relative to those which would be measured in their rest frame. Therefore the fields associated with the low-momentum constituents also evolve on this long time scale.
The low-momentum constituents are therefore glassy: their time evolution scale is unnaturally long compared to their natural time scale. Hence the name colour glass condensate.

There is also a typical scale associated with the colour glass condensate: the saturation momentum. This is the typical momentum scale where the phase-space density of gluons becomes $\rho \leq 1/\alpha_s$.

At very high momentum, the fields associated with the colour glass condensate can be treated as classical fields, like the fields of electricity and magnetism. Since they arise from fast moving partons, they are plane polarized, with mutually orthogonal colour electric and magnetic fields perpendicular to the direction of motion of the hadron. They are also random in two dimensions. This is shown in Fig. 13.

**Fig. 12:** The increasing density of wee partons as the energy increases

**Fig. 13:** The colour glass condensate as a high density of random gluon fields on a two-dimensional sheet travelling near the speed of light
2.3.2 *Why is the colour glass condensate important?*

Like nuclei and electrons compose atoms, and nucleons and protons compose nuclear matter, the colour glass condensate is the fundamental matter of which high-energy hadrons are composed. The colour glass condensate has the potential to allow for a first-principles description of the gross or typical properties of matter at high energies. For example, the total cross-section at high energies for proton–proton scattering, as shown in Fig. 14, has a simple form but for over 40 years has resisted simple explanation. (It has perhaps been understood recently in terms of the colour glass condensate or saturation ideas [8–11].)

![Fig. 14: The total cross-section for high-energy proton–proton interactions](image)

The colour glass condensate forms the matter in the quantum mechanical state which describes a nucleus. In the earliest stages of nucleus–nucleus collisions, the matter must not be changed much from these quantum mechanical states. The colour glass condensate therefore provides the initial conditions for the quark–gluon plasma to form in these collisions. A space–time picture of nucleus–nucleus collisions is shown in Fig. 15. At very early times, the colour glass condensate evolves into a distribution of gluons. Later these gluons thermalize and may eventually form a quark–gluon plasma. At even later times, a mixed phase of plasma and hadronic gas may form.

![Fig. 15: A space–time diagram for the evolution of matter produced in heavy-ion collisions](image)
2.4 The glasma

In collisions of nuclei, the matter inside a hadron is liberated. The matter immediately after the collision is called the glasma because it interpolates between the colour glass condensate in the initial nuclear wavefunction and the quark–gluon plasma. It has distinctive features which make it different from either the CGC or the QGP. In the colour glass condensate description, this may be thought of as the collision of two sheets of coloured glass, [12, 13], as shown in Fig. 16. When one writes down the equations which describe the evolution of the classical fields associated with the sheets of coloured glass, one finds that there is a discontinuity at $\tau = 0$. Immediately after the collision, the fields change their character, and instead of transverse colour electric and magnetic fields, with zero longitudinal fields, the fields transform into longitudinal colour electric and magnetic fields with zero transverse fields.

In low-energy, strong-interaction physics, one pictured collisions as generating a flux tube of colour electric field. In the high-energy limit, it seems one generates both longitudinal electric and magnetic fields. This is a consequence of the $E \leftrightarrow B$ duality symmetry of QCD.

One can see a consequence of a non-zero value of $E \cdot B$. Consider ordinary electricity and magnetism. An electric field accelerates an electron in one direction and $B$ makes it spiral around with a specific handedness. For a position, the acceleration is in the opposite direction, and so is the handedness. They therefore both generate the same chirality for an electron or a positron. In QCD, this works roughly the same way, and one can show that for strong values of $E \cdot B$, one can change the chirality of a system by spontaneous helicity flip induced by the field. $E \cdot B$ is proportional to a topological charge, which counts the number of such flips. Note that even though QCD in the limit of massless quarks has chiral symmetry, the topological charge can generate helicity flip processes. The topological charge is therefore related to anomalous processes, that is processes which are forbidden by the classical equations of motion.

In electroweak theory, a topological charge is generated in the Big Bang. This topological charge is associated with baryon number violation. It is possible that this topological charge generation is responsible for generating the baryon number of the universe.

When one writes down the Yang–Mills equations for these classical fields, the fields evolve in time. An example is the equation for the electric field,

$$D^0 E = \vec{D} \times \vec{B}$$

(14)

where $D^0 = \partial^0 - igA^0$ is the covariant derivative operator. If there was only an electric field and no magnetic field, then the electric field could not decay classically, and would have to decay by pair creation. For the glasma, the fields can decay from classical dynamics.

Particles are produced when the fields become of low enough density so that they can be treated as linear wave equations. This is just like in electrodynamics where far away from sources, the equations are free equations and describe radiation.

These equations have been solved to describe the production of particles for high-energy nuclear collisions [12]. They are, however, unstable with respect to small fluctuations [14]. What this means is that if there are small fluctuations due to, for example, the quantum mechanical wavefunction of the initial state, these fluctuations form the seeds for chaotic evolution of the classical fields. Such chaotic configurations can thermalize the system, and some have speculated that this may be the origin of the rapid thermalization at RHIC.

3 Lecture II: Ultrarelativistic nuclear collisions

Heavy-ion collisions at ultrarelativistic energies are visualized in Fig. 16 as the collision of two sheets of coloured glass [12].

At ultrarelativistic energies, these sheets pass through one another. In their wake is left glasma, which eventually materializes as quarks and gluons. These quarks and gluons would naturally form in
their rest frame on some natural microphysics time scale. For the saturated colour glass, this time scale is of the order of the inverse saturation momentum (again, we convert momentum into time by appropriate uses of Planck’s constant and the speed of light), in the rest frame of the produced particle. When a particle has a large momentum along the beam axis, this time scale is Lorentz dilated. This means that the slow particles are produced first towards the centre of the collision regions and the fast particles are produced later further away from the collision region.

This correlation between space and momentum is similar to what happens to matter in Hubble expansion in cosmology. The stars which are further away have larger outward velocities. This means that this system, like the universe in cosmology is born expanding. This is shown in Fig. 17.

As this system expands, it cools. Presumably at some time the produced quarks and gluons thermalize. They then expand as a quark–gluon plasma and eventually as some mixture of hadrons and quarks and gluons. Finally, they may become a gas of only hadrons before they stop interacting and fly off to detectors.

In the last lecture, we shall describe the results from nucleus–nucleus collisions at RHIC in some detail. Before proceeding there, we need to learn a little bit more about the properties of high-energy hadrons. It is useful to introduce some kinematic variables which are useful in what will follow.
The light cone momenta are defined as
\[ P^\pm = \frac{1}{\sqrt{2}} (E \pm p_z) \] (15)
and light cone coordinates are
\[ X^\pm = \frac{1}{\sqrt{2}} (t \pm z) . \] (16)
The metric in these variables is
\[ p \cdot x = p^+ x^- + p^- x^+ - p_T \cdot x_T . \] (17)
Conjugate variables are \( x^\pm (-) p^\mp \). The square of the four momentum is
\[ p^2 = 2P^+ P^- - P_T^2 = M^2 . \] (18)
The uncertainty principle is
\[ \Delta x^\pm \Delta p^\mp \geq 1 . \] (19)

Light cone variables are useful because in a high-energy collision, a left-moving particle has \( p_z \sim E \), so that \( p^+ \sim \sqrt{2} E \), but \( p^- \sim m_i^2 / p_z \sim 0 \). For the right-moving particles, it is \( p^- \) which is big and \( p^+ \) which is very small.

Light cone variables scale by a constant under Lorentz transformations along the collision axis. Ratios of light cone momentum are therefore invariant under such Lorentz boosts. The light cone momentum fraction \( x = p_i^+ / P^+ \), where \( P^+ \) is that of the particle we probe and \( p_i^+ \) is that of the constituent of the probed hadron, satisfies \( 0 \leq x \leq 1 \). It is the same as Bjorken \( x \), and for a fast-moving hadron, it is almost Feynman \( x_{\text{Feynman}} = E_i / E \). This is the \( x \) variable one is using when one describes deep inelastic scattering. In this case the label \( i \) corresponds to a quark or gluon constituent of a hadron.

One can also define a rapidity variable:
\[ y = \frac{1}{2} \ln \left( \frac{p_i^+}{p_i^+} \right) \sim \ln(2E_i / M_T) . \] (20)
Up to mass effects, the rapidity is in the range \(-y_{\text{proj}} \leq y \leq y_{\text{proj}}\). When particles, like pions, are produced in high-energy hadronic collisions, one often plots them in terms of the rapidity variable. Distributions tend to be slowly varying functions of rapidity.

### 3.1 Is there simple behaviour at high energy?

A hint of the underlying simplicity of high-energy hadronic interactions comes from studying the rapidity distributions of produced particles for various collision energies. In Fig. 18, a generic plot of the rapidity distribution of produced pions is shown for two different energies. The rapidity distribution at lower energies has been cut in half and the particles associated with each of the projectiles have been displaced in rapidity so that their staring points in rapidity are the same. It is remarkable that, except for the slowest particles in the centre-of-mass frame, those located near \( y \sim 0 \), the distributions are almost identical [15]. This is shown for the data from RHIC in Fig. 19.

We conclude from this that going to higher energy adds in new degrees of freedom, the small-\( x \) part of the hadron wavefunction. At lower energies, these degrees of freedom are not kinematically relevant as they can never be produced. On the other hand, going to higher energy leaves the fast degrees of freedom of the hadron unchanged.

This suggests that there should be a renormalization group description of the hadrons. As we go to higher energy, the high-momentum degrees of freedom remain fixed. Integrating out the previous small-\( x \) degrees of freedom should incorporate them into what are now the high-energy degrees of freedom at the higher energy. This process generates an effective action for the new low-momentum degrees of freedom. Such a process, when done iteratively, is a renormalization group.
Fig. 18: The rapidity distributions of particles at two different energies

Fig. 19: Experimental evidence from the PHOBOS experiment at RHIC on limiting fragmentation

3.2 A single hadron

A plot of the rapidity distribution of the constituents of a hadron, the gluons, is shown generically in Fig. 20. I have used $y = y_{\text{hadron}} - \ln(1/x)$ as my definition of rapidity. This distribution is similar in shape to that of the half of the rapidity distribution shown for hadron–hadron interactions in the centre-of-mass frame which has positive rapidity. The essential difference is that this distribution is for constituents where the hadron–hadron collision is for produced particles, mainly pions.

In the high-energy limit, as discussed in the previous section, the density of gluons grows rapidly. This suggests we introduce a density scale for the partons

$$\Lambda^2 = \frac{1}{\pi R^2} \frac{dN}{dy}. \quad (21)$$

One usually defines a saturation momentum to be $Q_{\text{sat}}^2 \sim \alpha_s \Lambda^2$, since this will turn out to be the typical momentum of particles in this high-density system. In fact, $\alpha_s$ is slowly varying compared to the variation of $\Lambda$, so that for the purposes of the estimates we make here, whether or not there is a factor of $\alpha_s$ will not be so important. Note that $\alpha_s$ evaluated at the saturation scale will be $\alpha_s \ll 1$. The typical
The rapidity distribution of the constituents of a hadron particle transverse momenta are of order $p_T^2 \sim Q^2_{sat} \gg 1/R^2_{had}$. Therefore it is consistent to think of the parton distribution as a high-density, weakly coupled system which is localized in the transverse plane. The high-momentum partons, the degrees of freedom which should be frozen, can be thought of as sitting on an infinitesimally thin sheet. We shall study this system with a resolution size scale which is $\Delta x \ll 1/\Lambda_{QCD}$, so that we may use weak coupling methods. Such a thin sheet is shown in Fig. 21.

It is useful to discuss different types of rapidity variables before proceeding. The typical momentum space rapidity is

$$y = \frac{1}{2} \ln \left( \frac{p^+}{p^-} \right)$$
Here $M_T$ is a particle transverse mass, and we have made approximations which ignore overall shifts in rapidity by of order one unit. Within these approximations, the momentum space rapidity used to describe the production of particles is the same as that used to describe the constituents of hadrons.

Oftentimes a coordinate-space rapidity is introduced. With $\tau = \sqrt{t^2 - z^2}$,

$$y = \frac{1}{2} \ln \left( \frac{x^+}{x^-} \right) = \ln \left( 2 \tau / x^- \right).$$

(23)

Taking $\tau$ to be a time scale of order $1/M_T$, and using the uncertainty principle $x^+ \sim 1/p^-$, we find that up to shifts in rapidity of order one, all the rapidities are the same. This implies that coordinate space and momentum space are highly correlated, and that one can identify momentum-space and coordinate-space rapidity with some uncertainty of order one unit.

If we plot the distribution of particles in a hadron in terms of the rapidity variable, the longitudinal dimension of the sheet is spread out. This is shown in Fig. 22. The longitudinal position is correlated with the longitudinal momentum. The highest-rapidity particles are the fastest. In ordinary coordinate space, this means the fastest particles are those most Lorentz contracted. If we now look down a tube of transverse size $\Delta x \ll 1/\Lambda_{\text{QCD}}$, we intersect the various constituents of the hadron only occasionally. The colour charge probed by this tube should therefore be random, until the transverse size scale becomes large enough so that it can probe the correlations. If the beam energy is large enough, or $x$ is small enough, there should be a large amount of colour charge in each tube of fixed size $\Delta x$. One can therefore treat the colour charge classically.

The physical picture we have generated is that there should be classical sources of to a good approximation random charges on a thin sheet. The current for this is

$$J^\mu_a = \delta^\mu + \delta(x^-) \rho_a(x_T).$$

(24)

The delta function approximation should be good for many purposes, but it may also be useful in some circumstances to insert the longitudinal structure

$$J^\mu_a = \delta^\mu + \rho_a(y, x_T)$$

(25)

and to remember that the support of the source is for very large $y$. 

---

**Fig. 22:** The distribution of particles in a hadron in terms of rapidity variables
3.3 The colour glass condensate

We now know how to write down a theory to describe the colour glass condensate. It is given by the path integral
\[ \int [dA][dp] \exp (iS[A, \rho] - W[\rho]). \] (26)

Here \( S[A, \rho] \) is the Yang–Mills action in the presence of a source current as described above. The function \( W \) weights the various configurations of colour charge. In the simplest version of the colour glass condensate, this can be taken to be a Gaussian
\[ W = \frac{1}{2} \int dy d^2 x_T \frac{\rho^2(y, x_T)}{\mu^2(y)}. \] (27)

In this ansatz, \( \mu^2(y) \) is the colour charge squared density per unit area per unit \( y \) scaled by \( 1/N_c^2 - 1 \). The theory can be generalized to less trivial forms of the weight function, but this form works at small transverse resolution scales, \( \Delta x \ll 1/Q_{\text{sat}} \). As one increases the transverse resolution scale one needs a better determination of \( W \). It turns out that at resolution scales of order \( 1/Q_{\text{sat}} \ll \Delta x \ll 1/\Lambda_{\text{QCD}} \), a Gaussian form is still valid.

The averaging over an external field makes the theory of the colour glass condensate similar to that of spin glasses. The solutions of the classical field equations also have \( F^2 \sim 1/\alpha \), so the gluon fields are strong and have high occupation number, hence the word condensate.

The theory described above has an implicit longitudinal momentum cutoff scale. Particles with momentum above this scale are treated as sources, and those below it as fields. One computes physical quantities by first computing the classical fields and then averaging over sources \( \rho \). This is a good approximation so long as the longitudinal momentum in the field is not too far below the longitudinal momentum cutoff \( \Lambda^+ \). If one computes quantum corrections, the expansion parameter is
\[ \alpha_s \ln(\Lambda^+/p^+) . \] (28)

To generate a theory at smaller momenta \( \Lambda^+ \) one first requires that \( \alpha_s \ln(\Lambda^+/\Lambda^+) \ll 1 \). Then one computes the quantum corrections in the presence of the background field. This turns out to change only the weight function \( W \). Therefore the theory maps into itself under a change of scale. This is a renormalization group, and it determines the weight function \( W \) [6], [16, 17].

3.4 Colour glass fields

The form of the classical fields is easily inferred. On either side of the sheet the fields are zero. They have no time dependence, and in light cone gauge \( A^+ = 0 \). It is plausible to look for a solution which is purely transverse. On either side of the sheet, we have fields which are gauge transformations of zero field. It can be a different gauge transformation of zero field on different sides of the sheet. Continuity requires that \( F^{ij} = 0 \). \( F^{+} \) is zero because of light cone time \( x^+ \) independence, and the assumption that \( A^- = 0 \). \( F^{\pm} \) is non-zero \( \sim \delta(x^-) \) because of the variation in \( x^- \) as one crosses the sheet. This means that \( F^{00} \sim -F^{tz} \), or that \( E \perp B \perp \vec{z} \). These are transversely polarized Weiszacker–Williams fields. They are random in the two-dimensional plane because the source is random. This is shown in Fig. 13. The intensity of these fields is of order \( 1/\alpha_s \), and they are not at all stringlike.

3.5 The gluon distribution and saturation

The gluon distribution function is given by computing the expectation value of the number operator \( \langle a^\dagger(p)a(p) \rangle \) and can be found from computing the gluon field expectation value \( \langle A(p)A(-p) \rangle \). This is
left as an exercise for the student. At large $p_T$, the distribution function scales as
\[ \frac{dN}{dy d^2p_T} \sim \frac{1}{\alpha_s} \frac{Q_{\text{sat}}^2}{p_T^4}, \] (29)
which is typical of a bremsstrahlung spectrum. At small $p_T$, the solution is
\[ \sim \ln\left(\frac{Q_{\text{sat}}^2}{p_T^2}\right)/\alpha_s. \]
The reason for this softer behaviour at smaller $p_T$ is easy to understand. At small distances corresponding to large $p_T$, one sees point sources of charge, but at smaller $p_T$, the charges cancel one another and lead to a dipole field. The dipole field is less singular at large $r$, and when transformed into momentum space, one loses two powers of momentum in the distribution function. In terms of the colour field, the saturation phenomenon is almost trivial to understand. (It is very difficult to understand if the gluons are treated as incoherently interacting particles.)

Now $Q_{\text{sat}}^2$ can grow with energy. In fact it turns out that $Q_{\text{sat}}^2$ never stops growing. The intrinsic transverse momentum grows without bound. Physically what is happening is that the low-momentum degrees of freedom below the saturation momentum grow very slowly, like $\ln(Q_{\text{sat}}^2)$ because repulsive gluon interactions prevent more filling. On the other hand, one can always add more gluons at high momentum since the phase space is not filled there.

How is this consistent with unitarity? Unitarity is a statement about cross-sections at fixed $Q^2$. If $Q^2$ is above the saturation momentum, then the gluon distribution function grows rapidly with energy, as $Q_{\text{sat}}^2$. On the other hand, once the saturation momentum becomes larger than $Q^2$, the number of gluons one can probe
\[ xG(x,Q^2) \sim \pi R^2 \int_0^{Q^2} d^2p_T \frac{dN}{d^2p_T dy} \] (30)
varies only logarithmically. The number of gluons scales as the surface area. (At high $Q^2$, it is proportional to $R^2 Q_{\text{sat}}^2$, and one expects that $Q_{\text{sat}}^2 \sim A^{1/3}$ so that $xG(x,Q^2) \sim A$.)

### 3.6 Hadron collisions

In Fig. 16, the collision of two hadrons is represented as that of two sheets of coloured glass. Before the collisions, the left-moving hadron has fields
\[ F^{i+} \sim \delta(x^-) \]
\[ F^{ij} \sim 0 \]
\[ F^{i-} \sim 0 \] (31)
and that of the right-moving fields is analogous to that of the above save that $\pm \to \mp$ in the indices and coordinates of all fields. The fields are of course different in each nucleus. We shall consider impact-parameter-zero head-on collisions in what follows.

These fields are plane polarized and have random colours. A solution of the classical Yang–Mills equation can be constructed by requiring that the fields be two-dimensional gauge transforms of vacuum everywhere but in the forward light cone. At the edges of the light cone, and at its tip $t = z = 0$, the equations are singular, and a global solution requires that the fields carry non-trivial energy and momentum in the forward light cone. At short times, these fields are highly non-linear. In a time of order $\tau \sim 1/Q_{\text{sat}}$, the fields linearize. When they linearize, we can identify the particle content of the classical radiation field.

The solution to this classical field theory problem is the glasma. The fields have a discontinuity at $t = 0$ where the transverse fields are converted into longitudinal fields. These fields have non-zero $\vec{E} \cdot \vec{B}$ and, as described at the end of the last section, carry a non-zero topological charge. During the
time evolution of the glasma, there are both gluonic and quark degrees of freedom, and large coherent
colour electric and magnetic fields.

This situation is much different than the case for quantum electrodynamics. Because of the gluon
self-interaction, it is possible to classically convert the energy in the incident nuclei into radiation. In
quantum electrodynamics, the charged particles are fermions, and they cannot be treated classically. Ra-
diation is produced by annihilation or bremsstrahlung as quantum corrections to the initial value problem.

The solution to the field equation in the forward light cone is approximately boost invariant over
an interval of rapidity of order $\Delta y \ll 1/\alpha_s$. At large momentum, the field equations can be solved in
perturbation theory and the distribution is like that of a bremsstrahlung spectrum
\[
\frac{dN}{dy d^2p_T} \sim \frac{1}{\alpha_s} \pi R^2 \frac{Q_{sat}^4}{p_T^4}. \tag{32}
\]

It can be shown that such a spectrum matches smoothly onto the result for high-momentum-
transfer jet production.

One of the outstanding problems of particle production is computing the total multiplicity of pro-
duced gluons. In the CGC description, this problem is solved. When $p_T \leq Q_{sat}$, non-linearities of
the field equations become important, and the field stops going as $1/p_T^4$. Instead it becomes of order
\[
\frac{dN}{dy d^2p_T} \sim \frac{1}{\alpha_s} \pi R^2. \tag{33}
\]
The total multiplicity is therefore of order
\[
N \sim \frac{1}{\alpha_s} \pi R^2 Q_{sat}^2. \tag{34}
\]

If $Q_{sat}^2 \sim A^{1/3}$, then the total multiplicity goes as $A$, the high-$p_T$ differential multiplicity goes as
$A^{4/3}$, as we naively expect for hard processes since they should be incoherent, and the low-momentum
differential multiplicity goes as $A^{2/3}$, since these particles arise from the region where the hadrons are
black disks and the emission should take place from the surface.

In Fig. 23, the various kinematic regions for production of gluons are shown. In Fig. 24, the results

![Fig. 23: A cartoon representation of the various kinematic regions of gluon production](image)

of numerical simulation of gluon production are shown. At small $p_T$, it is amusing that the distribution
is well described by a two-dimensional Bose–Einstein distribution. This is presumably a numerical
accident, and in this computation has absolutely nothing to do with thermalized distributions.
3.7 Thermalization

As shown in Fig. 17, in a heavy-ion collision, the slow particles are produced first near the collision point and the slow particles are produced later far from the collision point. This introduces a gradient into the initial matter distribution, and the typical comoving volume element expands like $1/\tau$. To understand the factor of $1/\tau$ in the above equation, note that if we convert $dN/dz = dN/dy$, $dy/dz = dN/dy 1/t$, where we used our previous definition of space–time rapidity, and where we evaluated at $z = 0$. This is the physical rest frame density at $z = 0$.

If entropy is conserved, as is the case for a thermalized system with expansion time small compared to collision time,

$$S \sim T^3\tau R^2$$

is fixed so that $T \sim 1/\tau^{1/3}$. Therefore for a thermalized system, the energy density decreases as $\epsilon \sim 1/\tau^{4/3}$ for a system with no scattering so that the typical transverse momentum does not change, $\epsilon \sim 1/\tau$.

For the initial conditions typical of a colour glass condensate, thermalization is not so easy to do [18]. At the earliest times, the typical transverse momentum is large, of order of the saturation momentum. At this scale, the coupling is weak $\alpha_s(Q_{sat}) \ll 1$, at least for asymptotically large energy. It is believed by some that thermalization might be achieved in the glasma due to instabilities of the classical equations of motion [14]. The simple boost invariant solution to the classical equations described above is unstable if one allows small rapidity dependent fluctuations. These fluctuations grow with time, generating large fields. These fields strongly influence the motion of the gluon and quark degrees of freedom. A chaotic or turbulent fluid is generated, and this fluid might thermalize itself through the interactions of the unstable modes with the gluon and quark degrees of freedom.

The details of how this thermalization occurs have not been fully worked out in detail. Current estimates of the time of thermalization matter produced in heavy-ion collisions at RHIC energies range from $0.3 \leq \tau \leq 3$ fm/c.

4 Lecture III: What we have learned from RHIC

In this lecture, I review results from RHIC and describe what we have learned so far about the production of new forms of matter in heavy-ion collisions. I shall make the case that we have produced matter of extremely high energy density, so high that it is silly not to think of it as composed of quarks and gluons. I shall also argue that this matter is strongly interacting with itself. The issue of the properties of this matter is still largely unresolved. For example, whether the various quantities measured are more
properly described as arising from a colour glass condensate or from a quark–gluon plasma, although we can easily understand in most cases which form of matter should be most important.

The data presented here are taken from the RHIC whitepapers [15]. For references to the original publications, please look there.

4.1 The energy density is big

The particle multiplicity as a function of energy has been measured at RHIC, as shown in Fig. 25. Combining the multiplicity data together with the measurements of transverse energy or of typical particle transverse momenta, one can determine the energy density of the matter when it decouples [15]. One can then extrapolate backwards in time. We extrapolate backwards using one-dimensional expansion, since decoupling occurs when the matter first begins to expand three dimensionally. We can extrapolate backwards until the matter has melted from a colour glass.

To do this extrapolation we use that the density of particles falls as $\frac{N}{V} \sim \frac{1}{t}$ during one-dimensional expansion. If the particles expand without interaction, then the energy per particle is constant. If the particles thermalize, then $\frac{E}{N} \sim T$, and since $\frac{N}{V} \sim T^3$ for a massless gas, the temperature falls as $T \sim t^{-1/3}$. For a gas which is not quite massless, the temperature falls somewhere in the range $T_o > T > T_o(t_o/t)^{1/3}$, that is the temperature is bracketed by the value corresponding to no interaction and to that of a massless relativistic gas. This one-dimensional expansion continues until the system begins to feel the effects of finite size in the transverse direction, and then rapidly cools through three-dimensional expansion. Very close to when three-dimensional expansion begins, the system decouples and particles free-stream to detectors without further interaction. We shall take a conservative overestimate of this time to be of order $t_{melt} \sim 0.3 \text{ fm/c}$. The extrapolation backwards is bounded by $\epsilon_f(t_f/t) < \epsilon(t) < \epsilon_f(t_f/t)^{4/3}$. The lower bound is that assuming that the particles do not thermalize...
and their typical energy is frozen. The upper bound assumes that the system thermalizes as an ideal massless gas. We argued above that the true result is somewhere in between. When the time is as small as the melting time, then the energy density begins to decrease as time is further decreased.

This bound on the energy density is shown in Fig. 26. On the left axis is the energy density and on the bottom axis is time. The system begins as a colour glass condensate, then melts to quark–gluon matter which eventually thermalizes to a quark–gluon plasma. At a time of a few fm/c, the plasma becomes a mixture of quarks, gluons, and hadrons which expand together.

At a time of about 10 fm/c, the system falls apart and decouples. At a time of $t \sim 1$ fm/c, the estimate we make is identical to the Bjorken energy density estimate, and this provides a lower bound on the energy density achieved in the collision. (All estimates agree that by a time of order 1 fm/c, matter has been formed.) The upper bound corresponds to assuming that the system expands as a massless thermal gas from a melting time of 0.3 fm/c. (If the time was reduced, the upper bound would be increased yet further.) The bounds on the energy density are therefore

$$2–3 \text{ GeV/fm}^3 < \epsilon < 20–30 \text{ GeV/fm}^3$$

where we included a greater range of uncertainty in the upper limit because of the uncertainty associated with the formation time. The energy density of nuclear matter is about 0.15 GeV/fm$^3$, and even the lowest energy densities in these collisions is in excess of this. At late times, the energy density is about that of the cores of neutron stars, $\epsilon \sim 1$ GeV/fm$^3$.

At such extremely high energy densities, the matter is most simply described in terms of its quark and gluon degrees of freedom.

4.2 The gross features of multiplicity distributions are consistent with coloured glass

As argued by Kharzeev and Nardi [19], the density of produced particles per nucleon which participates in the collision $N_{\text{part}}$ should be proportional to $1/\alpha_s(Q_{\text{sat}})$, and $Q_{\text{sat}}^2 \sim N_{\text{part}}$. This dependence follows from the $1/\alpha_s$ which characterizes the density of the colour glass condensate. In Fig. 27, we show the
Fig. 27: The CGC description of the participant dependence of the multiplicity of produced particles

Fig. 28: Colour glass condensate fits to the rapidity density measured in the PHOBOS and BRAHMS experiments

data from PHENIX and PHOBOS [15]. The Kharzeev–Nardi form fits the data well. Other attempts such as HIJING [20], or the so-called saturation model of Eskola–Kajantie–Ruuskanen–Tuominen [21] are also shown in the figure. Kharzeev and Levin have recently argued that the rapidity distributions as a function of centrality can be computed from the colour glass description [22]. This is shown in Fig. 28.

4.3 The CGC describes features of deep inelastic scattering

The colour glass condensate provides a theory of the hadron wavefunction at very small values of $x$. As such, it should describe features not only of high-energy nucleus–nucleus scattering, but also electron–hadron scattering. This includes inclusive scattering and diffraction. It indeed appears that there is such a successful phenomenology [23].
In these notes, I shall describe only one aspect of this phenomenology, geometric scaling [24]–[27]. The basic idea is that the cross-section for virtual photon scattering from a hadron should be, up to some trivial overall scale factor, a dimensionless function. If the saturation momentum is the only scale in the problem and the properties of the matter probed depend only upon the density of the matter, then

\[ \sigma_{\gamma p} \sim F(Q^2/Q^2_{\text{sat}}(x)) \]  

(37)

and is not an independent function of \( x \) and \( Q^2 \). The dependence of the saturation momentum on \( x \) can be computed [28], or can be determined from data. In Fig. 29, this cross-section is plotted as a function of \( \tau = Q^2/Q^2_{\text{sat}} \) for values of \( x \leq 10^{-2} \). Indeed, there appears to be such scaling.

While it is easy to understand this scaling for \( Q^2 \leq Q^2_{\text{sat}} \), it is perhaps a little surprising that it works to much larger values of \( Q^2 \). One can show that one expects approximate scaling up to \( Q^2 \sim Q^4_{\text{sat}}/\Lambda^2_{\text{QCD}} \). However, one should and can compute scaling violations [27].

### 4.4 The CGC provides a theory of shadowing

The naive expectation for the production of hard particles from a nucleus is that they should be generated by incoherent scattering. This is, however, modified because of multiparticle scattering, and because the gluon distribution function itself acquires a non-trivial dependence upon the nuclear baryon number. The colour glass condensate provides a theory of this modification [29–34].
One can understand how this works by first considering the effects of multiple elastic scattering. Such scattering does not change the number of particles. At very high \( p_T \), the effects of multiple scattering should be small, since cross-sections are small. At intermediate \( p_T \), the \( p_T \) distribution in a nucleus should broaden relative to that of incoherent scattering from nucleons. By conservation of probability, this requires a suppression at low \( p_T \). By similar reasoning, one expects that these effects will be accentuated as one goes from peripheral to more central collisions. The results of one such computation of multiple scattering are shown in Fig. 30 [35].

One also expects that the effects of multiple scattering will be larger at small values of \( x \) because there are more degrees of freedom to scatter from.

Such multiple-scattering effects are included in the CGC description of the hadron collisions, but there is another effect which is larger at very small \( x \). This is the quantum evolution of the hadron wavefunction. Because the saturation momentum is larger in nuclei than it is in protons, it is more difficult to produce glue at small \( x \). Therefore as one goes to smaller values of \( x \), there should be fewer particles at small \( x \) relative to the expectation from incoherent scattering. In Fig. 31, \( p_T \) distributions as a function of \( x \) are shown for the ratio of hadron–nucleus collisions to incoherent scattering. At large values of \( x \) there is a clear Cronin enhancement. At small values of \( x \), there is a suppression as predicted by quantum evolution in the CGC. There is a similar suppression as the centrality of the collisions increases in distinction from the effects of multiple elastic scattering.

In the BRAHMS experiment, dAu collisions were used to study this effect. The results are shown in Fig. 32 [15]. Similar results have been found by STAR and PHENIX [15]. The results are qualitatively in accord with the CGC expectations, and also exhibit semi-quantitative agreement [36].

4.5 Matter has been produced which interacts strongly with itself

In off-zero-impact-parameter heavy-ion collisions, the matter which overlaps has an asymmetry in density relative to the reaction plane. This is shown in the left-hand side of Fig. 33. Here the reaction plane is along the \( x \) axis. In the region of overlap there is an \( x-y \) asymmetry in the density of matter which overlaps. This means that there will be an asymmetry in the spatial gradients which will eventually transmute itself into an asymmetry in the momentum space distribution of particles, as shown in the right-hand side of Fig. 33.

This asymmetry is called elliptic flow and is quantified by the variable \( v_2 \). In all attempts to theoretically describe this effect, one needs very strong interactions among the quarks and gluons at very early times in the collision [37]. In Fig. 34, two different theoretical descriptions are fit to the data by STAR and PHOBOS [15]. On the left-hand side, a hydrodynamical model is used [38]. It is roughly of the correct order of magnitude and has roughly the correct shape to fit the data. This was not the case at lower energy. On the right-hand side are preliminary fits assuming colour glass [39]. Again it has roughly the correct shape and magnitude to describe the data. In the colour glass, the interactions are very strong essentially from \( t = 0 \), but unlike the hydrodynamic models it is field pressure rather than particle pressure which converts the spatial anisotropy into a momentum space-anisotropy.

Probably the correct story for describing flow is complicated and will involve both the quark–gluon plasma and the colour glass condensate. Either description requires that matter be produced in the collisions and that it interact strongly with itself. In the limit of single scatterings for the partons in the two nuclei, no flow is generated.

Recent data on charm particles show that they too flow with the produced matter [40, 41]. Charm is a very heavy particle, and as such it requires many collisions with other particles before it can flow with the surrounding matter. The amount of flow seen experimentally exceeds the wildest expectations of theorists.
Fig. 30: The expectations of multiple scattering in dAu collisions in a multiple-scattering computation

Fig. 31: The $p_T$ distributions in hadron–nucleus collisions relative to incoherent scattering. Different curves correspond to different values of $x$. 

103
Fig. 32: The measurements from BRAHMS of the ratio of dAu high-\(p_T\) particle production to that of incoherent scattering as a function of \(x\) and centrality.

\[
V_2 = \langle \cos(2\phi) \rangle \quad \tan(\phi) = p_y/p_x
\]

Fig. 33: The asymmetry in the distribution of matter in an off-centre collision is converted to an asymmetry of the momentum space distribution.

Fig. 34: (a) A hydrodynamic fit to \(v_2\). (b) The coloured glass fit.
4.6 How strongly does the quark–gluon plasma interact?

4.6.1 Jets are quenched

One of the most interesting results from the RHIC experiments is the so-called ‘jet quenching’ [15], [42–45]. In Fig. 35(a), the single-particle hadron spectrum is scaled by the same result in pp collisions and scaled by the number of collisions. The number of collisions is the number of nucleon–nucleon interactions, which for central collisions should be almost all of the nucleons. One is assuming hard scattering in computing this number, so that a single nucleon can hard-scatter a number of times as it penetrates the other nucleus. The striking feature of this plot is that the ratio does not approach one at large $p_T$. This would be the value if these particles arose from hard scattering which produced jets of quarks and gluons, and the jets subsequently decayed.

The explanation for this is shown in Fig. 35(b). Here a pair of jets is produced in a hard collision and they propagate through the surrounding matter.

This suppression has been conclusively shown to be a final-state effect. One can measure the corresponding suppression in dA collisions, and at the central rapidity values where one sees strong suppression in jet production in AuAu collisions, there is little suppression or even enhancement seen in dA collisions. One can also look at the correlation in azimuthal angle of high-$p_T$-produced particles, as shown in Fig. 36. In pp collisions, if there is a high-$p_T$ particle produced, then at an azimuthal angle 180 degrees away, one expects to see an excess of hard particles. This was done and verified in STAR. In central AuAu collisions, one looks in the backward direction, and the peak has disappeared.

The essential problem with jet quenching is that it is much stronger than expected from naive QCD computation. Jet quenching apparently persists out to 20 GeV. For charmed particles, the observed jet quenching defies realistic QCD computations.

One of the reasons why jet quenching is so important for the RHIC programme is that it gives a good measure of the degree of thermalization in the collisions. If jets are strongly quenched by transverse momenta of 4 GeV, then because cross-sections go like $1/E^2$ for quarks and gluons, this would be strong evidence for thermalization at the lower energies typical of the emitted particles.

4.6.2 The matter flows and is well approximated by perfect fluid hydrodynamics

One can look for evidence of thermalization directly from the measured $p_T$ distributions. Here one can do a hydrodynamic computation and in so far as it agrees with the results, one is encouraged to believe that there is thermalization. On the other hand, these distributions may have their origin in the
initial conditions for the collision, the coloured glass. In reality, one will have to understand the tradeoff between both effects. The hydrodynamic models do a good job in describing the data for $p_T \leq 2$ GeV. Here there is approximate $m_T$ scaling, a characteristic feature of hydrodynamic computations. This scaling arises naturally because hydrodynamic distributions are produced by flowing matter which has a characteristic transverse flow velocity with a well-defined local temperature. Particles with the same $m_T$ should have arisen from regions with the same transverse-flow velocity and temperature.

Hydrodynamical models successfully describe the data on $m_T$ distributions [46]. In Fig. 37 the results of the simulation by Shuryak and Teaney are shown compared to the STAR and PHENIX data [15]. The shape of the curve is a prediction of the hydrodynamic model, and is parametrized somewhat by the nature of the equation of state. Notice that the STAR data include protons near threshold, and here the $m_T$ scaling breaks down. The hydrodynamic fits get this dependence correctly, and this is one of the best tests of this description. The hydrodynamic models do less well on fits to the more peripheral collisions. In general, a good place to test the hydrodynamic models’ predictions is with massive particles close to

---

**Fig. 36:** The forward–backward correlation for high-$p_T$ particles produced in STAR

**Fig. 37:** The hydrodynamical model fits to the $m_T$ spectra for the PHENIX and STAR data
threshold, since here one deviates in a computable way from the $m_T$ scaling curve, which is arguably determined from parametrizing the equation of state.

If one can successfully argue that there is thermalization in the RHIC collisions, then the hydrodynamic computations become compelling. One should remember that hydrodynamics requires an equation of state plus initial conditions, and these initial conditions are determined by coloured glass. Presumably, a correct description will require the inclusion of both types of effects [47].

At present, hydrodynamical models do an excellent job of describing data on distributions of particles with $p_T \leq 2$ GeV. This uses perfect fluid hydrodynamics with no viscosity. This was not the case at CERN [48]. Estimates of the viscosity which is consistent with the experimental data give numbers which are quite small, leading some to conclude that the quark–gluon plasma is the most perfect fluid yet measured. There are of course some uncertainties in these conclusions, largely associated with the initial conditions for the hydrodynamic equations, uncertainty in the equation of state, and dispersion in the treatment of the matter at late times when the hydrodynamic description must break down. Nevertheless, the fact that the hydrodynamic computations seem to work well, and the existence of strong jet quenching, lead me to conclude that at a minimum, the matter produced is reasonably well approximated as a thermal system, and is remarkably strongly self-interacting. This means that I believe that the semi-quantitative conclusions drawn from hydrodynamic simulation have substance.

4.7 Confinement and chiral symmetry restoration

We would like to know whether or not deconfinement has occurred in dense matter or whether chiral symmetry restoration has changed particle masses.

![Fig. 38: The CERES data on low-mass electron–positron pairs. The expected contribution from ordinary hadrons is shown by the solid line. The data points are for the measured electron–positron pairs.](image)

This can be studied in principle by measuring the spectrum of dileptons emitted from the heavy-ion collision. These particles probe the interior of the hot matter since electromagnetically interacting particles are not significantly attenuated by the hadronic matter. For electron–positron pairs, the mass distribution has been measured in the CERES experiment at CERN [49], and is shown in Fig. 38. Shown
in the plot is the distribution predicted from extrapolating from pA collisions. There should be a clear $\rho$ and $\phi$ peak, which has disappeared. This has been interpreted as a resonance mass shift [50], enhanced $\eta'$ production [51], but is most probably collisional broadening of the resonances in the matter produced in the collisions [52]. Nevertheless, if one makes a plot such as this and the energy density is very high and there are no resonances at all, then this would be strong evidence that the matter is not hadronic, i.e., the hadrons have melted.

The resolution in the CERES experiment is unpleasantly large, making it difficult to unambiguously interpret the result. Whether or not such an experiment could be successfully run at RHIC, much less whether the resolution could be improved, is the subject of much internal debate among the RHIC experimentalists.

It has also been pointed out recently that the matter in the early stages after the collisions has remarkable properties [13]. At the earliest times, there is both longitudinal colour electric and colour magnetic fields. These fields evolve towards a thermalized system as the longitudinal fields evaporate into gluons. I call this early matter the glasma. The fields have a non-zero colour $\vec{E} \cdot \vec{B}$. This is an unusual situation and generates an anomaly in the axial vector current. This means that even very energetic quarks will flip their helicities in the presence of such fields, and generate chiral symmetry breaking. It has been conjectured that such helicity flip may ultimately be responsible for mass generation in QCD. The idea of the glasma with its anomalous fields is recent, and it is not yet possible to assess the experimental implications.

### 4.8 Confinement and $J/\Psi$ suppression

The NA(50) data for $J/\Psi$ production is shown in Fig. 39 [53]. In the first figure, the ratio of $J/\Psi$ production cross-section to that of Drell–Yan is shown as a function of $E_T$, the transverse energy, for the lead–lead collisions at CERN. There is a clear suppression at large $E_T$ which is greater than the hadronic absorption model calculations which are plotted with the data [54]. In the next figure, the theoretically expected $J/\Psi$ suppression based on hadronic absorption models is compared to that measured as a function of the Bjorken energy density for various targets and projectiles. There appears to be a sharp increase in the amount of suppression for central lead–lead collisions.

![Fig. 39:](image-url) (a) The ratio of produced $J/\Psi$ pairs to Drell–Yan pairs as a function of transverse energy $E_T$ at CERN energy. (b) The measured compared to the theoretically expected $J/\Psi$ suppression as a function of the Bjorken energy density for various targets and projectiles.

absorption model calculations which are plotted with the data [54]. In the next figure, the theoretically expected $J/\Psi$ suppression based on hadronic absorption models is compared to that measured as a function of the Bjorken energy density for various targets and projectiles. There appears to be a sharp increase in the amount of suppression for central lead–lead collisions.

---

L. McLerran

108
Is this suppression due to Debye screening of the confinement potential in a high-density quark–gluon plasma [55–57]. Might it be due to higher twists, enhanced rescattering, or changes in the gluon distribution function [58, 59]. Might the $J/\psi$ suppression be changed into an enhancement at RHIC energies and result from the recombination in the produced charm particles, many more of which are produced at RHIC energy [60–63].

These various descriptions can be tested by using the capability at RHIC to do pp and pA as well as AA collisions. Issues related to multiple scattering, higher twist effects, and changes in the gluon distribution function can be explored. A direct measurement of open charm will be important if final-state recombination of the produced open charm makes a significant amount of $J/\Psi$’s.

The data from the PHENIX experiment show roughly the same pattern of suppression as seen at CERN [64]. This is a surprise since one naively expected that there should be more suppression at higher energy densities. This had led some to speculate that there may be significant recombination effects in the final state [60], [63]. This will be resolved after measurements of resonant states decaying into charm, $J/\Psi$ flow, and more, as the programme at RHIC continues.

### 4.8.1 Direct photons

One of the first suggestions for a signal of the quark–gluon plasma was thermal radiation due to photons [65–69]. Produced photons are penetrating, and in principle can measure the properties of the hot matter at early times in the collision. The problem is that there are huge backgrounds from resonance decays.

At very large $p_T$, the resonance backgrounds are small, but there the dominant process for making photons is hard scattering of a gluon and a quark, and reflects the initial parton distribution functions. This has been measured at RHIC, and agrees with perturbative computations. More recently, there has been a claimed measurement from PHENIX of an excess seen at intermediate $p_T$ [70].

The excess is surely interesting, but it is an excess relative to a theoretical computation, and these computations must be checked against the pp and dA data before one can be too excited about the result seen in AA. Of course, one has to check against all possible sources of photons for AA before one concludes that these photons arise from thermal radiation. As the results are new, these checks are not yet complete.

### 4.8.2 The lifetime and size of the matter produced

The measurement of correlated pion pairs, the so-called HBT pion interferometry, can measure properties of the space–time volume from which the hadronic matter emerges in heavy-ion collisions [71]. The quantities $R_{long}$, $R_{side}$ and $R_{out}$ shown in Fig. 40 measure the transverse size of the matter at decoupling and the decoupling time.

![Fig. 40: The various radii used for HBT pion interferometry](image-url)
In Fig. 41, the data from STAR and PHENIX are shown [15]. There is only a weak dependence on energy, and the results seem to be more or less what was observed at CERN energies. This is a surprise, since a longer time for decoupling is expected at RHIC. Perhaps the most surprising result is that $R_{\text{out}}/R_{\text{side}}$ is close to 1, whereas most theoretical expectations give a value of about $R_{\text{out}}/R_{\text{side}} \sim 2$ [72, 73]. Perhaps this is due to greater than expected opacity of the emitting matter? At this time, there is no consistent theoretical description of the HBT data at RHIC. Is there something missing in our space–time picture?

![Fig. 41](image-url)

**Fig. 41:** (a) The various HBT radii measured in heavy-ion experiments including the new data from STAR. (b) The correlation functions which determine the radii as a function of the pair momenta measured in PHENIX.

### 4.8.3 The flavour composition of the quark–gluon plasma

The first signal proposed for the existence of a quark–gluon plasma in heavy-ion collisions was enhanced strangeness production [74]. This has led to a comprehensive programme in heavy-ion collisions to measure

the ratios of abundances of various flavours of particles [75]. In Fig. 42(a), the ratios of flavour abundances are compared to a thermal model for the particle abundances [76–78]. The fit is quite good. In Fig. 42(b), the temperature and baryon chemical potential extracted from these fits is shown for experiments at various energies and with various types of nuclei. It seems to agree reasonably well with what might be expected for a phase boundary between hadronic matter and a quark–gluon plasma.
This would appear to be a compelling case for the production of a quark–gluon plasma. The problem is that the fits done for heavy ions to particle abundances work even better in $e^+ e^-$ collisions. One definitely expects no quark–gluon plasma in $e^+ e^-$ collisions. There is a deep theoretical question to be understood here: How can thermal models work so well for non-thermal systems? Is there some simple saturation of phase space? The thermal model description can eventually be made compelling for heavy-ion collisions once the degree of thermalization in these collisions is understood.

Acknowledgements

I gratefully acknowledge conversations with Dima Kharzeev, Robert Pisarski and Raju Venugopalan on the subject of this talk.

This manuscript has been authorized under Contract No. DE-AC02-98H10886 with the U.S. Department of Energy.

References


[42] J. D. Bjorken, FERMILAB-PUB-82-059-THY.


Flavour dynamics and CP violation in the Standard Model: A crucial past—and an essential future

I.I. Bigi
University of Notre Dame, Notre Dame, IN 46556, U.S.A.

Abstract
Our knowledge of flavour dynamics has undergone a ‘quantum jump’ since just before the turn of the millenium: direct CP violation was firmly established in $K_L \to \pi\pi$ decays in 1999; the first CP asymmetry outside $K_L$ decays was discovered in 2001 in $B_d \to \psi K_S$, followed by $B_d \to \pi^+\pi^-$, $\eta' K_S$ and $B \to K^\pm\pi^\mp$, the last one establishing direct CP violation also in the beauty sector. Furthermore CKM dynamics allows a description of CP insensitive and sensitive $B$, $K$ and $D$ transitions that is impressively consistent also on the quantitative level. Theories of flavour dynamics that could serve as alternatives to CKM were ruled out. Yet these novel successes of the Standard Model (SM) do not invalidate any of the theoretical arguments for the incompleteness of the SM. In addition we have also more direct evidence for New Physics, namely neutrino oscillations, the observed baryon number of the Universe, dark matter and dark energy. While the New Physics anticipated at the TeV scale is not likely to shed any light on the SM’s mysteries of flavour, detailed and comprehensive studies of heavy flavour transitions will be essential in diagnosing salient features of that New Physics. Strategic principles for such studies will be outlined.

In my lecture series I shall sketch the past evolution of central concepts of the Standard Model (SM), which are of particular importance for its flavour dynamics. The reason is not primarily of a historical nature. I hope these sketches will illuminate the main message I want to convey, namely that we find ourselves in the midst of a great intellectual adventure: even with the recent novel successes of the SM the case for New Physics at the TeV (and at higher scales) is as strong as ever. While there is a crowd favourite for the TeV scale New Physics, namely some implementation of Supersymmetry (SUSY)—an expectation I happen to share—we should allow for many diverse scenarios. To deduce which one is realized in Nature we shall need all the experimental information we can get, including the impact of the New Physics on flavour dynamics. Yet based on the present successes of the SM, we cannot count on that impact being numerically massive. I shall emphasize general principles for designing search strategies for New Physics over specific and detailed examples. For at a school like this we want to help you prepare yourself for a future leadership role; that requires that you do your own thinking rather than ‘outsource’ it.

The outline of my three lectures is as follows:

- **Lecture I: Flavour dynamics in the second millenium** (→ 1999) – Basics of flavour dynamics and CP violation, CKM theory, $K^0$ and $B^0$ oscillations, the SM ‘Paradigm of large CP violation in $B$ decays’.
- **Lecture II: Flavour dynamics 2000–2006** – Verifying the SM ‘Paradigm of large CP violation in $B$ decays’, praising EPR correlations and hadronization, heavy quark theory, extracting CKM parameters and CKM triangle fits.
- **Lecture III: Probing the flavour paradigm of the emerging new Standard Model** – Indirect searches for New Physics, ‘King Kong’ scenarios (EDMs, charm, $\tau$ leptons) vs. precision probes (beauty), the case for a super-flavour factory and a new generation of kaon experiments in HEP’s future landscape.
To a large degree I shall follow the historical development, because it demonstrates best why it is advantageous to listen to predictions from theory—but also go against it at times!

1 Lecture I: Flavour dynamics in the second millennium (→ 1999)

Memento $\Delta S \neq 0$ dynamics:

- The ‘$\theta$ – $\tau$ puzzle’—the observation that two particles decaying into final states of opposite parity ($\theta \rightarrow 2\pi$, $\tau \rightarrow 3\pi$) exhibited the same mass and lifetime—led to the realization that parity was violated in weak interactions, and actually to a maximal degree in charged currents.
- The observation that the production rate of strange hadrons exceeded their decay rates by many orders of magnitude—a feature that gave rise to the term ‘strangeness’—was attributed to ‘associate production’ meaning the strong and electromagnetic forces conserve this new quantum number ‘strangeness’, while weak dynamics do not. Subsequently it gave rise to the notion of quark families.
- The great suppression of flavour changing neutral currents as evidenced by the tiny rates for $K_L \rightarrow \mu^+\mu^-$, $\gamma\gamma$ and the minute size for $\Delta M_K$, led some daring spirits to postulate the existence of a new quantum number for quarks, namely charm.
- The observation of $K_L \rightarrow \pi^+\pi^-$ established that CP invariance was not fully implemented in nature and induced two other daring spirits to postulate the existence of yet another, the third, quark family, with the top quark, as we learnt later, being two hundred times heavier than kaons.

All these features, which are pillars of the Standard Model now, represented ‘New Physics’ then!

1.1 On the uniqueness of the SM

A famous American Football coach once declared: “Winning is not the greatest thing—it is the only thing!” This quote provides some useful criteria for sketching the status of the different components of the Standard Model (SM). It can be characterized by the carriers of its strong and electroweak forces that are described by gauge dynamics and the mass matrices for its quarks and leptons as follows:

$$\text{SM}^* = SU(3)_C \times SU(2)_L \times U(1) \oplus '\text{CKM}'(\oplus '\text{PMNS}')$$ \hspace{1cm} (1)

I have attached the asterisk to ‘SM’ to emphasize that the SM contains a very peculiar pattern of fermion mass parameters that is not illuminated at all by its gauge structure. Next I shall address the status of these components.

1.1.1 QCD— the ‘only’ thing

1.1.1.1 ‘Derivation’ of QCD

While it is important to subject QCD again and again to quantitative tests as the theory for the strong interactions, one should note that these serve more as tests of our computational control over QCD dynamics than of QCD itself. For its features can be inferred from a few general requirements and basic observations. A simplified list reads as follows:

- Our understanding of chiral symmetry as a spontaneously realized one—which allows treating pions as Goldstone bosons implying various soft pion theorems—requires vector couplings for the gluons.
- The ratio $R = \sigma(e^+e^- \rightarrow \text{had.})/\sigma(e^+e^- \rightarrow \mu^+\mu^-)$ and the branching ratios for $\pi^0 \rightarrow \gamma\gamma$, $\tau^- \rightarrow e^-\bar{\nu}_e\nu_\tau$ and $B \rightarrow l\nu X_c$ point to the need for three colours.

- Colour has to be implemented as an unbroken symmetry. Local gauge theories are the only known way to couple massless spin-one fields in a Lorentz invariant way. The basic challenge is easily stated: $4 \neq 2$; i.e., while Lorentz covariance requires four component to describe a spin-one field, the latter contains only two physical degrees of freedom for massless fields. (For massive vector fields one can go to their rest frame to reduce and project out one component in a Lorentz invariant way to arrive at the three physical degrees of freedom.)

- Combining confinement with asymptotic freedom requires a non-Abelian gauge theory.

In summary: for describing the strong interactions QCD is the unique choice among local quantum field theories. A true failure of QCD would thus create a genuine paradigm shift, for one had to adopt an intrinsically non-local description. It should be remembered that string theory was first put forward for describing the strong interactions.

### 1.1.1.2 ‘Fly in the ointment’: the strong CP problem of QCD

A theoretical problem arises for QCD from an unexpected quarter that is very relevant for our context here: QCD does not automatically conserve P, T and CP. To reflect the nontrivial topological structure of QCD’s ground state one employs an effective Lagrangian containing an additional term to the usual QCD Lagrangian [1]:

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_{\text{QCD}} + \theta \frac{g_\pi^2}{32\pi^2} G_{\mu\nu} \tilde{G}^{\mu\nu}, \quad \tilde{G}_{\mu\nu} = \frac{i}{2} \epsilon_{\mu\nu\rho\sigma} G^{\rho\sigma}. \quad (2)$$

Since $G_{\mu\nu} \tilde{G}^{\mu\nu}$ is a gauge-invariant operator, its appearance in general cannot be forbidden, and what is not forbidden has to be considered allowed in a quantum field theory. It represents a total divergence, yet in QCD—unlike in QED—it cannot be ignored on account of the topological structure of the ground state.

Since under parity P and time reversal T one has

$$G_{\mu\nu} \tilde{G}^{\mu\nu} \rightarrow -G_{\mu\nu} \tilde{G}^{\mu\nu}, \quad (3)$$

the last term in Eq. (2) violates P as well as T. Since $G_{\mu\nu} \tilde{G}^{\mu\nu}$ is flavour-diagonal, it generates an electric dipole moment (EDM) for the neutron. From the upper bound on the latter

$$d_N < 0.63 \cdot 10^{-25} \text{ e cm} \quad (4)$$

one infers [1]

$$\theta < 10^{-9}. \quad (5)$$

Being the coefficient of a dimension-four operator, $\theta$ can be renormalized to any value, even zero. Yet the modern view of renomalization is more demanding: requiring the renormalized value to be smaller than its ‘natural’ one by orders of magnitude is frowned upon, since it requires finetuning between the loop corrections and the counterterms. This is what happens here. For purely within QCD the only intrinsically ‘natural’ scale for $\theta$ is unity. If $\theta \sim 0.1$ or even 0.01 were found, one would not be overly concerned. Yet the bound of Eq. (5) is viewed with great alarm as very unnatural—unless a symmetry can be called upon. If any quark were massless—most likely the $u$ quark—chiral rotations representing symmetry transformations in that case could be employed to remove $\theta$ contributions. Yet a considerable phenomenological body rules against such a scenario.

A much more attractive solution would be provided by transforming $\theta$ from a fixed parameter into the manifestation of a dynamical field—as is done for gauge and fermion masses through the Higgs–Kibble mechanism, see below—and imposing a Peccei–Quinn symmetry would lead naturally to $\theta \ll \ldots$
Alas—this attractive solution does not come ‘for free’: it requires the existence of axions. Those have not been observed despite great efforts to find them.

This is a purely theoretical problem. Yet I consider the fact that it remains unresolved a significant chink in the SM’s armour. I still have not given up hope that ‘victory can be snatched from the jaws of defeat’: establishing a Peccei–Quinn-type solution would be a major triumph for theory.

1.1.1.3 Theoretical technologies for QCD

True theorists tend to think that by writing down, say, a Lagrangian one has defined a theory. Yet to make contact with experiment one needs theoretical technologies to infer observable quantities from the Lagrangian. That is the task that engineers and plumbers like me have set for themselves. Examples for such technologies are

– perturbation theory;
– chiral perturbation;
– QCD sum rules;
– heavy quark expansions (which will be described in some detail in Lecture II).

Except for the first one they incorporate the treatment of nonperturbative effects.

None of these can claim universal validity; i.e., they are all ‘protestant’ in Nature. There is only one ‘catholic’ technology, namely lattice gauge theory\(^1\):

– It can be applied to nonperturbative dynamics in all domains (with the possible practical limitation concerning strong final-state interactions).
– Its theoretical uncertainties can be reduced in a systematic way.

Chiral perturbation theory is QCD at low energies describing the dynamics of soft pions and kaons. The heavy quark expansions treating the nonperturbative effects in heavy flavour decays through an expansion in inverse powers of the heavy quark mass are tailor-made for describing B decays; to which degree their application can be extended down to the charm scale is a more ‘iffy’ question. Different formulations of lattice QCD can approach the nonperturbative dynamics at the charm scale from below as well as from above. The degree to which they yield the same results for charm provides an essential cross-check on their numerical reliability. In that sense the study of charm decays serves as an important bridge between our understanding of nonperturbative effects in heavy and light flavours.

1.1.2 \(SU(2)_L \times U(1) – \text{not even the greatest thing}\)

1.1.2.1 Prehistory

It was recognized from early on that the four-fermion coupling of Fermi’s theory for the weak forces yields an effective description only that cannot be extended to very high energies. The lowest order contribution violates unitarity around 250 GeV. Higher order contributions cannot be called upon to remedy the situation, since owing to the theory being non-renormalizable those come with more and more untamable infinities. Introducing massive charged vector bosons softens the problem, yet does not solve it. Consider the propagator for a massive spin-one boson carrying momentum \(k\):

\[
\frac{-g_{\mu\nu} + \frac{k_\mu k_\nu}{M_W^2}}{k^2 - M_W^2}.
\]  

(6)

The second term in the numerator has great potential to cause trouble. For it can act like a coupling term with dimension \(1/(\text{mass})^2\); this is quite analogous to the original ansatz of Fermi’s theory and amounts

\(^1\) I hasten to add that lattice gauge theory—while catholic in substance—exhibits a different sociology: it has not developed an inquisition and deals with heretics in a rather gentle way.
to a non-renormalizable coupling. It is actually the longitudinal component of the vector boson that is at the bottom of this problem.

This potential problem is neutralized if these massive vector bosons couple to conserved currents. To guarantee the latter property, one needs a non-Abelian gauge theory, which implies the existence of neutral weak currents.

1.1.2.2 Strong points

The requirements of unitarity, which is nonnegotiable, and of renormalizability, which is to some degree, severely restrict possible theories of the electroweak interactions. It makes the generation of mass a highly nontrivial one, as sketched below. There are other strong points as well:

- Since there is a single $SU(2)_L$ group, there is a single set of gauge bosons. Their self-coupling controls also, how they couple to the fermion fields. As explained later in more detail, this implies the property of ‘weak universality’.
- The SM truly predicted the existence of neutral currents characterized by one parameter, the weak angle $\theta_W$, and the masses of the $W$ and $Z$ bosons.
- Most remarkably the $SU(2)_L \times U(1)$ gauge theory combines QED with a pure parity-conserving vector coupling to a massless neutral force field with the weak interactions, where the charged currents exhibit maximal parity violation due to their $V-A$ coupling and a very short range due to $M_Z > M_W \gg m_\pi$.

1.1.2.3 Generating mass

A massive spin-one field with momentum $k_\mu$ and spin $s_\mu$ has four (Lorentz) components. Going into its rest frame one realizes that the Lorentz-invariant constraint $k \cdot s = 0$ can be imposed, which leaves three independent components, as it has to be.

A massless spin-one field is still described by four components, yet has only two physical degrees of freedom. It needs another physical degree of freedom to transform itself into a massive field. This is achieved by having the gauge symmetry realized spontaneously. For the case at hand this is implemented through an ansatz that should be—although rarely is—referred to as Higgs–Brout–Englert–Guralnik–Hagen–Kibble mechanism (HBEGHK). Suffice it to consider the simplest case of a complex scalar field $\phi$ with a potential invariant under $\phi(x) \to e^{i\alpha(x)}\phi(x)$, since this mechanism has been described in great detail in Pich’s lectures [2]:

$$V(\phi) = \lambda|\phi|^4 - \frac{m^2}{2}|\phi|^2.$$ (7)

Its minimum is obviously not at $|\phi| = 0$, but at $\sqrt{m^2/4\lambda}$. Thus rather than having a unique ground state with $|\phi| = 0$ one has an infinity of different, yet equivalent ground states with $|\phi| = \sqrt{m^2/4\lambda}$. To understand the physical content of such a scenario, one considers oscillations of the field around the minimum of the potential: oscillations in the radial direction of the field $\phi$ represent a scalar particle with mass; in the polar direction (i.e. the phase of $\phi$) the potential is at its minimum, i.e., flat, and the corresponding field component constitutes a massless field.

It turns out that this massless scalar field can be combined with the two transverse components of a $M = 0$ spin-one gauge field to take on the role of the latter’s longitudinal component leading to the emergence of a massive spin-one field. Its mass is thus controlled by the nonperturbative quantity $\langle 0|\phi|0 \rangle$.

Applying this generic construction to the SM one finds that a priori both $SU(2)_L$ doublet and triplet Higgs fields could generate masses for the weak vector bosons. The ratio observed for the $W$ and $Z$ masses is fully consistent with only doublets contributing. Intriguingly enough such doublet fields can eo ipso generate fermion masses as well.
In the SM one adds a single, complex, scalar doublet field to the mix of vector boson and fermion fields. Three of its four components slip into the role of the longitudinal components of $W^\pm$ and $Z^0$; the fourth one emerges as an independent physical field—`the' Higgs field. Fermion masses are then given by the product of the single vacuum expectation value (VEV) $\langle 0 | \phi | 0 \rangle$ and their Yukawa couplings—a point we shall return to.

1.1.2.4 Triangle or ABJ anomaly

The diagram with an internal loop of only fermion lines, to which three external axial vector (or one axial vector and two vector) lines are attached, generates a `quantum anomaly'\(^2\): it removes a classical symmetry as expressed through the existence of a conserved current. In this specific case it affects the conservation of the axial vector current $J_5^\mu$. Classically we have $\partial_\mu J_5^\mu = 0$ for massless fermions; yet the triangle anomaly leads to

$$\partial_\mu J_5^\mu = \frac{g_2^2}{16\pi^2} G \cdot \tilde{G} \neq 0$$

even for massless fermions; $G$ and $\tilde{G}$ denote the gluonic field strength tensor and its dual, respectively, as introduced in Eq. (2).

While by itself it yields a finite result on the right-hand side of Eq. (8), it destroys the renormalizability of the theory. It cannot be `renormalized away’ (since in four dimensions it cannot be regularized in a gauge invariant way). Instead it has to be neutralized by requiring that adding up this contribution from all types of fermions in the theory yields a vanishing result.

For the SM this requirement can be expressed very concisely—all electric charges of the fermions of a given family have to add up to zero. This imposes a connection between the charges of quarks and leptons, yet does not explain it.

1.1.2.5 Theoretical deficiencies

With all the impressive, even amazing successes of the SM, it is natural to ask why the community is not happy with it. There are several drawbacks:

– Since the gauge group is $SU(2)_L \times U(1)$, only partial unification has been achieved.
– The HBEHK mechanism is viewed as providing merely an ‘engineering’ solution, in particular since the physical Higgs field has not been observed yet. Even if or when it is, theorists in particular will not feel relieved, since scalar dynamics induce quadratic mass renormalization and are viewed as highly ‘unnatural’, as exemplified through the gauge hierarchy problem. This concern has led to the conjecture of New Physics entering around the TeV scale, which has provided the justification for the LHC and the motivation for the ILC.
– maximal violation of parity is implemented for the charged weak currents ‘par ordre du mufti’\(^3\), i.e., based on the data with no deeper understanding.
– Likewise neutrino masses had been set to zero ‘par ordre du mufti’.
– The observed quantization of electric charge is easily implemented and is instrumental in neutralizing the triangle anomaly—yet there is no understanding of it.

One might say these deficiencies are merely ‘warts’ that hardly detract from the beauty of the SM. Alas—there is the whole issue of family replication!

---

\(^2\)It is referred to as ‘triangle’ anomaly owing to the form of the underlying diagram or A(dler) B(ell) J(ackiw) anomaly due to the authors that identified it [3].

\(^3\)A French saying describing a situation where a decision is imposed on someone with no explanation and no right of appeal.
1.1.3 The family mystery

The twelve known quarks and leptons are arranged into three families. Those families possess identical gauge couplings and are distinguished only by their mass terms, i.e., their Yukawa couplings. We do not understand this family replication or why there are three families. It is not even clear whether the number of families represents a fundamental quantity or is due to the more or less accidental interplay of complex forces one encounters when analysing the structure of nuclei. The only hope for a theoretical understanding we can spot on the horizon is superstring or M theory—which is merely a euphemistic way of saying we have no clue.

Yet the circumstantial evidence that we miss completely a central element of Nature’s ‘Grand Design’ is even stronger in view of the strongly hierarchical pattern in the masses for up- and down-type quarks, charged leptons and neutrinos and the CKM parameters as discussed later.

1.2 Basics of P, C, T, CP and CPT

1.2.1 Definitions

Parity transformations flip the sign of position vectors \( \vec{r} \) while leaving the time coordinate \( t \) unchanged:

\[
(\vec{r}, t) \xrightarrow{P} (-\vec{r}, t) .
\]  

(9)

Momenta change their signs as well, yet orbital and other angular momenta do not:

\[
\vec{p} \xrightarrow{P} -\vec{p} \quad \text{vs.} \quad \vec{l} \equiv \vec{r} \times \vec{p} \xrightarrow{P} \vec{l} .
\]  

(10)

Parity odd vectors—\( \vec{r} \), \( \vec{p} \)—and parity even ones—\( \vec{l} \)—are referred to as polar and axial vectors, respectively. Likewise one talks about scalars \( S \) and pseudoscalars \( P \) with \( S \xrightarrow{P} S \) and \( P \xrightarrow{P} -P \). Examples are \( S = \vec{p}_1 \cdot \vec{p}_2, \vec{l}_1 \cdot \vec{l}_2 \) and \( P = \vec{l}_1 \cdot \vec{p}_2 \). Parity transformations are equivalent to mirror transformations followed by a rotation. They are described by a linear operator \( P \).

Charge conjugation exchanges particles and antiparticles and thus flips the sign of all charges like electric charge, hyper-charge etc. It is also described by a linear operator \( C \).

Time reversal is operationally defined as a reversal of motion

\[
(\vec{p}, \vec{l}) \xrightarrow{T} - (\vec{p}, \vec{l}) ,
\]  

(11)

which follows from \( (\vec{r}, t) \xrightarrow{T} (\vec{r}, -t) \). While the Euclidean scalar \( \vec{l}_1 \cdot \vec{p}_2 \) is invariant under the time reversal operator \( T \), the triple correlations of (angular) momenta are not:

\[
\vec{v}_1 \cdot (\vec{v}_2 \times \vec{v}_3) \xrightarrow{T} -\vec{v}_1 \cdot (\vec{v}_2 \times \vec{v}_3) \quad \text{with} \quad \vec{v} = \vec{p}, \vec{l} .
\]  

(12)

The expectation value of such triple correlations accordingly are referred to as \( T \) odd moments.

In contrast to \( P \) or \( C \) the \( T \) operator is antilinear:

\[
T (\alpha |a\rangle + \beta |b\rangle) = \alpha^* T |a\rangle + \beta^* T |b\rangle .
\]  

(13)

This property of \( T \) is enforced by the commutation relation \([X, P] = i\hbar\), since

\[
\]  

(14)

\[
T^{-1}i\hbar T = -i\hbar .
\]  

(15)

The anti-linearity of \( T \) implies three important properties:
\[
T^{-1}(\mathcal{L}_{\text{eff}}\Delta t + \frac{i}{2}(\mathcal{L}_{\text{eff}}\Delta t)^2 + \ldots) = \mathcal{L}_{\text{eff}}\Delta t - \frac{i}{2}(\mathcal{L}_{\text{eff}}\Delta t)^2 + \ldots
\]

\[\mathcal{L}_{\text{eff}}\Delta t + \frac{i}{2}(\mathcal{L}_{\text{eff}}\Delta t)^2 + \ldots \neq \mathcal{L}_{\text{eff}}\Delta t + \frac{i}{2}(\mathcal{L}_{\text{eff}}\Delta t)^2 + \ldots
\]

(16)

1.2.2 Macroscopic T violation or ‘arrow of time’

Let us consider a simple example from classical mechanics: the motion of billiard ball(s) across a billiard table in three different scenarios.

i Watching a movie showing a single ball role around and bounce off the walls of the table one could not decide whether one was seeing the events in the actual time sequence or in the reverse order, i.e., whether one was seeing the movie running backwards. For both sequences are possible and equally likely.

ii Seeing one ball move in and hit another ball at rest leading to both balls moving off in different directions is a possible and ordinary sequence. The reverse—two balls moving in from different directions, hitting each other with one ball coming to a complete rest and the other one moving off in a different direction—is still a possible sequence yet a rather unlikely one since it requires fine tuning between the momenta of the two incoming billiard balls.

iii One billiard ball hitting a triangle of ten billiard balls at rest and scattering them in all directions is a most ordinary sequence for anybody but the most inept billiard player. The reverse sequence—eleven billiard balls coming in from all different directions, hitting each other in such a way that ten come to rest in a neatly arranged triangle while the eleventh one moves off—is a practically impossible one, since it requires a most delicate fine tuning of the initial conditions.

There are countless other examples of one time sequence being ordinary while the reversed one is (practically) impossible: \( \beta \) decay \( n \to pe^-\bar{\nu} \), the scattering of a plane wave off an object leading to an outgoing
spherical wave in addition to the continuing plane wave or the challenge of parking a car in a tight spot compared with the relative ease to get out of it. These daily experiences do not tell us anything about $T$ violation in the underlying dynamics; they reflect asymmetries in the macroscopic initial conditions, which are of a statistical nature.

Yet a central message of my lectures is that microscopic $T$ violation has been observed, i.e., $T$ violation that resides in the basic dynamics of the SM. It is conceivable though that in a more complete theory it reflects an asymmetry in the initial conditions in some higher sense.

1.3 The very special role of $CP$ invariance and its violation

While the discovery of $P$ violation in the weak dynamics in 1957 caused a well-documented shock in the community, even the theorists quickly recovered. Why then was the discovery of $CP$ violation in 1964 not viewed as a ‘déjà vu, all over again’ in the language of Yogi Berra? There are several reasons for that as illustrated by the following statements:

- Let me start with an analogy from politics. In my days as a student—at a time long ago and a place far away—politics was hotly debated. One of the subjects drawing out the greatest passions was the question of what distinguished the ‘left’ from the ‘right’. If you listened to it, you quickly found out that people almost universally defined ‘left’ and ‘right’ in terms of ‘positive’ and ‘negative’. The only problem was they could not quite agree who were the good guys and who the bad guys. There arises a similar conundrum when considering decays like $\pi^+ \rightarrow e^+\nu$. When saying that a pion decay produces a left-handed charged lepton one had $\pi^- \rightarrow e_L^\pm \bar{\nu}$ in mind. However, $\pi^+ \rightarrow e_R^\pm \nu$ yields a right-handed charged lepton. ‘Left’ is thus defined in terms of ‘negative’. No matter how much $P$ is violated, $CP$ invariance imposes equal rates for these $\pi^\pm$ modes, and it is untrue to claim that Nature makes an absolute distinction between ‘left’ and ‘right’. The situation is analogous to the saying that ‘the thumb is left on the right hand’—a correct, yet useless statement, since circular. $CP$ violation is required to define ‘matter’ vs. ‘antimatter’, ‘left’ vs. ‘right’, ‘positive’ vs. ‘negative’ in a convention-independent way.

- Owing to the almost unavoidable $CPT$ symmetry, violation of $CP$ implies one of $T$.

- It is the smallest observed violation of a symmetry as expressed through

$$\text{Im}M_{12}^K \simeq 1.1 \cdot 10^{-8} \text{ eV} \leftrightarrow \frac{\text{Im}M_{12}^K}{M_K} \simeq 2.2 \cdot 10^{-17}.$$ (17)

- It is one of the key ingredients in the Sakharov conditions for baryogenesis [5]: to obtain the observed baryon number of our Universe as a dynamically generated quantity rather than an arbitrary initial condition, one needs baryon-number-violating transitions with $CP$ violation to occur in a period where our Universe had been out of thermal equilibrium.

1.4 Flavour dynamics and the CKM ansatz

1.4.1 The GIM mechanism

A striking feature of (semi)leptonic kaon decays is the huge suppression of strangeness-changing neutral current modes:

$$\frac{\Gamma(K^+ \rightarrow \pi^+ e^+ e^-)}{\Gamma(K^+ \rightarrow \pi^0 e^+ \nu)} \sim 6 \cdot 10^{-6}, \quad \frac{\Gamma(K_L \rightarrow \mu^+ \mu^-)}{\Gamma(K^+ \rightarrow \mu^+ \nu)} \sim 3 \cdot 10^{-9}.$$ (18)

Embedding weak charged currents with their Cabibbo couplings

$$J^{(\pm)}_{\mu} = \cos \theta_C \bar{d}_L \gamma_\mu u_L + \sin \theta_C \bar{s}_L \gamma_\mu u_L$$
\( J_{\mu}^{(-)} = \cos \theta C \overline{u}_L \gamma_\mu d_L + \sin \theta C \overline{u}_L \gamma_\mu s_L \) (19)

into an SU(2) gauge theory to arrive at a renormalizable theory requires neutral currents of a structure as obtained from the commutator of \( J_{\mu}^{(+)} \) and \( J_{\mu}^{(-)} \). Using for the latter the expressions of Eq. (19) one arrives unequivocally at

\[ J_{\mu}^{(0)} = \ldots + \cos \theta C \sin \theta C (\overline{s}_L \gamma_\mu d_L + \overline{d}_L \gamma_\mu s_L) , \]

(20)
i.e., strangeness-changing neutral currents. Yet their Cabibbo suppression is not remotely sufficient to make them compatible with these observed super-tiny branching ratios\(^4\). The huge discrepancy between observed and expected branching ratios led some daring spirits [6] to postulate a fourth quark\(^5\) with quite specific properties to complete a second quark family in such a way that no strangeness-changing neutral currents arise at \textit{tree} level. The name ‘charm’ derives from this feature of warding off the evil of strangeness-changing neutral currents rather than an anticipated relation to beauty.

Yet I remember there was great scepticism felt in the community maybe best expressed by the quote: “Nature is smarter than Shelley (Glashow)—she can do without charm quarks.”\(^6\). These remarks can indicate how profound a shift in paradigm was begun by the observation of scaling in deep inelastic lepton–nucleon scattering and completed by the discovery of the \( J/\psi \) in 1974 and its immediate aftermath.

### 1.4.2 Quark masses and CP violation

Let us consider the mass terms for the up- and down-type quarks as expressed through matrices \( M_{U/D} \) and vectors of quark fields \( U^F = (u, c, t)^F \) and \( D^F = (d, s, b)^F \) in terms of the \textit{flavour} eigenstates denoted by the superscript \( F \):

\[ \mathcal{L}_M \propto \overline{U}^F_L M_{U} U^F_R + \overline{D}^F_L M_{D} D^F_R . \]

(21)

\( A \textit{priori} \) there is no reason why the matrices \( M_{U/D} \) should be diagonal. Yet applying bi-unitary rotations \( \mathcal{J}_{U/D,L} \) will allow one to diagonalize them

\[ M_{U/D}^{\text{diag}} = \mathcal{J}_{U/D,L} M_{U,D} \mathcal{J}_{U/D,R}^\dagger \]

(22)

and obtain the \textit{mass} eigenstates of the quark fields:

\[ U^{m}_{L/R} = \mathcal{J}_{U,L/R} U^F_{L/R} , \quad D^{m}_{L/R} = \mathcal{J}_{D,L/R} D^F_{L/R} . \]

(23)

The eigenvalues of \( M_{U/D} \) represent the masses of the quark fields. The measured values exhibit a very peculiar pattern that seems unlikely to be accidental being so hierarchical for up- and down-type quarks, charged and neutral leptons.

Yet again, there is much more to it. Consider the neutral current coupling

\[ \mathcal{L}^{U[D]}_{NC} \propto \bar{g}_Z U^F \gamma_\mu [D^F] Z^\mu . \]

(24)

It keeps its form when expressed in terms of the mass eigenstates

\[ \mathcal{L}^{U[D]}_{NC} \propto \bar{g}_Z U^{m} \gamma_\mu [D^{m}] Z^\mu ; \]

(25)

\(^4\)The observed huge suppression of strangeness-changing neutral currents actually led to some speculation that also flavour-\textit{conserving} neutral currents are greatly suppressed.

\(^5\)A fourth quark had been originally introduced by Glashow and Bjorken to regain quark–lepton correspondence by completing the second quark family.

\(^6\)The fact that Nature needed charm after all does not prove the inverse of this quote, of course.
i.e., there are no flavour-changing neutral currents on the tree level. This important property is referred to as the ‘generalized’ GIM mechanism [6].

However, for the charged currents the situation is quite different:

\[ \mathcal{L}_{CC} \propto \bar{g}_W \bar{U}_L^{F \gamma_\mu} D^F W^\mu = \bar{g}_W \bar{U}_L^{m \gamma_\mu} V_{CKM} D^m W^\mu \]  

(26)

with

\[ V_{CKM} = J_{U,D,L,R} \]  

(27)

There is no reason why the matrix \( V_{CKM} \) should be the identity matrix or even diagonal\(^7\). It means the charged-current couplings of the mass eigenstates will be modified in an observable way. In which way and by how much this happens requires further analysis since the phases of fermion fields are not necessarily observables. Such an analysis was first given by Kobayashi and Maskawa [7].

Consider \( N \) families. \( V_{CKM} \) then represents an \( N \times N \) matrix that has to be unitary based on two facts:

- The transformations \( J_{U,D,L,R} \) are unitary by construction.
- As long as the carriers of the weak force are described by a single local gauge group—\( SU(2)_L \) in this case—they have to couple to all other fields in a way fixed by their self-coupling. This was already implied by Eq. (26), when writing the weak coupling \( g_W \) as an overall factor.

The unitarity of \( V_{CKM} \) implies weak universality, as addressed later in more detail. There are actually \( N \) such relations characterized by

\[ \sum_j |V(ij)|^2 = 1, \quad i = 1, \ldots, N. \]  

(28)

These relations are important, yet insensitive to weak phases; thus they provide no direct information on CP violation.

Violations of weak universality can be implemented by adding dynamical layers to the SM. So-called horizontal gauge interactions, which differentiate between families and induce flavour-changing neutral currents, will do it. Another admittedly ad hoc possibility is to introduce a separate \( SU(2)_L \) group for each quark family while allowing the gauge bosons from the different \( SU(2)_L \) groups to mix with each other. This mixing can be set up in such a way that the lightest mass eigenstates couple to all fermions with approximately universal strength. Weak universality thus emerges as an approximate symmetry. Flavour-changing neutral currents are again induced, and they can generate electric dipole moments.

After this aside on weak universality let us return to \( V_{CKM} \). There are \( N^2 - N \) orthogonality relations:

\[ \sum_j V^{*}(ij)V(jk) = 0, \quad i \neq k. \]  

(29)

Those are very sensitive to complex phases and tell us directly about CP violation.

An \( N \times N \) complex matrix contains \( 2N^2 \) real parameters; the unitarity constraints reduce it to \( N^2 \) independent real parameters. Since the phases of quark fields like other fermion fields can be rotated freely, \( 2N - 1 \) phases can be removed from \( \mathcal{L}_{CC} \) (a global phase rotation of all quark fields has no impact on \( \mathcal{L}_{CC} \)). Thus we have \((N-1)^2\) independent physical parameters. Since an \( N \times N \) orthogonal matrix has \( N(N-1)/2 \) angles, we conclude that an \( N \times N \) unitary matrix contains also \((N-1)(N-2)/2\) physical phases. This was the general argument given by Kobayashi and Maskawa. Accordingly:

\(^7\)Even if some speculative dynamics were to enforce an alignment between the \( U \) and \( D \) quark fields at some high scales causing their mass matrices to get diagonalized by the same bi-unitary transformation, this alignment would probably get upset by renormalization down to the electroweak scales.
– For $N = 2$ families we have one angle—the Cabibbo angle—and zero phases.
– For $N = 3$ families we obtain three angles and one irreducible phase; i.e., a three-family ansatz can support CP violation with a single source—the ‘CKM phase’. PDG suggests a ‘canonical’ parametrization for the $3 \times 3$ CKM matrix:

$$V_{CKM} = \begin{pmatrix}
V(ud) & V(us) & V(ub) \\
V(cd) & V(cs) & V(cb) \\
V(td) & V(ts) & V(tb)
\end{pmatrix}$$

$$= \begin{pmatrix}
c_{12}c_{13} & s_{12}c_{23} & s_{13}e^{-i\delta_{13}} \\
-s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta_{13}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta_{13}} & c_{13}s_{23} \\
s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{13}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & c_{13}c_{23}
\end{pmatrix}$$

(30)

where

$$c_{ij} \equiv \cos \theta_{ij} , \quad s_{ij} \equiv \sin \theta_{ij}$$

with $i,j = 1, 2, 3$ being generation labels.

This is a completely general, yet not unique parametrization: a different set of Euler angles could be chosen; the phases can be shifted around among the matrix elements by using a different phase convention.

– For even more families we encounter a proliferation of angles and phases, namely six angles and three phases for $N = 4$.

These results obtained by algebraic means can be visualized graphically:

– For $N = 2$ we have two weak universality conditions and two orthogonality relations:

$$V^*(ud)V(us) + V^*(cd)V(cs) = 0$$
$$V^*(us)V(ud) + V^*(cs)V(cd) = 0 .$$

(32)

While the CKM angles can be complex, there can be no nontrivial phase ($\neq 0, \pi$) between their observable combinations; i.e., there can be no CP violation for two families in the SM.

– For three families the orthogonality relations read

$$\sum_{j=1}^{j=3} V^*(ij)V(jk) = 0 , \quad i \neq k .$$

(33)

There are six such relations, and they represent triangles in the complex plane with in general nontrivial relative angles.

– While these six triangles can and will have quite different shapes, as we shall describe later in detail, they all have to possess the same area, namely [8]

$$\text{area(every triangle)} = \frac{1}{2} J$$

$$J = \text{Im}[V(ud)V(cs)V^*(us)V^*(cd)] .$$

(34)

If $J = 0$, one has obviously no nontrivial angles, and there is no CP violation. The fact that all triangles have to possess the same area reflects the fact that for three families there is but a single CKM phase.

– Only the angles, i.e., the relative phases matter, but not the overall orientation of the triangles in the complex plane. That orientation merely reflects the phase convention for the quark fields.
If any pair of up-type or down-type quarks were mass degenerate, then any linear combination of those two would be a mass eigenstate as well. Forming different linear combinations thus represents symmetry transformations, and with this additional symmetry one can further reduce the number of physical parameters. For \( N = 3 \) it means CP violation could still not occur.

The CKM implementation of CP violation depends on the form of the quark mass matrices \( M_U, D \), not so much on how those are generated. Nevertheless, something can be inferred about the latter: within the SM all fermion masses are driven by a single VEV; to obtain an irreducible relative phase between different quark couplings thus requires such a phase in quark Yukawa couplings; this means that in the SM CP violation arises in dimension-four couplings, i.e., it is ‘hard’.

### 1.4.3 ‘Maximal’ CP violation?

As already mentioned, charged-current couplings with their \( V-A \) structure break parity and charge conjugation maximally. Since owing to CPT invariance CP violation is expressed through couplings with complex phases, one might say that maximal CP violation is characterized by complex phases of 90°. However, this would be fallacious: for by changing the phase convention for the quark fields one can change the phase of a given CKM matrix element and even rotate it away; it will of course re-appear in other matrix elements. For example \( |s⟩ \to e^{iδ_s} |s⟩ \) leads to \( V_{qs} \to e^{iδ_s} V_{qs} \) with \( q = u, c, t \). In that sense the CKM phase is like the ‘Scarlet Pimpernel’: “Sometimes here, sometimes there, sometimes everywhere.”

One can actually illustrate with a general argument, why there can be no straightforward definition for maximal CP violation. Consider neutrinos: Maximal P violation means there are \( ν_L \) and \( \bar{ν}_R \), yet no \( ν_R \) or \( \bar{ν}_L \). Likewise for C: there are \( ν_L \) and \( \bar{ν}_R \), but not \( \bar{ν}_L \) or \( ν_R \). One might then suggest that maximal CP violation means that \( ν_L \) exists, but \( \bar{ν}_R \) does not. Alas—CPT invariance already enforces the existence of both.

Similarly—and maybe more obviously—it is not clear what maximal T violation would mean although some formulations have entered daily language like the ‘no future generation’, the ‘woman without a past’, or the ‘man without a future’.

### 1.4.4 Some historical remarks

CP violation was discovered in 1964 through the observation of \( K_L \to π^+π^- \), yet it was not realized for a number of years that dynamics known at that time could not generate it. We should not be too harsh on our predecessors for that oversight: as long as one did not have a renormalizable theory for the weak interactions and thus had to worry about infinities in the calculated rates, one can be excused for ignoring a seemingly marginal rate with a branching ratio of \( 2 \cdot 10^{-3} \). Yet even after the emergence of the renormalizable Glashow–Salam–Weinberg model its phenomenological incompleteness was not recognized right away. There is a short remark by Mohapatra in a 1972 paper invoking the need for right-handed currents to induce CP violation.

It was the 1973 paper by Kobayashi and Maskawa [7] that fully stated the inability of even a two-family SM to produce CP violation and that explained what had to be added to it: right-handed charged currents, extra Higgs doublets—or (at least) a third quark family. Of the three options Kobayashi and Maskawa listed, their name has been attached only to the last one as the CKM description. They were helped by the ‘genius loci’ of Nagoya University:

- Since it was the home of the Sakata school and the Sakata model of elementary particles, quarks were viewed as physical degrees of freedom from the start.

\(^{8}\)To be more precise: \( ν_L \) and \( \bar{ν}_R \) couple to weak gauge bosons, \( ν_R \) or \( \bar{ν}_L \) do not.
– It was also the home of Professor Niu who in 1971 had observed a candidate for a charm decay in emulsion exposed to cosmic rays and actually recognized it as such. The existence of charm, its association with strangeness and thus of two complete quark families were thus taken for granted at Nagoya.

1.5 Meson–antimeson oscillations—on the power of quantum mysteries

After the conceptual exposition of the SM I return to the historical development. With respect to meson–antimeson oscillations, Nature has treated us like a patient teacher with somewhat dense students: she has provided us not with one, but with three meson systems that exhibit oscillations, namely the $K^0 - \bar{K}^0$, $B_d - \bar{B}_d$ and $B_s - \bar{B}_s$ complexes; as we shall discuss in some detail, those three systems present complementary perspectives on oscillations. I would like to add that these phenomena by and large followed theoretical predictions—yet the most revolutionary feature, namely the first manifestation of CP violation in particle decays, was outside the ‘theoretical’ horizon before 1964.

As already mentioned, ‘strange’ hadrons obtained their name from the observation that their production rate exceeds their decay rate by many orders of magnitude. This feature was explained by assigning them an internal quantum number strangeness $S = \pm 1$ and postulating that only the weak interactions can produce $\Delta S \neq 0$ transitions. One has two different neutral kaons: $K^0$ and $\bar{K}^0$ with $S = 1$ and $S = -1$, respectively. Then the question arises: How does one verify it experimentally?

The answer to this challenge came in the form of oscillations and represents one of the glory pages of particle physics. Symmetry considerations allow one to derive many essential features of oscillations without solving any equations explicitly. Without weak interactions $K^0$ and $\bar{K}^0$ have, owing to CPT invariance, equal masses and lifetimes (the latter being infinite at this point). With the weak $\Delta S \neq 0$ forces ‘switched on’ the two neutral kaon mass eigenstates will be linear combinations of $K^0$ and $\bar{K}^0$ and thus carry no definite strangeness. CP invariance implies the mass eigenstates to be CP eigenstates as well. With the definition

$$\text{CP} \, |K^0\rangle = |\bar{K}^0\rangle$$

one has for the CP even and odd states

$$|K_{\pm}\rangle = \frac{1}{\sqrt{2}}(|K^0\rangle \pm |\bar{K}^0\rangle) \quad \text{with} \quad \Delta M_K \equiv M(K_-) - M(K_+) \neq 0 \neq \Delta \Gamma_K = \Gamma(K_+) - \Gamma(K_-). \quad (36)$$

CP symmetry also constrains the decay modes

$$|K_+\rangle \to 2\pi, \quad 2\pi \neq |K_-\rangle \to 3\pi,$$  

(37)

since $\pi^+\pi^-$ and $2\pi^0$ are CP even, whereas $\pi^+\pi^-\pi^0$ can be CP odd and $3\pi^0$ has to be. (With $M_K < 4m_\pi\, K \to 4\pi$ cannot occur.) Such difference leads to $\tau(K_+) \neq \tau(K_-)$. A kinematical ‘accident’ intervenes at this point: since the kaon mass is barely above the three pion threshold and thus $K_- \to 3\pi$ greatly suppressed by phase space, its lifetime is much longer than for $K_+$. Their lifetime ratio is actually as large as 570; accordingly one refers to them as $K_L$ and $K_S$ with the subscripts $L$ and $S$ referring to ‘long’- and ‘short’-lived. Thus one predicts the following nontrivial scenario: if one starts with a pure beam of, say, $K^0$, one finds different components in the decay rate evolution depending on the nature of the final state:

- In $K^\text{neut} \to$ pions two distinct components will emerge, namely $K^\text{neut} \to 2\pi$ and $K^\text{neut} \to 3\pi$ following two separate exponential functions in (proper) time controlled by the lifetimes $\tau(K_S)$ and $\tau(K_L)$, respectively.
- Tracking the flavour-specific (semi)leptonic modes instead, one encounters a considerably more complex situation not described by simple exponential functions in time. The mathematics involved is rather straightforward though. Using Eq. (36) and the fact that $K_{\pm}$ are mass eigenstates

128
Fig. 1: The probabilities of finding a $K^0$ and a $\bar{K}^0$ in an initial $K^0$ beam as a function of time (in the limit of CP invariance), we obtain for the time evolution of the amplitude of an initially pure $K^0$ beam

$$|K^0(t)| = \frac{1}{\sqrt{2}} [ |K_+(t)| + |K_-(t)| ]$$

$$= \frac{1}{\sqrt{2}} e^{(iM(K_+)-\frac{1}{2}\Gamma_+)^t} [ |K_+\rangle + e^{(i\Delta M_K+\frac{1}{2}\Delta \Gamma_K)^t} |K_+\rangle ]$$

$$= \frac{1}{2} e^{(iM(K_+)-\frac{1}{4}\Gamma_+)^t} \left[ \left( 1 + e^{(i\Delta M_K+\frac{1}{2}\Delta \Gamma_K)^t} \right) |K^0\rangle + \left( 1 - e^{(i\Delta M_K+\frac{1}{2}\Delta \Gamma_K)^t} \right) |\bar{K}^0\rangle \right].$$

The probability for the initial $K^0$ to decay as a $K^0$ or a $\bar{K}^0$ is then given by

$$\text{Prob}(K^0 \rightarrow K^0; t) = \frac{1}{4} e^{-\Gamma_+^t} \left( 1 + e^{\Delta \Gamma_K^t + \frac{1}{2} \Delta \Gamma_K^t \cos \Delta M_K t} \right).$$

$$\text{Prob}(K^0 \rightarrow \bar{K}^0; t) = \frac{1}{4} e^{-\Gamma_+^t} \left( 1 - e^{\Delta \Gamma_K^t + \frac{1}{2} \Delta \Gamma_K^t \cos \Delta M_K t} \right).$$

The phenomenon that a state that is initially absent in a beam travelling through vacuum re-emerges, Eq. (40), is often called ‘spontaneous regeneration’.

These expressions are shown in Fig. 1: the decay rate for the ‘right-sign’ leptons $K^0 \rightarrow l^+\nu\pi^+$ at first drops off faster than follows from $e^{-\Gamma_0^t}$, an exponential dependence on the time of decay, then bounces back up etc., i.e., ‘oscillates’—hence the name. The rate for the ‘wrong-sign’ transitions $K^0 \rightarrow l^-\nu\pi^-$, which has to start out at zero for $t = 0$, rises quickly, yet turns around dropping down, before bouncing back up again etc. It provides the complement for $K^0 \rightarrow l^+\nu\pi^-$, i.e., the rate for the sum of both modes should exhibit a simple exponential behaviour.

These predictions given by Gell-Mann and Pais, first assuming C conservation and relaxing it later to CP symmetry, were verified experimentally with impressive numerical sensitivity [10]:

$$\Delta M_K \equiv M_{K_L} - M_{K_S} = (3.483 \pm 0.006) \cdot 10^{-12} \text{ MeV}.$$  (41)

(English speakers can rely on a simple mnemonic to remember which state is heavier: ‘L’ stands for larger mass and longer lifetime, whereas ‘S’ denotes smaller and shorter.) This number is a striking demonstration for the sensitivity reached when quantum mechanical interference can be tracked over
macroscopic distances, i.e., flight paths of metres or even hundreds of metres. Using the kaon mass as yardstick one can re-express Eq.(41)

$$\frac{\Delta M_K}{M_K} = \frac{M_{K_L} - M_{K_S}}{M_K} = 7.7 \cdot 10^{-15},$$

(42)

which is obviously a most striking number. A hard-nosed reader can point out that Eq. (42) vastly overstates the point since the kaon mass generated largely by the strong interactions has no intrinsic connection to \(\Delta M_K\) generated by the weak interactions and that calibrating \(\Delta M_K\) by, say, the mass of an elephant is not truly more absurd.

The more relevant yardstick for the oscillation rates is indeed provided by the weak decay rate [10]

$$x_K = \frac{\text{oscillation rate}}{\text{decay rate}} = \frac{\Delta M_K}{\Gamma_K} \simeq 0.945 \pm 0.003.$$  

(43)

$$y_K = \frac{\Delta \Gamma_K}{2\bar{\Gamma}_K} \approx 0.996 \quad \text{with} \quad \bar{\Gamma}_K = \frac{1}{2}(\Gamma_{K_S} + \Gamma_{K_L}).$$  

(44)

1.5.1 The shock of 1964—CP violation surfaces

1964 was an excellent year for high-energy physics: (i) The Higgs mechanism for the ‘spontaneous realization’ of a symmetry was first developed. (ii) The quark model (and the first elements of current algebra) were first suggested. (iii) The charm quark was first introduced to implement quark–lepton symmetry. (iv) The nonrelativistic \(SU(6)\) symmetry combining \(SU(3)_{FL}\) with the spin \(SU(2)\) was proposed for hadron spectroscopy. (v) The first \(e^+e^-\) storage ring was inaugurated in Frascati. (vi) The \(\Omega^-\) baryon was found at Brookhaven National Laboratory which was viewed as essential validation for the ‘Eightful Way’ of \(SU(3)_{FL}\) symmetry. The modern perspective on it has changed: being composed of three strange quarks it exhibits rather directly the need for colour as a new internal degree of freedom—together with other observables like \(R\) and \(\Gamma(\pi^0 \rightarrow 2\gamma)\) as already mentioned in Section 1.1.1.1. (vii) CP violation was discovered at the same laboratory through the observation that \(K_L\) mesons can decay both into three- and two-pion final states, albeit the latter with the tiny branching ratio of 0.23% only.

The theoretical concepts listed under items (i)–(iii) and the experimental tool of item (v) turn out to be crucial for the subject matter of these lectures.

It is a fact of life that if one wants to see what moves physicists, one should not focus on what they say (rarely a good indicator for scientists in general), but on what they do. Point in case: how much this discovery shook the HEP community is best gauged by noting the efforts made to reconcile the observation of \(K_L \rightarrow \pi^+\pi^-\) with CP invariance:

- To infer that \(K_L \rightarrow \pi\pi\) implies CP violation one has to invoke the superposition principle of quantum mechanics. One can introduce [11] nonlinear terms into the Schrödinger equation in such a way as to allow \(K_L \rightarrow \pi^+\pi^-\) with CP invariant dynamics. While completely \textit{ad hoc}, it is possible in principle. Such efforts were ruled out by further data, most decisively by \(\Gamma(K^0(t) \rightarrow \pi^+\pi^-) \neq \Gamma(K^0(t) \rightarrow \pi^+\pi^-)\).

- One can try to emulate the success of Pauli’s neutrino hypothesis. An apparent violation of energy–momentum conservation had been observed in \(\beta\) decay \(n \rightarrow pe^-\), since the electron exhibited a continuous momentum spectrum. Pauli postulated that the reaction actually was

$$n \rightarrow pe^- \bar{\nu}$$

(45)

with \(\bar{\nu}\) a neutral and light particle that had escaped direct observation, yet led to a continuous spectrum for the electron; i.e., Pauli postulated a new particle—and a most whimsical one at that—to save a symmetry, namely the one under translations in space and time responsible for the conservation of energy and momentum. Likewise it was suggested that the real reaction was

$$K_L \rightarrow \pi^+\pi^-U$$

(46)
with $U$ a neutral and light particle with odd intrinsic CP parity: i.e., a hitherto unseen particle was introduced to save a symmetry. This attempt at evasion was also soon rejected experimentally (see Homework # 1). This represents an example of the ancient Roman saying:

“Quod licet Jovi, non licet bovi.”

“What is allowed Jupiter, is not allowed a bull.”

That is, we mere mortals cannot get away with speculations like ‘Jupiter’ Pauli.

**Homework # 1**

What was the conclusive argument to rule out the reaction of Eq. (46) taking place even for a very tiny $U$ mass?

**End of Homework # 1**

Notwithstanding these attempts at evasion, the findings of the Fitch–Cronin experiment—namely that $K_L \rightarrow \pi^+\pi^-$ does occur—were soon widely accepted, since, in the words of Pram Pais the ‘perpetrators’ were considered ‘real pros’. Yet they induced a feeling of a certain frustration. Parity emerged as violated ‘maximally’ in the charged weak currents that involve left-handed, but no right-handed neutrinos; thus it followed Luther’s dictum “Peccont fortiter!”, i.e., “Sin boldly!”. In contrast, CP violation, while having an even more profound impact on Nature’s basic design as indicated above, appeared as a ‘near-miss’ as suggested by the rarity of the observed transition: $BR(K_L \rightarrow \pi^+\pi^-) \approx 0.0023$. Actually we do not know how to give an unambiguous definition of ‘maximal’ CP violation, as explained in Section 1.4.3.

As already mentioned, from the discovery in 1964 till the 1973 Kobayashi–Maskawa paper there was no theory of CP violation. Worse still, it was not even recognized—apart from a short remark in a paper by Mohapatra—that the dynamics known at that time were insufficient to implement CP violation. It should be noted that Wolfenstein’s ‘Superweak Model’, which will be sketched below, is not a theory, not even a model—it is a classification scheme, not more and not less.

Yet despite the lack of a true theoretical underpinning, the relevant phenomenology was quickly developed.

### 1.5.2 Phenomenology of CP violation, Part I

The discussion here will be given in terms of strangeness $S$, yet can be generalized to any other flavour quantum number $F$ like beauty, charm, etc.

Weak dynamics can drive $\Delta S = 1\&2$ transitions, i.e., decays and oscillations. While the underlying theory has to account for both, it is useful to differentiate between them on the phenomenological level. The interplay between $\Delta S = 1\&2$ affects also CP violation and how it can manifest itself. Consider $K_L \rightarrow \pi\pi$: while $\Delta S = 2$ dynamics transform the flavour eigenstates $K^0$ and $\bar{K}^0$ into mass eigenstates $K_L$ and $K_S$, $\Delta S = 1$ forces produce the decays into pions.

$$[K^0 \xrightarrow{\Delta S=2} \bar{K}^0] \Rightarrow K_L \xrightarrow{\Delta S=1} \pi\pi.$$  (47)

Both of these reactions can exhibit CP violation, which is usually expressed as follows:

$$\eta_{+--[00]} \equiv \frac{T(K_L \rightarrow \pi^+\pi^-[\pi^0\pi^0])}{T(K_S \rightarrow \pi^+\pi^-[\pi^0\pi^0])}, \quad \eta_{++} \equiv \epsilon_K + \epsilon', \quad \eta_{00} \equiv \epsilon_K - 2\epsilon'.$$  (48)

Both $\eta_{+--}, \eta_{00} \neq 0$ signal CP violation; $\epsilon_K$ is common to both observables and reflects the CP properties of the state mixing, i.e., in $\Delta S = 2$ dynamics; $\epsilon'$ on the other hand differentiates between the two
final states and parametrizes CP violation in $\Delta S = 1$ dynamics. With an obvious lack in Shakespearean flourish $\epsilon_K \neq 0$ is referred to as ‘indirect’ or ‘superweak’ CP violation and $\epsilon'$ as ‘direct’ CP violation. As long as CP violation is seen only through a single mode of a neutral meson—in this case either $K_L \to \pi^+\pi^0$ or $K_L \to \pi^0\pi^0$—the distinction between direct and indirect CP violation is somewhat arbitrary, as explained later for $B_d$ decays.

Five types of CP-violating observables have emerged through $K^0 - \bar{K}^0$ oscillations:

1. Existence of a transition: $K_L \to \pi^+\pi^-, \pi^0\pi^0$;
2. An asymmetry due to the initial state: $K^0 \to \pi^+\pi^-$ vs. $\bar{K}^0 \to \pi^+\pi^-$;
3. An asymmetry due to the final state: $K_L \to l^+\nu\pi^-$ vs. $K_L \to l^-\bar{\nu}\pi^+$, $K_L \to \pi^+\pi^-$ vs. $K_L \to \pi^0\pi^0$;
4. A microscopic T asymmetry: rate($K^0 \to \bar{K}^0$) $\neq$ rate($\bar{K}^0 \to K^0$);
5. A T-odd correlation in the final state: $K_L \to \pi^+\pi^-e^+\bar{e}^-$.  

We know now that all these observables except $|\eta_{-+}| \neq |\eta_{00}|$ are predominantly (or even exclusively) given by $\epsilon_K$, i.e., indirect CP violation. The asymmetry in semileptonic $K_L$ decays has been measured to be

$$\delta_L \equiv \frac{\Gamma(K_L \to l^+\nu\pi^-) - \Gamma(K_L \to l^-\bar{\nu}\pi^+)}{\Gamma(K_L \to l^+\nu\pi^-) + \Gamma(K_L \to l^-\bar{\nu}\pi^+)} = (3.27 \pm 0.12) \cdot 10^{-3},$$

(49)

averaged over electrons and muons. This measurement provides a convention-independent definition of ‘$+$’ vs. ‘$-$’, hence of ‘matter’$\to l^-$—vs. ‘antimatter’$\to l^+$—and of ‘left’$\to l_L^+$—vs. ‘right’$\to l_R^+$.

To describe oscillations in the presence of CP violation one turns to solving a nonrelativistic Schrödinger equation, which I formulate for the general case of a pair of neutral mesons $P^0$ and $\bar{P}^0$ with flavour quantum number $F$; it can denote a $K^0$, $\bar{D}^0$ or $B^0$ [12]:

$$\frac{d}{dt} \left( \begin{array}{c} P^0 \\ \bar{P}^0 \end{array} \right) = \left( \begin{array}{cc} M_{11} - \frac{i}{2} \Gamma_{11} & M_{12} - \frac{i}{2} \Gamma_{12} \\ M_{12} + \frac{i}{2} \Gamma_{12} & M_{22} - \frac{i}{2} \Gamma_{22} \end{array} \right) \left( \begin{array}{c} P^0 \\ \bar{P}^0 \end{array} \right).$$

(50)

CPT invariance imposes

$$M_{11} = M_{22}, \quad \Gamma_{11} = \Gamma_{22}.$$ 

(51)

Homework # 2

Which physical situation is described by an equation analogous to Eq. (50) where, however, the two diagonal matrix elements differ without violating CPT?

End of Homework # 2

The subsequent discussion might strike the reader as overly technical, yet I hope she or he will bear with me since these remarks will lay important groundwork for a proper understanding of CP asymmetries in $B$ decays as well.

The mass eigenstates obtained through diagonalizing this matrix are given by (for details see Ref. [1])

$$|P_A\rangle = \frac{1}{\sqrt{|p|^2 + |q|^2}} (p|P^0\rangle + q|\bar{P}^0\rangle),$$

(52)

$$|P_B\rangle = \frac{1}{\sqrt{|p|^2 + |q|^2}} (p|P^0\rangle - q|\bar{P}^0\rangle),$$

(53)

---

*This definition can be communicated to a far-away civilization using unpolarized radio signals. Such a communication is of profound academic as well as practical value: when meeting such a civilization in outer space, one had better find out whether they are made of matter or antimatter; otherwise the first handshake might also be the last.
with eigenvalues
\[ M_A - \frac{i}{2} \Gamma_A = M_{11} - \frac{i}{2} \Gamma_{11} + \frac{q}{p} \left( M_{12} - \frac{i}{2} \Gamma_{12} \right), \tag{54} \]
\[ M_B - \frac{i}{2} \Gamma_B = M_{11} - \frac{i}{2} \Gamma_{11} - \frac{q}{p} \left( M_{12} - \frac{i}{2} \Gamma_{12} \right), \tag{55} \]
as long as
\[ \left( \frac{q}{p} \right)^2 = \frac{M_{12}^* - \frac{i}{2} \Gamma_{12}^*}{M_{12} - \frac{i}{2} \Gamma_{12}}. \tag{56} \]
holds. I am using letter subscripts \( A \) and \( B \) for labelling the mass eigenstates rather than numbers 1 and 2 as is usually done. For I want to avoid confusing them with the matrix indices 1, 2 in \( M_{ij} - \frac{i}{2} \Gamma_{ij} \).

Equations (55) yield for the differences in mass and width
\[ \Delta M \equiv M_B - M_A = -2 \text{Re} \left[ \frac{q}{p} (M_{12} - \frac{i}{2} \Gamma_{12}) \right], \tag{57} \]
\[ \Delta \Gamma \equiv \Gamma_A - \Gamma_B = -2 \text{Im} \left[ \frac{q}{p} (M_{12} - \frac{i}{2} \Gamma_{12}) \right]. \tag{58} \]

Note that the subscripts \( A, B \) have been swapped in going from \( \Delta M \) to \( \Delta \Gamma \). This is done to have both quantities positive for kaons.

In expressing the mass eigenstates \( P_A \) and \( P_B \) explicitly in terms of the flavour eigenstates—Eqs. (53)—one needs \( \frac{q}{p} \). There are two solutions to Eq. (56):
\[ \frac{q}{p} = \pm \sqrt{\frac{M_{12}^* - \frac{i}{2} \Gamma_{12}^*}{M_{12} - \frac{i}{2} \Gamma_{12}}}. \tag{59} \]

There is actually a more general ambiguity than this binary one. For antiparticles are defined up to a phase only:
\[ \text{CP}|P^0(\eta)|P^0(\eta) \quad \text{with} \quad |\eta| = 1. \tag{60} \]

Adopting a different phase convention will change the phase for \( M_{12} - \frac{i}{2} \Gamma_{12} \) as well as for \( q/p \):
\[ |P^0\rangle \rightarrow e^{i\xi}|P^0\rangle \quad \Rightarrow \quad (M_{12}, \Gamma_{12}) \rightarrow e^{i\xi}(M_{12}, \Gamma_{12}) \& \frac{q}{p} \rightarrow e^{-i\xi} \frac{q}{p}, \tag{61} \]
yet leave \( (q/p)(M_{12} - \frac{i}{2} \Gamma_{12}) \) invariant—as it has to be since the eigenvalues, which are observables, depend on this combination, see Eq. (55). Also \( \left| \frac{q}{p} \right| \) is an observable; its deviation from unity is one measure of CP violation in \( \Delta F = 2 \) dynamics.

By convention most authors pick the positive sign in Eq. (59)
\[ \frac{q}{p} = + \sqrt{\frac{M_{12}^* - \frac{i}{2} \Gamma_{12}^*}{M_{12} - \frac{i}{2} \Gamma_{12}}}. \tag{62} \]

Up to this point the two states \( |P_{A,B}\rangle \) are merely labelled by their subscripts. Indeed \( |P_A\rangle \) and \( |P_B\rangle \) switch places when selecting the minus rather than the plus sign in Eq. (59).

One can define the labels \( A \) and \( B \) such that
\[ \Delta M \equiv M_B - M_A > 0 \tag{63} \]
is satisfied. Once this convention has been adopted, it becomes a sensible question whether
\[ \Gamma_B > \Gamma_A \quad \text{or} \quad \Gamma_B < \Gamma_A \tag{64} \]
holds, i.e., whether the heavier state is shorter or longer lived.

One can write the general mass eigenstates in terms of the CP eigenstates as well:

$$|P_A\rangle = \frac{1}{\sqrt{1 + |\bar{\epsilon}|^2}} (|P_+\rangle + \bar{\epsilon}|P_-\rangle) \quad CP|P_\pm\rangle = \pm|P_\pm\rangle$$ (65)

$$|P_B\rangle = \frac{1}{\sqrt{1 + |\bar{\epsilon}|^2}} (|P_-\rangle + \bar{\epsilon}|P_+\rangle) \quad (66)$$

$\bar{\epsilon} = 0$ means that the mass and CP eigenstates coincide, i.e., CP is conserved in $\Delta F = 2$ dynamics driving $P - \bar{P}$ oscillations. With the phase between the orthogonal states $|P_+\rangle$ and $|P_-\rangle$ arbitrary, the phase of $\bar{\epsilon}$ can be changed at will and is not an observable; $\bar{\epsilon}$ can be expressed in terms of $\frac{q}{p}$, yet in a way that depends on the convention for the phase of antiparticles. For $CP|P\rangle = \pm|P\rangle$ one has

$$|P_+\rangle = \frac{1}{\sqrt{2}} (|P^0\rangle \pm |\bar{P}^0\rangle)$$ (67)

$$|P_-\rangle = \frac{1}{\sqrt{2}} (|P^0\rangle \mp |\bar{P}^0\rangle)$$ (68)

$$\bar{\epsilon} = \frac{1 \mp \frac{q}{p}}{1 \pm \frac{q}{p}}.$$ (69)

### Homework # 3

While $\langle \bar{P}^0|P^0\rangle = 0$ holds—e.g., $\langle \bar{K}^0|K^0\rangle = 0$—, one has $\langle K_L|K_S\rangle \neq 0$ and in general $\langle P_B|P_A\rangle \neq 0$. Calculate it and interpret your result.

### End of Homework # 3

Later we shall discuss how to evaluate $M_{12}$ and thus also $\Delta M$ within a given theory for the $P - \bar{P}$ complex. The examples just listed illustrate that some care has to be applied in interpreting such results. For expressing mass eigenstates explicitly in terms of flavour eigenstates involves some conventions. Once adopted we have to stick with a convention; yet our original choice cannot influence observables.

Let me recapitulate the relevant points:

- The labels of the two mass eigenstates $P_A$ and $P_B$ can be chosen such that

$$M_{P_B} > M_{P_A}$$ (70)

holds.

- Then it becomes an empirical question whether $P_A$ or $P_B$ are longer lived:

$$\Gamma_{P_A} > \Gamma_{P_B} \quad \text{or} \quad \Gamma_{P_A} < \Gamma_{P_B} ?$$ (71)

- In the limit of CP invariance one can also raise the question whether it is the CP-even or the CP-odd state that is heavier.

- We shall see later that within a given theory for $\Delta F = 2$ dynamics one can calculate $M_{12}$, including its sign, if phase conventions are treated consistently. To be more specific: adopting a phase convention for $\frac{q}{p}$ and having $\mathcal{L}(\Delta F = 2)$ one can calculate $\frac{q}{p} (M_{12} - \frac{i}{2} \Gamma_{12}) = \frac{q}{p} \langle P^0|\mathcal{L}(\Delta F = 2)|P^0\rangle$. Then one assigns the labels $B$ and $A$ such that $\Delta M = M_B - M_A = -2Re \frac{q}{p} (M_{12} - \frac{i}{2} \Gamma_{12})$ turns out to be positive.
1.6 CKM—From a general ansatz to a specific theory

Electroweak forces can be dealt with perturbatively. Consider the $\Delta S = 1$ four-fermion transition operator: $(\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L)$. It constitutes a dimension-six operator. Yet placing such an operator—or any other operator with dimension larger than four—into the Lagrangian creates nonrenormalizable interactions. What happened is that we have started out from a renormalizable Lagrangian

$$\mathcal{L}_{CC} = g_W \bar{q}_L^{(i)} \gamma_\mu q_L^{(j)} W^\mu, \quad (72)$$

iterated it to second order in $g_W$ with $(q^{(i)}, q^{(j)}) = (u, s), (u, d)$, and then ‘integrated out’ the heavy field, namely in this case the vector boson field $W^\mu$. That way one arrives at an effective Lagrangian containing only light quarks as ‘active’ fields.

Such effective field theories have experienced a veritable renaissance in the last ten years. Constructing them in a self-consistent way is greatly helped by adopting a Wilsonian prescription:

- First one defines a field theory $\mathcal{L}(\Lambda_{UV})$ at a high ultraviolet scale $\Lambda_{UV} \gg$ germane scales of theory like $M_W, m_Q$ etc.
- For applications characterized by physical scales $\Lambda_{phys}$ one renormalizes the theory from the cutoff $\Lambda_{UV}$ down to $\Lambda_{phys}$. In doing so one integrates out the heavy degrees of freedom, i.e., with masses exceeding $\Lambda_{phys}$—like $M_W$—to arrive at an effective low energy field theory using the operator product expansion (OPE) as a tool:

$$\mathcal{L}(\Lambda_{UV}) \Rightarrow \mathcal{L}(\Lambda_{phys}) = \sum_i c_i(\Lambda_{phys}, \Lambda_{UV}, M_W, ...) \mathcal{O}_i(\Lambda_{phys}). \quad (73)$$

- The local operators $\mathcal{O}_i(\Lambda_{phys})$ contain the active dynamical fields, i.e., those with frequencies below $\mathcal{O}_i(\Lambda_{phys})$.
- Their number coefficients $c_i(\Lambda_{phys}, \Lambda_{UV}, M_W, ...)$ provide the gateway for heavy degrees of freedom with frequencies exceeding $\mathcal{O}_i(\Lambda_{phys})$ to enter. They are shaped by short-distance dynamics and therefore usually computed perturbatively.

- Lowering the value of $\mathcal{O}_i(\Lambda_{phys})$ in general changes the form of the Lagrangian: $\mathcal{L}(\Lambda_{phys}^{(1)}) \neq \mathcal{L}(\Lambda_{phys}^{(2)})$ for $\Lambda_{phys}^{(1)} \neq \Lambda_{phys}^{(2)}$. In particular, integrating out heavy degrees of freedom will induce higher-dimensional operators to emerge in the Lagrangian. In the example above integrating the $W$ field from the dimension-four term in Eq. (72) produces dimension-six four-quark operators.
- As a matter of principle observables cannot depend on the choice of $\Lambda_{phys}$: the latter primarily provides just a demarcation line:

$$\text{short distances } \ll 1/\Lambda_{phys} < \text{long distances}. \quad (74)$$

In practice, however, its value must be chosen judiciously owing to limitations of our (present) computational abilities: on the one hand we want to be able to calculate radiative corrections perturbatively and thus require $\alpha_S(\Lambda_{phys}) < 1$. Taken by itself it would suggest to choose $\Lambda_{phys}$ as large as possible. Yet on the other hand we have to evaluate hadronic matrix elements; there $\Lambda_{phys}$ can provide an UV cutoff on the momenta of the hadronic constituents. Since the tails of hadronic wave functions cannot be obtained from, say, quark models in a reliable way, one wants to pick $\Lambda_{phys}$ as low as possible. More specifically, for heavy-flavour hadrons, one can expand their matrix elements in powers of $\Lambda_{phys}/m_Q$. Thus one encounters a Scylla and Charybdis situation. A reasonable middle course can be steered by picking $\Lambda_{phys} \sim 1$ GeV, and hence I shall denote this quantity and this value by $\mu$.

Some concrete examples might illuminate these remarks.
Iterating the coupling of Eq. (72) leads to an effective current–current coupling \((\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L)\) at low energies, i.e., scales below \(M_W\). QCD radiative corrections have to be included: they affect the strength of these effective weak transition operators significantly, since they represent an expansion in \(\alpha_S\) multiplied by a numerically large logarithm \(\log(M_W/\mu)\) rather than merely \(\alpha_S\); they also create different types of such operators. On the tree graph level there is one \(\Delta S = 1\) operator, namely \((\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L)\). Including one-loop diagrams where a gluon is exchanged between quark lines, one obtains \(\mathcal{O}(\alpha_S)\) contributions to the original \((\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L)\) operator—and to the new coupling \((\bar{u}_L \gamma^\mu t_L)(\bar{d}_L \gamma^\mu t_L)\), where the \(t^i\) denote the generators of colour \(SU(3)\). That is, the two operators \(O^{1\times 1} = (\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L)\) and \(O^{8\times 8} = (\bar{u}_L \gamma^\mu t^i s_L)(\bar{d}_L \gamma^\mu t^i u_L)\), where the former [latter] represents the product of two colour-singlet [octet] currents, mix under QCD renormalization on the one-loop level:

\[
(\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L) \xrightarrow{\text{QCD 1-loop renormalization}} c_{1\times 1} O^{1\times 1} + c_{8\times 8} O^{8\times 8}
\]  

(75)

with \(c_{1\times 1} = 1 + \mathcal{O}(\alpha_S)\), whereas \(c_{8\times 8} = \mathcal{O}(\alpha_S)\). Since some of these \(\alpha_S\) corrections are actually enhanced by numerically sizeable \(\log(M_W/\mu)\) factors, they are quite significant. Therefore one wants to identify the multiplicatively renormalized transition operators with

\[
\tilde{\mathcal{O}} \xrightarrow{\text{QCD 1-loop renormalization}} \tilde{c} \tilde{\mathcal{O}} .
\]

(76)

This can be done even without brute-force computations by relying on isospin arguments: consider the weak scattering process between quarks

\[
s_L + u_L \rightarrow u_L + d_L
\]

(77)

proceeding in an \(S\) wave. It can be driven by two \(\Delta S = 1\) operators, namely

\[
O_{\pm} = \frac{1}{2} \left[ (\bar{u}_L \gamma^\mu s_L)(\bar{d}_L \gamma^\mu u_L) \pm (\bar{d}_L \gamma^\mu s_L)(\bar{u}_L \gamma^\mu u_L) \right] .
\]

(78)

The operator \(O_{+}[O_{-}]\) produces an \(ud\) pair in the final state that is [anti]symmetric in isospin and thus carries \(I = 1[0]\); since the initial \(su\) pair carries \(I = 1/2\), \(O_{+}[O_{-}]\) generates \(\Delta I = 1/2\) transitions. With QCD conserving isospin, its radiative corrections cannot mix the operators \(O_{\pm}\), which therefore are multiplicatively renormalized:

\[
O_{+}[O_{-}] \xrightarrow{\text{QCD 1-loop renormalization}} c_{+} O_{+} + c_{-} O_{-} ,
\]

(79)

and therefore

\[
\mathcal{L}_{\text{eff}}^{(0)}(\Delta S = 1) = O_{+} + O_{-} \xrightarrow{\text{QCD 1-loop ren}} \mathcal{L}_{\text{eff}}(\Delta S = 1) = c_{+} O_{+} + c_{-} O_{-}
\]

(80)

with \(c_{\pm} = 1 + \mathcal{O}(\alpha_S)\). Integrating out those loops containing a \(W\) line in addition to the gluon line and two quark lines yields terms \(\propto \alpha_S \log(M_W^2/\mu^2)\), which are not necessarily small. Using the renormalization group equation to sum those terms one finds on the leading log level

\[
c_{\pm} = \left[ \frac{\alpha_S(M_W^2)}{\alpha_S(\mu^2)} \right]^{\gamma_{\pm}}, \quad \gamma_+ = \frac{6}{33 - 2N_F} = -\frac{1}{2} \gamma_- .
\]

(81)

That is\(^{10}\),

\[
c_{-} > 1 > c_{+}, \quad c_{-} c_{+}^2 = 1 .
\]

(82)

\(^{10}\)The expressions of Eq. (81) hold in the ‘leading log approximation’; including terms \(\propto \alpha_S^{n+1} \log^n(M_W^2/\mu^2)\) modifies them, yet \(c_{-} > 1 > c_{+}\) and \(c_{-} c_{+}^2 \simeq 1\) still hold.
That means that QCD radiative corrections provide a quite sizeable $\Delta I = 1/2$ enhancement. Corresponding effects arise for $\mathcal{L}_{\text{eff}}(\Delta C/B = 1)$. QCD radiative corrections create yet another effect, namely they lead to the emergence of `penguin' operators. Without the gluon line their diagram would decompose into two disconnected parts and thus not contribute to a transition operator. These penguin diagrams can drive only $\Delta I = 1/2$ modes. Furthermore in the loop all three quark families contribute; the diagram thus contains the irreducible CKM phase, i.e., it generates direct CP violation in strange decays. Similar effects arise in beauty, but not necessarily in charm decays.

- Consider $K^0 - \bar{K}^0$ oscillations, which represent $\Delta S = 2$ transitions. As explained in Section 1.5 those are driven by the off-diagonal elements of a `generalized mass matrix':

$$
\mathcal{M}_{12} = M_{12} + i/2 \Gamma_{12} = (K^0)|\mathcal{L}_{\text{eff}}(\Delta S = 2)|\bar{K}^0). \quad (83)
$$

The observables $\Delta M_K$ and $\epsilon_K$ are given in terms of $\text{Re}M_{12}$ and $\text{Im}M_{12}$, respectively. In the SM $\mathcal{L}_{\text{eff}}(\Delta S = 2)$, which generates $M_{12}$, is produced by iterating two $\Delta S = 1$ operators:

$$
\mathcal{L}_{\text{eff}}(\Delta S = 2) = \mathcal{L}(\Delta S = 1) \otimes \mathcal{L}(\Delta S = 1). \quad (84)
$$

This leads to the well-known quark box diagrams, which generate a local $\Delta S = 2$ operator. The contributions that do not depend on the mass of the internal quarks cancel against each other owing to the GIM mechanism, which leads to highly convergent diagrams. Integrating over the internal fields, namely the $W$ bosons and the top and charm quarks\(^{11}\) then yields a convergent result:

$$
\mathcal{L}_{\text{eff}}^{\text{box}}(\Delta S = 2, \mu) = \left(\frac{G_F}{4\pi}\right)^2 \cdot

\left[\xi_c^2 E(x_c)\eta_{cc} + \xi_t^2 E(x_t)\eta_{tt} + 2\xi_c \xi_t E(x_c, x_t)\eta_{ct}\right] + \frac{\alpha_s(\mu^2)}{27} \left(\bar{d}\gamma_\mu(1 - \gamma_5)s\right)^2 + h.c. \quad (85)
$$

with $\xi_i$ denoting combinations of KM parameters

$$
\xi_i = V (is)V^* (id), \quad i = c, t; \quad (86)
$$

$E(x_i)$ and $E(x_c, x_t)$ reflect the box loops with equal and different internal quarks, respectively [13]:

$$
E(x_i) = x_i \left(\frac{1}{4} + \frac{9}{4(1 - x_i)} - \frac{3}{2(1 - x_i)^2}\right) = \frac{3}{2} \left(\frac{x_i}{1 - x_i}\right)^3 \log x_i \quad (87)
$$

$$
E(x_c, x_t) = x_c x_t \left[\left(\frac{1}{4} + \frac{3}{2(1 - x_t)} - \frac{3}{4(1 - x_t)^2}\right) \left(\log x_t - \log x_c\right) - \left(\frac{x_c}{x_t - x_c}\right) + (x_c \leftrightarrow x_t) - \left(\frac{3}{4(1 - x_c)(1 - x_t)}\right)\right] \quad (88)
$$

$$
x_i = \frac{m_i^2}{M_W^2}. \quad (89)
$$

The $\eta_{ij}$ represent the QCD radiative corrections from evolving the effective Lagrangian from $M_W$ down to the internal quark mass. The factor $|\alpha_s(\mu^2)|^{-6/27}$ reflects the fact that a scale $\mu$ must be introduced at which the four-quark operator $(\bar{s}\gamma_\mu(1 - \gamma_5)d)^2$ is defined. This dependence on the auxiliary variable $\mu$ drops out when one takes the matrix element of this operator (at least when one does it correctly). Including next-to-leading log corrections one finds (for $m_t \simeq 180$ GeV) [14]:

$$
\eta_{cc} \simeq 1.38 \pm 0.20, \quad \eta_{tt} \simeq 0.57 \pm 0.01, \quad \eta_{cc} \simeq 0.47 \pm 0.04. \quad (90)
$$

\(^{11}\)The up quarks act merely as a subtraction term here.
The dominant contributions for $\Delta M(K)$ and $\epsilon_K$ are produced when (in addition to the $W^\pm$ pair) the internal quarks are charm and top, respectively. In either case the internal quarks are heavier than the external ones: $m_d, m_s \ll m_c, m_t$, and evaluating the Feynman diagrams indeed corresponds to integrating out the heavy fields. The situation is qualitatively very similar for $\Delta M(B^0)$, and in some sense even simpler: for within the SM by far the leading contribution is due to internal top quarks. Evaluating the quark box diagram with internal $W$ and top quark lines corresponds to integrating those heavy degrees of freedom out in a straightforward way leading to:

$$L_{\text{eff}}^{\text{box}}(\Delta B = 2, \mu) \simeq \left( \frac{G_F}{4\pi} \right)^2 M_W^2 \cdot \xi_t^2 E(x_t) \eta_t (\bar{q} \gamma_\mu (1 - \gamma_5) b)^2 + \text{h.c.} \quad (91)$$

with $q = d, s$.

**Homework # 4**

When one calculates $\Delta M(B)$ as a function of the top mass employing the quark box diagram, one finds, see Eq. (87)

$$\Delta M(B) \propto \left( \frac{m_t}{M_W} \right)^2 \quad \text{for} \quad m_t \gg M_W \quad (92)$$

The factor on the right-hand side reflects the familiar GIM suppression for $m_t \ll M_W$; yet for $m_t \gg M_W$ it constitutes a (huge) enhancement! It means that a low-energy observable, namely $\Delta M(B)$, is controlled more and more by a state or field at asymptotically high scales. Does this not violate decoupling theorems and even common sense? Does it violate decoupling—and if so, why is it allowed to do so—or not?

**End of Homework # 4**

While quark box diagrams contribute also to $\Gamma_{12}(\Delta S = 2)$, it would be absurd to assume they are significant. For $\Gamma_K$ is dominated by the impact of hadronic phase space causing $\Gamma(K_{\text{neat}} \rightarrow 2\pi) \gg \Gamma(K_{\text{neat}} \rightarrow 3\pi)$. Yet even beyond that it is unlikely that such a computation would make much sense: to contribute to $\Delta \Gamma_K$ the internal quark lines in the quark box diagram have to be $u$ and $\bar{u}$ quarks, i.e., lighter than the external quarks $s$ and $\bar{s}$. That means calculating this Feynman diagram does not correspond to integrating out the heavy degrees of freedom. For the same reason (and others as explained later in more detail) computing quark box diagrams tells us little of value concerning $D^0 - \bar{D}^0$ oscillations, since the internal quarks on the leading CKM level—$s$ and $\bar{s}$—are lighter than the external charm quarks.

A new and more intriguing twist concerning quark box diagrams occurs when addressing $\Delta \Gamma$ for $B^0$ mesons. Those diagrams again do not generate a local operator, since the internal charm quarks carry less than half the mass of the external $b$ quarks. Nevertheless it can be conjectured that the on-shell $\Delta B = 2$ transition operator generating $\Delta \Gamma_B$ is largely shaped by short-distance dynamics.

The main message of these more technical considerations was to show that while QCD conserves flavour, it has a highly nontrivial impact on flavour transitions by not only affecting the strength of the bare weak operator, but also inducing new types of weak transition operators on the perturbative level. In particular, QCD creates a source of direct CP violation in strange decays naturally, albeit with a significantly reduced strength.

### 1.7 The SM paradigm of large CP violation in $B$ decays

#### 1.7.1 Basics

As pointed out in Section 1.2.1, for an observable CP asymmetry to emerge in a decay one needs two different, yet coherent amplitudes to contribute. In 1979 it was pointed out that $B^0 - \bar{B}^0$ oscillations

---

12There is also a non-local $\Delta S = 2$ operator generated from the iteration of $\mathcal{L}(\Delta S = 1)$. While it presumably provides a major contribution to $\Delta m_K$, it is not sizeable for $\epsilon_K$ within the KM ansatz, as inferred from the observation that $|\epsilon'/\epsilon_K| \ll 0.05$. 

---
are well suited to satisfy this requirement for final states $f$ that can be fed both by $B^0$ and $\bar{B}^0$ decays, in particular since those oscillation rates were expected to be sizeable [15]:

$$B^0 \Rightarrow \bar{B}^0 \rightarrow f \leftarrow B^0 \hspace{1cm} \text{vs.} \hspace{1cm} \bar{B}^0 \Rightarrow B^0 \rightarrow \bar{f} \leftarrow B^0.$$ (93)

In 1980 it was predicted [16] that in particular $B_d \rightarrow \psi K_S$ should exhibit such a CP asymmetry larger by two orders of magnitude than the corresponding one in $K^0 \rightarrow 2\pi$ vs. $\bar{K}^0 \rightarrow 2\pi$, if CKM theory provides the main driver of $K_L \rightarrow \pi^+\pi^-$; even values close to 100% were suggested as conceivable. The analogous mode $B_s \rightarrow \psi\phi$ should, however, show an asymmetry not exceeding the few per cent level.

It was also suggested that in rare modes like $B_d \rightarrow K^-\pi^+$ sizeable direct CP violation could emerge due to intervention of ‘penguin’ operators [17].

We now know that these predictions were rather prescient. It should be noted that at the time of these predictions very little was known about $B$ mesons. While their existence had been inferred from the discovery of the $\Upsilon(1S - 4S)$ family at FNAL in 1977, none of their exclusive decays had been identified, and their lifetime was unknown as were a fortiori their oscillation rates. Yet the relevant formalism for CP asymmetries involving $B^0 - \bar{B}^0$ oscillations was already fully given.

Decay rates for CP conjugate channels can be expressed as follows:

$$\text{rate}(B(t) \rightarrow f) = e^{-\Gamma_B t} G_f(t)$$
$$\text{rate}(\bar{B}(t) \rightarrow \bar{f}) = e^{-\Gamma_B t} \bar{G}_f(t)$$ (94)

where CPT invariance has been invoked to assign the same lifetime $\Gamma_B^{-1}$ to $B$ and $\bar{B}$ hadrons. Obviously if

$$\frac{G_f(t)}{\bar{G}_f(t)} \neq 1$$ (95)

is observed, CP violation has been found. Yet one should keep in mind that this can manifest itself in two (or three) qualitatively different ways:

1. 

$$\frac{G_f(t)}{\bar{G}_f(t)} \neq 1 \text{ with } \frac{d}{dt} \frac{G_f(t)}{\bar{G}_f(t)} = 0;$$ (96)

i.e., the asymmetry is the same for all times of decay. This is true for direct CP violation; yet, as explained later, it also holds for CP violation in the oscillations.

2. 

$$\frac{G_f(t)}{\bar{G}_f(t)} \neq 1 \text{ with } \frac{d}{dt} \frac{G_f(t)}{\bar{G}_f(t)} \neq 0;$$ (97)

here the asymmetry varies as a function of the time of decay. This can be referred to as CP violation involving oscillations.

A straightforward application of quantum mechanics with its linear superposition principle yields [1] for $\Delta \Gamma = 0$, which holds for $B^\pm$ and $\Lambda_b$ exactly and for $B_d$ to a good approximation$^{13}$:

$$G_f(t) = \lvert T_f \rvert^2 \left[ 1 + \frac{1}{q^2} \lvert \rho_f \rvert^2 \right] + \left[ 1 - \frac{1}{q^2} \lvert \rho_f \rvert^2 \right] \cos \Delta M_B t - 2 \lvert \sin \Delta M_B t \rvert \text{Im} \frac{q}{q^2} \rho_f \right]$$
$$\bar{G}_f(t) = |T_f|^2 \left[ 1 + \frac{1}{q^2} \lvert \rho_f \rvert^2 \right] + \left[ 1 - \frac{1}{q^2} \lvert \rho_f \rvert^2 \right] \cos \Delta M_B t - 2 \lvert \sin \Delta M_B t \rvert \text{Im} \frac{q}{q^2} \rho_f \right].$$ (98)

$^{13}$Later I shall address the scenario with $B_s$, where $\Delta \Gamma$ presumably reaches a measurable level.
The amplitudes for the instantaneous \( \Delta B = 1 \) transition into a final state \( f \) are denoted by \( T_f = T(B \rightarrow f) \) and \( \bar{T}_f = T(\bar{B} \rightarrow f) \); the more intriguing case arises when one considers a transition that requires oscillations to take place.

\[
\bar{\rho}_f = \frac{\bar{T}_f}{T_f}, \quad \rho_f = \frac{T_f}{\bar{T}_f}, \quad \frac{q}{p} = \sqrt{\frac{M_{12}^2 - \frac{1}{2}\Gamma_{12}^2}{M_{12}^2 - \frac{1}{4}\Gamma_{12}^2}}.
\]  

(99)

Staring at the general expression is not always very illuminating; let us therefore consider three limiting cases:

- \( \Delta M_B = 0 \), i.e., no \( B^0 - \bar{B}^0 \) oscillations:

\[
G_f(t) = 2|T_f|^2, \quad \bar{G}_f(t) = 2|\bar{T}_f|^2 \Rightarrow \frac{\bar{G}_f(t)}{G_f(t)} = \left| \frac{\bar{T}_f}{T_f} \right|^2, \quad \frac{d}{dt} G_f(t) \equiv 0 \equiv \frac{d}{dt} \bar{G}_f(t).
\]  

(100)

This is explicitly what was referred to above as direct CP violation.

- \( \Delta M_B \neq 0 \) and \( f \) a flavour-specific final state with no direct CP violation; i.e., \( T_f = 0 = \bar{T}_f \) and \( \bar{T}_f = T_f \):

\[
G_f(t) = \left| \frac{q}{p} \right|^2 |T_f|^2 (1 - \cos \Delta M_B t), \quad \bar{G}_f(t) = \left| \frac{\bar{p}}{\bar{q}} \right|^2 |\bar{T}_f|^2 (1 - \cos \Delta M_B t) \Rightarrow \frac{\bar{G}_f(t)}{G_f(t)} = \left| \frac{\bar{p}}{\bar{q}} \right|^4, \quad \frac{d}{dt} G_f(t) \equiv 0, \quad \frac{d}{dt} \bar{G}_f(t) \neq 0 \neq \frac{d}{dt} G_f(t).
\]  

(101)

This constitutes CP violation in the oscillations. For the CP conserving decay into the flavour-specific final state is used merely to track the flavour identity of the decaying meson. This situation can therefore be denoted also in the following way:

\[
\frac{\text{Prob}(B^0 \Rightarrow B^0; t) - \text{Prob}(\bar{B}^0 \Rightarrow B^0; t)}{\text{Prob}(B^0 \Rightarrow B^0; t) + \text{Prob}(\bar{B}^0 \Rightarrow B^0; t)} = \frac{|q/p|^2 - |p/q|^2}{|q/p|^2 + |p/q|^2} = 1 - \frac{|p/q|^4}{1 + |p/q|^4}.
\]  

(102)

- \( \Delta M_B \neq 0 \) with \( f \) now being a flavour-nonspecific final state—a final state common to \( B^0 \) and \( \bar{B}^0 \) decays—of a special nature, namely a CP eigenstate—\( |\bar{f}\rangle = \text{CP}|f\rangle = \pm |f\rangle \)—without direct CP violation—\( |\bar{\rho}_f| = 1 = |\rho_f| \):

\[
G_f(t) = 2|T_f|^2 \left[ 1 - (\sin \Delta M_B t) \cdot \text{Im} \frac{q}{p} \bar{\rho}_f \right], \quad \bar{G}_f(t) = 2|\bar{T}_f|^2 \left[ 1 + (\sin \Delta M_B t) \cdot \text{Im} \frac{\bar{p}}{\bar{q}} \bar{\rho}_f \right] \Rightarrow \frac{\bar{G}_f(t) - G_f(t)}{G_f(t) + \bar{G}_f(t)} = (\sin \Delta M_B t) \cdot \text{Im} \frac{q}{p} \bar{\rho}_f
\]  

(103)

is the concrete realization of what was called CP violation involving oscillations. For \( f \) still denoting a CP eigenstate, yet with \( |\bar{\rho}_f| \neq 1 \) one has the more complex asymmetry expression

\[
\frac{\bar{G}_f(t) - G_f(t)}{G_f(t) + \bar{G}_f(t)} = S_f \cdot (\sin \Delta M_B t) - C_f \cdot (\cos \Delta M_B t)
\]  

with

\[
S_f = \frac{2 \text{Im} \frac{q}{p} \bar{\rho}_f}{1 + \left| \frac{q}{p} \bar{\rho}_f \right|^2}, \quad C_f = \frac{1 - \left| \frac{q}{p} \bar{\rho}_f \right|^2}{1 + \left| \frac{q}{p} \bar{\rho}_f \right|^2}.
\]  

(104)

(105)

\[^{14}\text{For a flavour-specific mode one has in general } T_f \cdot \bar{T}_f = 0; \text{ the more intriguing case arises when one considers a transition that requires oscillations to take place.}\]
For the decays of neutral mesons the following general statement is relevant, at least conceptually.

**Theorem:**

Consider a beam with an arbitrary combination of neutral mesons $B^0$ and $\bar{B}^0$ decaying into a final state $f$ that is a CP eigenstate. If the decay rate evolution in (proper) time $t$ is not described by a single exponential, i.e.,

$$\text{rate}(B^0/\bar{B}^0(t) \to f) \neq Ke^{-\Gamma t} \quad \text{with} \quad \frac{d}{dt}K \equiv 0$$

for any real $\Gamma$, then CP invariance is violated.

**Homework # 5**

Prove this theorem.

**End of Homework # 5**

An obvious, yet still useful criterion for CP observables is that they must be ‘re-phasing’ invariant under $|\bar{B}^0 \to e^{-i\xi}|\bar{B}^0\rangle$. The expressions above show there are three classes of such observables:

1. An asymmetry in the instantaneous transition amplitudes for CP conjugate modes:

   $$|T(B \to f)| \neq |T(\bar{B} \to \bar{f})| \quad \iff \quad \Delta B = 1 \ .$$  \hspace{1cm} (107)

   It reflects pure $\Delta B = 1$ dynamics and thus amounts to direct CP violation. Those modes are most likely to be nonleptonic; in the SM they practically have to be.

2. CP violation in $B^0 - \bar{B}^0$ oscillations:

   $$|q| \neq |p| \quad \iff \quad \Delta B = 2 \ .$$ \hspace{1cm} (108)

   It requires CP violation in $\Delta B = 2$ dynamics. The theoretically cleanest modes here are semileptonic ones due to the SM $\Delta Q = \Delta B$ selection rule.

3. CP asymmetries involving oscillations\textsuperscript{15}:

   $$\text{Im} \frac{q}{p} \bar{\rho}(f) \neq 0 \ ; \ \bar{\rho}(f) = \frac{T(B \to f)}{T(\bar{B} \to \bar{f})} \quad \iff \quad \Delta B = 1 & 2 \ .$$ \hspace{1cm} (109)

Such an effect requires the interplay of $\Delta B = 1 & 2$ forces.

While $C_f \neq 0$ unequivocally signals direct CP violation in Eq. (104), the interpretation of $S_f \neq 0$ is more complex. (i) As long as one has measured $S_f$ only in a single mode, the distinction between direct and indirect CP violation—i.e., CP violation in $\Delta B = 1$ and $\Delta B = 2$ dynamics—is convention dependent, since a change in phase for $\bar{B}^0 \to e^{-i\xi}|\bar{B}^0\rangle$—leads to $\bar{\rho}_f \to e^{-i\xi}\bar{\rho}_f$ and $(q/p) \to e^{i\xi}(q/p)$, i.e., can shift any phase from $(q/p)$ to $\bar{\rho}_f$ and back while leaving unchanged. However, once $S_f$ has been measured for two different final states $f$, then the distinction becomes truly meaningful independent of theory: $S_{f_1} \neq S_{f_2}$ implies $(q/p)\bar{\rho}_{f_1} \neq (q/p)\bar{\rho}_{f_2}$ and thus $\bar{\rho}_{f_1} \neq \bar{\rho}_{f_2}$, i.e., CP violation in the $\Delta B = 1$ sector. One should note that this direct CP violation might not generate a $C_f$ term, since the conditions laid out below in Section 2.1.6 might not be satisfied. For $\bar{\rho}_{f_1} = e^{i\delta_1}$ and $\bar{\rho}_{f_2} = e^{i\delta_2}$ causing $S_{f_1} \neq S_{f_2}$ would both lead to $C_{f_1} = 0 = C_{f_2}$.

\textsuperscript{15}This condition is formulated for the simplest case of $f$ being a CP eigenstate.
Once the final state consists of more than two pseudoscalar or one pseudoscalar and one vector meson, it contains more dynamical information than expressed through the decay width into it, as can be described through a Dalitz plot.

- Accordingly one can have a CP asymmetry in final-state distributions of $B$ mesons, as discussed later. There is a precedent for such an effect, namely a $T$-odd correlation that has been observed between the $\pi^+ - \pi^-$ and $e^+ - e^-$ planes in the rare mode $K_L \to \pi^+ \pi^- e^+ e^-$, the size of which can be inferred from $K_L \to \pi^+ \pi^-$.  

1.7.2 The first central pillar of the paradigm: long lifetimes

Beauty, the existence of which had been telegraphed by the discovery of the $\tau$ as the third charged lepton, was indeed observed exhibiting a surprising feature: starting in the early 1980s its lifetime was found to be about $10^{-12}$ s. This was considered ‘long’. For one can get an estimate for $\tau(B)$ by relating it to the muon lifetime:

$$
\tau(B) \approx \tau_{\mu} \sim \tau(\mu) \left( \frac{m(\mu)}{m(b)} \right)^5 \frac{1}{9} \frac{1}{|V(cb)|^2} \simeq 3 \cdot 10^{-14} \left| \sin \theta_C \right|^2 s.
$$

(110)

One had expected $|V(cb)|$ to be suppressed, since it represents an out-of-family coupling. Yet one had assumed without deeper reflection that $|V(cb)| \sim \sin \theta_C$ — what else could it be? The measured value for $\tau(B)$ however pointed to $|V(cb)| \sim |\sin \theta_C|^2$. By the end of the millenium one had obtained a rather accurate value: $\tau(B_d) = (1.55 \pm 0.04) \cdot 10^{-12}$ s. Now the data have become even more precise:

$$
\tau(B_d) = (1.530 \pm 0.009) \cdot 10^{-12} s, \quad \tau(B^\pm_0)/\tau(B_d) = 1.071 \pm 0.009 .
$$

(111)

The lifetime ratio, which reflects the impact of hadronization, had been predicted [18] successfully, well before data of the required accuracy was available.

1.7.3 Oscillations of $B_d$ and $B_s$ mesons – exactly like for kaons, only different

The general phenomenology of $B^0 - B^\dag_0$ oscillations posed no mystery from the beginning, since it follows a close qualitative—though not quantitative—analogy with kaon oscillations described above. One obvious difference arises in the lifetime ratios of the two mass eigenstates: the huge disparity in the $K_L$ and $K_S$ lifetimes—$\tau(K_L) \sim 600 \tau(K_S)$—is due to the kinematical ‘accident’ that the kaon is barely above the three-pion threshold; this does not have an analogue for the heavier mesons, where one expects on general grounds $\Delta \Gamma \ll 1$, to be quantified below.

The most general observable signature of oscillations is the apparent violation of some selection rule. In the SM one has

$$
l^- \bar{\nu} X^+_e \not\to B^0 \to l^+ \nu X^-_e \not\to \bar{B}^0 \to l^- \bar{\nu} X^+_e .
$$

(112)

Yet oscillations can circumvent it in the following way:

$$
B^0 \Longrightarrow \bar{B}^0 \to l^- \nu + X^+_e , \quad \bar{B}^0 \Longrightarrow B^0 \to l^+ \nu X^-_e ,
$$

(113)

where “$\Longrightarrow$” and “$\not\to$” denote the $\Delta B = 2$ oscillation and $\Delta B = 1$ direct transitions, respectively. This apparent violation of the selection rule exhibits a characteristic dependence on the time of decay analogous to that of Eq. (40) where $\Delta \Gamma = 0$ has been set for simplicity:

$$
\text{rate}(B^0 \to l^- X^+_e ; t) \propto \frac{1}{2} e^{-\Gamma_B t} (1 - \cos \Delta M_B t)
$$

(114)

$$
\text{rate}(B^0 \to l^+ X^-_e ; t) \propto \frac{1}{2} e^{-\Gamma_B t} (1 + \cos \Delta M_B t) .
$$

(115)
Integrating over all times of decay one finds for the ratio of wrong- to right-sign leptons and for the probability of wrong-sign leptons

\[ r_B = \frac{\Gamma(B^0 \to l^+\nu_X) - \Gamma(B^0 \to l^-\bar{\nu}_X)}{2 \Gamma(B^0 \to l^-\bar{\nu}_X) + \Gamma(B^0 \to l^+\nu_X)} \quad \text{and} \quad \chi_B = \frac{\Gamma(B^0 \to l^+\nu_X)}{\Gamma(B^0 \to l^-\bar{\nu}_X)} = \frac{r_B}{1 + r_B} \]  

The quantities \( r_B \) and \( \chi_B \) thus represent the violation of the selection rule of Eq. (112) ‘on average’. Maximal oscillations can be defined as \( x \gg 1 \) and thus \( r \to 1 \) and \( \chi \to 1/2 \).

1.7.3.1 \( B_d - \bar{B}_d \) oscillations

Present data yield for \( B_d \) mesons:

\[ x_d = x_{B_d} = 0.776 \pm 0.008 \quad \text{and} \quad \chi_d = 0.188 \pm 0.003 \]  

Huge samples of beauty mesons can be obtained in \( p\bar{p} \) or \( pp \) collisions at high energies, which yield incoherent pairs of \( B \) mesons. Two cases have to be distinguished:

\[ p\bar{p} \to B_d + X/B^-B_d + X \]  

leading to a single beam of neutral \( B \) mesons, for which Eq. (116) applies.

\[ p\bar{p} \to B_d\bar{B}_d + X \]  

when both \( B \) mesons can oscillate—actually into each other—leading to like-sign di-leptons

\[ p\bar{p} \to B_d\bar{B}_d + X \implies B_dB_d/\bar{B}_d\bar{B}_d + X \to l^\pm l'^\pm + X' \]  

Its relative probability can be expressed as follows

\[ \frac{\text{rate}(p\bar{p} \to B_d\bar{B}_d + X \to l^\pm l'^\pm + X')}{\text{rate}(p\bar{p} \to B_dB_d + X \to ll + X')} = 2\chi_d(1 - \chi_d) \]  

meaning that like-sign di-leptons require one \( B \) meson to have oscillated into its antiparticle at its time of decay, while the other one has not.

- In

\[ e^+e^- \to B_d\bar{B}_d \]  

one encounters the coherent production of two neutral beauty mesons. As discussed in detail in Section 2.1.3 EPR correlations combine with the requirement of Bose–Einstein statistics to make the pair act as a single oscillating system leading to [16]

\[ \frac{\text{rate}(e^+e^- \to B_d\bar{B}_d \to l^\pm l'^\pm + X)}{\text{rate}(e^+e^- \to B_dB_d \to ll + X')} = \chi_d \]  

For the measured value of \( x_d \) the two expressions in Eqs. (121) and (123) yield

\[ \chi_d = 0.188 \pm 0.003 \quad \text{vs.} \quad 2\chi_d(1 - \chi_d) \approx 0.305 \]  

i.e., the two ratios of like-sign dileptons to all dileptons emerging from the decays of a coherently and incoherently produced \( B_d\bar{B}_d \) pair differ by a factor of almost two owing to EPR correlations as explained below.
One predicts on rather general grounds that $L(\Delta B = 2)$ is dominated by short distance dynamics and more specifically by the quark box diagram to a higher degree than $L(\Delta S = 2)$. It is often stated that $B_d - \bar{B}_d$ oscillations were found to proceed much faster than predicted. It is correct—yet one should note the main reason for it. The prediction for $x_B$ depends very much on the value of the top quark mass $m_t$, see Eq. (92) for a rough scaling law. In the early 1980s there had been the experimental claim by the UA1 Collaboration that top quarks had been discovered in $p\bar{p}$ collisions with a mass $m_t = 40 \pm 10$ GeV. With $x_B \propto m_t^2$ and $r_B \propto x_B^2 \propto m_t^4$ for moderate values of $x_B$, one finds $r_B$ increases by more than one order of magnitude when going from $m_t = 40$ GeV to 170 GeV! Once the ARGUS Collaboration discovered $B_d - \bar{B}_d$ oscillations with $x_d \sim 0.7$, theorists quickly concluded that top quarks had to be much heavier than previously considered, namely $m_t > 100$ GeV. This was the first indirect evidence for top quarks being ‘super-heavy’. A second and more accurate indirect piece of evidence came later from studying electroweak radiative corrections at LEP.

Since $x = \Delta M/B$ denotes the ratio between the oscillation and decay rates, $x = 1$ represents the optimal realization of the scenario sketched in Eq. (93) for obtaining a CP asymmetry, namely to rely on oscillations to provide a second coherent amplitude of a comparable effective strength. This statement can be made more quantitative by integrating the asymmetry of Eq. (103) over all times of decay $t$:

\[
\begin{align*}
\text{rate} \left( B^0(t) \to f \right) &= Ke^{-\Gamma_B t} \left( 1 - A \cdot \sin(x_\text{B}\Gamma_B t) \right), \quad x = \frac{\Delta M_B}{\Gamma_B} \\
\int_0^\infty dt \text{rate} \left( B^0(t) \to f \right) &= K \frac{1}{\Gamma_B} \left( 1 - A \cdot \frac{x}{1 + x^2} \right).
\end{align*}
\]

The oscillation induced factor $x/(1 + x^2)$ is maximal for $x = 1$; i.e., with Eq. (117) nature has given us an almost optimal stage for observing CP violation in $B_d$ decays.

1.7.3.2 The ‘hot’ news: $B_s - \bar{B}_s$ oscillations

For a moment I shall deviate considerably from the historical sequence by presenting the ‘hot’ news of the resolution of $B_s - \bar{B}_s$ oscillations.

Nature actually provided us with an ‘encore’ in $B^0$ oscillations. It had been recognized from the beginning that within the SM one predicts $\Delta M_{B_s} \gg \Delta M_{B_d}$, i.e., that $B_s$ mesons oscillate much faster than $B_d$ mesons. Both receive their dominant contributions from $t\bar{t}$ quarks in the quark box diagram making their ratio depend on the CKM parameters and the hadronic matrix element of the relevant four-quark operator only:

\[
\frac{\Delta M_{B_s}}{\Delta M_{B_d}} \simeq \frac{B_s f^2_{B_s}}{B_d f^2_{B_d}} \frac{|V(ts)|^2}{|V(td)|^2}.
\]

This relation also exhibits the phenomenological interest in measuring $\Delta M_{B_s}$, namely to obtain an accurate value for $|V(td)|$. Lattice QCD is usually invoked to gain theoretical control over the first ratio of hadronic quantities. Taking its findings together with the CKM constraints on $|V(ts)/V(ts)|$ yields the following SM prediction:

\[
\Delta M_{B_s}|_{SM} = (18.3^{+6.5}_{-1.5}) \text{ ps}^{-1} \equiv (1.20^{+0.43}_{-0.10}) \cdot 10^{-2} \text{ eV} \quad \text{CKM fit}.
\]

Those rapid oscillations have been resolved now by CDF [19] and D0 [20]:

\[
\begin{align*}
\Delta M_{B_s} &= \begin{cases} 
(19 \pm 2) \text{ ps}^{-1} & \text{D0} \\
(17.77 \pm 0.10 \pm 0.07) \text{ ps}^{-1} & \text{CDF}
\end{cases} \\
x_s &= \frac{\Delta M_{B_s}}{\Gamma_{B_s}} \simeq 25.
\end{align*}
\]

These findings represent another triumph of CKM theory even more impressive than a mere comparison of the observed and predicted values of $\Delta M(B_s)$, as explained later.
There is also marginal evidence for $\Delta \Gamma_{B_s} \neq 0$ [21]  
\[
\frac{\Delta \Gamma_{B_s}}{\bar{\Gamma}_{B_s}} = 0.31 \pm 0.13 .
\] (131)

My heart wishes that $\Delta \Gamma_{B_s}$ were indeed as large as 0.5 or even larger. For it would open up a whole new realm of CP studies in $B_s$ decays with a great potential to identify New Physics. Yet my head tells me that values exceeding 0.25 or so are very unlikely; it would point at a severe limitation in our theoretical understanding of $B$ lifetimes. For only on-shell intermediate states $f$ in $B^0 \to f \to \bar{B}^0$ can contribute to $\Delta \Gamma(B)$, and for $B^0 = B_s$ these are predominantly driven by $b \to c\bar{c}s$. Let $R(b \to c\bar{c}s)$ denote their fraction of all $B_s$ decays. If these transitions contribute only to $\Gamma(B_s(CP = +))$ one has $\Delta \Gamma_s/\bar{\Gamma}_s = 2R(b \to c\bar{c}s)$. Of course this is actually an upper bound quite unlikely to be even remotely saturated. With the estimate $R(b \to c\bar{c}s) \approx 25\%$, which is consistent with the data on the charm content of $B_{u,d}$ decays this upper bound reads 50%. More realistic calculations have yielded considerably smaller predictions:

\[
\frac{\Delta \Gamma_s}{\bar{\Gamma}_s} = \begin{cases} 
0.22 \cdot \left( \frac{f(B_u)}{220 \text{ MeV}} \right)^2 \
0.12 \pm 0.05
\end{cases} ;
\] (132)

where the two predictions are taken from Refs. [22] and [23], respectively. A value as high as 0.20–0.25 is thus not out of the question theoretically, and Eq. (131) is still consistent with it. One should note that invoking New Physics would actually “backfire” since it leads to a lower prediction. If, however, a value exceeding 0.25 were established experimentally, we would have to draw at least one of the following conclusions: (i) $R(b \to c\bar{c}s)$ actually exceeds the estimate of 0.25 significantly. This would imply at the very least that the charm content is higher in $B_s$ than $B_{u,d}$ decays by a commensurate amount and the $B_s$ semileptonic branching ratio lower. (ii) Such an enhancement of $R(b \to c\bar{c}s)$ would presumably—though not necessarily—imply that the average $B_s$ width exceeds the $B_d$ width by more than the predicted 1–2% level. That means in analysing $B_s$ lifetimes one should allow $\bar{\tau}(B_s)$ to float freely. (iii) If in the end one found the charm content of $B_s$ and $B$ decays to be quite similar and $\bar{\tau}(B_s) \approx \tau(B_d)$, yet $\Delta \Gamma_s/\bar{\Gamma}_s$ to exceed 0.25, we would have to concede a loss of theoretical control over $\Delta \Gamma$. This would be disappointing, yet not inconceivable: the a priori reasonable ansatz of evaluating both $\Delta \Gamma_B$ and $\Delta M_B$ from quark box diagrams—with the only manifest difference being that the internal quarks are charm in the former and top in the latter case—obscures the fact that the dynamical situation is actually different. In the latter case the effective transition operator is a local one involving a considerable amount of averaging over off-shell transitions; the former is shaped by on-shell channels with a relatively small amount of phase space: for the $B_s$ resides barely 1.5 GeV above the $D_s\bar{D}_s$ threshold. To say it differently: the observable $\Delta \Gamma_s$ is more vulnerable to limitations of quark–hadron duality than $\Delta M_s$ and even beauty lifetimes.\footnote{These are all dominated by nonleptonic transitions, where duality violations can be significantly larger than for semileptonic modes.}

In summary: establishing $\Delta \Gamma_s \neq 0$ amounts to important qualitative progress in our knowledge of beauty hadrons; it can be of great practical help in providing us with novel probes of CP violations in $B_s$ decays, and it can provide us theorists with a reality check concerning the reliability of our theoretical tools for nonleptonic $B$ decays.

1.7.4 Large CP asymmetries in $B$ decays without ‘plausible deniability’

The above-mentioned observation of a long $B$ lifetime pointed to $|V(cb)| \sim \mathcal{O}(\lambda^2)$ with $\lambda = \sin \theta_C$. Together with the expected observation $|V(ub)| \ll |V(cb)|$ and coupled with the assumption of three-family unitarity this allows one to expand the CKM matrix in powers of $\lambda$, which yields the following
most intriguing result through order $\lambda^5$, as first recognized by Wolfenstein:

$$V_{CKM} = \begin{pmatrix}
1 - \frac{1}{2} \lambda^2 & \lambda & A \lambda^3 (\rho - i \eta + i \frac{7}{2} \eta \lambda^2) \\
-\lambda & 1 - \frac{1}{2} \lambda^2 - i \eta A \lambda^2 & A \lambda^2 (1 + i \eta \lambda^2) \\
A \lambda^2 (1 - \rho - i \eta) & -A \lambda^2 & 1
\end{pmatrix}.$$  \hspace{1cm} (133)

The three Euler angles and one complex phase of the representation given in Eq. (30) is taken over by

$$V_{CKM} = \begin{pmatrix}
1 & O(\lambda) & O(\lambda^3) \\
O(\lambda) & 1 & O(\lambda^2) \\
O(\lambda^3) & O(\lambda^2) & 1
\end{pmatrix}, \quad \lambda = \sin \theta_C.$$  \hspace{1cm} (134)

As explained in Section 1.4.2, we know this matrix has to be unitary. Yet in addition it is almost the

identity matrix, almost symmetric and the moduli of its elements shrink with the distance from the
diagonal. It has to contain a message from nature—albeit in a highly encoded form.

My view of the situation is best described by a poem by the German poet Joseph von Eichendorff

from the late romantic period$^{17}$:

Schläft ein Lied in allen Dingen,  There sleeps a song in all things
die da träumen fort und fort,  that dream on and on,
und die Welt hebt an zu singen,  and the world will start to sing,
findst Du nur das Zauberwort.  if you find the magic word.

The six triangle relations obtained from the unitarity condition fall into three categories:

1. $K^0$ triangle:

$$V^*(ud)V(us) + \frac{V^*(cd)V(cs)}{O(\lambda)} + \frac{V^*(td)V(ts)}{O(\lambda^5)} = \delta_{ds} = 0;$$  \hspace{1cm} (135)

$D^0$ triangle:

$$V^*(ud)V(cd) + \frac{V^*(us)V(cs)}{O(\lambda)} + \frac{V^*(ub)V(cb)}{O(\lambda^5)} = \delta_{uc} = 0,$$  \hspace{1cm} (136)

where below each product of matrix elements I have noted their size in powers of $\lambda$. These two

triangles are extremely ‘squashed’: two sides are of order $\lambda$, the third one of order $\lambda^5$ and their

ratio of order $\lambda^4 \simeq 2.3 \cdot 10^{-3}$; Eq. (135) and Eq. (136) control the situation in strange and charm
decays; the relevant weak phases there are obviously tiny.

2. $B_s$ triangle:

$$V^*(us)V(ub) + \frac{V^*(cs)V(cb)}{O(\lambda^4)} + \frac{V^*(ts)V(tb)}{O(\lambda^2)} = \delta_{sb} = 0.$$  \hspace{1cm} (137)

$tc$ triangle:

$$V^*(td)V(cd) + \frac{V^*(ts)V(cs)}{O(\lambda^4)} + \frac{V^*(tb)V(cb)}{O(\lambda^2)} = \delta_{ct} = 0.$$  \hspace{1cm} (138)

The third and fourth triangles are still rather squashed, yet less so: two sides are of order $\lambda^2$ and

the third one of order $\lambda^4$.

$^{17}$I have been told that early romantic writers would have used the term ‘symmetry’ instead of ‘song’.
The discovery of $B_d - \bar{B}_d$ oscillations defined the ‘CKM paradigm of large CP violation in $B$ decays’ that had been anticipated in 1980:

- A host of nonleptonic $B$ channels has to exhibit sizeable CP asymmetries.

Let the reader be reminded that all six triangles, despite their very different shapes, have the same area, see Eq. (34), reflecting the single CKM phase for three families.

Some comments on notation might not be completely useless. The BABAR collaboration and its followers refer to the three angles of the CKM unitarity triangle as $\alpha$, $\beta$ and $\gamma$; the BELLE collaboration instead has adopted the notation $\phi_1$, $\phi_2$ and $\phi_3$. While it poses no problem to be conversant in both languages, the latter has not only historical priority on its side [24], but is also more rational. For the angles $\phi_i$ in the ‘bd’ triangle of Eq. (139) are always opposite the side defined by $V^{*}(id)V(ib)$. Furthermore this classification scheme can readily be generalized to all six unitarity triangles; those triangles can be labelled by $kl$ with $k \neq l = d, s, b$ or $k \neq l = u, c, t$, see Eqs. (135)–(140). Its 18 angles can then be unambiguously denoted by $\phi^{kl}_i$: it is the angle in triangle $kl$ opposite the side $V^{*}(ik)V(il)$ or $V^{*}(ki)V(il)$, respectively. Therefore I view the notation $\phi^{(kl)}_i$ as the only truly Cartesian one.

The discovery of $B_d - \bar{B}_d$ oscillations defined the ‘CKM paradigm of large CP violation in $B$ decays’ that had been anticipated in 1980:

- A host of nonleptonic $B$ channels has to exhibit sizeable CP asymmetries.
– For $B_d$ decays to flavour-nonspecific final states (like CP eigenstates) the CP asymmetries depend on the time of decay in a very characteristic manner; their size should typically be measured in units of 10% rather than 0.1%.

– There is no plausible deniability for the CKM description, if such asymmetries are not found.

– For $m_t \geq 150$ GeV the SM prediction for $\epsilon_K$ is dominated by the top quark contribution like $\Delta M_{B_d}$. It thus drops out of their ratio, and $\sin 2\phi_1$ can be predicted within the SM irrespective of the (superheavy) top quark mass. In the early 1990s, i.e., before the direct discovery of top quarks, it was predicted \[ \epsilon_K \Delta M_{B_d} \propto \sin 2\phi_1 \sim 0.6 - 0.7 \] with values for $B_B f_B^2$ inserted as now estimated by LQCD.

– The CP asymmetry in the Cabibbo favoured channels $B_s \to \psi \phi/\psi \eta$ is Cabibbo suppressed, i.e., below 4%, for reasons very specific to CKM theory, as pointed out already in 1980 [16].

In 1974 top quarks were finally observed directly with a mass fully consistent with the indirect estimates given above; the most recent analyses from CDF and D0 list

$$ m_t = 172.7 \pm 2.9 \text{ GeV} \, . \tag{143} $$

### 1.7.5 Data in 1998

CP violation had been observed only in the decays of neutral kaons, and all its manifestations—$K_L \to \pi^+\pi^-$, $\pi^0\pi^0$, $K^0 \to \pi^+\pi^-$ vs. $K^0 \to \pi^-\pi^+$, $K_L \to l^+\nu\pi^-$ vs. $K_L \to l^-\bar{\nu}\pi^+$—could be described for 35 years with a single real number, namely $|\eta_{+-}|$ or $\Phi(\Delta S = 2) = \arg(M_{12}/\Gamma_{12})$.

There was intriguing, though not conclusive evidence for direct CP violation:

$$ \frac{\epsilon'}{\epsilon_K} = \left\{ \begin{array}{ll} (2.30 \pm 0.65) \cdot 10^{-3} & \text{NA31} \\
(0.74 \pm 0.59) \cdot 10^{-3} & \text{E731} \end{array} \right. \, . \tag{144} $$

These measurements were made in the 1980s and had been launched by theory guestimates suggesting values for $\epsilon'$ that would be within the reach of these experiments. Theory, however, had ‘moved on’ favouring values $\leq 10^{-3}$—or so it was claimed.

### 1.8 Completion of a heroic era

Direct CP violation was unequivocally established in 1999. The present world average dominated by the data from NA48 and KTeV reads as follows [26]:

$$ \langle \epsilon'/\epsilon_K \rangle = (1.63 \pm 0.22) \cdot 10^{-3} \, . \tag{145} $$

Quoting the result in this way does not do justice to the experimental achievement, since $\epsilon_K$ is a very small number itself. The sensitivity achieved becomes more obvious when quoted in terms of actual widths [26]:

$$ \frac{\Gamma(K^0 \to \pi^+\pi^-) - \Gamma(K^0 \to \pi^+\pi^-)}{\Gamma(K^0 \to \pi^+\pi^-) + \Gamma(K^0 \to \pi^+\pi^-)} = (5.04 \pm 0.82) \cdot 10^{-6} \, ! \tag{146} $$

This represents a discovery of the very first rank.\(^\dagger\) Its significance does not depend on whether the SM can reproduce it or not—which is the most concise confirmation of how important it is. The HEP community can take pride in this achievement; the tale behind it is a most fascinating one about imagination.

\(^\dagger\)As a consequence of Eq. (146) I am not impressed by CPT tests falling short of the $10^{-6}$ level.
and perseverance. The two groups and their predecessors deserve our respect; they have certainly earned my admiration.

The experimental findings are consistent with CKM theory on the qualitative level, since the latter does not represent a superweak scenario even for strange decays due to the existence of penguin operators. It is not inconsistent with it even quantitatively. One should keep in mind that within the SM $\epsilon'/\epsilon_K$ has to be considerably suppressed. For $\epsilon'$ requires interference between $\Delta I = 1/2$ and $3/2$ amplitudes and is thus reduced by the ‘$\Delta I = 1/2$ rule’: $|T(\Delta I = 3/2)/T(\Delta I = 1/2)| \sim 1/20$. Furthermore $\epsilon'$ is generated by loop diagrams—as is $\epsilon_K$; yet the top quark mass enhances $\epsilon_K$ powerlike—$|\epsilon_K| \propto m_t^2/M_W^2$—whereas $\epsilon'$ only logarithmically. When there is only one weak phase—as is the case for CKM theory—one has $|\epsilon'/\epsilon_K| \propto \log m_t^2/m_t^2$, i.e., greatly reduced again for superheavy top quarks (revisit Homework # 4).

CKM theory can go beyond such semiquantitative statements, but one should not expect a precise prediction from it in the near future. For the problem of uncertainties in the evaluation of hadronic matrix elements is compounded by the fact that the two main contributions to $\epsilon'$ are similar in magnitude, yet opposite in sign [27].

1.8.1 CKM theory at the end of the second millenium

It is indeed true that large fractions of the observed values for $\Delta M_K$, $\epsilon_K$, and $\Delta M_B$ and even most of $\epsilon'$ could be due to New Physics given the limitations in our theoretical control over hadronic matrix elements. Equivalently constraints from these and other data translate into ‘broad’ bands in plots of the unitarity triangle, see Fig. 3.

The problem with this statement is that it is not even wrong—it misses the real point. Let me illustrate it by a local example first. If you plot the whereabouts of the students at this school on a local map, you would find a seemingly broad band stretching from Aronsborg to Stockholm and Uppsala; however when you look at the ‘big’ picture—say a map of Europe—you realize these students are very closely bunched together in one tiny spot on the map. This cannot be by accident, there has to be a good reason for it, which, I hope, is obvious in this specific case. Likewise for the problem at hand: observables like $\Gamma(B \to \nu X_{c,u})$, $\Gamma(K \to \nu \nu)$, $\Delta M_K$, $\Delta M_B$, $\epsilon_K$ and $\sin 2\phi_1$ etc., represent very different dynamical regimes that proceed on time-scales that span several orders of magnitude. The very fact that CKM theory can accommodate such diverse observables always within a factor two or better and relate them in such a manner that its parameters can be plotted as meaningful constraints on

\[Fig. 3: \text{The CKM Unitarity Triangle fit}\]
a triangle is highly nontrivial and—in my view—must reflect some underlying, yet unknown dynamical layer. Furthermore the CKM parameters exhibit an unusual hierarchical pattern—\( |V(ud)| \sim |V(cs)| \sim |V(ts)| \sim 1, |V(us)| \simeq |V(cd)| \simeq \lambda, |V(cb)| \sim |V(ts)| \sim \mathcal{O}(\lambda^2), |V(ub)| \sim |V(td)| \sim \mathcal{O}(\lambda^3) \)—as do the quark masses culminating in \( m_t \simeq 175 \text{ GeV} \). Picking such values for these parameters would have been seen as frivolous at best—had they not been forced upon us by (independent) data. Thus I view it already as a big success for CKM theory that the experimental constraints on its parameters can be represented through triangle plots in a meaningful way.

**Interlude: Singing the Praise of Hadronization**

Hadronization and nonperturbative dynamics in general are usually viewed as unwelcome complication, if not outright nuisances. A case in point was already mentioned: while I view the CKM predictions for \( \Delta M_K, \Delta M_B, \epsilon_K \) to be in remarkable agreement with the data, significant contributions from New Physics could be hiding there behind the theoretical uncertainties due to lack of computational control over hadronization. Yet *without* hadronization, bound states of quarks and antiquarks will not form; without the existence of kaons \( K^0 - \bar{K}^0 \) oscillations obviously cannot occur. It is hadronization that provides the ‘cooling’ of the (anti)quark degrees of freedom, which allows subtle quantum mechanical effects to add up coherently over macroscopic distances. Otherwise one would not have access to a super-tiny energy difference \( \text{Im} M_{12} \sim 10^{-8} \text{ eV} \), which is very sensitive to different layers of dynamics, and indirect CP violation could not manifest itself. The same would hold for \( B \) mesons and \( B^0 - \bar{B}^0 \) oscillations.

To express it in a more down-to-earth way:

- Hadronization leads to the formation of kaons and pions with masses exceeding greatly (current) quark masses. It is the hadronic phase space that suppresses the CP conserving rate for \( K_L \rightarrow 3\pi \) by a factor \( \sim 500 \), since the \( K_L \) barely resides above the three pion threshold.
- It rewards ‘patience’; i.e., one can ‘wait’ for a pure \( K_L \) beam to emerge after starting out with a beam consisting of \( K^0 \) and \( \bar{K}^0 \).
- It enables CP violation to emerge in the existence of a reaction, namely \( K_L \rightarrow 2\pi \) rather than an asymmetry; this greatly facilitates its observation.

For these reasons alone we should praise hadronization as the hero in the tale of CP violation rather than the villain it is all too often portrayed as.

**End of Interlude**

The first unequivocal manifestation of a penguin contribution surfaced in radiative \( B \) decays, first the exclusive channel \( B \rightarrow \gamma K^* \) and subsequently the inclusive one \( B \rightarrow \gamma X_s \). These transitions represent flavour-changing neutral currents and as such represent a one-loop, i.e., quantum process.

By the end of the second millenium a rich and diverse body of data on flavour dynamics had been accumulated, and CKM theory provided a surprisingly successful description of it. This prompted some daring spirits to perform detailed fits of the CKM triangle to infer a rather accurate prediction for the CP asymmetry in \( B_d \rightarrow \psi K_S \) [28]:

\[
\sin 2\phi_1 = 0.72 \pm 0.07 .
\]

(147)

### 1.9 Summary of Lecture I

The status of flavour dynamics in general and of CKM theory in particular just before the turn of the millenium can be summarized as follows:
“Never underestimate Nature’s ability to come up with an unexpected trick.” Physicists thought they had seen it all after the shock of parity violation in 1957—and then the ‘earthquake’ of CP violation struck in 1964.

“CKM theory—all it does, it works.” We have only an ‘engineering’ solution for the generation of masses (the Higgs mechanism), yet no deeper understanding in particular of fermion masses and family replication. Yet CKM theory, which is based on a set of mass related parameters (fermion masses, CKM parameters) that any sober person would view as frivolous—were they not forced upon us by data—successfully describes a vast and very diverse array of transitions characterized by dynamical scales that a priori span several orders of magnitude.

While no accurate CKM prediction for $\epsilon'/\epsilon_K$ is available now (and presumably for some time to come), it is highly nontrivial that the predictions match to data to better than an order of magnitude. It provides some understanding why direct CP violation is so feeble in kaon decays: it is greatly reduced by the $\Delta I = 1/2$ rule—i.e., $T(S = 1; \Delta I = 3/2)/T(S = 1; \Delta I = 1/2) \simeq 1/20$—and the unexpectedly large top-quark mass.

The SM has to produce a host of truly large CP asymmetries in $B$ decays—there is no plausible deniability. This is far from trivial: based on a tiny CP impurity in the $K^0 - \bar{K}^0$ system one predicts an almost maximal CP asymmetry in $B_d$ decays:

$$\text{few } \times 0.001 \text{ CP asymm. in } K^0 - \bar{K}^0 \implies \text{few } \times 0.1 \text{ CP asymm. in } B_d - \bar{B}_d;$$

i.e., an effect two orders of magnitude larger.

While it is quite possible, or even likely, that New Physics will affect CP asymmetries in $B$ decays, we cannot expect it to create a numerically massive impact except for some special cases.


As explained in the previous lecture, within CKM theory one is unequivocally led to a paradigm of large CP violation in $B$ decays. This realization became so widely accepted that two $B$ factories employing $e^+e^- \to \Upsilon(4S) \to BB$ were constructed—one at KEK in Japan and one in Stanford in the US—together with specialized detectors, around which two collaborations gathered, the Belle and BaBar Collaborations, respectively.

2.1 Establishing the CKM ansatz as a theory—CP violation in $B$ decays

The three angles $\phi_{1,2,3}$ in the CKM unitarity triangle (see Fig. 2 for notation) can be determined through CP asymmetries in $B_d(t) \to \psi K_S; \pi^+ \pi^-$ and $B_d \to K^+ \pi^-$—in principle. In practice the angle $\phi_3$ can be extracted from $B^\pm \to D^{\text{neut}} K^\pm$ with better theoretical control, and $B \to 3\pi$, $4\pi$ offer various experimental advantages over $B \to 2\pi$. These issues will be addressed in five acts plus two interludes.

2.1.1 Act I: $B_d(t) \to \psi K_S$ and $\phi_1$ (a.k.a. $\beta$)

The first published result on the CP asymmetry in $B_d \to \psi K_S$ was actually obtained by the OPAL Collaboration at LEP I [29]:

$$\sin 2\phi_1 = 3.2^{+1.8}_{-2.0} \pm 0.5,$$

(149)

where the ‘unphysical’ value of $\sin 2\phi_1$ is made possible, since a large background subtraction has to be performed. The first value inside the physical range was obtained by CDF [30]:

$$\sin 2\phi_1 = 0.79 \pm 0.44.$$

(150)

In 2000 the two $B$ factory collaborations BaBar and Belle presented their first measurements [21]:

$$\sin 2\phi_1 = \begin{cases} 0.12 \pm 0.37 \pm 0.09 & \text{BaBar ’00} \\ 0.45 \pm 0.44 \pm 0.09 & \text{Belle ’00}. \end{cases}$$

(151)
Fig. 4: CKM unitarity triangle from $|V(ub)/V(cb)|$ and $\Delta M_{B_d}/\Delta M_{B_s}$ on the left and compared to constraints from $\epsilon_K$ and $\sin 2\phi_1/\beta$ on the right (courtesy of M. Pierini)

One year later these inconclusive numbers turned into conclusive ones, and the first CP violation outside the $K^0 - \bar{K}^0$ complex was established:

$$\sin 2\phi_1 = \begin{cases} 0.59 \pm 0.14 \pm 0.05 & \text{BaBar '01} \\ 0.99 \pm 0.14 \pm 0.06 & \text{Belle '01} \end{cases} \quad (152)$$

By 2003 the numbers from the two experiments had well converged

$$\sin 2\phi_1 = \begin{cases} 0.741 \pm 0.067 \pm 0.03 & \text{BaBar '03} \\ 0.733 \pm 0.057 \pm 0.028 & \text{Belle '03} \end{cases} \quad (153)$$

allowing one to state just the world averages, which is actually a BaBar/Belle average [21]:

$$\sin 2\phi_1 = \begin{cases} 0.726 \pm 0.037 & \text{WA '04} \\ 0.685 \pm 0.032 & \text{WA '05} \\ 0.675 \pm 0.026 & \text{WA '06} \end{cases} \quad (154)$$

The CP asymmetry in $B_d \rightarrow \psi K_S$ is there, is huge and as expected even quantitatively. For CKM fits based on constraints from $|V(ub)/V(cb)|$, $B^0 - \bar{B}^0$ oscillations and—as the only CP sensitive observable—$\epsilon_K$ yield [31]

$$\sin 2\phi_1^{CKM} = 0.755 \pm 0.039 \quad (155)$$

The CKM prediction has stayed within the $\sim 0.72-0.75$ interval for the last several years. Throughout 2005 it was in impressive agreement with the data. In 2006 a hint of a deviation emerged. It is not more than that, since it is not (yet) statistically significant and furthermore depends very much on the value extracted for $|V(ub)/V(cb)|$ and its uncertainty. The latter might very well be underestimated, as discussed later. This is illustrated by Fig. 3 showing these constraints. This figure actually obscures another impressive triumph of CKM theory: the CP insensitive observables $|V(ub)/V(cb)|$ and $\Delta M_{B_d}/\Delta M_{B_s}$—i.e., observables that do not require CP violation for acquiring a non-zero value—imply

- a non-flat CKM triangle and thus CP violation, see the left of Fig. 4,
- that is fully consistent with the observed CP sensitive observables $\epsilon_K$ and $\sin 2\phi_1$, see the right of Fig. 4.

2.1.2 CP violation in $K$ and $B$ decays—exactly the same, only different

There are several similarities between $K^0 - \bar{K}^0$ and $B_d - \bar{B}_d$ oscillations even on the quantitative level. Their values for $x = \Delta M/\Gamma$ and thus for $\chi$ are very similar. It is even more intriguing that also
Fig. 5: The observed decay time distributions for $K^0$ vs. $K^0$ from CPLEAR on the left and for $B_d$ vs. $\bar{B}_d$ from BaBar on the right

their pattern of CP asymmetries in $K^0(t)/\bar{K}^0(t) \rightarrow \pi^+\pi^-$ and $B_d(t)/\bar{B}_d(t) \rightarrow \psi K_S$ is very similar. Consider the two lower plots in Fig. 5, which show the asymmetry directly as a function of $\Delta t$: it looks intriguingly similar qualitatively and even quantitatively. The lower left plot shows that the difference between $K^0 \rightarrow \pi^+\pi^-$ and $\bar{K}^0 \rightarrow \pi^+\pi^-$ is actually measured in units of 10% for $\Delta t \sim (8-16)\tau_{K_S}$, which is the $K_S - K_L$ interference region.

Clearly one can find domains in $K \rightarrow \pi^+\pi^-$ that exhibit a truly large CP asymmetry. Nevertheless it is an empirical fact that CP violation in $B$ decays is much larger than in $K$ decays. For the mass eigenstates of neutral kaons are very well approximated by CP eigenstates, as can be read off from the upper left plot: it shows that the vast majority of $K \rightarrow \pi^+\pi^-$ events follow a single exponential decay law that coincides for $K^0$ and $\bar{K}^0$ transitions. This is in marked contrast to the $B_d \rightarrow \psi K_S$ and $\bar{B}_d \rightarrow \psi K_S$ transitions, which in no domain are well approximated by a single exponential law and do not coincide at all, except for $\Delta t = 0$, as it has to be, see Section 2.1.3.

2.1.3 Interlude: “Praise the Gods Twice for EPR Correlations”

The BaBar and Belle analyses are based on a glorious application of quantum mechanics and in particular EPR correlations [32]. The CP asymmetry in $B_d \rightarrow \psi K_S$ had been predicted to exhibit a peculiar dependence on the time of decay, since it involves $B_d - \bar{B}_d$ oscillations in an essential way:

$$\text{rate}(B_d(t)|\bar{B}_d(t)) \rightarrow \psi K_S) \propto e^{-t/\tau_B}(1 - [+]A \sin \Delta M_B t) .$$

At first it would seem that an asymmetry of the form given in Eq. (156) could not be measured for practical reasons. For in the reaction

$$e^+e^- \rightarrow \Upsilon(4S) \rightarrow B_d\bar{B}_d$$

the point where the $B$ meson pair is produced is ill determined on account of the finite size of the electron and positron beam spots: the latter amounts to about 1 mm in the longitudinal direction, while a $B$ meson typically travels only about a quarter of that distance before it decays. It would then seem that the length of the flight path of the $B$ mesons is poorly known and that averaging over this ignorance would greatly dilute or even eliminate the signal.
It is here where the existence of an EPR correlation comes to the rescue. While the two $B$ mesons in the reaction of Eq. (157) oscillate back and forth between a $B_d$ and $\bar{B}_d$, they change their flavour identity in a \textit{completely correlated} way. For the $B \bar{B}$ pair forms a $C$ \textit{odd} state; Bose statistics then tells us that there cannot be two identical flavour hadrons in the final state:

\[ e^+ e^- \rightarrow \Upsilon(4S) \rightarrow B_d \bar{B}_d \neq B_d B_d, \bar{B}_d \bar{B}_d. \tag{158} \]

Once one of the $B$ mesons decays through a flavour specific mode, say $B_d \rightarrow l^+ \nu X$ [$\bar{B}_d \rightarrow l^- \bar{\nu} X$], then we know unequivocally that the other $B$ meson was a $\bar{B}_d$ [$B_d$] at that time. The time evolution of $\bar{B}_d(t) |B_d(t) \rightarrow \psi K_S$ as described by Eq. (156) starts at that time as well; i.e., the relevant time parameter is the \textit{interval} between the two times of decay, not those times themselves. That time interval is related to—and thus can be inferred from—the distance between the two decay vertices, which is well defined and can be measured.

The great practical value of the EPR correlation is instrumental for another consideration as well, namely how to see directly from the data that $C P$ violation is matched by $T$ violation. Figure 6 shows two distributions, one for the interval $\Delta t$ between the times of decays $B_d \rightarrow l^+ X$ and $\bar{B}_d \rightarrow \psi K_S$ and the other one for the $C P$ conjugate process $\bar{B}_d \rightarrow l^- X$ and $B_d \rightarrow \psi K_S$. They are clearly different proving that $C P$ is broken. Yet they show more: the shape of the two distributions is actually the same (within experimental uncertainties) the only difference being that the average of $\Delta t$ is \textit{positive} for $(l^- X)_{\bar{B}}(\psi K_S)$ and \textit{negative} for $(l^+ X)_B(\psi K_S)$ events. That is, there is a (slight) preference for $B_d \rightarrow \psi K_S$ [$\bar{B}_d \rightarrow \psi K_S$] to occur \textit{after} [\textit{before}] and thus more [less] slowly (rather than just more rarely) than $\bar{B} \rightarrow l^- X$ [$B \rightarrow l^+ X$]. Invoking $C P T$ invariance merely for semileptonic $B$ decays—but not for nonleptonic transitions—synchronizes the starting point of the $B$ and $\bar{B}$ decay ‘clocks’, and the EPR correlation keeps them synchronized. We thus see that $C P$ and $T$ violation are ‘just’ different sides of the same coin. As explained above, EPR correlations are essential for this argument!

The reader can be forgiven for feeling that this argument is of academic interest only, since $C P T$ invariance of all processes is based on very general arguments. Yet the main point to be noted is that EPR correlations, which represent some of quantum mechanics’ most puzzling features, serve as an essential precision tool, which is routinely used in these measurements. I feel it is thus inappropriate to refer to EPR correlations as a paradox.
2.1.4 Act II: $B_d(t) \to$ pions and $\phi_2$ (a.k.a. $\alpha$)

2.1.4.1 $B \to 2\pi$

The situation is theoretically more complex than for $B_d(t) \to \psi K_S$ for two reasons:

- While both final states $\pi\pi$ and $\psi K_S$ are CP eigenstates, the former unlike the latter is not reached through an isoscalar transition. The two pions can form an $I = 0$ or $I = 2$ configuration (similar to $K \to 2\pi$), which in general will be affected differently by the strong interactions.

- For all practical purposes $B_d \to \psi K_S$ is described by two tree diagrams representing the two effective operators $(\bar{c}_L \gamma_\mu b_L)(\bar{s}_L \gamma_\mu c_L)$ and $(\bar{c}_L \gamma_\mu \lambda_i b_L)(\bar{s}_L \gamma^\mu \lambda_i c_L)$ with the $\lambda_i$ representing the $SU(3)_C$ matrices. Yet for $B \to \pi\pi$ we have effective operators $(\bar{d}_L \gamma_\mu \lambda_i b_L)(\bar{q}^\mu \lambda_i q)$ generated by the Cabibbo-suppressed penguin loop diagrams in addition to the two tree operators $(\bar{u}_L \gamma_\mu b_L)(\bar{d}_L \gamma^\mu u_L)$ and $(\bar{u}_L \gamma_\mu \lambda_i b_L)(\bar{d}_L \gamma^\mu \lambda_i u_L)$.

This greater complexity manifests itself in the phenomenological description of the time-dependent CP asymmetry:

$$\frac{R_+(-\Delta t) - R_-(\Delta t)}{R_+(-\Delta t) + R_-(\Delta t)} = S \sin(\Delta M_d \Delta t) - C \cos(\Delta M_d \Delta t), \quad S^2 + C^2 \leq 1$$

(159)

where $R_{+[-]}(\Delta t)$ denotes the rate for $B^{tag}(t)\bar{B}_d(t + \Delta)[\bar{B}^{tag}(t)B_d(t + \Delta)]$ and

$$S = \frac{2 \text{ Im } q^{\mu} \rho_{\pi^+\pi^-}}{1 + \left| \frac{q^{\mu} \rho_{\pi^+\pi^-}}{p} \right|^2}, \quad C = 1 - \left| \frac{q^{\mu} \rho_{\pi^+\pi^-}}{p} \right|^2$$

(160)

As before, on account of the EPR correlation between the two neutral $B$ mesons, it is the relative time interval $\Delta t$ between the two $B$ decays that matters, not their lifetime. The new feature is that one has also a cosine dependence on $\Delta t$.

BaBar and Belle find

$$S = \begin{cases} -0.53 \pm 0.14 \pm 0.02 & \text{BaBar '06} \\ -0.61 \pm 0.10 \pm 0.04 & \text{Belle '06} \\ -0.59 \pm 0.09 & \text{HFAG} \end{cases}$$

(161)

$$C = \begin{cases} -0.16 \pm 0.11 \pm 0.03 & \text{BaBar '06} \\ -0.55 \pm 0.08 \pm 0.05 & \text{Belle '06} \\ -0.39 \pm 0.07 & \text{HFAG} \end{cases}$$

(162)

While BaBar and Belle agree nicely on $S$ making the HFAG average straightforward, their findings on $C$ indicate different messages making the HFAG average more iffy.

$S \neq 0$ has been established and thus CP violation also in this channel. While Belle finds $C \neq 0$ as well, BaBar’s number is still consistent with $C = 0$. $C \neq 0$ obviously represents direct CP violation. Yet it is often overlooked that $S$ also can reveal such CP violation. For if one studies $B_d$ decays into two CP eigenstates $f_a$ and $f_b$ and finds

$$S(f_a) \neq \eta(f_a)\eta(f_b)S(f_b)$$

(163)

with $\eta_i$ denoting the CP parities of $f_i$, then one has established direct CP violation. For the case under study that means even if $C(\pi\pi) = 0$, yet $S(\pi^+\pi^-) \neq -S(\psi K_S)$, one has observed unequivocally direct CP violation. One should note that such direct CP violation might not necessarily induce $C \neq 0$. For the latter requires, as explained below in Section 2.1.6 [see Eq. (171)], that two different amplitudes contribute coherently to $B_d \to f_b$ with non-zero relative weak as well as strong phases. $S(f_a) \neq 0$
\( \eta(f_a)\eta(f_b)S(f_b) \) on the other hand only requires that the two overall amplitudes for \( B_d \rightarrow f_a \) and \( B_d \rightarrow f_b \) possess a relative phase. This can be illustrated with a familiar example from CKM dynamics: if there were no penguin operators for \( B_d \rightarrow f_a \) and \( B_d \rightarrow f_b \), one would expect \( C(\pi^+\pi^-) = 0 \), yet at the same time \( S(\psi K_S) = \sin(2\phi_1) \) together with \( S(\pi^+\pi^-) = \sin(2\phi_2) \neq -\sin(2\phi_1) \). That is, without direct CP violation one would have to find \( C = 0 \) and \( S = -\sin 2\phi_1 \) [33]. Yet since the measured value of \( S \) is within one sigma of \( -\sin 2\phi_1 \), this distinction is mainly of academic interest at the moment.

Once the categorical issue of whether there is direct CP violation has been settled, one can take up the challenge of extracting a value for \( \phi_2 \) from the data\(^{19}\). This can be done in a model-independent way by analysing \( B_d(t) \rightarrow \pi^+\pi^-, \pi^0\pi^0 \) and \( B_d^+ \rightarrow \pi^+\pi^0 \) transitions and performing an isospin decomposition. For the penguin contribution cannot affect \( B_d(t) \rightarrow [\pi\pi]_{l=2} \) modes. Unfortunately there is a serious experimental bottle-neck, namely to study \( B_d(t) \rightarrow \pi^0\pi^0 \) with sufficient accuracy. Therefore alternative decays have been suggested, in particular \( B \rightarrow \rho\pi \) and \( \rho\rho \).

2.1.4.2 \( B \rightarrow 3\pi/4\pi \)

The final states in \( B \rightarrow 3\pi \) and \( 4\pi \) are largely of the \( \rho\pi \) and \( \rho\rho \) form, respectively. For those one can also undertake an isospin decomposition to disentangle the penguin contribution. These channels are less challenging experimentally than \( B_{d,u} \rightarrow 2\pi \), yet they pose some complex theoretical problems.

For going from the experimental starting point \( B \rightarrow 3\pi \) to \( B \rightarrow \rho\pi \) configurations is quite nontrivial. There are other contributions to the three-pion final state like \( \sigma\pi \), and cutting on the dipion mass provides a rather imperfect filter because of the large \( \rho \) width. It hardly matters in this context whether the \( \sigma \) is a bona fide resonance or some other dynamical enhancement. This actually leads to a further complication, namely that the \( \sigma \) structure cannot be described adequately by a Breit–Wigner shape. As analysed first in Ref. [34] and then in more detail in Ref. [35] ignoring such complications can induce a significant systematic uncertainty in the extracted value of \( \phi_2 \).

The modes \( B_{d,u} \rightarrow \rho\rho \) contain even more theoretical complexities, since they have to be extracted from \( B \rightarrow 4\pi \) final states, where one has to allow for \( \sigma\rho, 2\sigma, \rho2\pi \) etc. in addition to \( 2\rho \).

My point here is one of caution rather than of agnosticism. The concerns sketched above might well be more academic than practical with the present statistics. My main conclusions are the following: (i) I remain unpersuaded that averaging over the values for \( \phi_2 \) obtained so far from the three methods listed above provides a reliable value, since I do not think that the systematic uncertainties have been analysed sufficiently. (ii) It will be mandatory to study those in a comprehensive way before we can make full use of the even larger data sets that will become available in the next few years. As I have emphasized repeatedly, our aim has to be to reduce the uncertainty down to at least the 5% level in a way that can be defended. (iii) In the end we shall need

- to perform time-dependent Dalitz plot analyses (and their generalizations) and
- involve the expertise that already exists or can be obtained concerning low-energy hadronization processes like final-state interactions among low-energy pions and kaons; valuable information can be gained on those issues from \( D(s) \rightarrow \pi \)’s, kaons etc. as well as \( D(s) \rightarrow l\nu K\pi/\pi\pi/KK \), in particular when analysed with state-of-the-art tools of chiral dynamics.

2.1.5 Act III, first version: \( B_d \rightarrow K^+\pi^- \)

It was pointed out in a seminal paper [36] that (rare) transitions like \( B_d \rightarrow K^+\pi^- \) have the ingredients for sizeable direct CP asymmetries stated in Section 2.1.6:

---

\(^{19}\)The complications due to the presence of the penguin contribution are all too often referred to as ‘penguin pollution’. Personally I find it quite unfair to blame our lack of theoretical control on water fowls rather than on the guilty party, namely us.
Two different amplitudes can contribute coherently, namely the highly CKM suppressed tree diagram with $b \to u \bar{u}s$ and the penguin diagram with $b \to s \bar{q}q$.

The tree diagram contains a large weak phase from $V(ub)$.

The penguin diagram with an internal charm quark loop exhibits an imaginary part, which can be viewed—at least qualitatively—as a strong phase generated by the production and subsequent annihilation of a $c \bar{c}$ pair (the diagram with an internal $u$ quark loop acts merely as a subtraction point allowing a proper definition of the operator).

While the penguin diagram with an internal top quark loop is actually not essential, the corresponding effective operator can be calculated quite reliably, since integrating out first the top quarks and then the $W$ boson leads to a truly local operator. Determining its matrix elements is, however, another matter.

To translate these features into accurate numbers represents a formidable task that we have not yet mastered. In Ref. [37] an early and detailed effort was made to treat theoretic with the following results:

$$\text{BR}(\bar{B}_d \to K^- \pi^+) \sim 10^{-5}, \quad A_{\text{CP}} \sim -0.10. \quad (164)$$

Those numbers turn out to be rather prescient, since they are in gratifying agreement with the data

$$\text{BR}(\bar{B}_d \to K^- \pi^+) = (1.85 \pm 0.11) \cdot 10^{-5}$$

$$A_{\text{CP}} = \begin{cases} -0.133 \pm 0.030 \pm 0.009 & \text{BaBar} \\ -0.113 \pm 0.021 & \text{Belle} \end{cases} \quad (165)$$

Cynics might point out that the authors in Ref. [37] did not give a specific estimate of the theoretical uncertainties in Eq. (164). More recent authors have been more ambitious—with somewhat mixed success. I list the predictions inferred from pQCD [38] and QCD factorization [39] and the data for the three modes $\bar{B}_d \to K^- \pi^+$ and $B^- \to K^- \pi^0, \bar{K}^0 \pi^-$:

$$A_{\text{CP}}(B_d \to K^- \pi^+) = \begin{cases} -0.133 \pm 0.030 \pm 0.009 & \text{BaBar} \\ -0.113 \pm 0.021 & \text{Belle} \\ -0.09 \pm 0.08 \pm 0.06 & \text{pQCD} \\ +0.05 \pm 0.09 & \text{QCD fact.} \end{cases} \quad (166)$$

$$A_{\text{CP}}(B^- \to K^- \pi^0) = \begin{cases} +0.06 \pm 0.06 \pm 0.01 & \text{BaBar} \\ +0.04 \pm 0.04 \pm 0.02 & \text{Belle} \\ -0.01 \pm 0.05 & \text{pQCD} \\ +0.07 \pm 0.09 & \text{QCD fact.} \end{cases} \quad (167)$$

$$A_{\text{CP}}(B^- \to \bar{K}^0 \pi^-) = \begin{cases} -0.09 \pm 0.05 \pm 0.01 & \text{BaBar} \\ +0.05 \pm 0.05 & \text{Belle} \\ +0.00 & \text{pQCD} \\ +0.01 \pm 0.01 & \text{QCD fact.} \end{cases} \quad (168)$$

As explained next, the size of these asymmetries depends very much on hadronization effects, namely hadronic matrix elements and strong phase shifts. While for the observed asymmetry in $B_d \to K \pi$ with CKM expectations, we do not have an accurate prediction.

### 2.1.6 Interlude: On final-state interactions and CPT invariance

Owing to CPT invariance, CP violation can be implemented only through a complex phase in some effective couplings. For it to become observable, two different, yet coherent amplitudes have to contribute to an observable. There are two types of scenarios for implementing this requirement:
1. When studying a final state \( f \) that can be reached by a \( \Delta B = 1 \) transition from \( B^0 \) as well as \( \bar{B}^0 \), then \( B_0 - \bar{B}_0 \) oscillations driven by \( \Delta B = 2 \) dynamics provide the second amplitude, the weight of which varies with time. This is what happens in \( B_d \rightarrow \psi K_S, \pi^+ \pi^- \).

2. Two different \( \Delta B = 1 \) amplitudes \( M_{a,b} \) of fixed ratio—distinguished by, say, their isospin content—exist leading coherently to the same final state:

\[
T(B \rightarrow f) = \lambda_a M_a + \lambda_b M_b.
\] (169)

I have factored out the weak couplings \( \lambda_{a,b} \) while allowing the amplitudes \( M_{a,b} \) to be still complex due to strong or electromagnetic FSI. For the CP conjugate reaction one has

\[
T(\bar{B} \rightarrow \bar{f}) = \lambda_a^* M_a + \lambda_b^* M_b.
\] (170)

It is important to note that the reduced amplitudes \( M_{a,b} \) remain unchanged, since strong and electromagnetic forces conserve CP. Therefore we find

\[
\Gamma(\bar{B} \rightarrow \bar{f}) - \Gamma(B \rightarrow f) = \frac{2 \text{Im} \lambda_a \lambda_b^* \cdot \text{Im} M_a M_b^*}{|\lambda_a|^2 |M_a|^2 + |\lambda_b|^2 |M_b|^2 + 2 \text{Re} \lambda_a \lambda_b^* \cdot \text{Re} M_a M_b^*},
\] (171)

i.e., for a CP asymmetry to become observable, two conditions have to be satisfied simultaneously irrespective of the underlying dynamics:

- Im \( \lambda_a \lambda_b^* \neq 0 \), i.e., there has to be a relative phase between the weak couplings \( \lambda_{a,b} \).
- Im \( M_a M_b^* \neq 0 \), i.e., final-state interactions (FSI) have to induce a phase shift between \( M_{a,b} \).

It is often not fully appreciated that CPT invariance places constraints on the phases of the \( M_{a,b} \). For it implies much more than equality of masses and lifetimes of particles and antiparticles. It tells us that the widths for subclasses of transitions for particles and antiparticles have to coincide already, either identically or at least practically. Just writing down strong phases in an equation like Eq. (169) does not automatically satisfy CPT constraints.

I shall illustrate this feature first with two simple examples and then express it in more general terms.

- CPT invariance already implies \( \Gamma(K^- \rightarrow \pi^- \pi^0) = \Gamma(K^+ \rightarrow \pi^+ \pi^0) \) up to small electromagnetic corrections, since in that case there are no other channels with which it can rescatter.
- While \( \Gamma(K^0 \rightarrow \pi^+ \pi^-) \neq \Gamma(\bar{K}^0 \rightarrow \pi^+ \pi^-) \) and \( \Gamma(K^0 \rightarrow \pi^0 \pi^0) \neq \Gamma(\bar{K}^0 \rightarrow \pi^0 \pi^0) \) one has, \( \Gamma(K^0 \rightarrow \pi^+ \pi^- + \pi^0 \pi^0) = \Gamma(\bar{K}^0 \rightarrow \pi^+ \pi^- + \pi^0 \pi^0) \).
- Let us now consider a scenario where a particle \( P \) and its antiparticle \( \bar{P} \) can each decay into two final states only, namely \( a, b \) and \( \bar{a}, \bar{b} \), respectively [40, 41]. Let us further assume that strong (and electromagnetic) forces drive transitions among \( a \) and \( b \)—and likewise for \( \bar{a} \) and \( \bar{b} \)—as described by an S matrix \( S \). The latter can then be decomposed into two parts

\[
S = S_{\text{diag}} + S_{\text{off-diag}},
\] (172)

where \( S_{\text{diag}} \) contains the diagonal transitions \( a \rightarrow a, b \rightarrow b \)

\[
S_{\text{diag}}^{ss} = e^{2i\delta_s}, s = a, b
\] (173)

and \( S_{\text{off-diag}} \) the off-diagonals ones \( a \rightarrow b, b \rightarrow a \):

\[
S_{ab}^{\text{off-diag}} = 2i T_{ab}^{\text{resc}} e^{i(\delta_a + \delta_b)}
\] (174)

with

\[
T_{ab}^{\text{resc}} = T_{ba}^{\text{resc}} = (T_{ab}^{\text{resc}})^*,
\] (175)
since the strong and electromagnetic forces driving the rescattering conserve CP and T. The resulting S matrix is unitary to first order in $T_{ab}^{\text{resc}}$. CPT invariance implies the following relation between the weak decay amplitude of $P$ and $\bar{P}$:

$$T(P \to a) = e^{i\delta_a} [T_a + T_b^{\text{resc}}]$$  \hspace{1cm} (176)

$$T(\bar{P} \to \bar{a}) = e^{i\delta_a} [T_a + T_b^{\text{resc}}]$$  \hspace{1cm} (177)

and thus

$$\Delta\gamma(a) \equiv |T(\bar{P} \to \bar{a})|^2 - |T(P \to a)|^2 = 4T_{ab}^{\text{resc}} \text{Im} T_a^* T_b^* ;$$  \hspace{1cm} (178)

likewise

$$\Delta\gamma(b) \equiv |T(\bar{P} \to \bar{b})|^2 - |T(P \to b)|^2 = 4T_{ab}^{\text{resc}} \text{Im} T_b^* T_a^*$$  \hspace{1cm} (179)

and therefore as expected

$$\Delta\gamma(b) = -\Delta\gamma(b).$$  \hspace{1cm} (180)

Some further features can be read off from Eq. (178):

1. If the two channels that rescatter have comparable widths—$\Gamma(P \to a) \sim \Gamma(P \to b)$—one would like the rescattering $b \leftrightarrow a$ to proceed via the usual strong forces; for otherwise the asymmetry $\Delta\Gamma$ is suppressed relative to these widths by the electromagnetic coupling.

2. If on the other hand the channels command very different widths—say $\Gamma(P \to a) \gg \Gamma(P \to b)$—then a large relative asymmetry in $P \to b$ is accompanied by a tiny one in $P \to a$.

This simple scenario can easily be extended to two sets $A$ and $B$ of final states so that for all states $a$ in set $A$ the transition amplitudes have the same weak coupling and likewise for states $b$ in set $B$. One then finds

$$\Delta\gamma(a) = 4 \sum_{b \in B} T_{ab}^{\text{resc}} \text{Im} T_a^* T_b^* .$$  \hspace{1cm} (181)

The sum over all CP asymmetries for states $a \in A$ cancels the corresponding sum over $b \in B$:

$$\sum_{a \in A} \Delta\gamma(a) = 4 \sum_{b \in B} T_{ab}^{\text{resc}} \text{Im} T_a^* T_b^* = - \sum_{b \in B} \Delta\gamma(b).$$  \hspace{1cm} (182)

These considerations tell us that the CP asymmetry averaged over certain classes of channels defined by their quantum numbers has to vanish. Yet these channels can still be very heterogeneous, namely consisting of two- and quasi-two-body modes, three-body channels and other multi-body decays. Hence we can conclude:

- If one finds a direct CP asymmetry in one channel, one can infer—based on rather general grounds—which other channels have to exhibit the compensating asymmetry as required by CPT invariance. Observing them would enhance the significance of the measurements very considerably.

- Typically there can be several classes of rescattering channels. The SM weak dynamics select a subclass of those where the compensating asymmetries have to emerge. QCD frameworks like generalized factorization can be invoked to estimate the relative weight of the asymmetries in the different classes. Analysing them can teach us important lessons about the inner workings of QCD.

- If New Physics generates the required weak phases (or at least contributes significantly to them), it can induce rescattering with novel classes of channels. The pattern in the compensating asymmetries then can tell us something about the features of the New Physics involved.

I want to end this interlude by adding that penguins are rather smart beings: they know about these CPT constraints. For when one considers the imaginary parts of the penguin diagrams, which are obtained by cutting the internal quark lines, namely the up and charm quarks (top quarks do not contribute there, since they cannot reach their mass shell in $b$ decays), one realizes that CP asymmetries in $B \to K + \pi$'s are compensated by those in $B \to D\bar{D}_s + \pi$'s.
2.1.7 Act III, second version: $\phi_3$ from $B^+ \to D_{neut}K^+$ vs. $B^- \to D_{neut}K^-$

As first mentioned in 1980 [42], then explained in more detail in 1985 [43], and further developed in Ref. [44], the modes $B^\pm \to D_{neut}K^{\pm}$ should exhibit direct CP violation driven by the angle $\phi_3$ if the neutral $D$ mesons decay to final states that are common to $D^0$ and $\bar{D}^0$. Based on simplicity the original idea was to rely on two-body modes like $K_S\pi^0$, $K^+K^-$, $\pi^+\pi^-$, $K^\pm\pi^\mp$. One drawback of that method are the small branching ratios and low efficiencies.

A new method was pioneered by Belle and then implemented also by BaBar, namely to employ $D_{neut} \to K_S\pi^+\pi^-$ and perform a full Dalitz plot analysis. This requires a very considerable analysis effort—yet once this initial investment has been made, it will pay handsome profits in the long run. For obtaining at least a decent description of the full Dalitz plot population provides considerable cross-checks concerning systematic uncertainties and thus a high degree of confidence in the results. Belle and BaBar find:

$$\phi_3 = \begin{cases} 
53^{+15}_{-18}(\text{stat}) \pm 3(\text{syst}) \pm 9(\text{model})^\circ & \text{Belle} \\
92^\circ \pm 41(\text{stat}) \pm 11(\text{syst}) \pm 12^\circ(\text{model}) & \text{BaBar} 
\end{cases}$$

(183)

At present these studies are severely statistics limited; one should also note that with more statistics one will be able to reduce in particular the model dependence. I view this method as the best one to establish confidence in the accuracy of our results.

2.1.8 Act IV: $\phi_1$ from CP violation in $B_d \to 3$ kaons—snatching victory from the jaws of defeat or defeat from the jaws of victory

Analysing CP violation in $B_d \to \phi K_S$ decays is a most promising way to search for New Physics. For the underlying quark-level transition $b \to s\bar{s}s$ represents a pure loop-effect in the SM, it is described by a single $\Delta B = 1$ & $\Delta I = 0$ operator (a ‘penguin’), a reliable SM prediction exists for it [45]—$
\sin 2\phi_1(B_d \to \psi K_S) \simeq \sin 2\phi_1(B_d \to \phi K_S)$—and the $\phi$ meson represents a narrow resonance.

Great excitement was created when Belle reported a large discrepancy between the predicted and observed CP asymmetry in $B_d \to \phi K_S$ in the summer of 2003:

$$\sin 2\phi_1(B_d \to \phi K_S) = \begin{cases} 
-0.96 \pm 0.5 \pm 0.10 & \text{Belle }'03 \\
0.45 \pm 0.43 \pm 0.07 & \text{BaBar }'03 
\end{cases}$$

(184)

Based on more data taken, this discrepancy has shrunk considerably: the BaBar/Belle average for 2005 yields [21]

$$\sin 2\phi_1(B_d \to \psi K_S) = 0.685 \pm 0.032$$

(185)

versus

$$\sin 2\phi_1(B_d \to \phi K_S) = \begin{cases} 
0.44 \pm 0.27 \pm 0.05 & \text{Belle }'05 \\
0.50 \pm 0.25^{+0.07}_{-0.04} & \text{BaBar }'05 
\end{cases}$$

(186)

while the 2006 values read as follows:

$$\sin 2\phi_1(B_d \to \psi K_S) = 0.675 \pm 0.026$$

(187)

compared to

$$\sin 2\phi_1(B_d \to \phi K_S) = \begin{cases} 
0.50 \pm 0.21 \pm 0.06 & \text{Belle }'06 \\
0.12 \pm 0.31 \pm 0.10 & \text{BaBar }'06 \\
0.39 \pm 0.18 & \text{HFAG }'06 
\end{cases}$$

(188)

I summarize the situation as follows:
Performing dedicated CP studies in channels driven mainly or even predominantly by $b \rightarrow sq\bar{q}$ to search for New Physics signatures makes eminent sense since the SM contribution, in particular from the one-loop penguin operator, is greatly suppressed.

The experimental situation is far from settled, as can be seen also from how the central value has moved over the years. It is tantalizing to see that the $S$ contributions for all the modes in this category—$B_d \rightarrow \pi^0K_S, \rho^0K_S, \omega K_S, f_0K_S$—are all low compared to the SM expectation Eq. (187). Yet none of them is significantly lower; furthermore none of these modes, a non-zero CP asymmetry, has been established except for

$$\sin 2\phi_1(B_d \rightarrow \eta' K_S) = \begin{cases} 0.64 \pm 0.10 \pm 0.04 & \text{Belle '06} \\ 0.58 \pm 0.10 \pm 0.03 & \text{BaBar '06} \\ 0.61 \pm 0.07 & \text{HFAG '06} \end{cases} \quad (189)$$

Obviously there is considerable space still for significant deviations from SM predictions. It is ironic that such a smaller deviation would actually be more believable as signalling an incomplete-ness of the SM than the large one originally reported by Belle. While it is tempting to average over all these hadronic transitions, I would firmly resist this temptation for the time being, till several modes exhibit a significant asymmetry.

One complication has to be studied, though, in particular if the observed value of $\sin 2\phi_1(B_d \rightarrow \phi K_S)$ falls below the predicted one by a moderate amount only. For one is actually observing $B_d \rightarrow K^+K^-K_S$. If there is a single weak phase like in the SM one finds

$$\sin 2\phi_1(B_d \rightarrow \phi K_S) = -\sin 2\phi_1(B_d \rightarrow 'f_0(980)' K_S) \quad (190)$$

where $'f_0(980)'$ denotes any scalar $K^+K^-$ configuration with a mass close to that of the $\phi$, be it a resonance or not. A smallish pollution by such a $'f_0(980)'K_S$—by, say, 10% in amplitude—can thus reduce the asymmetry assigned to $B_d \rightarrow \phi K_S$ significantly—by 20% in this example.

In the end it is therefore mandatory to perform a full time-dependent Dalitz plot analysis for $B_d \rightarrow K^+K^-K_S$ and compare it with that for $B_d \rightarrow 3K_S$ and $B^+ \rightarrow K^+K^-K^+$, $K^+K_SK_S$ and also with $D \rightarrow 3K$. BaBar has presented such a preliminary study. This is a very challenging task, but in my view essential. There is no ‘royal way’ to fundamental insights.

An important intermediate step in this direction is given by one application of Bianco’s Razor [46], namely to analyse the CP asymmetry in $B_d \rightarrow [K^+K^-]_M K_S$ as a function of the cut $M$ on the $K^+K^-$ mass.

All of this might well lead to another triumph of the SM, when its predictions agree with accurate data in the future even for these rare transition rates dominated by loop-contributions, i.e., pure quantum effects. It is equally possible—personally I think it actually more likely—that future precision data will expose New Physics contributions. In that sense the SM might snatch victory from the jaws of defeat—or defeat from the jaws of victory. For those of us who are seeking indirect manifestations of New Physics it is the other way round.

In any case the issue has to be pursued with vigour since these reactions provide such a natural portal to New Physics on the one hand and possess such an accurate yardstick from $B_d \rightarrow \psi K_S$.

2.1.9 The beginning of Act V—CP violation in charged $B$ decays

So far CP violation has not been established in the decays of charged mesons, which is not surprising, since meson–antimeson oscillations cannot occur there and it has to be purely direct CP violation. Now

---

20The ruler of a Greek city in southern Italy once approached the resident sage with the request to be educated in mathematics, but in a ‘royal way’, since he was very busy with many obligations. Whereupon the sage replied with admirable candor: “There is no royal way to mathematics.”
Belle [47] has found strong evidence for a large $CP$ asymmetry in charged $B$ decays with a 3.9 sigma significance, namely in $B^\pm \to K^\pm \rho^0$ observed in $B^\pm \to K^\pm \pi^\pm \pi^\mp$:

$$A_{CP}(B^\pm \to K^\pm \rho^0) = (30 \pm 11 \pm 2.0^{+11}_{-4}) \% .$$

(191)

I find it a most intriguing signal since a more detailed inspection of the mass peak region shows a pattern as expected for a genuine effect. Furthermore a similar signal is seen in BaBar’s data, and it would make sense to make to undertake a careful average over the two data sets.

I view Belle’s and BaBar’s analyses of the Dalitz plot for $B^\pm \to K^\pm \pi^\pm \pi^\mp$ as important pilot studies from which one can infer important lessons about the strengths and pitfalls of such studies in general.

### 2.2 Loop-induced rare $B_{u,d}$ transitions

Processes that require a loop diagram to proceed—i.e., are classically forbidden—provide a particularly intriguing stage to probe fundamental dynamics.

It marked a tremendous success for the SM when radiative $B$ decays were measured, first in the exclusive mode $B \to \gamma K^*$ and subsequently also inclusively: $B \to \gamma X$. Both the rate and the photon spectrum are in remarkable agreement with SM prediction; they have been harnessed to extract heavy quark parameters, as explained below.

More recently the next, i.e., even rarer level, has been reached with transitions to final states containing a pair of charged leptons:

$$\text{BR}(B \to l^+ l^-) = \begin{cases} (6.2 \pm 1.1 \pm 1.5) \cdot 10^{-6} & \text{BaBar/Belle} \\ (4.7 \pm 0.7) \cdot 10^{-6} & \text{SM} \end{cases}.$$  (192)

Again the data are consistent with the SM prediction [48], yet the present experimental uncertainties are very sizeable. We are just at the beginning of studying $B \to l^+ l^- X$, and it has to be pursued in a dedicated and comprehensive manner as discussed in Lecture III.

The analogous decays with a $\nu \bar{\nu}$ instead of the $l^+ l^-$ pair is irresistibly attractive to theorists—although quite resistibly so to experimentalists:

$$\text{BR}(B \to \nu \bar{\nu} X) \begin{cases} \leq 7.7 \cdot 10^{-4} & \text{ALEPH} \\ 3.5 \cdot 10^{-5} & \text{SM} \end{cases};$$

$$\text{BR}(B \to \nu \bar{\nu} K) \begin{cases} \leq 7.0 \cdot 10^{-5} & \text{BaBar} \\ (3.8^{+1.2}_{-0.6}) \cdot 10^{-6} & \text{SM} \end{cases},$$

(193)

(194)

where the SM predictions are taken from Refs. [49] and [50], respectively.

### 2.3 Other rare decays

There are some relatively rare $B$ decays that could conceivably reveal New Physics, although they proceed already on the tree level. Semileptonic decays involving $\tau$ leptons are one example and will be discussed in Lecture III. The most topical example is $B^+ \to \tau \nu$ which has been pursued vigorously since it provides information on the decay constant $f_B$ and is sensitive to contributions from charged Higgs fields. A first signal has been found by Belle with a 3.5 sigma significance:

$$\text{BR}(B^- \to \tau^- \bar{\nu}) = (1.79^{+0.56}_{-0.49} \pm 0.36) \cdot 10^{-4} .$$

(195)

Hence one extracts

$$f_B |V(ub)| = (10.1^{+1.6}_{-1.4} +1.3) \cdot 10^{-4} \text{ GeV} .$$

(196)
Table 1: The 2005 values \cite{51} of $b$ and $c$ quark masses and of $|V_{cb}|$ compared to the Cabibbo angle

<table>
<thead>
<tr>
<th>Heavy quark parameter</th>
<th>Value as of 2005</th>
<th>Relative uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_b$ (1 GeV)</td>
<td>$(4.59 \pm 0.04)$ GeV</td>
<td>$\pm 1.0%$</td>
</tr>
<tr>
<td>$m_c$ (1 GeV)</td>
<td>$(1.14 \pm 0.06)$ GeV</td>
<td>$\pm 5.3%$</td>
</tr>
<tr>
<td>$m_b$ (1 GeV) $-0.67m_c$ (1 GeV)</td>
<td>$(3.82 \pm 0.017)$ GeV</td>
<td>$\pm 0.5%$</td>
</tr>
<tr>
<td>$</td>
<td>V_{cb}</td>
<td>$</td>
</tr>
<tr>
<td>$</td>
<td>V_{us}</td>
<td>_{KTeV}$</td>
</tr>
</tbody>
</table>

2.4 Adding high accuracy to high sensitivity

As mentioned before and addressed in more detail in Lecture III, we cannot count on a numerically massive impact of New Physics in heavy flavour transitions. Therefore it no longer suffices to rely on the high sensitivity that loop processes like $B^0 - \bar{B}^0$ oscillations or radiative $B$ decays possess to New Physics; we have to strive also for high accuracy.

The spectacular success of the $B$ factories and the emerging successes of CDF and D0 to obtain high quality data on beauty transitions in a hadronic environment give us confidence that even greater experimental precision can be achieved in the future. However, this would be of little help if it could not be matched by a decrease in the theoretical uncertainties. I shall describe now why I think that theory will be able to hold up its side of the bargain as well and what the required elements for such an undertaking have to be.

The question is: Can we answer the challenge of $\sim \%$ accuracy? One guiding principle will be in Lenin’s concise words:

“Trust is good—control is better!”

Table 1 provides a sketch of the theoretical control we have achieved over some aspects of $B$ decays. I hope it will excite the curiosity of the reader and fortify her/him to read the following more technical discussion; let me add that one can skip this Section 2.4 at the first reading and continue with Section 2.5.

2.4.1 Heavy quark theory

While QCD is the only candidate among local quantum field theories to describe the strong interactions, as explained in Lecture I in Sections 1.1.1 and 1.1.2, $SU(2)_L \times U(1)$ is merely the minimal theory for the electroweak forces. Obtaining reliable information about the latter is, however, limited by our lack of full calculational control over the former.

It had been conjectured for more than thirty years that the theoretical treatment of heavy flavour hadrons should be facilitated when the heavy quark mass greatly exceeds the nonperturbative scale of QCD:\footnote{A striking prediction has been that super-heavy top quarks—i.e., with $m_t \geq 150$ GeV—would decay before they could hadronize \cite{52} thus bringing top quarks under full theoretical control. For then the decay width of top quarks is of order 1 GeV and provides an infrared cutoff for QCD corrections. This feature comes with a price, though, in so far as CP studies are concerned: without hadronization as a ‘cooling’ mechanism, the degree of coherence between different transition amplitudes—a necessary condition for CP violation to become observable—will be rather tiny.}

$$m_Q \gg \Lambda_{QCD}.$$ (197)

This conjecture has been transformed into a reliable theoretical framework only in the last fifteen years, as far as beauty hadrons are concerned. I refer to it as Heavy Quark Theory (mentioned already in Section 1.1.1.3); comprehensive reviews with references to the original literature can be found in Refs. [53] and [54]. Its goal is to treat nonperturbative dynamics quantitatively, as it affects heavy flavour
hadrons, in full conformity with QCD and without model assumptions. It has achieved this goal already for several classes of beauty meson transitions with a reliability and accuracy that before would have seemed unattainable.

Heavy quark theory is based on a two-part strategy analogous to the one adopted in chiral perturbation theory—another theoretical technology to deal reliably with nonperturbative dynamics in a special setting. Like there, heavy quark theory combines two basic elements, namely an asymptotic symmetry principle and a dynamical treatment telling us how the asymptotic limit is approached:

1. The symmetry principle is Heavy Quark Symmetry stating that all sufficiently heavy quarks behave identically under the strong interactions without sensitivity to their spin. This can easily be illustrated with the Pauli Hamiltonian describing the interaction of a quark of mass \( m_Q \) with a gauge field \( A_\mu = (A_0, \vec{A}) \):

\[
H_{\text{Pauli}} = -A_0 + \frac{(i\vec{D} - \vec{A})^2}{2m_Q} + \frac{\vec{\sigma} \cdot \vec{B}}{2m_Q} \Rightarrow -A_0 \text{ as } m_Q \to \infty ;
\]

i.e., in the infinite mass limit, quarks act like static objects without spin dynamics and subject only to colour Coulomb fields.

This simple consideration illustrates a general feature of heavy quark theory, namely that the spin of the heavy quark \( Q \) decouples from the dynamics in the heavy quark limit. Hadrons \( H_Q \) can therefore be labelled by the angular momentum \( j_q \) carried by its ‘light’ components—light valence quarks, gluons and sea quarks—in addition to its total spin \( S \). The \( S \) wave pseudoscalar and vector mesons—\( B & B^* \) and \( D & D^* \)—then form the ground-state doublet of heavy quark symmetry with \( [S, j_q] = [0, \frac{1}{2}], [1, \frac{1}{2}] \); a quartet of \( P \) wave configurations form the first excited states with \( [S, j_q] = [0, \frac{1}{2}], [1, \frac{1}{2}], [1, \frac{3}{2}], [2, \frac{3}{2}] \).

Heavy quark symmetry can be understood in an intuitive way: consider a hadron \( H_Q \) containing a heavy quark \( Q \) with mass \( m_Q \gg \Lambda_{QCD} \) surrounded by a ‘cloud’ of light degrees of freedom carrying quantum numbers of an antiquark \( \bar{q} \) or diquark \( qq \)\(^{22} \). This cloud has a rather complex structure: in addition to \( \bar{q} \) (for mesons) or \( qq \) (for baryons), it contains an indefinite number of \( q\bar{q} \) pairs and gluons that are strongly coupled to and constantly fluctuate into each other. There is, however, one thing we know: since typical frequencies of these fluctuations are \( \sim \mathcal{O}(\text{few}) \times \Lambda_{QCD} \), the normally dominant soft dynamics allow the heavy quark to exchange momenta of order few times \( \Lambda_{QCD} \) only with its surrounding medium. \( Q\bar{Q} \) pairs then cannot play a significant role, and the heavy quark can be treated as a quantum mechanical object rather than a field theoretic entity requiring second quantization. This provides a tremendous computational simplification even while maintaining a field theoretic description for the light degrees of freedom. Furthermore, techniques developed long ago in QED can profitably be adapted here.

2. We can go further and describe the interactions between \( Q \) and its surrounding light degrees of freedom through an expansion in powers of \( 1/m_Q \)—the Heavy Quark Expansion (HQE). This allows us to analyse pre-asymptotic effects, i.e., effects that fade away like powers of \( 1/m_Q \) as \( m_Q \to \infty \).

Let me anticipate the lessons we have learned: we have

- identified the sources of the non-perturbative corrections;
- found them to be smaller than they could have been;
- succeeded in relating the basic quantities of the heavy quark theory—KM parameters, masses and kinetic energy of heavy quarks, etc.—to various a priori independent observables with a considerable amount of redundancy;

\(^{22}\)This cloud is often referred to—somewhat disrespectfully—as ‘brown muck’, a phrase coined by the late Nathan Isgur.
– developed a better understanding of incorporating perturbative and nonperturbative corrections without double-counting.

In the following I shall sketch the concepts on which the heavy quark expansions are based, the techniques employed, the results obtained, and the problems encountered. It will not constitute a self-sufficient introduction to this vast and ever expanding field. My intent is to provide a guide through the literature for the committed student.

2.4.2 \textit{H(eavy) Q(uark) E(xpansions), fundamentals}

In describing weak decays of heavy flavour \textit{hadrons} one has to incorporate perturbative as well as non-perturbative contributions in a self-consistent and complete way. The only known way to tackle such a task invokes the Operator Product Expansion à la Wilson involving an effective Lagrangian. Further conceptual insights as well as practical results can be gained by analysing \textit{sum rules}; in particular they shed light on various aspects and formulations of quark–hadron duality.

2.4.2.1 Operator Product Expansion (OPE) for inclusive weak decays

Similar to the well-known case of $\sigma(e^+e^- \rightarrow \text{had})$ one invokes the optical theorem to describe the decay into a sufficiently inclusive final state $f$ through the imaginary part of the forward scattering operator evaluated to second order in the weak interactions

$$T(Q \rightarrow f) = \text{Im} \int d^4x i \{\mathcal{L}_W(x)\mathcal{L}_W^\dagger(0)\}_T$$

with the subscript $T$ denoting the time-ordered product and $\mathcal{L}_W$ the relevant weak Lagrangian\textsuperscript{23}. The expression in Eq. (199) represents in general a non-local operator with the space–time separation $x$ being fixed by the inverse of the energy release. If the latter is large compared to typical hadronic scales, then the product is dominated by short-distance physics, and one can apply a Wilsonian OPE, which yields an infinite series of local operators of increasing dimension\textsuperscript{24}. The width for the decay of a hadron $H_Q$ containing $Q$ is then obtained by taking the $H_Q$ expectation value of the operator $\hat{T}$:

$$\frac{\langle H_Q | \text{Im} \hat{T}(Q \rightarrow f \rightarrow Q) | H_Q \rangle}{2M_{H_Q}} \propto \Gamma(H_Q \rightarrow f) = \frac{G_F m_Q^5(\mu)}{192\pi^3} |V_{CKM}|^2 \cdot$$

$$\left[ c_3(\mu) \langle H_Q | Q | Q | H_Q \rangle(\mu) + \frac{c_5(\mu)}{m_Q^2} \langle H_Q | Q | Q \cdot Q | H_Q \rangle(\mu) + \right.$$\n
$$\left. + \sum_i \frac{c_{6,i}(\mu)}{m_Q^3} \langle H_Q | q \Gamma_i q | H_Q \rangle(\mu) + O(1/m_Q^4) \right].$$

Equation (200) exhibits the following important features:

– An auxiliary scale $\mu$ has been introduced to consistently separate short- and long-distance dynamics:

\begin{equation}
\text{short distance} < \mu^{-1} < \text{long distance}
\end{equation}

\textsuperscript{23}There are two qualitative differences from the case of $e^+e^- \rightarrow \text{had}$: in describing weak decays of a hadron $H_Q$, (i) one employs the weak rather than the electromagnetic Lagrangian, and (ii) one takes the expectation value between the $H_Q$ state rather than thehadronic vacuum.

\textsuperscript{24}I shall formulate the expansion in powers of $1/m_Q$, although it has to be kept in mind that it is really controlled by the inverse of the energy release. While there is no fundamental difference between the two for $b \rightarrow c/u\bar{u}\nu$ or $b \rightarrow c/u\bar{d}d$, since $m_b/m_c \gg \Lambda_{QCD}$, the expansion becomes of somewhat dubious reliability for $b \rightarrow c\bar{c}s$. It actually would break down for a scenario $Q_2 \rightarrow Q_1 l\bar{\nu}$ with $m_{Q_2} \simeq m_{Q_1}$—in contrast to HQET!
with the former entering through the coefficients and the latter through the effective operators; their matrix elements will thus depend on $\mu$.

In principle the value of $\mu$ does not matter: it reflects merely our computational procedure rather than how nature goes about its business. The $\mu$ dependence of the coefficients thus has to cancel against that of the corresponding matrix elements.

In practice, however, there are competing demands on the choice of $\mu$:

- On the one hand one has to choose

$$\mu \gg \Lambda_{\text{QCD}} ;$$

otherwise radiative corrections cannot be treated within perturbative QCD.

- On the other hand many computational techniques for evaluating matrix elements—among them the heavy quark expansions—require

$$\mu \ll m_b .$$

The choice

$$\mu \sim 1 \text{ GeV}$$

satisfies both of these requirements. It is important to check that the obtained numerical results do not exhibit a significant sensitivity to the exact value of $\mu$ when varying the latter in a reasonable range.

- Short-distance dynamics shape the $c$ number coefficients $c_i^{(f)}$. In practice they are evaluated in perturbative QCD. It is quite conceivable, though, that also nonperturbative contributions arise there; yet they are believed to be fairly small in beauty decays [55]. By the same token these short-distance coefficients provide also the portals through which New Physics can enter in a natural way.

- Nonperturbative contributions on the other hand enter through the expectation values of operators of dimension higher than three—$\bar{Q} \frac{1}{2} \sigma \cdot GQ$ etc.—and higher order corrections to the expectation value of the leading operator $\bar{Q}Q$, see below.

- In practice we cannot go beyond evaluating the first few terms in this expansions. More specifically we are limited to contributions through order $1/m_Q^3$; those are described in terms of six heavy quark parameters, namely two quark masses—$m_{b,c}$—, two expectation values of dimension-five operators—$\mu_2^2$ and $\mu_1^2$—and of dimension-six operators—the Darwin and ‘LS’ terms, $\rho_3^D$ and $\rho_3^LS$, respectively.\(^{25}\)

- This small and universal set of nonperturbative quantities describes a host of observables in $B$ transitions. Therefore their values can be determined from some of these observables and still leave a large number of predictions.

- It opens the door to a novel symbiosis of different theoretical technologies for heavy flavour dynamics—in particular between HQE and lattice QCD. For the HQP can be inferred from lattice studies. This enhances the power of and confidence in both technologies by

  * increasing the range of applications and
  * providing more validation points.

I shall give some examples later on.

- Expanding the expectation value of the leading operator $\bar{Q}Q$ for a pseudoscalar meson $P_Q$ with quantum number $Q$ in powers of $1/m_Q$ yields

$$\frac{1}{2M_{P_Q}} \langle P_Q | \bar{Q}Q | P_Q \rangle = 1 - \frac{\mu_\pi^2}{2m_\pi^2} + \frac{\mu_{GQ}^2}{2m_Q^2} + O(1/m_Q^3) ;$$

\(^{25}\)For simplicity I ignore here so-called ‘Intrinsic Charm’ contributions, see Ref. [56].
\( \mu_2^2(\mu) \) and \( \mu_G^2(\mu) \) denote the expectation values of the kinetic and chromomagnetic operators, respectively:

\[
\mu_2^2(\mu) \equiv \frac{1}{2M_{H_Q}} \langle H_Q | \bar{Q} \pi^2 Q | H_Q \rangle_{(\mu)} , \quad \mu_G^2(\mu) \equiv \frac{1}{2M_{H_Q}} \langle H_Q | \bar{Q} \gamma^i \cdot GQ | H_Q \rangle_{(\mu)} ;
\]

for short they are often called the kinetic and chromomagnetic moments.

Equation (205) implies that one has \( \langle H_Q | \bar{Q} Q | H_Q \rangle_{(\mu)} / 2M_{H_Q} = 1 \) for \( m_Q \rightarrow \infty \); i.e., the free quark model expression emerges asymptotically for the total width.

- The leading nonperturbative corrections arise at order \( 1/m_Q^2 \) only. That means they are rather small in beauty decays since \( (\mu/m_Q)^2 \sim \) few per cent for \( \mu \lesssim 1 \text{ GeV} \).
- This smallness of nonperturbative contributions explains a posteriori, why partonic expressions when coupled with a ‘smart’ perturbative treatment often provide a decent approximation.
- These nonperturbative contributions which are power suppressed can be described only if considerable care is applied in treating the parametrically larger perturbative corrections.
- Explicitly flavour-dependent effects arise in order \( 1/m_Q^3 \). They mainly drive the differences in the lifetimes of the various mesons of a given heavy flavour.
- An important practical distinction from the OPE treatment of \( e^+e^- \rightarrow \text{had} \) or deep-inelastic lepton nucleon scattering is the fact that the weak width depends on the fifth power of the heavy quark mass, see Eq. (200), and thus requires particular care in dealing with the delicate concept of quark masses.

One general, albeit subtle point has to be kept in mind here: while everybody these days invokes the OPE it is often not done employing Wilson’s prescription with the auxiliary scale \( \mu \), and different definitions of the relevant operators have been suggested. While results from one prescription can be translated into another one, order by order, great care has to be applied. I shall adopt here the so-called ‘kinetic scheme’ with \( \mu \simeq 1 \text{ GeV} \). It should be noted that the quantities \( \mu_2^2(\mu) \) and \( \mu_G^2(\mu) \) are quite distinct from the so-called HQET parameters \( \lambda_1 \) and \( \lambda_2 \) although the operators look identical. Furthermore the fact that perturbative corrections are rather smallish in the kinetic scheme generally does not hold in other schemes.

The absence of corrections of order \( 1/m_Q \) [57] is particularly noteworthy and intriguing since such corrections do exist for hadronic masses—\( M_{H_Q} = m_Q(1 + \bar{\Lambda}/m_Q + \mathcal{O}(1/m_Q^2)) \)—and those control the phase space. Technically this follows from the fact that there is no independant dimension-four operator that could emerge in the OPE. This result can be illuminated in more physical terms as follows. Bound-state effects in the initial state like mass shifts do generate corrections of order \( 1/m_Q \) to the total width; yet so does hadronization in the final state. Local colour symmetry demands that those effects cancel each other out. It has to be emphasized that the absence of corrections linear in \( 1/m_Q \) is an unambiguous consequence of the OPE description. If their presence were forced upon us, we would have encountered a qualitative change in our QCD paradigm. A discussion of this point has arisen recently phrased in the terminology of quarkhadron duality. I shall return to this point later.

2.4.2.2 Sum rules

There are classes of sum rules derived from QCD proper that relate the heavy quark parameters appearing in the OPE for inclusive \( B \rightarrow l\nu X_c \)–like \( \mu_2^2, \mu_G^2 \) etc.—with restricted sums over exclusive channels. They provide rigorous definitions, inequalities and experimental constraints [58]; for example:

\[
\mu_2^2(\mu)/3 = \sum_n \epsilon_n^2 \left| \tau_{1/2}^{(n)}(1) \right|^2 + 2 \sum_m \epsilon_m^2 \left| \tau_{3/2}^{(m)}(1) \right|^2 ,
\]

\textsuperscript{26}The operator \( \bar{Q}u\bar{D}Q \) can be reduced to the leading operator \( \bar{Q}Q \) through the equation of motion.
\[ \mu_G^2(\mu)/3 = -2 \sum_n \epsilon_n^2 |\tau_{1/2}^{(n)}(1)|^2 + 2 \sum_m \epsilon_m^2 |\tau_{3/2}^{(m)}(1)|^2 \] (208)

where \( \tau_{1/2}, \tau_{3/2} \) are the amplitudes for \( B \to \nu \ell D(j_q) \) with \( D(j_q) \) a hadronic system beyond the \( D \) and \( D^* \), \( j_q = 1/2 \) and \( 3/2 \) the angular momentum carried by the light degrees of freedom in \( D(j_q) \), as explained in the paragraph below Eq. (198), and \( \epsilon_m \) the excitation energy of the \( m \)th such system above the \( D \) with \( \epsilon_m \leq \mu \).

These sum rules have become of great practical value. I want to emphasize here one of their conceptual features: they show that the heavy quark parameters in the kinetic scheme are observables themselves.

### 2.4.2.3 Quark–hadron duality

The concept of quark–hadron duality (or duality for short), which goes back to the early days of the quark model, refers to the notion that a quark-level description should provide a good description of transition rates that involve hadrons, if one sums over a sufficient number of channels. This is a rather vague formulation: How many channels are ‘sufficiently’ many? How good an approximation can one expect? How process dependent is it? Yet it is typical in the sense that no precise definition of duality had been given for a long time, and the concept has been used in many different incarnations. A certain lack of intellectual rigour can be of great euristic value in the ‘early going’—but not forever.

A precise definition requires theoretical control over perturbative as well as nonperturbative dynamics. For limitations to duality have to be seen as effects over and beyond uncertainties due to truncations in the perturbative and nonperturbative expansions. To be more explicit: duality violations are due to corrections not accounted for due to

- truncations in the expansion and
- limitations in the algorithm employed.

One important requirement is to have an OPE treatment of the process under study, since otherwise we have no unambiguous and systematic inclusion of nonperturbative corrections. This is certainly the case for inclusive semileptonic and radiative \( B \) decays.

While we have no complete theory for duality and its limitations, we have certainly moved beyond the folkloric stage in the last few years. We have developed a better understanding of the physics effects that can generate duality violations—the presence of production thresholds for example—and have identified mathematical portals through which duality violations can enter. The fact that we construct the OPE in the Euclidean range and then have to extrapolate it to the Minkowskian domain provides such a gateway.

The problem with the sometimes heard statement that duality represents an additional ad hoc assumption is that it is not even wrong—it just misses the point.

More details on this admittedly complex subject can be found in Ref. [59] and for the truly committed student in Ref. [60]. Suffice it here to say that it had been predicted that duality violation in \( \Gamma_{SL}(B) \) can safely be placed below 0.5% [60]. The passion in the arguments over the potential size of duality violations in \( B \to \nu \ell X \) has largely faded away, since, as I discuss later on, the experimental studies of it have shown no sign of such limitations.

### 2.4.2.4 Heavy quark parameters

Through order \( 1/m_Q^3 \) there are six heavy quark parameters (HQPs) which fall into two different classes:

1. The heavy quark masses \( m_b \) and \( m_c \); they are ‘external’ to QCD; i.e., they can never be calculated by lattice QCD without experimental input.
2. The expectation values of the dimension five and six operators: $\mu_\pi^2$, $\mu_G^2$, $\rho_D^3$ and $\rho_{LS}^3$. They are ‘intrinsic’ to QCD, i.e., can be calculated by lattice QCD without experimental input.

Since weak decay widths depend on the fifth power of the heavy quark mass, great care has to be applied in defining this somewhat elusive entity in a way that can pass full muster by quantum field theory. To a numerically lesser degree this is true for the other HQPs as well. Their dependence on the auxiliary scale $\mu$ has to be carefully tracked.

- **Quark masses:** There is no quark mass per se—one has to specify the renormalization scheme used and the scale at which the mass is to be evaluated. The pole mass—i.e., the position of the pole in the perturbative Green function—has the convenient features that it is gauge invariant and infrared finite in perturbation theory. Yet in the complete theory it is infrared unstable [53] due to ‘renormalon’ effects. Those introduce an irreducible intrinsic uncertainty into the quark mass: $m_Q(1 + \delta(m_Q)/m_Q)$, with $\delta(m_Q)$ being roughly $\Lambda_{QCD}$. For the weak width it amounts to an uncertainty $\delta(m_Q^2) \sim 5\delta(m_Q)/m_Q$; i.e., it is parametrically larger than the power suppressed terms $\sim O(1/m_Q^2)$ one is striving to calculate. The pole mass is thus ill suited when including nonperturbative contributions. Instead one needs a running mass with an infrared cut-off $\mu$ to ‘freeze out’ renormalons.

The $\overline{MS}$ mass, which is a rather ad hoc expression convenient in perturbative computations rather than a parameter in an effective Lagrangian, would satisfy this requirement. It is indeed a convenient tool for treating reactions where the relevant scales exceed $m_Q$ in production processes like $Z^0 \rightarrow b\bar{b}$. Yet in decays, where the relevant scales are necessarily below $m_Q$, the $\overline{MS}$ mass is actually inconvenient or even inadequate. For it has a hand-made infrared instability:

$$\overline{m}_Q(\mu) = m_Q(m_Q) \left[ 1 + \frac{2\alpha_S}{\pi} \log \frac{m_Q}{\mu} \right] \rightarrow \infty \text{ as } \frac{\mu}{m_Q} \rightarrow 0.$$  \hspace{1cm} (209)

It is much more advantageous to use the ‘kinetic’ mass instead with

$$\frac{dm_Q(\mu)}{d\mu} = -\frac{16\alpha_S(\mu)}{3\pi} - \frac{4\alpha_S(\mu)}{3\pi} \frac{\mu}{m_Q} + \ldots,$$  \hspace{1cm} (210)

which has a linear scale dependence in the infrared. It is this kinetic mass I shall use in the following. Its value had been extracted from

$$e^+e^- \rightarrow \Upsilon(4S) \rightarrow H_bH'_bX$$  \hspace{1cm} (211)

before 2002 by different authors with better than about 2% accuracy [61] based on an original idea of M. Voloshin. Their findings expressed in terms of the kinetic mass can be summarized as follows:

$$\langle m_b(1 \text{ GeV})|_{\Upsilon(4S)-b\bar{b}} = 4.57 \pm 0.08 \text{ GeV}. \quad (212)$$

Charmonium sum rules yield

$$m_c(m_c) \simeq 1.25 \pm 0.15 \text{ GeV}. \quad (213)$$

The HQE allows one to relate the difference $m_b - m_c$ to the ‘spin averaged’ beauty and charm meson masses and the higher order HQPs [53]:

$$m_b - m_c = \langle M_B \rangle - \langle M_D \rangle + \left( \frac{1}{2m_c} - \frac{1}{2m_b} \right) \mu_\pi^2 + \ldots \simeq 3.50 \text{ GeV} + 40 \text{ MeV} \cdot \frac{\mu_\pi^2 - 0.5 \text{ (GeV)}^2}{0.1 \text{ (GeV)}^2} \ldots$$  \hspace{1cm} (214)

Yet this relation is quite vulnerable since it is dominantly an expansion in $1/m_c$ rather than $1/m_b$ and nonlocal correlators appear in order $1/m_b^2$. Therefore one is ill-advised to impose this relation a priori. One is of course free to consider it a posteriori.
– **Chromomagnetic moment**: Its value can be inferred quite reliably from the hyperfine splitting in the $B^*$ and $B$ masses:

$$
\mu_G^2(1 \text{ GeV}) \approx \frac{3}{2} \left[ M^2(B^*) - M^2(B) \right] \approx 0.35 \pm 0.03 \text{ (GeV)}^2 .
$$

(215)

– **Kinetic moment**: The situation here is not quite so definite. We have a rigorous lower bound from the SV sum rules [62]:

$$
\mu^2(\mu) \geq \mu_G^2(\mu)
$$

(216)

for any $\mu$; QCD sum rules yield

$$
\mu^2(1 \text{ GeV}) \approx 0.45 \pm 0.1 \text{ (GeV)}^2.
$$

(217)

– **Darwin and LS terms**: The numbers are less certain still for those. The saving grace is that their contributions are reduced in weight, since they represent $O(1/m^3_Q)$ terms.

$$
\rho^3_D(1 \text{ GeV}) \sim 0.1 \text{ (GeV)}^3 , -\rho^3_{LS}(\mu) \leq \rho^3_D(\mu).
$$

(218)

2.4.3 **First tests: weak lifetimes and SL branching ratios**

Let me begin with three general statements:

– Within the SM the semileptonic widths have to coincide for $D^0$ and $D^+$ mesons and for $B_d$ and $B_s$ mesons up to small isospin violations, since the semileptonic transition operators for $b \to l\nu c$ and $c \to l\nu s$ are isosinglets. The ratios of their semileptonic branching ratios are therefore equal to their lifetime ratios to a very good approximation:

$$
\frac{\text{BR}_{SL}(B^+)}{\text{BR}_{SL}(B_d)} = \frac{\tau(B^+)}{\tau(B_d)} + O \left( \left| \frac{V_{ub}}{V_{cb}} \right|^2 \right) ,
\frac{\text{BR}_{SL}(D^+)}{\text{BR}_{SL}(D^0)} = \frac{\tau(D^+)}{\tau(D^0)} + O \left( \left| \frac{V_{cd}}{V_{cs}} \right|^2 \right). \quad (219)
$$

For dynamical rather than symmetry reasons such a relation can be extended to $B_s$ and $D_s$ mesons [63]:

$$
\frac{\text{BR}_{SL}(B_s)}{\text{BR}_{SL}(B_{d})} \approx \frac{\tau(B_s)}{\tau(B_d)}, \quad (220)
$$

where $\tau(B_s)$ denotes the average of the two $B_s$ lifetimes.

– Yet the semileptonic widths of heavy flavour baryons will *not* be universal for a given flavour. The ratios of their semileptonic branching ratios will therefore not reflect their lifetime ratios. In particular for the charmed baryons one predicts large differences in their semileptonic widths [64].

– It is more challenging for theory to predict the absolute value of a semileptonic branching ratio than the ratio of such branching ratios.

2.4.3.1 **Charm lifetimes**

The lifetimes of all seven $C = 1$ charm hadrons have been measured now with the FOCUS experiment being the only one that has contributed to all seven lifetimes. In Table 2 the predictions based on the HQE (together with brief theory comments) are juxtaposed to the data [65]. While *a priori* the HQE might be expected to fail even on the semiquantitative level since $\mu_{had}/m_c \sim 1/2$ is an uncomfortably large expansion parameter, it works surprisingly well in describing the lifetime ratios even for baryons except for $\tau(\Xi_c^+)$ being about 50% longer than predicted. This agreement should be viewed as quite
nontrivial, since these lifetimes span more than an order of magnitude between the shortest and longest: $\tau(D^+)/\tau(\Omega_c) \simeq 14$. It provides one of the better arguments for charm acting like a heavy quark, at least in cases when the leading nonperturbative correction is of order $1/m_c^2$ rather than $1/m_c$.

The SELEX Collaboration has reported candidates for weakly decaying double charm baryons. It is my judgement that those candidates cannot be $C = 2$ baryons since their reported lifetimes are too short and do not show the expected hierarchy [65].

### 2.4.3.2 Beauty lifetimes

Theoretically one is on considerably safer ground when applying the HQE to lifetime ratios of beauty hadrons, since the expansion parameter $\mu_{had}/m_b \sim 1/7$ is small compared to unity. The HQE provided predictions in the old-fashioned sense; i.e., it produced them before data with significant accuracy were known.

Several comments are in order to interpret the results:

- The $B^+ - B_d$ lifetime ratio has been measured now with better than 1% accuracy – and the very first prediction based on the HQE was remarkably on target [18].
- The most dramatic deviation from a universal lifetime for $B = 1$ hadrons has emerged in $B_c$ decays. Their lifetime is only a third of the other beauty lifetimes—again in full agreement with the HQE prediction. That prediction is actually less obvious than it might seem. For the observed $B_c$ lifetime is close to the charm lifetime as given by $\tau(D^0)$, and that is what one would expect.

### Table 2: The weak lifetime ratios of $C = 1$ hadrons

<table>
<thead>
<tr>
<th>$1/m_c$ expect.</th>
<th>Theory comments</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau(D^+)/\tau(D^0)$</td>
<td>$1 + \left(\frac{f_D}{200 \text{ MeV}}\right)^2 \sim 2.4$</td>
<td>PI dominant</td>
</tr>
<tr>
<td>$\tau(D^+)/\tau(D^0)$</td>
<td>$0.9 - 1.3{1.0 - 1.07}$</td>
<td>With [Without] WA</td>
</tr>
<tr>
<td>$\tau(\Lambda_c^+)/\tau(D^0)$</td>
<td>$1.3 - 1.7$</td>
<td>Quark model matrix elements</td>
</tr>
<tr>
<td>$\tau(\Xi_c^+)/\tau(D^0)$</td>
<td>$1.6 - 2.2$</td>
<td>Quark model matrix elements</td>
</tr>
<tr>
<td>$\tau(\Xi^+)/\tau(D^0)$</td>
<td>$2.8$</td>
<td>Quark model matrix elements</td>
</tr>
<tr>
<td>$\tau(\Xi^+)/\tau(D^0)$</td>
<td>$4$</td>
<td>Quark model matrix elements</td>
</tr>
<tr>
<td>$\tau(\Xi^+)/\tau(D_r^0)$</td>
<td>$1.4$</td>
<td>Quark model matrix elements</td>
</tr>
</tbody>
</table>

### Table 3: The weak lifetime ratios of $B = 1$ hadrons

<table>
<thead>
<tr>
<th>$1/m_b$ expect.</th>
<th>Theory comments</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau(B^+)/\tau(B_s)$</td>
<td>$1 + 0.05\left(\frac{f_B}{200 \text{ MeV}}\right)^2 \simeq 0.92 [18]$</td>
<td>PI in $\tau(B^+)$</td>
</tr>
<tr>
<td>$1.06 \pm 0.02 [23]$</td>
<td>fact. at low scale 1 GeV</td>
<td></td>
</tr>
<tr>
<td>$\tau(B_s)$</td>
<td>$0.90 \pm 0.30 [66]$</td>
<td></td>
</tr>
<tr>
<td>$\tau(\Lambda_b)/\tau(B_s)$</td>
<td>$0.94 \pm 0.96 [68, 69]$</td>
<td>Quark model</td>
</tr>
<tr>
<td>$\tau(B_c) \sim (0.3 - 0.7) \text{ ps} [94, 96, 97]$</td>
<td>Largest lifetime diff. no $1/m_Q$ term crucial</td>
<td>$0.45 \pm 0.12 \text{ ps} [66]$</td>
</tr>
</tbody>
</table>

$$\frac{\Delta \tau(B^+)}{\tau(B^+)} \sim 22\% \cdot \left(\frac{f(B^+)}{200 \text{ MeV}}\right)^2 \simeq 0.87 [22]$$

$$12 \pm 5\% [66]$$

$\Delta M_{B_s}$ Less reliable $0.65 \pm 0.3 \text{ CDF}$

$\Delta M_{B_s}$ than $\Delta M_{B_s}$ $0.23 \pm 0.17 \text{ D0}$

171
already in a naive parton model treatment, where $\Gamma(b\bar{c}) \simeq \Gamma(c) \cdot [1 + \Gamma(b)/\Gamma(c)]$. However, it had been argued that inside such a tightly bound state the $b$ and $c$ quark masses had to be replaced by effective masses reduced by the (same) binding energy: $m_{b}^{eff} = m_{b} - B.E.$, $m_{c}^{eff} = m_{c} - B.E.$, with $B.E. \sim O(\Lambda_{QCD})$. This would prolong the weak lifetimes of the two quarks greatly, since those depend on the fifth power of the quark masses and would do so much more for the charm transition than for the beauty one. Yet such an effect would amount to a correction of order $1/m_{Q}$, which is not allowed by the OPE, as explained above at the end of Section 2.4.2.1; the more detailed argument can be found in Ref. [72].

- A veritable saga is emerging with respect to $\tau(\Lambda_{b})$. The first prediction stated [67] that $\tau(\Lambda_{b})/\tau(\Lambda_{d})$ could not fall below 0.9. A more detailed analysis led to two conclusions [68], namely that the HQE most likely leads to

$$\frac{\tau(\Lambda_{b})}{\tau(\Lambda_{d})} \simeq 0.94$$

(221)

with an uncertainty of a few per cent, while a lower bound had to hold

$$\frac{\tau(\Lambda_{b})}{\tau(\Lambda_{d})} \geq 0.88$$

(222)

A violation of this bound would imply that we need a new paradigm for evaluating at least baryonic matrix elements.

There are actually two questions one can ask concerning $\tau(\Lambda_{b})/\tau(\Lambda_{d})$:

1. What is theoretically the most likely value for $\tau(\Lambda_{b})/\tau(\Lambda_{d})$?
2. How much lower can one reasonably push it?

While there is a connection between those two questions, they clearly should be distinguished. Most theoretical analyses—employing quark models, QCD sum rules, or lattice studies—agree on the first question, namely that the ratio is predicted to lie above 0.90. Yet the data have for many years pointed to a significantly lower value $\sim 0.80$. This apparent discrepancy has given rise to the second question listed above. Reference [68] provided a carefully reasoned answer to it. Reference [73] stated a value of $0.86 \pm 0.05$, which is sometimes quoted as the theory prediction.

I object to viewing this value as the answer to the first question above; one might consider it as a response to the second question, although even then I remain sceptical of it.

The new CDF result seems to reshuffle the cards. The question is whether it is just a high fluctuation—implying a worrisome discrepancy between theory and experiment—or represents a new trend to be confirmed in the future, which would represent an impressive ‘comeback’ success for the HQE.

No matter what the final verdict will be on $\tau(\Lambda_{b})$, it is important to measure also $\tau(\Xi_{b}^{0})$ and $\tau(\Xi_{b}^{-})$—either to confirm success or diagnose failure. One expects [74]:

$$\tau(\Xi_{b}^{0}) \simeq \tau(\Lambda_{b}) < \tau(\Lambda_{d}) < \tau(\Xi_{b}^{-})$$

(223)

where the ‘$<$’ signs indicate an about 7% difference. If the $\Lambda_{b} - \Lambda_{d}$ lifetime difference were larger than predicted, one would like to know whether the whole lifetime hierarchy of Eq. (223) is stretched out—say ‘$<$’ in $\tau(\Lambda_{b}) < \tau(\Lambda_{d}) < \tau(\Xi_{b}^{-})$ represents differences of 10% or even more—or whether the splittings in the baryon lifetimes are as expected, yet their overall values reduced relative to $\tau(\Lambda_{d})$.

- The original prediction that $\tau(\Lambda_{d})/\tau(B_{s})$ is unity within 1–2% [63, 67] has been confirmed by subsequent authors. Yet the data have stubbornly remained somewhat low. This measurement deserves great attention and effort. While I consider the prediction to be on good footing, it is based on an evaluation of a complex dynamical situation rather than a theorem or even symmetry. Establishing a discrepancy between theory and experiment here would raise some very intriguing questions.
2.4.4 The $V(cb)$ ‘saga’ – a case study in accuracy

2.4.4.1 Inclusive semileptonic $B$ decays

The value of $|V(cb)|$ is extracted from $B \rightarrow \ell \nu X_c$ in two steps.

A: One expresses $\Gamma(B \rightarrow \ell \nu X_c)$ in terms of the HQPs—quark masses $m_b, m_c$ and the expectation values of local operators $\mu^2_x, \mu^2_G, \rho^3_D$ and $\rho^3_{LS}$—as accurately as possible, namely through $\mathcal{O}(1/m_b^2)$ and to all orders in the BLM treatment for the partonic contribution. Having precise values for these HQPs is not only of obvious use for extracting $|V(cb)|$ and $|V(ub)|$, but also yields benchmarks for how much numerical control lattice QCD provides us over nonperturbative dynamics.

B: The numerical values of these HQPs are extracted from the *shapes* of inclusive lepton distributions as encoded in their *normalized* moments. Two types of moments have been utilized, namely lepton energy and hadronic mass moments. While the former are dominated by the contribution from the ‘partonic’ term $\propto (B|\bar{b}b|B)$, the latter are more sensitive to higher nonperturbative terms $\mu^2_x$ and $\mu^2_G$ and thus have to form an integral part of the analysis.

Executing the first step in the so-called kinetic scheme and inserting the experimental number for $\Gamma(B \rightarrow \ell \nu X_c)$ one arrives at [75]

$$
\frac{|V(cb)|}{0.0417} = D_{exp} \cdot (1 + \delta_{th}) [1 + 0.3(a_S(m_b) - 0.22)] [1 - 0.66(m_b - 4.6) + 0.39(m_c - 1.15) \\
+ 0.013(\mu^2_x - 0.4) + 0.05(\mu^2_G - 0.35) + 0.09(\rho^3_D - 0.2) + 0.01(\rho^3_{LS} + 0.15)],
$$

$$
D_{exp} = \sqrt{\frac{BR_{SL}(B)}{0.105}} \frac{1.55 \text{ ps}}{\tau_B}
$$

(244)

where all the HQPs are taken at the scale 1 GeV and their ‘seed’ values are given in the appropriate power of GeV; the theoretical error at this point is given by

$$
\delta_{th} = \pm 0.5\%_{\text{pert}} \pm 1.2\%_{\text{hWc}} \pm 0.4\%_{\text{hpc}} \pm 0.3\%_{\text{IC}}
$$

(225)

reflecting the remaining uncertainty in the Wilson coefficient of the leading operator $\bar{b}b$, as yet uncalculated perturbative corrections to the Wilson coefficients of the chromomagnetic and Darwin operators, higher order power corrections including duality violations in $\Gamma_{SL}(B)$ and nonperturbative effects due to operators containing charm fields, respectively. Concerning the last item, in Ref. [75] an error of 0.7% was stated. A dedicated analysis of such IC effects allowed one to reduce this uncertainty down to 0.3% [56].

BaBar has performed the state-of-the-art analysis of several lepton energy and hadronic mass moments [76] obtaining an impressive fit with the following HQPs in the kinetic scheme [77]:

$$
m_b(1 \text{ GeV}) = (4.61 \pm 0.068) \text{ GeV}, \ m_c(1 \text{ GeV}) = (1.18 \pm 0.092) \text{ GeV}
$$

(226)

$$
m_b(1 \text{ GeV}) - m_c(1 \text{ GeV}) = (3.436 \pm 0.032) \text{ GeV}
$$

(227)

$$
\mu^2_x(1 \text{ GeV}) = (0.447 \pm 0.053) \text{ GeV}^2, \ \mu^2_G(1 \text{ GeV}) = (0.267 \pm 0.067) \text{ GeV}^2
$$

(228)

$$
\rho^3_D(1 \text{ GeV}) = (0.195 \pm 0.029) \text{ GeV}^3
$$

(229)

$$
|V(cb)|_{\text{incl}} = 41.390 \cdot (1 \pm 0.021) \times 10^{-3}.
$$

(230)

The DELPHI Collaboration have refined their pioneering study of 2002 [78] obtaining [79]:

$$
|V(cb)|_{\text{incl}} = 42.1 \cdot (1 \pm 0.014)_{\text{meas}} \pm 0.014)_{\text{fit}} \pm 0.015)_{\text{th}} \times 10^{-3}
$$

(231)
### Table 4: The 2005 values of the HQPs obtained from a comprehensive analysis of $B \to \ell \nu X_c$ and $B \to \gamma X$ [51] and compared with predictions

<table>
<thead>
<tr>
<th>Heavy Quark Parameter</th>
<th>Value from $B \to \ell \nu X_c/\gamma X$</th>
<th>Predict. from other observ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_b (1 \text{ GeV})$</td>
<td>$(4.59 \pm 0.025)_{\text{exp}} \pm 0.030</td>
<td>_{\text{HQE}} \text{ GeV}$</td>
</tr>
<tr>
<td>$m_c (1 \text{ GeV})$</td>
<td>$(1.142 \pm 0.037)_{\text{exp}} \pm 0.045</td>
<td>_{\text{HQE}} \text{ GeV}$</td>
</tr>
<tr>
<td>$[m_b - m_c] (1 \text{ GeV})$</td>
<td>$(3.446 \pm 0.025) \text{ GeV}$</td>
<td>$(3.46 \pm X) \text{ GeV, Eq. (214)}$</td>
</tr>
<tr>
<td>$[m_b - 0.67m_c] (1 \text{ GeV})$</td>
<td>$(3.82 \pm 0.017) \text{ GeV}$</td>
<td></td>
</tr>
<tr>
<td>$\mu^2_G (1 \text{ GeV})$</td>
<td>$(0.297 \pm 0.024)_{\text{exp}} \pm 0.046</td>
<td>_{\text{HQE}} \text{ GeV}^2$</td>
</tr>
<tr>
<td>$\mu^2_L (1 \text{ GeV})$</td>
<td>$(0.401 \pm 0.019)_{\text{exp}} \pm 0.035</td>
<td>_{\text{HQE}} \text{ GeV}^2$</td>
</tr>
<tr>
<td>$\rho^3_{D} (1 \text{ GeV})$</td>
<td>$(0.174 \pm 0.009)_{\text{exp}} \pm 0.022</td>
<td>_{\text{HQE}} \text{ GeV}^3$</td>
</tr>
<tr>
<td>$\rho^3_{LS} (1 \text{ GeV})$</td>
<td>$-(0.183 \pm 0.054)_{\text{exp}} \pm 0.071</td>
<td>_{\text{HQE}} \text{ GeV}^3$</td>
</tr>
</tbody>
</table>

A comprehensive analysis of all relevant data from $B$ decays, including from $B \to \gamma X$ yields the results listed in Table 4 [51], where they are compared to their predicted values. Some had already been given in Table 1. With these HQPs one arrives at

$$\langle |V(cb)|_{\text{incl}} \rangle = 41.96 \cdot (1 \pm 0.0055)_{\text{exp}} \pm 0.0083|_{\text{HQE}} \pm 0.014|_{\Gamma_{SL}} \times 10^{-3}.$$ \hspace{1cm} (232)

For a full appreciation of these results one has to note the following:

- With just these six parameters one obtains an excellent fit to several energy and hadronic mass moments even for different values of the lower cut on the lepton or photon energy. Varying those lower cuts also provides more direct information on the respective energy spectra beyond the moments.

- Even better, the fit remains very good when one ‘seeds’ two of these HQPs to their predicted values, namely $\mu^2_G (1 \text{ GeV}) = 0.35 \pm 0.03 \text{ GeV}^2$ as inferred from the $B^* - B$ hyperfine mass splitting and $\rho^3_{LS} = -0.1 \text{ GeV}^3$ allowing only the other four HQPs to float.

- These HQPs are treated as free fitting parameters. It could easily have happened that they assume unreasonable or even unphysical values. Yet they take on very special values fully consistent with all constraints that can be placed on them by theoretical means as well as other experimental input. To cite but a few examples:

  - The value for $m_b$ inferred from the weak decay of a $B$ meson agrees completely within the stated uncertainties with what has been derived from the electromagnetic and strong production of $b$ hadrons just above threshold.

  - The rigorous inequality $\mu^2_L > \mu^2_G$, which had not been imposed as a constraint, is satisfied.

  - $\mu^2_G$ indeed emerges with the correct value, as does $\mu^2_L$.

  - $m_b - m_c$ agrees very well with what one infers from the spin-averaged $B$ and $D$ meson masses. However this a posteriori agreement does not justify imposing it as an a priori constraint. For the mass relation involves an expansion in $1/m_c$, which is of less than sterling reliability. Therefore I have denoted its uncertainty by $X$.

  - The 1% error in $m_b$ taken at face value might suggest that it alone would generate more than a 2.5% uncertainty in $|V(cb)|$, i.e., by itself saturating the total error given in Eq. (232). The resolution of this apparent contradiction is as follows. The dependence of the total semileptonic width and also of the lowest lepton energy moments on $m_b$ and $m_c$ can be approximated by $m_b^4 (m_b - m_c)^2$ for the actual quark masses; for the leading contribution this can be written as $\Gamma_{SL}(B) \propto (m_b - \frac{2}{3}m_c)^5$. From the values for $m_b$ and $m_c$, Eq. (226), and their correlation given in Ref. [76] one derives

$$m_b (1 \text{ GeV}) - 0.67 m_c (1 \text{ GeV}) = (3.819 \pm 0.017) \text{ GeV} = 3.819 \cdot (1 \pm 0.45\%) \text{ GeV}.$$ \hspace{1cm} (233)
That is, it is basically this peculiar combination that is measured directly through $\Gamma_{SL}(B)$, and thus its error is so tiny. It induces an uncertainty of 1.1% into the value for $|V(cb)|$. Equation (233) has another important use in the future, namely to provide a very stiff validation challenge to lattice QCD’s determinations of $m_b$ and $m_c$.

With all these cross-checks we can defend the smallness of the stated uncertainties. The analysis of Ref. [80] arrives at similar numbers (although I cannot quite follow their error analysis).

More work remains to be done: (i) The errors on the hadronic mass moments are still sizeable; decreasing them will have a significant impact on the accuracy of $m_b$ and $\mu^2_{\pi}$. (ii) As discussed in more detail below, imposing high cuts on the lepton energy degrades the reliability of the theoretical description. Yet even so it would be instructive to analyse at which cut theory and data part ways. I shall return to this point below. (iii) As another preparation for $V(ub)$ extractions one can measure $q^2$ moments or mass moments with a $q^2$ cut to see how well one can reproduce the known $V(cb)$.

2.4.4.2 Exclusive semileptonic $B$ decays

While it is my judgement that the most precise value for $|V(cb)|$ can be extracted from $B \to l\nu X_c$, this does not mean that there is no motivation for analysing exclusive modes. On the contrary: the fact that one extracts a value for $|V(cb)|$ from $B \to l\nu D^*$ at zero recoil fully consistent within a smallish uncertainty represents a great success since the systematics experimentally as well as theoretically are very different:

$$|V(cb)|_{B \to D^*} = 0.0416 \cdot \left(1 \pm 0.022_{\text{exp}} \pm 0.06_{\text{th}}\right) \quad \text{for} \quad F_{B \to D^*}(0) = 0.90 \pm 0.05 . \quad (234)$$

It has been suggested [81] to treat $B \to l\nu D$ with the ‘BPS expansion’ based on $\mu^2_{\pi} \simeq \mu^2_{\pi}$ and extract $|V(cb)|$ with a theoretical error not larger than $\sim 2\%$. It would be most instructive to compare the formfactors and their slopes found in this approach with those of LQCD [82].

2.4.5 The adventure continues: $V(ub)$

There are several lessons we can derive from the $V(cb)$ saga: (i) Measuring various moments of $B \to l\nu X_u$ and extracting HQPs from them is a powerful tool to strengthen confidence in the analysis. Yet it is done for validation purposes only. For there is no need to ‘reinvent the wheel’: When calculating the width and (low) moments of $B \to l\nu X_u$ one has to use the values of the HQPs as determined in $B \to l\nu X_c$. (ii) $\Gamma(B \to l\nu X_u)$ is actually under better theoretical control than $\Gamma(B \to l\nu X_c)$ since the expansion parameter is smaller $-\frac{m_{had}}{m_b}$ vs. $-\frac{m_{had}}{m_b-m_c}$ and $O(\alpha_s^2)$ corrections are known exactly.

On the impact of cuts: In practice there arises a formidable complication: to distinguish $b \to u$ from the huge $b \to c$ background, one applies cuts on variables like lepton energy $E_l$, hadronic mass $M_X$, the lepton-pair invariant mass $q^2$. As a general rule the more severe the cut, the less reliable the theoretical calculation becomes. More specifically the imposition of a cut introduces a new dimensional scale called ‘hardness’ $Q$ [83]. Nonperturbative contributions emerge scaled by powers of $1/Q$ rather than $1/m_b$. If $Q$ is much smaller than $m_b$ such an expansion becomes unreliable. Furthermore the OPE cannot capture terms of the form $e^{-Q/\mu}$. While these are irrelevant for $Q \sim m_b$, they quickly gain relevance when $Q$ approaches $\mu$. Ignoring this effect would lead to a ‘bias’, i.e. a systematic shift of the HQPs away from their true values.

This impact has been studied for radiative $B$ decays with their simpler kinematics in a pilot study [83] and a detailed analysis [84] of the average photon energy and its variance. The first provides a measure mainly of $m_b/2$, the latter of $\mu^2_{\pi}/12$. These biases were found to be relevant down to $E_{cut} = 1.85$ GeV and increasing quickly above 2 GeV. While the existence of such effects is of a general nature, the estimate of their size involves model-dependent elements. Yet as long as those corrections are of moderate size, they can be considered reliable. Once they become large, we are losing theoretical control.
Figure 7 shows data for the average photon energy and its variance for different lower cuts on the photon energy from CLEO, BaBar and Belle compared to the OPE predictions without and with bias corrections. The comparison shows the need for those bias corrections and their being under computational control over a sizeable range of \( E_{\text{cut}} \). Even more important than providing us with possibly more accurate values for \( m_b \) and \( \mu_2^2 \), these studies enhance confidence in our theoretical tools.

These findings lead to the following conclusions: (i) As far as theory is concerned there is a high premium on keeping the cuts as low as possible. (ii) Such cuts introduce biases in the HQP values extracted from the truncated moments; yet within a certain range of the cut variables those biases can be corrected for and thus should not be used to justify inflating the theoretical uncertainties. (iii) In any case, measuring the moments as functions of the cuts provides powerful cross-checks for our theoretical control.

‘Let a thousand blossoms bloom’: Several suggestions have been made for cuts to suppress the \( b \to c \) background to manageable proportions. None provides a panacea. The most straightforward one is to focus on the lepton energy endpoint region; however, it captures merely a small fraction of the total \( b \to u \) rate, which can be estimated only with considerable model dependence. This model sensitivity can be moderated with information on the heavy quark distribution function inferred from \( B \to \gamma X \). Furthermore, weak annihilation contributes only in the endpoint region and with different weight in \( B_d \) and \( B_u \) decays [63]. Thus the lepton spectra have to be measured separately for charged and neutral \( B \) decays.

Measuring the hadronic recoil mass spectrum up to a maximal value \( M_X^{\text{max}} \) captures the lion’s share of the \( b \to u \) rate if \( M_X^{\text{max}} \) is above 1.5 GeV; yet it is still vulnerable to theoretical uncertainties in the very low \( q^2 \) region. This problem can be addressed in two different ways: adopting Alexander the Great’s treatment of the Gordian knot one can impose a lower cut on \( q^2 \) or one can describe the low \( q^2 \) region with the help of the measured energy spectrum in \( B \to \gamma X \) for \( 1.8 \text{ GeV} \leq E_\gamma \leq 2.0 \text{ GeV} \). Alternatively one can apply a combination of cuts. Studying \( B_d \) and \( B_u \) decays is still desirable, yet not as essential as for the previous case.
In any case one should not restrict oneself to a fixed cut, but vary the latter over some reasonable range and analyse to what degree theory can reproduce this cut dependence to demonstrate control over the uncertainties.

There is not a single ‘catholic’ path to the promised land of a precise value for $|V(ub)|$; presumably many paths will have to be combined [85]. Yet it seems quite realistic that the error can be reduced to about 5% over the next few years.

2.5 Summary of Lecture II

As explained in Lecture I, while CKM forces are generated by the exchange of gauge bosons, its couplings involve elements of the CKM matrix. Yet those originate in the elements of the up- and down-type quark mass matrices. Thus the CKM parameters are intrinsically connected with one of the central mysteries of the SM, namely the generation in particular of fermion masses and family replication. Furthermore the hierarchy in the quark masses and the likewise hierarchical pattern of the CKM matrix elements strongly hints at some deeper level of dynamics about which we are quite ignorant. Nevertheless CKM theory with its mysterious origins has proved itself to be highly successful in describing even quantitatively a host of phenomena occurring over a wide array of scales. It led to the ‘Paradigm of Large $CP$Violation in $B$ Decays’ as a prediction in the old-fashioned sense; i.e., predictions were made well before data of the required sensitivity existed. From the observation of a tiny and shy phenomenon—$CP$ violation in $K_L$ decays on the $O(10^{-3})$ level—it predicted without ‘plausible deniability’ almost ubiquitous manifestations of $CP$ violation about two orders of magnitude larger in $B$ decays. This big picture has been confirmed now in qualitative as well as impressively quantitative agreement with SM predictions:

- Two $CP$-insensitive observables, namely $|V(ub)/V(cb)|$ and $\Delta M_{B_s}/\Delta M_{B_s}$, imply that $CP$ violation has to exist and in a way that at present is fully consistent with the measurements of $\epsilon$ and $\sin 2\phi_1$ and others.
- Time-dependent $CP$ asymmetries in the range of several $\times 10\%$ have been established in $B_d \to \psi K_S, \pi^+\pi^-$ and $\eta'/K_S$ with several others on the brink of being found.
- Direct $CP$ violation of about 10% or even larger has been discovered in $B_d \to \pi^+\pi^-$ and $K^-\pi^+$.
- The first significant sign of $CP$ violation in a charged meson has surfaced in $B^+ \to K^+\rho^0$.
- The optimists among us might discern the first signs of tension between data and the predictions of CKM theory in $|V(ub)/V(cb)|$ and $\Delta M_{B_d}/\Delta M_{B_s}$ vs. $\sin 2\phi_1$ and in the $CP$ asymmetries in $b \to sqq$ vs. $b \to c\bar{c}s$ driven transitions.

For all these successes it is quite inappropriate to refer anymore to CKM theory as an ‘ansatz’ with the latter’s patronizing flavour. Instead I would characterize these developments as ‘the expected triumph of peculiar theory’. However, as explained in Lecture III, it makes great sense—actually it is mandatory to search for its phenomenological limitations in future even more sensitive data sets. This will require great advances in experimental sensitivity—I have no doubt about their feasibility—and further progress in our quantitative theoretical control over heavy flavour decays. I have presented some case studies which give reason for optimism in this area as well. An essential element there is the availability of a comprehensive set of high-quality data: among other things they provide the motivation for theorists to sharpen their tools, and they allow us to defend our estimates of uncertainties rather than merely state them.

I shall indulge myself in three more ‘cultural’ conclusions:

- The aforementioned ‘CKM Paradigm of Large $CP$ Violation in $B$ Decays’ is due to the confluence of several favourable, yet $a$ priori less than likely factors that must be seen as gifts from Nature who had

\footnote{The German ‘ansatz’ refers to an educated guess.}
– arranged for a huge top mass,
– a ‘long’ $B$ lifetime,
– the $\Upsilon (4S)$ resonance being above the $B\bar{B}$, yet below the $B\bar{B}^*$ thresholds, and
– regaled us previously with charm hadrons, which prompted the development of detectors with an effective resolution that is needed to track $B$ decays.

– ‘Quantum mysteries’ like EPR correlations with their intrinsic non-local features were essential for observing $CP$ violation involving $B_d - \bar{B}_d$ oscillations in $\Upsilon (4S) \rightarrow B_d \bar{B}_d$ and to establish that indeed there is $T$ violation commensurate with $CP$ violation.
– While hadronization is not easily brought under quantitative theoretical control, it greatly enhances observable $CP$ asymmetries and can provide most valuable cross-checks for our interpretation of data.

3 Lecture III: Probing the flavour paradigm of the *emerging new* Standard Model

3.1 On the incompleteness of the SM

As described in the previous lectures the SM has scored novel—i.e., qualitatively new—successes in the last few years in the realm of flavour dynamics. Owing to the very peculiar structure of the latter they have to be viewed as amazing. Yet even so, the situation can be characterized with a slightly modified quote from Einstein:

“We know a lot—yet understand so little.”

That is, these successes do not invalidate the general arguments in favour of the SM being *incomplete*—the search for New Physics is as mandatory as ever.

You have heard about the need to search for New Physics before and what the outcome has been of such efforts so far, have you not? And it reminds you of a quote by Samuel Beckett:

“Ever tried? Ever failed?
No matter.
Try again. Fail again. Fail better.”

Only an Irishman can express profound scepticism concerning the world in such a poetic way. Beckett actually spent most of his life in Paris, since Parisians like to listen to someone expressing such a world view, even while they do not share it. Being in the service of Notre Dame du Lac, the home of the ‘Fighting Irish’, I cannot just ignore such advice.

My colleague and friend Antonio Masiero likes to say: “You have to be lucky to find New Physics.” True enough—yet let me quote someone who just missed by one year being a fellow countryman of Masiero, namely Napoleon, who said: “Being lucky is part of the job description for generals.” Quite seriously I think that if you as an high-energy physicist do not believe that someday somewhere you will be a general—maybe not in a major encounter, but at least in a skirmish—then you are frankly in the wrong line of business.

My response to these concerns is: “Cheer up—we know there is New Physics—we will not fail forever!” I shall marshall the arguments—compelling ones in my judgement—that point to the existence of New Physics.

3.1.1 Theoretical shortcomings

These arguments have been given already in the beginning of Lecture I.
Quantization of electric charge: While electric charge quantization
\[ Q_e = 3Q_d = -\frac{3}{2}Q_u \] (235)
is an essential ingredient of the SM—it allows one to vitiate the ABJ anomaly—it does not offer any understanding. It would naturally be explained through Grand Unification at very high energy scales implemented through, e.g., SO(10) gauge dynamics. I call this the ‘guaranteed New Physics’ gNP.

Family replication and CKM structure: We infer from the observed width of \( Z^0 \) decays that there are three (light) neutrino species. The hierarchical pattern of CKM parameters as revealed by the data is so peculiar as to suggest that some other dynamical layer has to underlie it. I refer to it as ‘strongly suspected New Physics’ or ssNP. We are quite in the dark about its relevant scales. Saying we pin our hopes for explaining the family replication on super-string or M theory is a scholarly way of saying we hardly have a clue what that ssNP is.

Electroweak symmetry breaking and the gauge hierarchy: What are the dynamics driving the electroweak symmetry breaking of \( SU(2)_L \times U(1) \rightarrow U(1)_{QED} \)? How can we tame the instability of Higgs dynamics with its quadratic mass divergence? I find the arguments compelling that point to New Physics at the \( \sim 1 \) TeV scale—like low-energy SUSY; therefore I call it the ‘confidently predicted’ New Physics or cpNP.

Furthermore the more specific ‘Strong CP Problem’ of QCD has not been resolved. Similar to the other shortcomings listed above it is a purely theoretical problem in the sense that the offending coefficient for the \( P \) and \( CP \) odd operator \( \tilde{G} \cdot G \) can be fine-tuned to zero, see Section 1.1.1.2,—yet in my eyes that is not a flaw.

3.1.2 Experimental signs

Strong, albeit not conclusive (by itself) evidence for neutrino oscillations comes from the KamLAND and K2K experiments in Japan studying the evolution of neutrino beams on Earth.

Yet compelling experimental evidence for the SM being incomplete comes from ‘heavenly signals’, namely from astrophysics and cosmology.

The baryon number of the Universe: One finds only about one baryon per \( 10^9 \) photons with the latter being mostly in the cosmic background radiation; there is no evidence for primary antimatter.

Therefore New Physics has to exist.

The aforementioned New CP Paradigm tells us that CP violating phases can be large.

Dark matter: Analysis of the rotation curves of stars and galaxies reveals that there is a lot more ‘stuff’—i.e. gravitating agents—out there than meets the eye. About a quarter of the gravitating agents in the Universe are such dark matter, and they have to be mostly nonbaryonic.

The SM has no candidate for it.

Solar and atmospheric \( \nu \) anomalies: The sun has been ‘seen’ by Super-Kamiokande in the light of neutrinos, as shown in Fig. 8. Looking carefully one realizes that the sun looks paler than it should: more than half of the originally produced \( \bar{\nu}_e \) disappear on the way to the Earth by changing their identity. Muon neutrinos produced in the atmosphere perform a similar disappearance act. These disappearances have to be attributed predominantly to neutrino oscillations (rather than neutrino decays). This requires neutrinos to carry nondegenerate masses.

Dark energy: Type 1a supernovae are considered ‘standard candles’; i.e., considering their real light output known allows one to infer their distance from their apparent brightness. When in 1998 two teams of researchers studied them at distance scales of about five billion light years,
they found them to be fainter as a function of their redshift than what the conventional picture of the Universe’s decelerating expansion would yield. Unless gravitational forces are modified over cosmological distances, one has to conclude that the Universe is filled with an hitherto completely unknown agent accelerating the expansion. A tiny, yet non-zero cosmological constant would apparently ‘do the trick’—yet it would raise more fundamental puzzles.

These heavenly signals are unequivocal in pointing to New Physics, yet leave wide open the nature of this New Physics.

Thus we can be assured that New Physics exists ‘somehow’ ‘somewhere’, and quite likely even ‘nearby’, namely around the TeV scale; above I have called the latter cpNP. The LHC programme and the Linear Collider project are justified—correctly—to conduct campaigns for cpNP. That is unlikely to shed light on the ssNP, though it might. Likewise I would not count on a comprehensive and detailed programme of heavy flavour studies to shed light on the ssNP behind the flavour puzzle of the SM. Yet the argument is reasonably turned around: such a programme will be essential to elucidate salient features of the cpNP by probing the latter’s flavour structure and having sensitivity to scales of order 10 TeV. One should keep in mind the following: one very popular example of cpNP is supersymmetry; yet it represents an organizing principle much more than even a class of theories. I find it unlikely we can infer all required lessons by studying only flavour diagonal transitions. Heavy flavour decays provide a powerful and complementary probe of cpNP. Their potential to reveal something about the ssNP is a welcome extra not required for justifying efforts in that direction.

Accordingly I see a dedicated heavy flavour programme as an essential complement to the studies pursued at the high energy frontier at the TeVatron, the LHC and, it is to be hoped, the ILC. I shall illustrate this assertion in the remainder of this lecture.

3.2 $\Delta S \neq 0$—the ‘established hero’

The chapter on $\Delta S \neq 0$ transitions is a most glorious one in the history of particle physics, as sketched in Table 5. We should note that all these features, which now are pillars of the SM, were New Physics at that time!
Table 5: On the history of $\Delta S \neq 0$ studies

<table>
<thead>
<tr>
<th>Observation</th>
<th>Lesson learned</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau - \theta$ puzzle</td>
<td>$P$ violation</td>
</tr>
<tr>
<td>Production rate $\gg$ decay rate</td>
<td>Concept of families</td>
</tr>
<tr>
<td>Suppression of flavour-changing neutral currents</td>
<td>GIM mechanism and existence of charm</td>
</tr>
<tr>
<td>$K_L \to \pi \pi$</td>
<td>$CP$ violation and existence of top</td>
</tr>
</tbody>
</table>

3.2.1 *Future ‘bread-and-butter’ topics*

Detailed studies of radiative decays like $K \to \pi \gamma \gamma$ and $K \to \pi \pi \gamma$ will allow deeper probes of chiral perturbation theory. The lessons thus obtained might lead to a better treatment of long distance dynamics’ impact on the $\Delta I = 1/2$ rule, $\Delta M_K$, $\epsilon_K$ and $\epsilon'$. 

3.2.2 *The ‘dark horse’*

The $T$-odd moment (see Section 1.2)

$$Pol_\perp(\mu) \equiv \frac{\langle \hat{s}(\mu) \cdot (\hat{p}(\mu) \times \hat{p}(\pi)) \rangle}{|\hat{p}(\mu) \times \hat{p}(\pi)|}$$ (236)

measured in $K^+ \to \mu^+ \nu \pi^0$ would

- represent genuine $T$ violation (as long as it exceeded the order $10^{-6}$ level) and
- constitute *prima facie* evidence for $CP$ violation in *scalar* dynamics.

3.2.3 *Heresy’*

The large $T$-odd correlation found in $K_L \to \pi^+ \pi^- e^+ e^-$ for the relative orientation of the $\pi^+ \pi^-$ and $e^+ e^-$ decay planes, see the discussion below in Section 3.3.2.1, is fully consistent with a $T$ violation as inferred from the $CP$ violation expressed through $\epsilon_K$—yet it does not prove it [87]. In an unabashedly contrived scenario —something theorists usually avoid at great pains—one could reconcile the data on $K_L \to \pi^+ \pi^- e^+ e^-$ with $T$ invariance without creating a conflict with known data. Yet the $CPT$ violation required in this scenario would have to surface through [87]

$$\frac{\Gamma(K^+ \to \pi^+ \pi^0) - \Gamma(K^- \to \pi^- \pi^0)}{\Gamma(K^+ \to \pi^+ \pi^0) + \Gamma(K^- \to \pi^- \pi^0)} > 10^{-3}. \tag{237}$$

3.2.4 *The ‘Second Trojan War’: $K \to \pi \nu \bar{\nu}$*

According to Greek Mythology the Trojan War described in Homer’s *Iliad* was actually the second war over Troy. In a similar vein I view the heroic campaign over $K^0 - \bar{K}^0$ oscillations—$\Delta M_K$, $\epsilon_K$ and $\epsilon'$—as a first one to be followed by a likewise epic struggle over the two ultra-rare modes $K^+ \to \pi^+ \nu \bar{\nu}$ and $K_L \to \pi^0 \nu \bar{\nu}$. This campaign has already been opened through the observation of the first through three events very roughly as expected within the SM. The second one, which requires $CP$ violation for its mere existence, so far remains unobserved at a level well above SM predictions. These reactions are like ‘standard candles’ for the SM; their rates are functions of $V(td)$ with a theoretical uncertainty of about 5% and 2%, respectively, which is mainly due to the uncertainty in the charm quark mass.

While their rates could be enhanced by New Physics greatly over their SM expectation, I personally find that somewhat unlikely for various reasons. Therefore I suggest one should aim for collecting ultimately about 1000 events of these modes to extract the value of $V(td)$ and/or identify likely signals of New Physics.
3.3 The ‘King Kong’ scenario for New Physics searches

This scenario can be formulated as follows: “One is unlikely to encounter King Kong; yet once it happens one will have no doubt that one has come across something quite out of the ordinary!”

What it means can be best illustrated with the historical precedent of $\Delta S \neq 0$ studies sketched above: the existence of New Physics can unequivocally be inferred if there is a qualitative conflict between data and expectation; i.e., if a theoretically ‘forbidden’ process is found to proceed nevertheless—like in $K_L \to \pi\pi$—or the discrepancy between expected and observed rates amounts to several orders of magnitude—like in $K_L \to \mu^+\mu^-$ or $\Delta M_K$. This does not mean that the effects are large or straightforward to discover—only that they are much larger than the truly minute SM effects.

History might repeat itself in the sense that future measurements might reveal such qualitative conflicts, where the case for the manifestation of New Physics is easily made. This does not mean that such measurements will be easy—far from it, as will become obvious.

I have already mentioned one potential candidate for revealing such a qualitative conflict, namely the muon transverse polarization in $K_{\mu3}$ decays.

3.3.1 Electric dipole moments

The energy shift of a system placed inside a weak electric field can be expressed through an expansion in terms of the components of that field $\vec{E}$:

$$\Delta \mathcal{E} = d_i E_i + d_{ij} E_i E_j + \mathcal{O}(E^3) . \quad (238)$$

The coefficients $d_i$ of the term linear in the electric field form a vector $\vec{d}$, called an electric dipole moment (EDM). For a non-degenerate system—it does not have to be elementary—one infers from symmetry considerations that this vector has to be proportional to that system’s spin:

$$\vec{d} \propto \vec{s} . \quad (239)$$

Yet, since

$$E_i \xrightarrow{T} E_i , \quad s_i \xrightarrow{T} -s_i \quad (240)$$

under time reversal $T$, a non-vanishing EDM constitutes $T$ violation.

No EDM has been observed yet; the upper bounds of the neutron and electron EDM read as follows [10]:

$$d_N < 5 \cdot 10^{-26} \text{ e cm} \quad \text{[from ultracold neutrons]} . \quad (241)$$

$$d_e < 1.5 \cdot 10^{-27} \text{ e cm} \quad \text{[from atomic EDM]} . \quad (242)$$

The experimental sensitivity achieved can be illustrated as follows: (i) A neutron EDM of $5 \cdot 10^{-26}$ e cm of an object with a radius $r_N \sim 10^{-13}$ cm scales to a displacement of about 7 $\mu m$ i.e., less than the width of human hair, for an object of the size of the Earth. (ii) Expressing the uncertainty in the measurement of the electron’s magnetic dipole moment—$\delta((g - 2)/2) \sim 10^{-11}$ in analogy to its EDM, one finds a sensitivity level of $\delta(F_2(0)/2m_e) \sim 2 \cdot 10^{-22}$ e cm compared to $d_e < 2 \cdot 10^{-26}$ e cm.

Despite the tremendous sensitivity reached, these numbers are still several orders of magnitude above what is expected in CKM theory:

$$d_N^{CKM} \leq 10^{-30} \text{ e cm} \quad (243)$$

$$d_e^{CKM} \leq 10^{-36} \text{ e cm} \quad (244)$$

where in $d_N^{CKM}$ I have ignored any contribution from the strong CP problem. These numbers are so tiny for reasons very specific to CKM theory, namely its chirality structure and the pattern in the quark
and lepton masses. Yet New Physics scenarios with right-handed currents, flavour-changing neutral currents, a non-minimal Higgs sector, heavy neutrinos etc. are likely to generate considerably larger numbers: $10^{-28} - 10^{-26}$ e cm represents a very possible range there quite irrespective of whether these new forces contribute to $\epsilon_K$ or not. This range appears to be within reach in the foreseeable future. There is a vibrant multiprong programme going on at several places. Such experiments while being of the `table top’ variety require tremendous efforts, persistence and ingenuity—yet the insights to be gained by finding a nonzero EDM somewhere are tremendous.

3.3.2 Charm decays

Charm dynamics is often viewed as physics with a great past—it was instrumental in driving the paradigm shift from quarks as mathematical entities to physical objects and in providing essential support for accepting QCD as the theory of the strong interactions—yet one without a future since the electroweak phenomenology for $\Delta C \neq 0$ transitions is decidedly on the ‘dull’ side: ‘known’ CKM parameters, slow $D^0 - \bar{D}^0$ oscillations, small CP asymmetries, and extremely rare loop driven decays.

Yet more thoughtful observers have realized that the very ‘dullness’ of the SM phenomenology for charm provides us with a dual opportunity, namely to

- probe our quantitative understanding of QCD’s nonperturbative dynamics thus calibrating our theoretical tools for $B$ decays and
- perform almost ‘zero-background’ searches for New Physics.

However, the latter statement of ‘zero-background’ has to be updated carefully since experiments over the last ten years have bounded the oscillation parameters $x_D, y_D$ to fall below very few per cent and direct CP asymmetries below several per cent. While New Physics signals can still exceed SM predictions on CP asymmetries by orders of magnitude, they might not be large in absolute terms, as specified later [88].

**One should take note that charm is the only up-type quark allowing the full range of probes for New Physics, including flavour-changing neutral currents:** while top quarks do not hadronize [52], in the $u$ quark sector you cannot have $\pi^0 - \pi^0$ oscillations, and many CP asymmetries are already ruled out by CPT invariance. My basic contention is the following: Charm transitions are a unique portal for obtaining a novel access to flavour dynamics with the experimental situation being a priori favourable (except for the lack of Cabibbo suppression)!

I shall sketch such searches for New Physics in the context of $D^0 - \bar{D}^0$ oscillations and CP violation.

1. Like for $K^0$ and $B^0$ mesons the oscillations of $D^0$ mesons represent a subtle quantum mechanical phenomenon of practical importance: it provides a probe for New Physics, albeit an ambiguous one, and constitutes an important ingredient for CP asymmetries arising in $D^0$ decays due to New Physics.

In qualitative analogy to the $K^0$ and $B^0$ cases these phenomena can be characterized by two quantities, namely $x_D = \frac{\Delta M_{D}}{1_{D}}$ and $y_D = \frac{\Delta \Gamma_{D}}{2\Gamma_{D}}$. Oscillations are slowed down in the SM due to GIM suppression and $SU(3)_L$ symmetry. Comparing a conservative SM bound with the present data

$$x_D(SM), y_D(SM) < \mathcal{O}(0.01) \quad vs. \quad x_D|_{exp} < 0.03, \quad y_D|_{exp} = 0.01 \pm 0.005 \quad (245)$$

we conclude that the search has just now begun. There exists a considerable literature—yet typically with several *ad hoc* assumptions concerning the nonperturbative dynamics. It is widely understood that the usual quark box diagram is utterly irrelevant due to its untypically severe GIM suppression $(m_{u}/m_{c})^4$. A systematic analysis based on an OPE treatment has been given
in Ref. [89] in terms of powers of $1/m_c$ and $m_s$. Contributions from higher-dimensional operators with a much softer GIM reduction of $(m_s/\mu_{had})^2$ (even $m_s/\mu_{had}$ terms could arise) due to ‘condensate’ terms in the OPE yield

$$x_D(SM)|_{OPE}, \quad y_D(SM)|_{OPE} \sim \mathcal{O}(10^{-3}).$$ (246)

Reference [90] finds very similar numbers, albeit in a quite different approach. While one predicts similar numbers for $x_D(SM)$ and $y_D(SM)$, one should keep in mind that they arise in very different dynamical environments. $\Delta M_D$ is generated from off-shell intermediate states and thus is sensitive to New Physics, which could produce $x_D \sim \mathcal{O}(10^{-2})$. $\Delta T_D$ on the other hand is shaped by on-shell intermediate states; while it is hardly sensitive to New Physics, it involves much less averaging or ‘smearing’ than $\Delta M_D$ making it thus much more vulnerable to violations of quark–hadron duality. Observing $y_D \sim 10^{-3}$ together with $x_D \sim 0.01$ would provide intriguing, though not conclusive evidence for New Physics, while $y_D \sim 0.01 \sim x_D$ would pose a true conundrum for its interpretation.

2. Since the baryon number of the Universe implies the existence of New Physics in $\text{CP}$-violating dynamics, it would be unwise not to undertake dedicated searches for $\text{CP}$ asymmetries in charm decays, where the ‘background’ from known physics is small: within the SM the effective weak phase is highly diluted, namely $\sim \mathcal{O}(\lambda^4)$, and it can arise only in singly Cabibbo suppressed transitions, where one expects them to reach the 0.1% level; significantly larger values would signal New Physics. Any asymmetry in Cabibbo allowed or doubly suppressed channels requires the intervention of New Physics—except for $D^\pm \to K_S\pi^\pm$ [65], where the $\text{CP}$ impurity in $K_S$ induces an asymmetry of $3.3 \cdot 10^{-3}$. Several facts actually favour such searches: strong phase shifts required for direct $\text{CP}$ violation to emerge in partial widths are in general large as are the branching ratios into relevant modes; finally $\text{CP}$ asymmetries can be linear in New Physics amplitudes thus enhancing sensitivity to the latter. As said above, the benchmark scale for $\text{KM}$ asymmetries in singly Cabibbo suppressed partial widths is 0.1%. This does not exclude the possibility that $\text{KM}$ dynamics might exceptionally generate an asymmetry as ‘large’ as 1% in some special cases. It is therefore essential to analyse a host of channels.

Decays to final states of more than two pseudoscalar or one pseudoscalar and one vector meson contain more dynamical information than given by their widths; their distributions as described by Dalitz plots or $\text{T-odd}$ moments can exhibit $\text{CP}$ asymmetries that can be considerably larger than those for the width. Final-state interactions, while not necessary for the emergence of such effects, can fake a signal; yet that can be disentangled by comparing $\text{T-odd}$ moments for $\text{CP}$ conjugate modes. I view this as a very promising avenue, where we still have to develop the most effective analysis tools for small asymmetries.

$\text{CP}$ violation involving $D^0 - \bar{D}^0$ oscillations can be searched for in final states common to $D^0$ and $\bar{D}^0$ decays like $\text{CP}$ eigenstates—$D^0 \to K_S\phi, K^+K^-, \pi^+\pi^-$—or doubly Cabibbo suppressed modes—$D^0 \to K^+\pi^-$. The $\text{CP}$ asymmetry is controlled by $\sin \Delta M_D t \cdot \text{Im}(q/p)\bar{\rho}(D \to f)$; within the SM both factors are small, namely $\sim \mathcal{O}(10^{-3})$, making such an asymmetry unobservably tiny—unless there is New Physics! One should note that this observable is linear in $x_D$ rather than quadratic as for $\text{CP}$ insensitive quantities. $D^0 - \bar{D}^0$ oscillations, $\text{CP}$ violation and New Physics might thus be discovered simultaneously in a transition.

One wants to reach the level at which $\text{SM}$ effects are likely to emerge, namely down to time-dependent $\text{CP}$ asymmetries in $D^0 \to K_S\phi, K^+K^-, \pi^+\pi^- [K^+\pi^-]$ down to $10^{-5} [10^{-4}]$ and direct $\text{CP}$ asymmetries in partial widths and Dalitz plots down to $10^{-5}$.

### 3.3.2.1 $\text{CP}$ asymmetries in final-state distributions

So far $\text{CP}$ violation has surfaced in time-integrated or time-dependent partial widths with one notable exception. A large $\text{T-odd}$ moment was found in the rare $K_L$ mode—$\text{BR}(K_L \to \pi^+\pi^-e^+e^-) = (3.32 \pm
$0.14 \pm 0.28) \cdot 10^{-7}$: with $\phi$ defined as the angle between the planes spanned by the two pions and the two leptons in the $K_L$ restframe:

$$\phi \equiv \angle(\vec{n}_l, \vec{n}_\pi)$$

$$\vec{n}_l = \vec{p}_{e^+} \times \vec{p}_{e^-} / |\vec{p}_{e^+} \times \vec{p}_{e^-}|, \quad \vec{n}_\pi = \vec{p}_{\pi^+} \times \vec{p}_{\pi^-} / |\vec{p}_{\pi^+} \times \vec{p}_{\pi^-}|$$

(247)

one analyses the decay rate as a function of $\phi$:

$$\frac{d\Gamma}{d\phi} = \Gamma_1 \cos^2 \phi + \Gamma_2 \sin^2 \phi + \Gamma_3 \cos \phi \sin \phi.$$ 

(248)

Since

$$\cos \phi \sin \phi = (\vec{n}_l \times \vec{n}_\pi) \cdot (\vec{p}_{\pi^+} + \vec{p}_{\pi^-})(\vec{n}_l \cdot \vec{n}_\pi)/|\vec{p}_{\pi^+} + \vec{p}_{\pi^-}|$$

(249)

one notes that

$$\cos \phi \sin \phi \xrightarrow{\text{CP}} - \cos \phi \sin \phi$$

(250)

under both $T$ and $CP$ transformations; i.e., the observable $\Gamma_3$ represents a $T$ - and $CP$ -odd correlation. It can be projected out by comparing the $\phi$ distribution integrated over two quadrants:

$$A = \frac{\int_0^{\pi/2} d\phi \frac{d\Gamma}{d\phi} - \int_{\pi/2}^{\pi} d\phi \frac{d\Gamma}{d\phi}}{\int_0^{\pi} d\phi \frac{d\Gamma}{d\phi}} = \frac{2\Gamma_3}{\pi(\Gamma_1 + \Gamma_2)}.$$ 

(251)

It was first measured by KTeV and then confirmed by NA48 [10]:

$$A = (13.7 \pm 1.5)\%.$$ 

(252)

$A \neq 0$ is induced by $\epsilon_K$, the $CP$ violation in the $K^0 - \bar{K}^0$ mass matrix, leading to the prediction [91]

$$A = (14.3 \pm 1.3)\%.$$ 

(253)

The observed value for the $T$-odd moment $A$ is fully consistent with $T$ violation. Yet $A \neq 0$ by itself does not establish $T$ violation [87].

One should note that this sizeable forward–backward asymmetry is driven by the tiny quantity $|\eta_{+-}| \approx 0.0023$, which can be understood. For $K_L \rightarrow \pi^+ \pi^- e^+ e^-$ is driven by the two sub-processes

$$K_L \xrightarrow{CP \& \Delta S = 1} \pi^+ \pi^- \xrightarrow{E1} \pi^+ \pi^- \gamma^* \rightarrow \pi^+ \pi^- e^+ e^-$$

$$K_L \xrightarrow{M1 \& \Delta S = 1} \pi^+ \pi^- \gamma^* \rightarrow \pi^+ \pi^- e^+ e^-,$$

(254)

where the first reaction is suppressed, since it requires $CP$ violation in $K_L \rightarrow 2\pi$, and the second one, since it involves an $M1$ transition. Those two $a priori$ very different suppression mechanisms happen to yield comparable amplitudes, which thus generate sizeable interference. The price one pays is the small branching ratio.

$D$ decays can be treated in an analogous way. Consider the Cabibbo suppressed channel

$$(-) \mathrm{D} \rightarrow K\bar{K}\pi^+\pi^-$$

(256)

and define by $\phi$ now the angle between the $K\bar{K}$ and $\pi^+\pi^-$ planes. Then one has

$$\frac{d\Gamma}{d\phi}(D \rightarrow K\bar{K}\pi^+\pi^-) = \Gamma_1 \cos^2 \phi + \Gamma_2 \sin^2 \phi + \Gamma_3 \cos \phi \sin \phi$$

(257)

$$\frac{d\Gamma}{d\phi}(\bar{D} \rightarrow K\bar{K}\pi^+\pi^-) = \Gamma_1 \cos^2 \phi + \Gamma_2 \sin^2 \phi + \Gamma_3 \cos \phi \sin \phi.$$ 

(258)

This mode can exhibit direct $CP$ violation even within the SM.
As before, the partial width for $D[\bar{D}] \rightarrow K\bar{K}\pi^+\pi^-$ is given by $\Gamma_{1,2}[\bar{\Gamma}_{1,2}]$; $\Gamma_1 \neq \bar{\Gamma}_1$ or $\Gamma_2 \neq \bar{\Gamma}_2$ represents direct CP violation in the partial width. $\Gamma_3$ and $\Gamma_{\bar{3}}$ constitute T-odd correlations. By themselves they do not necessarily indicate CP violation, since they can be induced by strong final-state interactions. However,

$$\Gamma_3 \neq \bar{\Gamma}_3 \implies \text{CP violation!} \quad (259)$$

It is quite possible or even likely that a difference in $\Gamma_3$ vs. $\bar{\Gamma}_3$ is significantly larger than in $\Gamma_1$ vs. $\bar{\Gamma}_1$ or $\Gamma_2$ vs. $\bar{\Gamma}_2$. Furthermore one can expect that differences in detection efficiencies can be handled by comparing $\Gamma_3$ with $\Gamma_{1,2}$ and $\bar{\Gamma}_3$ with $\Gamma_{1,2}$. A pioneering search for such an effect has been undertaken by FOCUS [92].

### 3.3.3 CP violation in the lepton sector

I find the conjecture that baryogenesis is a secondary phenomenon driven by primary leptogenesis a most intriguing and attractive one also for philosophical reasons\(^{29}\). Yet then it becomes mandatory to search for CP violation in the lepton sector in a most dedicated fashion.

In Section 3.3.1 I have sketched the importance of measuring electric dipole moments as accurately as possible. The electron’s EDM is a most sensitive probe of CP violation in leptodynamics. Comparing the present experimental and CKM upper bounds, respectively

$$d_e^{exp} \leq 1.5 \cdot 10^{-27} \text{ e cm vs. } d_e^{CKM} \leq 10^{-36} \text{ e cm} \quad (260)$$

we see there is a wide window of several orders of magnitude, where New Physics could surface in an unambiguous way. This observation is reinforced by the realization that New Physics scenarios can naturally generate $d_e > 10^{-28}$ e cm, while of only secondary significance in $\epsilon_K$, $\epsilon'$ and $\sin 2\phi_t$.

The importance that at least part of the HEP community attributes to finding CP violation in leptodynamics is best demonstrated by the efforts contemplated for observing CP asymmetries in neutrino oscillations. Clearly hadronization will be the least of the concerns, yet one has to disentangle genuine CP violation from matter enhancements, since the neutrino oscillations can be studied only in a matter, not an antimatter environment. Our colleagues involved in such endeavours will rue their previous complaints about hadronization and remember the wisdom of an ancient Greek saying:

“When the dogs want to really harm you, they fulfil your wishes.”

### 3.3.4 The decays of $\tau$ leptons—the next ‘hero candidate’

Like charm hadrons the $\tau$ lepton is often viewed as a system with a great past, but hardly a future. Again I think this is a very misguided view and I shall illustrate it with two examples.

Searching for $\tau^\pm \rightarrow \mu^\pm \mu^+\mu^-$ (and its variants)—processes forbidden in the SM—is particularly intriguing, since it involves only ‘down-type’ leptons of the second and third family and is thus the complete analogy of the quark lepton process $b \rightarrow s\bar{s}s$, driving $B_s \rightarrow \phi K_S$, which has recently attracted such strong attention. Following this analogy literally one guestimates $\text{BR}(\tau \rightarrow 3\mu) \sim 10^{-8}$ to be compared with the present bound from Belle

$$\text{BR}(\tau \rightarrow 3\mu) \leq 2 \cdot 10^{-7} \quad (261)$$

It would be very interesting to know what the $\tau$ production rate at the hadronic colliders is and whether they could be competitive with or even superior to the $B$ factories in such a search.

---

\(^{29}\)For it would complete what is usually called the Copernican Revolution [93]: first our Earth was removed from the centre of the Universe, then in due course our Sun, our Milky Way and local cluster; few scientists believe life exists only on our Earth. Realizing that the stuff we are mostly made out of—protons and neutrons—is just a cosmic ‘afterthought’ fits this pattern, which culminates in the dawning realization that even our Universe is just one among innumerable others, albeit a most unusual one.
In my judgement $\tau$ decays—together with electric dipole moments for leptons and possibly $\nu$ oscillations referred to above—provide the best stage to search for manifestations of CP-breaking lepton-dynamics.

The most promising channels for exhibiting CP asymmetries are $\tau \rightarrow \nu K\pi$, since due to the heaviness of the lepton and quark flavours they are most sensitive to nonminimal Higgs dynamics, and they can show asymmetries also in the final-state distributions rather than integrated rates [94].

There is also a unique opportunity in $e^+e^- \rightarrow \tau^+\tau^-$: since the $\tau$ pair is produced with its spins aligned, the decay of one $\tau$ can ‘tag’ the spin of the other $\tau$. That is, one can probe spin-dependent CP asymmetries with unpolarized beams. This provides higher sensitivity and more control over systematic uncertainties.

I feel these features are not sufficiently appreciated even by proponents of Super-B factories. It has been recently pointed out [95] that based on known physics one can actually predict a CP asymmetry:

$$\frac{\Gamma(\tau^+ \rightarrow K_S\pi^+\nu) - \Gamma(\tau^- \rightarrow K_S\pi^-\nu)}{\Gamma(\tau^+ \rightarrow K_S\pi^+\nu) + \Gamma(\tau^- \rightarrow K_S\pi^-\nu)} = (3.27 \pm 0.12) \times 10^{-3}$$

(262)

ing owing to $K_S$’s preference for antimatter.

3.4 Future studies of $B_{u,d}$ decays

The successes of CKM theory to describe flavour dynamics do not tell us at all that New Physics does not affect $B$ decays; the message is that typically we cannot count on a numerically massive impact there. Shifting an asymmetry by, say, ten percentage points—for example from 40% to 50%—might already be on the large side. Thus we have to aim for uncertainties that do not exceed a few per cent.

The discussion given in Lecture II shows that an integrated luminosity of 1 ab$^{-1}$ at the $B$ factories will fall short of such a goal for $B_d \rightarrow \pi\pi, B^\pm \rightarrow D^{\text{neut}}K^\pm$ and in particular also for the modes driven by $b \rightarrow sq\bar{q}$. Even ten times that statistics would not suffice in view of the ‘big picture’, i.e., when one includes other rare transitions. Of course we are in the very fortunate situation that one of the LHC experiments, namely LHCb, is dedicated to undertaking precise measurements of the weak decays of beauty hadrons. Thus we can expect a stream of high quality data to be forthcoming over the next several years. I shall briefly address different classes of rare decays with different motivations and requirements.

3.4.1 Radiative $B$ decays

3.4.1.1 $B \rightarrow \gamma X$

As already mentioned in Lecture II

- the branching ratio for $B \rightarrow \gamma X_s$ has been measured with good accuracy and in agreement with the SM prediction;
- the photon energy spectrum has been determined down to $E_\gamma = 1.9$ GeV or even 1.8 GeV; its moments have provided important information on the heavy quark parameters, in particular the $b$ quark mass $m_b$.

There is another more subtle observable, for which the SM makes a rather accurate prediction, namely the photon polarization: the SM electroweak penguin operator produces mostly left-handed photons. New Physics on the other hand can generate right-handed photons as well. They would hardly be noticeable in the total rate: since left- and right-handed photons cannot interfere, the rate would be quadratic in the New Physics amplitude. The gluonic counterpart to such a New Physics $b \rightarrow s\gamma\bar{q}$ could, however, contribute linearly in amplitude to the CP asymmetry in $B_d \rightarrow \phi K_S$ and other $b \rightarrow sq\bar{q}$ modes and thus become significant there.
Rather than measure the photon polarization, which seems hardly feasible, one can infer it from measuring angular correlations in $B \to \gamma K^{*0} \to \gamma(K\pi\pi)$ modes [96].

It has been suggested to distinguish $B \to \gamma X_d$ against $B \to \gamma X_s$ to extract $V(td)/V(ts)$ or to probe for New Physics using a value for $V(td)/V(ts)$ extracted from $\Delta M_{B_d}/\Delta M_{B_s}$ or the overall CKM fit. This does not seem to be a hopeless undertaking—at a Super-B factory.

3.4.1.2 $B \to l^+ l^- X$

We are just at the beginning of studying $B \to l^+ l^- X$, and it has to be pursued in a dedicated and comprehensive manner for the following reasons:

- With the final state being more complex than for $B \to \gamma X$, it is described by a larger number of observables: rates, spectra of the lepton pair masses and the lepton energies, their forward–backward asymmetries and CP asymmetries.
- These observables provide independent information, since there is a larger number of effective transition operators than for $B \to \gamma X$. By the same token there is a much wider window to find New Physics and even diagnose its salient features.
- It will take the statistics of a Super-B factory to mine this wealth of information on New Physics.
- Essential insights can be gained also by analysing the exclusive channel $B \to l^+ l^- K^*$ at hadronic colliders like the LHC, in particular the position of the zero in the lepton forward–backward asymmetry. For the latter appears to be fairly insensitive to hadronization effects in this exclusive mode [97]. It will be important to analyse quantitatively down to which level of accuracy this feature persists.

3.4.2 Semileptonic decays involving $\tau$ leptons

There are some relatively rare $B$ decays that could conceivably reveal New Physics, although they proceed already on the tree level. One well known example is $B^+ \to \tau \nu$ that is sensitive to charged Higgs fields. This applies also to semileptonic $B$ decays. As described in Section 2.4, the Heavy Quark Expansion (HQE) has provided a sturdy and accurate description of $B \to l\nu X_e$ that allowed one to extract $|V(cb)|$ with less than 2% uncertainty. With it and other heavy quark parameters determined with considerable accuracy, one can predict $\Gamma(B \to \tau \nu X)\bigg/\Gamma(B \to l\nu X)$ within the SM and compare it with the data. A discrepancy can be attributed to New Physics, presumably in the form of a charged Higgs field. Measuring also its hadronic mass moments can serve as a valuable cross-check. Such studies will probably require the statistics of a Super-B factory.

This is true also for studying the exclusive channel $B \to \tau \nu D$. As pointed out in Ref. [98], one could find that the ratio $\Gamma(B \to \tau \nu D)/\Gamma(B \to \mu \nu D)$ deviates from its SM value due to the exchange of a charged Higgs boson with a mass of even several hundred GeV. This is the case in particular for ‘large $tg\beta$ scenarios’ of two-Higgs-doublet models. There is a complication, though. Contrary to the suggestion in the literature the hadronic form factors do not drop out from this ratio. One should keep in mind that (i) the contribution from the second form factor $f_-$, which is proportional to the square of the lepton mass, cannot be ignored for $B \to \tau \nu D$ and (ii) the form factors are not taken at the same momentum transfer in the two modes.

These complications can be overcome by Uraltsev’s BPS approximation [81]. Relying on it one can extract $|V(cb)|$ from $B \to e/\mu \nu D$ and compare it with the ‘true’ value obtained from $\Gamma_{SL}(B)$. If this comparison is successful and our theoretical control over $B \to l\nu D$ thus validated, one can apply the BPS approximation to $B \to \tau \nu D$. Since, as mentioned above, the second form factor $f_-$ can be measured there, one has another cross-check.
3.5 $B_s$ Decays—an independent chapter in Nature’s book

When the programme for the $B$ factories was planned, it was thought that studying $B_s$ transitions would be required to construct the CKM triangle, namely to determine one of its sides and the angle $\phi_3$. As discussed above a powerful method has been developed to extract $\phi_3$ from $B^\pm \to D^{\text{neut}}K^\pm$ and a meaningful value for $|V(td)/V(ts)|$ has been inferred from the measured value of $\Delta M_{B_d}/\Delta M_{B_s}$. None of this, however, reduces the importance of a future comprehensive programme to study $B_s$ decays—on the contrary! With the basic CKM parameters fixed or to be fixed in $B_{u,d}$ decays, $B_s$ transitions can be harnessed as powerful probes for New Physics and its features.

In this context it is essential to think ‘outside the box’—pun intended. The point here is that several relations that hold in the SM (as implemented through quark box and other loop diagrams) are unlikely to extend beyond minimal extensions of the SM. In that sense $B_{u,d}$ and $B_s$ decays constitute two different and complementary chapters in Nature’s book on fundamental dynamics.

3.5.1 CP violation in non-leptonic $B_s$ decays

One class of nonleptonic $B_s$ transitions does not follow the paradigm of large CP violation in $B$ decays [16]:

$$A_{CP}(B_s(t) \to [\psi\phi]_{l=0}/\psi\eta) = \sin 2\phi(B_s) \sin \Delta M_{B_s} t$$

$$\sin 2\phi(B_s) = \text{Im} \left[ \frac{(V^{\ast}(tb)V(ts))^2}{(V^{\ast}(cb)V(cs))^2}\right] \approx 2\lambda^2 \eta \sim 0.02.$$  \hspace{1cm} (263)

This is easily understood: on the leading CKM level only quarks of the second and third families contribute to $B_s$ oscillations and $B_s \to \psi\phi$ or $\psi\eta$; therefore on that level there can be no CP violation making the CP asymmetry Cabibbo suppressed. Yet New Physics of various ilks can quite conceivably generate $\sin 2\phi(B_s) \sim$ several $\times 10\%$.

Analysing the decay rate evolution in proper time of

$$B_s(t) \to \phi\phi$$  \hspace{1cm} (264)

with its direct as well as indirect CP violation is much more than a repetition of the $B_d(t) \to \phi K_S$ saga:

- $\mathcal{M}_{12}(B_s)$ and $\mathcal{M}_{12}(B_d)$—the off-diagonal elements in the mass matrices for $B_s$ and $B_d$ mesons, respectively—provide in principle independent pieces of information on $\Delta B = 2$ dynamics.

- While the final-state $\phi K_S$ is described by a single partial wave, namely $l = 1$, there are three partial waves in $\phi\phi$, namely $l = 0, 1, 2$. Disentangling the three partial rates and their CP asymmetries—or at least separating $l = \text{even and odd contributions}$—provides a new diagnostics about the underlying dynamics.

3.5.2 Leptonic, semileptonic and radiative modes

The decays into a lepton pair and to ‘wrong-sign’ leptons should be studied also for $B_d$ mesons; however, here I discuss only $B_s$ decays, where one can expect more dramatic effects.

- The mode $B_s \to \mu^+\mu^-$ is necessarily very rare since it suffers from helicity suppression $\propto (m(\mu)/M(B_s))^2$ and ‘wave function suppression’ $\propto (f_B/M(B_s))^2$, which reflects the practically zero range of the weak interactions. Within the SM one predicts

$$\text{BR}(B_s \to \mu^+\mu^-)|_{SM} \sim 3 \cdot 10^{-9}.$$  \hspace{1cm} (265)

These tiny factors can be partially compensated in some large $\tan \beta$ SUSY scenarios, where an enhancement factor of $\tan^6 \beta$ arises [99], which could produce a rate at the experimental bound of $10^{-7}$.
● Owing to the rapid $B_s$ oscillations those mesons have a practically equal probability to decay into ‘wrong’ and ‘right’ sign leptons. One can then search for an asymmetry in the wrong sign rate for mesons that initially were $B_s$ and $\bar{B}_s$:

$$a_{SL}(B_s) = \frac{\Gamma(\bar{B}_s \to l^+ X) - \Gamma(B_s \to l^- X)}{\Gamma(\bar{B}_s \to l^+ X) + \Gamma(B_s \to l^- X)}.$$  

This observable is necessarily small; among other things it is proportional to $\frac{\Delta \Gamma_{B_s}}{\Delta M_{B_s}} \ll 1$. The theoretical CKM predictions are not very precise, yet certainly tiny [23]:

$$a_{SL}(B_s) \sim 2 \cdot 10^{-5}, \quad a_{SL}(B_d) \sim 4 \cdot 10^{-4};$$  

$a_{SL}(B_s)$ suffers a suppression quite specific to CKM dynamics; analogous to $B_s \to \psi \phi$ quarks of only the second and third family participate on the leading CKM level, which therefore cannot exhibit CP violation. Yet again, New Physics can enhance $a_{SL}(B_s)$, this time by two orders of magnitude up to the 1% level.

● As already emphasized $B_s \to \gamma X$ and $B_s \to l^+l^- X$ should be studied in a comprehensive manner.

3.6 Instead of a summary: on the future HEP landscape—a call to well-reasoned action

Originally I had intended to name this section ‘A call to arms’. Yet recent events have reminded us that when the drums of war sound, reason all too often is left behind.

The situation of the SM, as it enters the third millenium, can be characterized through several statements:

1. There is a new dimension due to the findings on $B$ decays: one has established the first CP asymmetries outside the $K^0 - \bar{K}^0$ complex in four $B_d$ modes—as predicted qualitatively as well as quantitatively by CKM dynamics:

$$B_d(t) \to \psi K_S;$$  

$$B_d(t) \to \pi^+ \pi^-;$$  

$$B_d(t) \to K^+ \pi^-;$$  

$$B_d(t) \to \eta' K_S.$$  

Taken together with the other established signals—$K^0(t) \to 2\pi$ and $|\eta_{++}| \neq |\eta_{00}|$—we see that in all these cases except for $B_d \to K^+ \pi^-$ the intervention of meson–antimeson oscillations was instrumental in CP violation becoming observable. This is why I write $B_d[K^0](t) \to f$. For practical reasons this holds even for $|\eta_{+\pm}| \neq |\eta_{00}|$.

For the first time strong evidence has emerged for CP violation in the decays of a charged state, namely in $B^\pm \to K^\pm \rho^0$.

The SM’s success here can be stated more succinctly as follows:

- From a tiny signal of $|\eta_{+\pm}| \simeq 0.0023$ one successfully infers CP asymmetries in $B$ decays two orders of magnitude larger, namely $2\phi_1 \simeq 0.7$ in $B_d(t) \to \psi K_S$.

- From the measured values of two CP insensitive quantities—$|V_{ub}/V_{cb}|$ in semileptonic $B$ decays and $|V_{td}/V_{ts}|$ in $B^0 - \bar{B}^0$ oscillations—one deduces the existence of CP violation in $K_L \to 2\pi$ and $B_d(t) \to \psi K_S$ even in quantitative agreement with the data.

We know now that CKM dynamics provides at least the lion’s share in the observed CP asymmetries. The CKM description thus has become a tested theory. Rather then searching for alternatives to CKM dynamics we hunt for corrections to it.
None of these novel successes of the SM invalidate the theoretical arguments for it being incomplete. There is also clean evidence of mostly heavenly origin for New Physics, namely

- neutrino oscillations,
- dark matter,
- presumably dark energy,
- probably the baryon number of our Universe, and
- possibly the Strong CP Problem.

Flavour dynamics has become even more intriguing owing to the emergence of neutrino oscillations. We do not understand the structure of the CKM matrix in any profound way—nor the PMNS matrix, its leptonic counterpart. Presumably we do understand why they look different, since only neutrinos can possess Majorana masses, which can give rise to the ‘see-saw’ mechanism. Sometimes it is thought that the existence of two puzzles makes their resolution harder. I feel the opposite way: having a larger set of observables allows us to direct more questions to Nature, if we are sufficiently persistent, and learn from her answers.

The next ‘Grand Challenge’ after studying the dynamics behind the electroweak phase transition is to find CP violation in the lepton sector—anywhere.

While the quantization of electric charge is an essential ingredient of the SM, it does not offer any understanding of it. It would naturally be explained through Grand Unification at very high energy scales. I refer to it as the ‘guaranteed New Physics’, see Section 3.1.1.

The SM’s success in describing flavour transitions is not matched by a deeper understanding of the flavour structure, namely the patterns in the fermion masses and CKM parameters. For those do not appear to be of an accidental nature. I have referred to the dynamics generating the flavour structure as the ‘strongly suggested’ New Physics (ssNP), see Section 3.1.1.

Discovering the cpNP that drives the electroweak phase transition has been the justification for the LHC programme, which will come online soon. Personally I am very partisan to the idea that the cpNP will be of the SUSY type. Yet SUSY is an organizing principle rather than a class of theories, let alone a theory. We are actually quite ignorant about how to implement the one empirical feature of SUSY that has been established beyond any doubt, namely that it is broken.

The LHC is likely, I believe, to uncover the cpNP, and I have not given up hope that the TeVatron will catch the first glimpses of it. Yet the LHC and a fortiori the TeVatron are primarily discovery machines. The ILC project is motivated as a more surgical probe to map out the salient features of that cpNP.

This cpNP is unlikely to shed light on the ssNP behind the flavour puzzle of the SM, although one should not rule out such a most fortunate development. On the other hand New Physics even at the ~ 10–100 TeV scale could well affect flavour transitions significantly through virtual effects. A comprehensive and dedicated programme of heavy flavour studies might actually elucidate salient features of the cpNP that could not be probed in any other way. Such a programme is thus complementary to the one pursued at the TeVatron, the LHC, and, it is to be hoped, at the ILC and—I firmly believe— actually a necessity rather than a luxury to identify the cpNP.

To put it in more general terms: Heavy flavour studies

- are of fundamental importance,
- many of its lessons cannot be obtained any other way, and
- they cannot become obsolete.
That is, no matter what studies of high $p_{\perp}$ physics at the LHC and ILC will or will not show—comprehensive and detailed studies of flavour dynamics will remain crucial in our efforts to reveal Nature’s Grand Design.

10. Yet a note of caution has to be expressed as well. Crucial manifestations of New Physics in flavour dynamics are likely to be subtle. Thus we have to succeed in acquiring data as well as interpreting them with high precision. Obviously this represents a stiff challenge—however, one that I believe we can meet, if we prepare ourselves properly as I have exemplified in Section 2.4.1.

One of three possible scenarios will emerge in the next several years.

1. The optimal scenario: New Physics has been observed in ‘high $p_{\perp}$ physics’, i.e., through the production of new quanta at the TeVatron and/or LHC. Then it is imperative to study the impact of such New Physics on flavour dynamics; even if it should turn out to have none, this is an important piece of information, no matter how frustrating it would be to my experimental colleagues. Knowing the typical mass scale of that New Physics from collider data will be of great help to estimate its impact on heavy flavour transitions.

2. The intriguing scenario: Deviations from the SM have been established in heavy flavour decays—like the $B \to \phi K_S$ CP asymmetry or an excess in $\Gamma(K \to \pi\nu\bar{\nu})$—without a clear signal for New Physics in high $p_{\perp}$ physics. A variant of this scenario has already emerged through the observations of neutrino oscillations.

3. The frustrating scenario: No deviation from SM predictions has been identified.

I am optimistic it will be the ‘optimal’ scenario, quite possibly with some elements of the ‘intriguing’ one. Of course one cannot rule out the ‘frustrating’ scenario; yet we should not treat it as a case for defeatism: a possible failure to identify New Physics in future experiments at the hadronic colliders (or the $B$ factories) does not—in my judgement—invalidate the persuasiveness of the theoretical arguments and experimental evidence pointing to the incompleteness of the SM. It ‘merely’ means we have to increase the sensitivity of our probes. I firmly believe a Super-flavour factory with a luminosity of order $10^{36}$ cm$^{-2}$ s$^{-1}$ or more for the study of beauty, charm and $\tau$ decays has to be an integral part of our future efforts towards deciphering Nature’s basic code. For a handful of even perfectly measured transitions will not be sufficient for the task at hand—a comprehensive body of accurate data will be essential. Likewise we need a new round of experiments that can measure the rates for $K \to \pi\nu\bar{\nu}$ accurately with sample sizes $\sim O(10^3)$ and mount another serious effort to probe the muon transverse polarization in $K_{\mu3}$ decays.

I shall finish with a poem I learned from T.D. Lee a number of years ago. It was written by A.A. Milne, who is best known as the author of Winnie-the-Pooh (1926):

Wind on the Hill

No one can tell me
Nobody knows
Where the wind comes from,
Where the wind goes.

But if I stopped holding
The string of my kite,
It would blow with the wind
For a day and a night.
And then when I found it,
Wherever it blew,
I should know that the wind
Had been going there, too.

So then I could tell them
Where the wind goes ...
But where the wind comes from
Nobody knows.

One message from the poem is clear: we have to let our ‘kite’ respond to the wind, i.e., we have to perform experiments. Yet the second message ‘... Nobody knows.’ is overly agnostic: Indeed experiments by themselves will not provide us with all these answers. It means one will still need ‘us’, the theorists, to figure out ‘where the wind comes from’.

In any case, we are at the beginning of an exciting adventure—and we are most privileged to participate.

Acknowledgments: I truly enjoyed the beautiful setting and warm atmosphere of the school, the kind help from Danielle Métral and Egil Lillestøl and the seemingly effortless, yet firm guidance from the ‘Boss’ Tord Ekelöf. I benefited from the discussions with the students about both physics and World Cup soccer, and I am grateful to them for their kindness in including me in their soccer teams. This work was supported by the NSF under the grant number PHY-0355098.

References
[12] The most general time evolution for the $P^0 - P^0$ complex, including its decays, is given by an infinite-component vector in Hilbert space reading for $P^0 = K^0$ $\Psi(t) = a(t)|K^0\rangle + b(t)|\bar{K}^0\rangle + c(t)|2\pi\rangle + d(t)|3\pi\rangle + e(t)|\pi l\nu_l\rangle + ...$, which is the solution of an infinite-component Schrödinger equation. We do not know how to solve the latter, which describes a strong interaction problem. Fortunately we do not have to and can solve Eq. (50) instead, since for our purposes we restrict ourselves to times much longer than typical strong interaction times as appropriate for the Weisskopf–Wigner approximation: V. F. Weisskopf and E. P. Wigner, Z. Phys. 63 (1930) 54; ibid. 65 (1930) 18.
[27] For a recent review, see, for example, A. Buras, ‘Flavour physics and CP violation’, hep-ph/0505175.
[73] CDF note 7867.


[93] The usual tale that the Dark Ages of the Middle Ages were overcome by the Copernican Revolution being born like the goddess Athena jumping out of the head of her father Zeus fully developed and in full armour is unfair to the Middle Ages. Yet more importantly it completely overlooks the immeasurable service to human culture rendered by Arab science. For the truly committed student I recommend reading: Ahmed Djebbar, Une histoire de la science arabe, Editions du Seuil, 2001.


Cosmology

V.A. Rubakov
Institute for Nuclear Research of the Russian Academy of Sciences, 117312 Moscow, Russia

Abstract
These lectures aim at a pedagogical introduction to cosmology from the viewpoint of a high-energy physicist. After introducing some theoretical background, a brief overview of cosmological data is presented, and the gross features of the late and present Universe are outlined. Then the main properties of the hot stage of the cosmological evolution are discussed, and inflationary theory with its predictions for cosmological perturbations is described.

1 Introduction
The evolution of the Universe is determined, to a large extent, by microscopic laws of physics—the same laws that govern particle interactions at high energies. Hence, discoveries in particle physics are of direct relevance to the theory of the Universe. Conversely, cosmology provides important insights for high-energy physics. Amazingly, many fundamental aspects of cosmology require dramatic extensions beyond known physics; it may even happen that some aspects of cosmology will be possible to understand only by invoking hints from string theory.

How well do we understand the present and the early Universe? Why are cosmologists so confident when inventing new physics, even though it has not been discovered yet by the high-energy physics community? What lessons should high-energy physicists learn from advances in cosmology? These lectures are an attempt to address these issues, which are at the core of high-energy physicists’ interest in cosmology.

There are many excellent books and reviews on cosmology and astrophysics. Here we mention a few [1–12]. When discussing specific subjects in these lectures, we shall mostly refer to review papers; an interested reader may find references to original literature there.

2 Expanding universe
2.1 Friedmann–Robertson–Walker metric
Two basic facts about the observable part of the Universe are that it is homogeneous and isotropic at large spatial scales, and that it expands.

There are three types of homogeneous and isotropic three-dimensional spaces, which are conventionally labelled by a parameter $\kappa = 0, +1, -1$. These are three-sphere ($\kappa = 1$), flat space ($\kappa = 0$) and three-hyperboloid ($\kappa = -1$). Were our space a three-sphere, the sum of angles of a triangle would be greater than $\pi$; for a three-hyperboloid it would be smaller than $\pi$, while for a flat three-space it would be equal to $\pi$. Accordingly, one speaks about a closed, a flat, and an open Universe ($\kappa = +1, 0$ and $-1$, respectively); in the latter two cases the spatial size of the Universe is infinite, whereas in the former the Universe is compact.

The homogeneity and isotropy of the Universe mean that its hypersurfaces of constant time are either three-spheres or three-planes or three-hyperboloids. The distances between points may (and in fact, do) depend on time, i.e., the interval has the form

$$ds^2 = dt^2 - a^2(t)dx^2$$

1\footnote{Strictly speaking, this statement is valid only locally; in principle, a homogeneous and isotropic Universe may have complex global properties. As an example, a spatially flat Universe may have the topology of a three-torus. There is some discussion of such a possibility in the literature, and fairly strong limits have been obtained by the analyses of cosmic microwave background [13].}
where \( dx^2 \) is the distance on unit three-sphere/plane/hyperboloid. It is straightforward to see that a point mass (e.g., a galaxy) put at some fixed value of \( x \) at zero velocity will stay at the same \( x \). For this reason the coordinates \( x \) are often called comoving coordinates. The physical distance between two such neighbouring points at a given moment of time is

\[
dt^2 = a^2 dx^2
\]

and thus it depends on time. Note that time \( t \) is the proper time of an observer staying at fixed \( x \); this is physical ‘cosmic time’. To an excellent approximation it is this time that we measure by our watches.

Metric (1) is usually called the Friedmann–Robertson–Walker (FRW) metric, and \( a(t) \) is called the scale factor. In our Universe

\[
\dot{a} \equiv \frac{da}{dt} > 0
\]

which means that the distance (2) between points of fixed spatial coordinates \( x \) grows. The Universe expands.

As the Universe expands, non-relativistic objects lose their velocities \( \dot{x} \), i.e., they get frozen in the comoving coordinate frame. Today the galaxies, to a good approximation, stay at certain fixed values of \( x \), so the physical distance between them grows according to (2).

### 2.2 Redshift and the Hubble law

Let us discuss the propagation of photons in an expanding Universe. Instead of dealing with a vector field, let us consider a massless scalar field; this study will be used in what follows for other purposes, while the main results are valid for photons and other massless particles. We are going to consider a spatially flat Universe (\( \kappa = 0 \)); it is clear that the results will be valid for a closed and an open Universe as well, provided that the radius of spatial curvature is much greater than the wavelengths of the waves of interest (this is certainly the case in the present Universe for photons of all relevant wavelengths).

To begin with, let us find out the action for the massless scalar field \( \varphi(t, x) \) in the expanding Universe. In Minkowski space–time, the action is

\[
S_{\text{Mink}} = \int dt \, d^3x \left[ \frac{1}{2} (\dot{\varphi})^2 - \frac{1}{2} \left( \nabla_i \varphi \right)^2 \right]
\]

where the dot denotes a time-derivative, and \( \nabla_i \) is a spatial gradient (hereafter \( i = 1, 2, 3 \) is a spatial index). To generalize this action to the expanding background (1), we note that the physical spatial gradient is

\[
\nabla = \frac{\partial \varphi}{\partial l}
\]

where \( l \) is a vector of the physical distance (2), so that

\[
\nabla_i \varphi = \frac{1}{a(t)} \frac{\partial \varphi}{\partial x^i} = \frac{1}{a(t)} \partial_i \varphi.
\]

Another modification is that the physical volume is

\[
d^3l = a^3(t) d^3x.
\]

Thus, we arrive at the following expression for the action in the expanding Universe

\[
S = \int dt \, d^3x \, a^3(t) \left[ \frac{1}{2} (\dot{\varphi})^2 - \frac{1}{2a^2(t)} (\partial_i \varphi)^2 \right].
\]

The generalization of the action of massless scalar field to an arbitrary geometry described by metric \( ds^2 = g_{\mu\nu} dx^\mu dx^\nu \) is

\[
S = \int d^3x \sqrt{-g} \frac{1}{2} g^{\mu\nu} \partial_\mu \varphi \partial_\nu \varphi.
\]

The expression (3) specifies the latter formula to the particular metric (1).
By varying this action with respect to $\varphi$ we obtain the field equation, generalizing the Klein–Gordon equation to the expanding background,

$$\ddot{\varphi} + 3\frac{\dot{a}}{a}\dot{\varphi} - \frac{1}{a^2}\partial_i\partial_i\varphi = 0 . \quad (4)$$

We note in the first place that solutions to this equation are collections of plane waves

$$\varphi(t, x) = \varphi_k(t)e^{-ikx} \quad (5)$$

where the coordinate wave vector (coordinate momentum) $k$ is independent of time. The coordinate wavelength $2\pi/|k|$ stays constant, but the physical wavelength

$$\lambda(t) = a(t)\frac{2\pi}{k}$$

increases. The physical momentum

$$p(t) = \frac{k}{a(t)}$$

decreases in time, i.e., the waves get red-shifted. The mode $\varphi_k(t)$ obeys the equation

$$\ddot{\varphi}_k + 3\frac{\dot{a}}{a}\dot{\varphi}_k - \frac{k^2}{a^2}\varphi_k = 0 . \quad (6)$$

The WKB solution to this equation is

$$\varphi_k(t) = \frac{\text{const}}{a(t)} e^{i\int \omega(t) dt} \quad (6)$$

where

$$\omega(t) = \frac{k}{a(t)}$$

is the physical frequency, which decreases in time in the same way as the physical momentum. All this is in accord with an intuitive understanding of the expansion of space: as it expands, all lengths, including wavelengths of photons, increase.

We shall always label the present values of time-dependent quantities by the subscript 0: the present physical wavelength of a photon is thus denoted by $\lambda_0$, the present time is $t_0$, the present value of the scale factor is $a_0 \equiv a(t_0)$, etc. If a photon was emitted at some moment of time $t_e$ in the past, and its physical wavelength at the moment of emission was $\lambda_e$ ($\lambda_e$ is fixed by physics of the source; say, it is the wavelength of a photon emitted by an excited hydrogen atom), then we receive today a photon whose physical wavelength is longer,

$$\frac{\lambda_0}{\lambda_e} = 1 + z = \frac{a_0}{a(t_e)} .$$

Here we introduced the redshift $z$ which, on the one hand, is directly measurable$^3$, and, on the other hand, is related to the time of emission, and hence to the distance to the source. To summarize, the relation between redshift and scale factor is

$$1 + z(t) = \frac{a_0}{a(t)} . \quad (7)$$

It is worth noting that in a spatially flat Universe, the scale factor does not have absolute meaning: at a given moment of time it may be set equal to 1 by rescaling the spatial coordinates. What is physically meaningful is the ratio of scale factors at different times; only this ratio enters Eq. (7), as well as all formulas for a flat Universe.

$^3$One identifies a series of emission or absorption lines, thus obtaining $\lambda_e$, and measures their actual wavelength $\lambda_0$. These spectroscopic measurements give very accurate values of $z$ even for distant sources.
Let us consider a ‘nearby’ source, for which

\[ z \ll 1. \]

This corresponds to relatively small \((t_0 - t_e)\). Expanding \(a(t_e)\), one writes

\[ a(t_e) = a_0 - \dot{a}(t_0)(t_0 - t_e) . \tag{8} \]

To the leading order in \(z\), the difference between the present time and the emission time is equal to the distance to the source \(r\) (the speed of light is set equal to 1). Let us define the Hubble parameter

\[ H(t) = \frac{\dot{a}(t)}{a(t)} \]

and denote its present value by \(H_0\). Then Eq. (8) takes the form

\[ a(t_e) = a_0(1 - H_0r) \]

and we get for the redshift, again to the leading non-trivial order in \(z\),

\[ 1 + z = \frac{1}{1 - H_0r} = 1 + H_0r . \]

In this way we obtain the Hubble law,

\[ z = H_0r , \quad z \ll 1 . \tag{9} \]

Traditionally, one tends to interpret the expansion of the Universe as a runaway of galaxies from each other, and the redshift as the Doppler effect. Then at small \(z\) one writes \(z = v\), where \(v\) is the radial velocity of the source with respect to the Earth, so \(H_0\) is traditionally measured in ‘velocity per distance’ units. Observational data, which we shall discuss briefly in Section 3, give

\[ H_0 \approx 70 \frac{\text{km}}{\text{s} \cdot \text{Mpc}} \tag{10} \]

with about five per cent precision. Traditionally, the present value of the Hubble parameter is written as

\[ H_0 = h \cdot 100 \frac{\text{km}}{\text{s} \cdot \text{Mpc}} \tag{11} \]

(1 Mpc \( \approx 3\) mln. light yrs. \( \approx 3 \cdot 10^{24}\) cm). Thus

\[ h \approx 0.7 . \]

We shall use this value in further estimates.

Let us point out that the interpretation of the redshift in terms of the Doppler shift is actually not adequate, at least for large enough \(z\). In fact, there is no need in this interpretation at all: the ‘radial velocity’ enters neither theory nor observations, so this notion may be safely dropped. A physically meaningful quantity is the redshift \(z\) itself.

A final comment is that \(H_0^{-1}\) has dimension of time, or length. Clearly, this quantity sets the cosmological scales of time and distance at the present epoch,

\[ H_0^{-1} \approx 14 \cdot 10^9 \text{ yrs} \approx 4300 \text{ Mpc} . \]

We shall discuss this point further in Section 2.5.
2.3 Hot Universe

Our Universe is filled with a cosmic microwave background (CMB). The cosmic microwave background as observed today consists of photons with an excellent black-body spectrum of temperature

$$T_0 = 2.725 \pm 0.001 \text{ K}. \quad (12)$$

The spectrum has been precisely measured by various instruments and does not show any deviation from the Planck spectrum. Using the Planck distribution, one calculates all properties of the photon gas; as an example, the present number density of CMB photons is

$$n_{\gamma,0} = \frac{2\zeta(3)}{\pi^2} \cdot T_0^3 = 410 \frac{1}{\text{cm}^3}. \quad (13)$$

The entropy density in the present Universe is of the same order of magnitude. The present energy density of the CMB photons is

$$\rho_{\gamma,0} = \frac{\pi^2}{15} \cdot T_0^4 = 2.7 \cdot 10^{-10} \text{ GeV cm}^{-3}. \quad (14)$$

Thus the present Universe is ‘warm’. The earlier Universe was warmer; it cooled down because of the expansion. While the CMB photons freely propagate today, it was not so at the early stage. When the Universe was hot, the usual matter (electrons and protons with a rather small admixture of light nuclei) was in the plasma phase. At that time electromagnetic interactions between electrons, photons and protons in the plasma were strong, so all these particles were in thermal equilibrium. As the Universe cooled down, electrons ‘recombined’ with protons into neutral hydrogen atoms, and the Universe became transparent to photons. The temperature scale of recombination is, very crudely speaking, determined by the ionization energy of hydrogen, which is of order 10 eV. In fact, recombination occurred at lower temperature

$$T_{\text{rec}} \approx 3000 \text{ K}.$$ 

An important point is that the recombination process lasted quite a bit less than the Hubble time at that epoch; to a reasonable approximation, recombination occurred instantaneously.

Another point is that even though after recombination photons were no longer in thermal equilibrium with anything, the shape of the photon distribution function did not change, except for the overall redshift. Indeed, the thermal distribution function for ultra-relativistic particles, the Planck distribution, depends only on the ratio of frequency to temperature,

$$f_{\text{Planck}}(p, T) = f \left( \frac{\omega_p}{T} \right), \quad \omega_p = |p|.$$ 

As the Universe expands, the frequency gets red-shifted, $$\omega_p \rightarrow \omega_p / (1 + z)$$, but the shape of the spectrum remains Planckian, with temperature $$T/(1 + z)$$. Hence, the Planckian form of the observed spectrum is no surprise. Generally speaking, this property does not hold for massive particles.

At even earlier times, the temperature of the Universe was even higher. The earliest time which has been observationally probed to date is the Big Bang Nucleosynthesis epoch (which we shall briefly discuss below), and corresponds to a temperature of order 1 MeV.

---

4The reason is that the number density of electrons and protons was small compared to the number density of photons, and that the thermal velocities of electrons and protons were small. At temperature $$T$$ considerably smaller than 10 eV, photons from the tail of thermal distribution still disintegrated hydrogen atoms into electrons and protons, while an electron had to travel for a relatively long time to find a proton to recombine with. It was only at $$T = T_{\text{rec}}$$ that the rate of disintegration became small enough.

5A similar property holds, however, for particles that decouple being non-relativistic (hydrogen, cold dark matter). At decoupling, they have a Maxwell–Boltzmann distribution function, which is a function of the ratio $$p^2/(2mT)$$. As the momentum gets red-shifted, $$p \rightarrow p/(1 + z)$$, the shape of this distribution function remains Maxwell–Boltzmann, with effective temperature $$T/(1 + z)^2$$. Incidentally, this explains the fact, mentioned above, that non-relativistic free massive particles get frozen into the comoving frame.
To summarize, the effective temperature of photons scales as

$$T(t) \propto \frac{1}{a(t)}$$

(15)

that is

$$\frac{T}{T_0} = \frac{1}{1 + z}.$$ 

This behaviour is characteristic of *ultra-relativistic* free species (at zero chemical potential) only. The same formula is valid for ultra-relativistic particles (at zero chemical potential) which are in thermal equilibrium. Thermal equilibrium means adiabatic expansion; during adiabatic expansion, the temperature of ultra-relativistic gas scales as the inverse size of the system, hence Eq. (15).

Both for free photons and for photons in thermal equilibrium, the number density behaves as follows,

$$n_\gamma = \text{const} \cdot T^3 \propto \frac{1}{a^3}$$

and the energy density is

$$\rho_\gamma = \frac{\pi^2}{30} \cdot 2 \cdot T^4 \propto \frac{1}{a^4}$$

(16)

where the factor 2 accounts for two photon polarizations.

Let us now turn to non-relativistic particles: baryons, massive neutrinos, possible ‘dark matter’ species, etc. If they are not destroyed during the evolution of the Universe (that is, they are stable and do not co-annihilate), their number density merely gets diluted,

$$n \propto \frac{1}{a^3}.$$ 

(17)

This means, in particular, that the baryon-to-photon ratio stays constant,

$$\eta \equiv \frac{n_B}{n_\gamma} = \text{const}.$$ 

(18)

The energy density of non-relativistic particles scales as

$$\rho(t) = m \cdot n(t) \propto \frac{1}{a^3(t)}$$

(19)

in contrast with more rapid fall off (16) characteristic to ultra-relativistic species.

As we shall discuss later, there exists strong evidence for *dark energy* in the Universe, whose density does not decrease in time as fast as in Eqs. (16) or (19). For the moment suffice it to mention that this property holds for vacuum, whose energy density stays constant (in a locally Lorentz frame),

$$\rho_{\text{vac}} = \text{const}$$

(20)

while the vacuum energy-momentum tensor in the arbitrary frame is, by general covariance,

$$T^{\text{vac}}_{\mu\nu} = \rho_{\text{vac}} g_{\mu\nu},$$

i.e., the vacuum has negative pressure [see Eq. (24)],

$$p_{\text{vac}} = -\rho_{\text{vac}}.$$ 

(21)

In this context, the vacuum energy density is the same thing as the cosmological constant, or $\Lambda$ term.
2.4 Friedmann equation

The basic equation governing the expansion rate of the Universe is the Friedmann equation

\[ H^2 \equiv \left( \frac{\dot{a}}{a} \right)^2 = \frac{8\pi}{3} G \rho - \frac{\kappa}{a^2} \]  

(22)

where the dot again denotes a derivative with respect to time \( t \), \( \rho \) is the total energy density in the Universe, the parameter \( \kappa \) has been introduced in Section 2.1 and distinguishes closed (\( \kappa = +1 \)), flat (\( \kappa = 0 \)) and open (\( \kappa = -1 \)) Universes, and \( G \) is Newton’s gravity constant. In natural units

\[ G = M_{\text{Pl}}^{-2} \]

where

\[ M_{\text{Pl}} = 1.2 \cdot 10^{19} \text{ GeV} \]

is the Planck mass. The first and second terms on the right-hand side of Eq. (22) may be viewed as the contributions of all kinds of matter/energy and spatial curvature, respectively, to the expansion rate. As we shall discuss in Section 3, observational data show that the curvature term is very close to zero today,

\[ \frac{1}{a_0^2} \frac{H_0^2}{H_0^2} < 0.02 , \]  

(23)

i.e., it currently makes up not more than 2% of the right-hand side of the Friedmann equation. In fact, inflationary theory suggests that the spatial curvature vanishes to an excellent approximation.

The Friedmann equation (22) is nothing but one of the Einstein equations of General Relativity specialized to homogeneous and isotropic space. It has to be supplemented by the law of evolution for the energy density \( \rho \). To obtain the latter, we use the first law of thermodynamics, and take into account the fact that the expansion of the Universe is slow, and hence adiabatic, at interesting temperatures. We apply the first law to a comoving volume \( V = a^3 \) and write, using entropy conservation in a comoving volume,

\[ d(\rho V) = -p dV \]

where \( p \) is pressure. We obtain

\[ d\rho \cdot V = -(p + \rho) dV \]

and get finally\(^6\)

\[ \frac{d\rho}{\rho + p} = -3 \frac{da}{a} . \]  

(25)

To close the system of equations, one needs also an equation of state

\[ p = p(\rho) . \]

The latter equation is governed by the composition of the Universe at a given moment of time.

Other Einstein equations are satisfied automatically for this simple geometry. The relations (16), (19) and (20) may in fact be viewed as consequences of Eq. (25), since \( p = \rho/3, p = 0 \) and \( p = -\rho \) for relativistic matter, non-relativistic particles and vacuum, respectively.

As we pointed out above, the Universe is spatially flat today to a good approximation; the curvature term \( \kappa/a^2 \) in Eq. (22) today is less than two per cent of the matter term. This can also be phrased in the following way. One defines the critical density \( \rho_c \) according to

\[ \frac{8\pi}{3} G \rho_c = H_0^2 . \]

(26)

\(^6\)This equation is in fact a consequence of the covariant conservation of energy-momentum, \( \nabla_\mu T^{\mu\nu} = 0 \). For the energy-momentum tensor of a fluid one has

\[ T^{\mu\nu}_f = \text{diag} \left( \rho, -p, -p, -p \right) . \]

(24)

In the background FRW metric the covariant conservation law reduces precisely to Eq. (25).
The meaning of this quantity is as follows. If the actual energy density $\rho$ of all forms of matter in the Universe (including vacuum, quintessence, etc.) is larger than $\rho_c$, then $\kappa = +1$, and the Universe is closed; if $\rho < \rho_c$, the Universe is open, and it is flat for $\rho = \rho_c$. From (23) we see that observationally,

$$\rho = (1 \pm 0.02)\rho_c.$$  \hspace{1cm} (27)

At earlier times, the curvature term was even less significant, so we shall neglect it in the study of the evolution of the Universe, and write the Friedmann equation as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3} G \rho.$$  \hspace{1cm} (28)

To get an idea of numerics, one plugs the present value of the Hubble parameter (11) into the definition (26) and converts $\rho_c$ into

$$\rho_c = h^2 \cdot 1 \cdot 10^{-5} \text{ GeV cm}^{-3} \approx 5 \cdot 10^{-6} \text{ GeV cm}^{-3},$$  \hspace{1cm} (29)

where we use our fiducial value $h = 0.7$.

2.5 Sample solutions

Solutions to the Friedmann equation (28) are most easily obtained in cases when matter of a definite type gives a dominant contribution to the energy density $\rho$. Let us present some solutions of this sort.

2.5.1 Matter-dominated Universe

If the dominant contribution to the energy density comes from non-relativistic particles, then $p = 0$, and from (25) one finds

$$\rho = \text{const} \cdot a^{-3}.$$  

Making use of this relation, one finds that the Friedmann equation reads

$$\frac{\dot{a}}{a} = \frac{\text{const}}{a^{3/2}}.$$  

The solution is

$$a = \text{const} \cdot t^{2/3}.$$  \hspace{1cm} (30)

Note that this solution describes a decelerating Universe,

$$\ddot{a} < 0.$$  

As $t \to 0$, the scale factor tends to zero, and the energy density tends to infinity. This is a cosmological singularity, ‘beginning of the Universe’ (‘Big Bang’), and $t$ is the lifetime of the Universe. Note that the lifetime is related to the Hubble parameter, since

$$H \equiv \frac{\dot{a}}{a} = \frac{2}{3t}.$$  

Until fairly recently, our Universe was indeed matter dominated, so this relation may be used to obtain a crude estimate of its present age,

$$t_0 \sim \frac{2}{3} H_0^{-1} \approx 1 \cdot 10^{10} \text{ yrs}$$  

(again with $h = 0.7$). In fact, the lifetime of 10 billion years was a bit of a problem about 10 years ago, as independent estimates of lifetimes of old objects in our Universe suggested that their lifetimes were close to, and sometimes even larger than 10 billion years. This was one of the reasons for suggesting,
even before the observational evidence for an accelerating Universe, that our Universe is not in the
classic dark-energy-dominated stage today—rather, it is in a dark-energy-dominated stage.

Let us make use of this solution to introduce another notion, the cosmological horizon. Suppose
that at \( t = 0 \), signals were emitted everywhere in space, and then propagated in the Universe with the
speed of light. We ask at what distance today are the sources of signals we receive now. This sphere is
precisely the cosmological horizon for the solution (30): the interior of this sphere is causally connected
to us, while the part of the Universe outside this sphere is causally disconnected from us. The world line
of a signal propagating with the speed of light obeys \( ds = 0 \), that is along this world line

\[
ds = a(t)dt.
\]

This implies that the coordinate distance to the horizon is

\[
r_H = \int_0^t \frac{dt'}{a(t')}.
\]

The physical distance to the horizon at the time \( t \) is thus

\[
l_H(t) = a(t)r_H = a(t) \int_0^t \frac{dt'}{a(t')}.
\]

For the solution (30) this distance is finite,

\[
l_H(t) = 3t = 2H^{-1}.
\]

Hence the present size of the visible part of the Universe is estimated as

\[
l_{H,0} \sim 2H_0^{-1} \approx 3 \cdot 10^{28} \text{ cm} \approx 10^4 \text{ Mpc}.
\]

Note that in a matter-dominated Universe, the integral in Eq. (31) is saturated at large \( t' \), so the relic
photons, which were actually emitted somewhat after the Big Bang, travelled almost the same distance
as \( l_{H,0} \).

Let us stress that the above estimates for the present age and horizon size are not quite correct,
since during a good part of the evolution, the expansion of the Universe was not dominated by non-
relativistic matter; rather, it was dark-energy dominated. Taking that into account, the present lifetime of
the Universe is

\[
t_0 = 14 \cdot 10^9 \text{ yrs}.
\]

We do not need the refined estimate of the present size of the horizon.

### 2.5.2 Radiation-dominated Universe

An elementary calculation shows that the equation of state for ultra-relativistic matter (not necessarily in
thermal equilibrium) is

\[
p = \frac{1}{3} \rho.
\]

If the dominant contribution to the energy density comes from ultra-relativistic particles, then from (25)
it follows that [see also Eq. (16)]

\[
\rho = \text{const} \cdot a^{-4}
\]

and the solution to the Friedmann equation (28) is

\[
a(t) = \text{const} \cdot t^{1/2}.
\]

Qualitative features of this solution are similar to the matter-dominated case: the Universe starts from
the singularity, its expansion decelerates, age and horizon size are finite at given \( t \).
2.5.3 Vacuum-dominated Universe

A qualitatively different solution occurs if the Universe is vacuum dominated. The vacuum energy density $\rho_{\text{vac}}$ is time-independent, so the solution to the Friedmann equation (28) is

$$a(t) = \text{const} \cdot e^{H_{\text{vac}} t}$$

(33)

where the time-independent Hubble parameter is determined by the vacuum energy density,

$$H_{\text{vac}} = \sqrt{\frac{8\pi}{3} G \rho_{\text{vac}}} .$$

One notices that the Universe accelerates, rather than decelerates,

$$\ddot{a} > 0 .$$

We note in passing that a spatially flat FRW metric with scale factor (33) describes (part of) the de Sitter space–time. Unlike other cosmological solutions, de Sitter geometry does not have a past singularity.

2.5.4 Equation of state $p = w \rho$

Let us now consider the general case of a fluid with the equation of state

$$p = w \rho$$

where $w$ is a constant. For definiteness, let us restrict ourselves to the case

$$w > -1 .$$

Then the solution to Eq. (25) is

$$\rho = \frac{\text{const}}{a^{3(1+w)}} .$$

With $\kappa = 0$ (spatially flat Universe) one finds from Eq. (22)

$$a = \text{const} \cdot t^\alpha$$

(34)

where

$$\alpha = \frac{2}{3} \frac{1}{1 + w} .$$

The behaviour of solutions is qualitatively different for $w > -1/3$ and $w < -1/3$, i.e., for $\alpha < 1$ and $\alpha > 1$:  

$$\begin{align*}
  w > -\frac{1}{3} : & \quad \ddot{a} < 0 , \quad \text{decelerated expansion} \\
  w < -\frac{1}{3} : & \quad \ddot{a} > 0 , \quad \text{accelerated expansion} .
\end{align*}$$

(35)

Thus, accelerated expansion of the Universe requires negative pressure.

It is worth noting that the two cases differ in another respect: in the former case there exists a cosmological horizon, while in the latter the entire Universe is causally connected. Indeed, we have seen that the cosmological horizon exists, if the following integral converges [see Eq. (31)],

$$\int_0^t \frac{dt'}{a(t')} .$$

This integral is convergent for $\alpha < 1$, i.e., $w > -1/3$. Otherwise this integral diverges, so the cosmological horizon is absent. Note that this observation has to do with early times, $t \to 0$; it is of relevance for inflation rather than for the present epoch.
2.6 Changing regimes
As we shall discuss in Sections 3 and 4, the present expansion of the Universe is dark-energy dominated; to a reasonable approximation, the expansion of the Universe today follows the exponential law (33). This was not so at earlier times. Indeed, the dark energy density is constant in time (or almost constant), while the energy density of non-relativistic particles scales like $a^{-3}(t)$. Hence, even though the present energy density of non-relativistic matter is smaller than dark energy, non-relativistic matter dominated at an earlier epoch, when the scale factor was smaller than today. Even earlier, the expansion of the Universe was dominated by ultra-relativistic matter. Indeed, the energy density of the latter scales as $a^{-4}(t)$, so at small enough $a(t)$ it was higher than the energy density of non-relativistic particles. The curvature term $\kappa/a^2$ never dominated the expansion of the Universe: at present it contributes much less than not only the dark energy density, but also the matter energy density. Its contribution was even less significant at earlier times, as it scales as $a^{-2}$, whereas the matter energy density scales as $a^{-3}$. This is why it is legitimate to neglect the curvature term for describing the entire evolution of the Universe.

We shall quantify this discussion later, when we have a better idea of the composition of the Universe.

3 Overview of cosmological data
In the last 10 to 15 years, cosmology has become a qualitative science. Detailed data on the present and earlier Universe are now available, and even more precise data are due to come. Before going into further theoretical discussions, let us briefly consider what kinds of data there are, and what gross features of the Universe they show.

3.1 Distribution of luminous matter
Distribution of luminous matter (galaxies, quasars) is obtained by deep surveys. The largest surveys (2dF, SDSS) measure angular positions of, and distances (redshifts) to hundreds of thousands of galaxies, with a depth of the order of 2000 Mpc (about 6 billion light years). This is a fairly large portion of the present Universe. An even larger part of the Universe is sampled by quasars. Thus, by now we have a ‘map’ of our ‘neighbourhood’, and can discuss matter distribution on various length scales.

At large scales, the Universe is homogeneous and isotropic, as illustrated in Fig. 1.

Fig. 1: Spatial distribution of galaxies (left panel) and quasars (right panel), according to SDSS survey [14]. Shown are samples of usual and brighter galaxies and quasars. The parameter $h$ is defined in the text.
At shorter scales, the Universe is of course inhomogeneous. The largest structures visible (superclusters of galaxies, giant voids) extend to several dozens of Mpc. The comparison of the observed structure to simulations, at scales ranging from a few kpc (size of a galaxy) and smaller, to thousands of Mpc, tells a lot about the primordial density perturbations in the early Universe, the composition of the Universe, and the history of its expansion.

### 3.2 Observation of ‘standard candles’

These are the objects whose absolute luminosity is believed to be known. What is measured is the visible luminosity $F$ and redshift $z$. At relatively short distances, the visible luminosity is related to the distance,

$$F = \frac{L}{4\pi r^2}$$  \hspace{1cm} (36)

where $L$ is the absolute luminosity. This relation becomes confusing for $z \sim 1$, yet one often defines the luminosity distance by Eq. (36), and talks about the redshift–distance relation (the notion of velocity is usually not used for large $z$).

At $z \ll 1$, the relation between $z$ and $r$ follows a linear Hubble law (9), as shown in Fig. 2. The present value of the Hubble parameter consistent with virtually all measurements is

$$H_0 = (73 \pm 3) \frac{\text{km}}{\text{s} \cdot \text{Mpc}}.$$  \hspace{1cm} (37)

![Hubble diagram for Supernovae 1a as standard candles, see Refs. [7, 15]. The straight line corresponds to the Hubble law.](image)

**Fig. 2:** Hubble diagram for Supernovae 1a as standard candles, see Refs. [7, 15]. The straight line corresponds to the Hubble law.

At large $z$, the linear relation (9) no longer holds. The relation between $z$ and $r$ tells us about the expansion rate of the Universe at a relatively late epoch. Currently, the data for large $z$ come from the observations of type 1a supernovae (SNe 1a), Figs. 3 and 4. Surprisingly, they show that the Universe undergoes accelerated expansion today (and at relatively small $z$, i.e., at late times), while at higher redshift the expansion was decelerating. We shall discuss the significance of this result shortly.

### 3.3 Cosmic-microwave-background radiation

The cosmic microwave background (CMB) is an extremely important source of information about the properties of the earlier Universe.
Fig. 3: Upper panel: stellar magnitude (proportional to $-\log$(visible luminosity)) of SNe 1a as a function of redshift, according to relatively old data [16]. The larger the stellar magnitude, the dimmer the object. For a definition of $\Omega$'s see the text. Lower panel: the same figure, with expectation from the model $\Omega_m=0.2$, $\Omega_{\Lambda}=0$ subtracted.

Fig. 4: More recent data [17] on SNe 1a; subtracted is the expectation from a model in which the Universe expands at constant velocity, $\dot{a} = \text{const}$
The CMB photons were last scattered/emitted at the recombination epoch, when the Universe was only about $3 \cdot 10^5$ years old; for comparison, the present age of the Universe is about $1.4 \cdot 10^{10}$ years. The observations of CMB give the photographic picture (literally) of the ‘young’ Universe, which had quite different properties than it has today. One of these properties is a much higher level of homogeneity and isotropy: photons coming from different directions in the sky have almost (but not exactly!) the same temperature. Crudely speaking, relative angular anisotropy of CMB temperature, $\delta T/T_0$, is of order $10^{-5}$. This means that the Universe was homogeneous and isotropic at the level of $10^{-5}$ when it was 300 thousand years old.

Yet the angular anisotropy of the CMB temperature exists, and has been measured at various angular scales. In a wide range of angular scales this anisotropy was accurately measured by the WMAP satellite, see Figs. 5–7, while at smaller angular scales data are available from measurements made by ground-based interferometers.

![Fig. 5: WMAP data [18]: temperature of photons coming from different directions in the sky. Darker regions correspond to lower temperatures. The average temperature $T_0$ and the dipole component are subtracted. The angular variation of the temperature is at the level of $\delta T \sim 100 \mu$K, i.e., $\delta T/T_0 \sim 10^{-4}$–$10^{-5}$.](image)

It is convenient to decompose the temperature, as a function of the direction $\vec{n}$, in spherical harmonics $Y_{lm}(\vec{n})$, which make a complete set of functions on a sphere. One writes

$$\delta T(\vec{n}) \equiv T(\vec{n}) - T_0 - \delta T_{\text{dipole}} = \sum_{l,m} C_{lm} Y_{lm}(\vec{n}) .$$

The angular momentum $l$ corresponds to fluctuations with typical angular scale $\pi/l$.

Figures 6 and 7 show the measured anisotropy as a function of $l$.

To understand what is shown, one notices that the data are grossly consistent with Gaussian fluctuations [19], for which $C_{lm}$ are statistically independent. For an isotropic Universe this means

$$\langle C_{lm} C_{l'm'}^* \rangle = C_l^2 \delta_{ll'} \delta_{mm'} .$$

The coefficients $C_l$ determine the correlation function

$$\langle \delta T(\vec{n}_1) \delta T(\vec{n}_2) \rangle = \sum_l \frac{2l + 1}{4\pi} C_l^2 P_l(\cos \theta) ,$$

---

7The dipole component is due to the motion of the Earth with respect to the rest frame of the CMB photon gas.
Fig. 6: Measured angular anisotropy of cosmic microwave background [18, 20, 21]. Shown are data points from the first-year WMAP observations and from ground-based observations by CBI and ACBAR.

Fig. 7: Recent three-year WMAP data [19], shown by solid data points, together with older one-year WMAP data (light grey data points). Different lines correspond to slightly different best-fit cosmological models.

where $P_l$ are the Legendre polynomials, and $\theta$ is the angle between $\vec{n}_1$ and $\vec{n}_2$. In particular, for mean fluctuation of the temperature one has

$$\langle \delta T^2 \rangle = \sum_l \frac{2l + 1}{4\pi} C_l^2 \approx \int \frac{l(l+1)}{2\pi} C_l^2 \, d\ln l.$$  

Hence the quantity

$$\delta T(l) \equiv \sqrt{\frac{l(l+1)}{2\pi} C_l}$$

is a measure of the contribution coming from angular momenta in a decimal interval of $l$. It is this quantity that is shown in Fig. 6, while Fig. 7 shows the square of this quantity.
The CMB anisotropy, and also CMB polarization, are sensitive to most of the properties of the Universe, such as the primordial spectrum of density perturbations, relic gravity waves, matter content of the Universe, its expansion history, spatial geometry, etc.

It is clear from Fig. 7 that the anisotropy as a function of $l$ has a fairly complex behavior. There are peaks (at least two of them are clearly visible) and dips. The physics beyond these features is roughly as follows. The CMB anisotropy has its origin in the density perturbations\(^8\), i.e., sound waves of all possible wavelengths. Waves of longer wavelengths are seen in the photographic picture at larger angles; this is the small-$l$ region in Fig. 6. Conversely, the region of large $l$ corresponds to shorter wavelengths of the sound waves. The properties of these waves are different for different modes. The modes with long wavelengths, and hence small frequencies had no time to oscillate by the time of recombination. Their amplitudes are determined by the initial spectrum of density perturbations, which, as the data show, is in some sense flat, i.e., independent of the wavelength. The effect of these modes on temperature is mostly due to the gravitational potential they produce (Sachs–Wolfe effect): roughly speaking, light from denser regions with stronger negative gravitational potentials has to climb up the gravitational well, so it gets more red-shifted as compared to light from less dense regions. This is the flat region of small $l$ in Fig. 6. For shorter waves there is another effect: the waves oscillate, i.e., the particles in plasma move; this causes the Doppler effect leading to the CMB anisotropy. The waves of ‘just right’ frequency (and hence wavelength) are in the phase of maximum motion at recombination. These are the waves seen at an angle of 0.7 degrees ($l \sim 200$), where $C_l$ has the first peak. Further peaks correspond to higher harmonics.

All physical processes involved—the expansion of the Universe during its first 300 thousand years, the evolution of density perturbations and the recombination itself—are very well understood, so the calculations of the CMB anisotropy are very reliable. The predictions depend, of course, on a number of parameters characterizing the early and present Universe, so the CMB data are used for extracting these parameters.

Let us point out two specific results coming from the analysis of the CMB anisotropy. The first one is that our Universe is spatially flat to rather high precision. The positions of the peaks in $C_l$ (in particular, of the first peak measured with good accuracy) are sensitive to the spatial curvature: the absolute wavelength of the corresponding sound waves is reliably calculable, while its angular size strongly depends on whether space is a 3-sphere, 3-hyperboloid or 3-plane. Quantitatively, the result is usually expressed in terms of the contribution of the curvature term to the right-hand side of the Friedmann equation (22): this contribution is less than about two per cent [we mentioned this already, see Eq. (27)]. In more physical terms, this means that the radius of the spatial curvature of our Universe $a$ is quite a bit greater than the length of the visible part (size of the cosmological horizon) $l_{H,0}$:

$$a > 4l_{H,0}.$$  \hspace{1cm} (38)

Another way to phrase this is to say that if the Universe were 3-sphere, its volume would still be a lot larger than the volume we can observe,

$$\frac{V_{\text{tot}}}{V_{\text{obs}}} = \frac{2\pi^2 a^3}{\frac{4}{3}\pi l_{H,0}^3} > 100 .$$

Hence even if the Universe has finite volume, we know from observations that we are able to observe not more than one per cent of it. It is worth stressing that Eq. (38) is an observational bound; it is likely that the actual radius of the spatial curvature of the Universe is many orders of magnitude greater than the horizon size.

The second result concerns the baryonic content of the Universe. The height of the second peak in $C_l$ is sensitive to the number density of electrons, which is equal to the number density of protons by

\(^8\)Another possible source is gravity waves; their effect, if any, is small, see Section 8.3.
electric neutrality. The analysis gives for the present number density of baryons
\[ n_{B,0} = 2.5 \times 10^{-7} \text{ cm}^{-3} \] (39)
with precision better than 10 per cent. In terms of time-independent parameter \( \eta \) defined in Eq. (18) this corresponds to
\[ \eta = 6 \times 10^{-10} . \] (40)
It is remarkable that the same value of this parameter comes from an entirely different set of observations, which we are about to discuss in brief.

Before doing that, let us stress that the information extracted from CMB data is by no means limited by these two examples. These data provide independent determinations of the present value of the Hubble parameter, the rate of acceleration of the cosmological expansion, the power spectrum of the primordial density perturbations, the amount of dark matter in the Universe, the bound on the primordial gravitational waves, etc. It is important that these data be consistent with virtually all other cosmological measurements.

### 3.4 Big Bang Nucleosynthesis

Theory of Big Bang Nucleosynthesis (BBN) and observations of primordial abundances of light elements probe the earliest epoch of the evolution of the Universe, accessible to observations today. This epoch corresponds to temperatures ranging from 1 MeV to (a few) \( \cdot 10 \) keV, and the age of the Universe from 1 to 200 s. At temperatures above 1 MeV, there is thermal equilibrium with respect to reactions
\[ p + e^- \leftrightarrow n + \nu_e . \] (41)
As the Universe cools down below \( T \approx 1 \) MeV, neutrons are no longer produced or destroyed; their concentration (relative to protons) ‘freezes out’. At temperatures of about 100 keV and somewhat lower, these neutrons combine with protons into light nuclei, mostly \(^4\)He, but also deuterium \(^2\)H, lithium \(^7\)Li and others. These elements remain in the Universe, so their primordial abundances are measurable today. The calculations of the thermonuclear reactions are again based on well-known physics, and the results are sensitive to the only unknown parameter\(^9\) \( \eta \). The results of the calculations and data are shown in Fig. 8.

It is worth pointing out that Big Bang Nucleosynthesis serves also as a source of constraints on particle physics exotica. The very fact that the temperature of the Universe reached at least 1 MeV or so, and that the expansion was described by known physics at that stage, constrains significantly some extensions of the Standard Model, like models with large extra dimensions [23] (for a review see Ref. [24]). More generally, constraints from BBN are important in models with stable or long-living new particles: these are produced at earlier stages and may contribute too much to the energy density at the nucleosynthesis epoch, modifying in this way the expansion rate, and hence the predictions of the BBN theory. Another example are particles that decay at the BBN epoch: these may destroy thermal equilibrium and therefore affect BBN; a situation of this sort occurs in some theories with light gravitinos. Any extension of the Standard Model has to be checked against cosmology, in particular against Big Bang Nucleosynthesis.

There are many other data of cosmological significance, like measurements of mass distributions in galaxies and galactic clusters. We shall briefly discuss some of them in appropriate places, and now proceed to immediate consequences.

---

\(^9\)Assuming the Standard Model particle content, see next paragraph.
Fig. 8: Big Bang Nucleosynthesis [22]: theoretical predictions (lines) versus observations (smaller boxes: 2\(\sigma\) errors, statistical only; larger boxes: 2\(\sigma\) with systematic uncertainties included). The vertical lines are the results from CMB anisotropy (the widths of bands depend on priors on other cosmological parameters).

4 Composition of the present Universe

As we shall discuss in this section, the cosmological data require a very surprising composition of the Universe.

Before proceeding, let us introduce a notion traditional in the analysis of the composition of the present Universe. For every type of matter \(i\) with the present energy density \(\rho_{i,0}\), one defines a parameter

\[
\Omega_i = \frac{\rho_{i,0}}{\rho_c}.
\]

Then Eq. (27) tells us that

\[
\sum_i \Omega_i = 1 \pm 0.02
\]

where the sum runs over all forms of energy. Let us now discuss contributions of different species to this sum.

We begin with baryons. The result (39), together with (13) gives

\[
\rho_{B,0} = m_B \cdot n_{B,0} \approx 2.5 \cdot 10^{-7} \frac{\text{GeV}}{\text{cm}^3}.
\]

Comparing this result with the value of \(\rho_c\) given in (29), one finds

\[
\Omega_B = 0.05.
\]
Thus baryons constitute a rather small fraction of the present energy density in the Universe. One point to note is that most of the baryons in our Universe are dark: direct measurements of the mass density of stars give an estimate

$$\Omega_{\text{stars}} \sim 0.005$$

which is about an order of magnitude smaller than $\Omega_B$. There is nothing particularly dramatic about this fact: baryons hide in dust and gas clouds (most likely), brown dwarfs and other non-luminous objects, etc.

**Photons** contribute an even smaller fraction, as is clear from (14):

$$\Omega_{\gamma} \approx 6 \cdot 10^{-4} .$$  \hspace{1cm} (46)

From electric neutrality, the number density of **electrons** is about\(^{10}\) the same as that of baryons, so electrons contribute a negligible fraction to the total mass density. The remaining known stable particles are **neutrinos**. Their number density is calculable in Hot Big Bang theory and these calculations are nicely confirmed by Big Bang Nucleosynthesis. The number density of each type of neutrinos is

$$n_{\nu\alpha} = 115 \frac{1}{\text{cm}^3}$$

where $\nu_\alpha = \nu_e, \nu_\mu, \nu_\tau$. A direct limit on the mass of the electron neutrino, $m_{\nu_e} < 2.6$ eV, together with the observations of neutrino oscillations suggest that every type of neutrino has a mass smaller than 2.6 eV (neutrinos with masses above 0.03 eV must be degenerate in mass, according to neutrino oscillation data). The energy density of all types of neutrinos is thus smaller than $\rho_c$:

$$\rho_{\nu,\text{total}} = \sum_\alpha m_{\nu\alpha} n_{\nu\alpha} < 3 \cdot 2.6 \text{ eV} \cdot 115 \frac{1}{\text{cm}^3} \sim 8 \cdot 10^{-7} \frac{\text{GeV}}{\text{cm}^3}$$

which means

$$\Omega_{\nu,\text{total}} < 0.16 .$$

This estimate does not make use of any cosmological data. In fact, cosmological observations give a stronger bound

$$\Omega_{\nu,\text{total}} < 0.01 .$$  \hspace{1cm} (47)

This bound is mostly due to the analysis of the structures at small length scales, and has to do with streaming of neutrinos from the gravitational potential wells at early times when neutrinos were ultra-relativistic. In terms of the neutrino masses the bound (47) reads [25]

$$\sum m_{\nu\alpha} < 0.4 \text{ eV}$$

so every neutrino must be lighter than 0.14 eV. On the other hand, atmospheric neutrino data and the K2K experiment tell us that the mass of at least one neutrino must be larger than 0.04 eV. Comparing these numbers, one sees that it may be feasible to measure neutrino masses by cosmological observations (!).

Coming back to our main topic here, we conclude that not more than 6% of the energy density in the present Universe is in the form of known particles; most energy in the present Universe must be in ‘something unknown’. Furthermore, there is strong evidence that this ‘something unknown’ has two components: clustered (dark matter) and unclustered (dark energy).

**Clustered dark matter** consists presumably of new stable massive particles. These make clumps of energy density which count for much of the mass of galaxies and most of the mass of galactic clusters. There are a number of ways of estimating the contribution of non-baryonic dark matter to the total energy density of the Universe (see Ref. [26] for details):

\(^{10}\)There are neutrons, whose number is smaller than the number of protons by roughly a factor of 7.
– The composition of the Universe affects the angular anisotropy of the cosmic microwave background. Quite accurate measurements of the CMB anisotropy, available today, enable one to estimate the total mass density of dark matter.

– The composition of the Universe, and especially the density of non-baryonic dark matter, is crucial for structure formation of the Universe. Comparison of the results of numerical simulations of structure formation with observational data gives a reliable estimate of the mass density of non-baryonic clustered dark matter.

The bottom line is that the non-relativistic component constitutes about 30 per cent of the total present energy density, which means that non-baryonic ‘cold dark matter’ contributes

$$\Omega_{\text{CDM}} \approx 0.25 .$$  \hspace{1cm} (48)

There is direct evidence that dark matter exists in the largest gravitationally bound objects—clusters of galaxies. There are various methods of determining the gravitating mass of a cluster, and even mass distribution in a cluster, which give consistent results. To name a few:

– One measures velocities of galaxies in galactic clusters, and makes use of the gravitational virial theorem,

$$\text{Kinetic energy} = \frac{1}{2} \text{Potential energy} .$$

In this way one obtains the gravitational potential, and thus the distribution of the total mass in a cluster.

– Another measurement of masses of clusters makes use of intracluster gas. Its temperature obtained from X-ray measurements is also related to the gravitational potential through the virial theorem.

– Fairly accurate reconstruction of mass distributions in clusters is obtained from the observations of gravitational lensing of background galaxies by clusters.

These methods enable one to measure the mass-to-light ratio in clusters of galaxies. Assuming that this ratio applies to all matter in the Universe\(^{11}\), one arrives at the estimate for the mass density of clumped matter in the present Universe. Remarkably, this estimate coincides with (48).

Finally, dark matter exists also in galaxies. Its distribution is measured by the observations of rotation velocities of distant stars and gas clouds around a galaxy. An example is shown in Fig. 9.

Thus cosmologists are confident that much of the energy density in our Universe consists of new stable particles. We shall see that natural candidates are particles which participate in weak interactions (or, more generally, particles whose annihilation cross-section is determined by a scale of the order of the electroweak scale, \(M_{\text{EW}} \sim 100 \text{ GeV}\)). Of course, this is only a hypothesis for the time being, and there are many other candidates for dark matter species.

Non-baryonic clustered dark matter is not the whole story. Besides the known and dark matter particles there is also unclustered dark energy, the most mysterious substance in the Universe. Making use of the above estimates, one obtains an estimate for the energy density of all particles,

$$\Omega_{\gamma} + \Omega_{\text{B}} + \Omega_{\nu,\text{total}} + \Omega_{\text{CDM}} \approx 0.3 .$$

We note in passing that the contribution of photons and possible massless neutrinos is very small here, so the left-hand side is the contribution of all non-relativistic matter, and is often denoted by \(\Omega_M\). Thus

$$\Omega_M = \Omega_{\text{B}} + \Omega_{\text{CDM}} \approx 0.3 .$$

\(^{11}\)This is a strong assumption, since only about 10 per cent of galaxies are in clusters.
Equation (43) implies then that 70 per cent of the energy density is unclustered.

All this fits nicely with the observations of SNe 1a. Indeed, we have seen that neither relativistic, nor non-relativistic matter can lead to the accelerated expansion of the Universe. So, the accelerated expansion requires energy stored in something dramatically different from conventional particles. Furthermore, we have seen that this ‘something’—dark energy—must have negative pressure. In fact the analysis of the entire set of cosmological data [19,25] in terms of dark energy with the phenomenological equation of state\(^\text{12}\)

\[ p = w \rho, \quad w = \text{const} \]

\(^{12}\)The data are consistent with time-independent \(w\), although allow for some time variation.

gives

\[ \Omega_\Lambda = 0.72 \pm 0.02 \]

(here subscript \(\Lambda\) traditionally refers to dark energy) and, see Fig. 10,

\[-1.1 < w < -0.9.\]

It is worth noting that the vacuum value, \(w = -1\) is right in the middle of the allowed region.

To conclude, the composition of the present Universe is fairly complex. Most of the energy density comes from species with which particle physicists are unfamiliar: vacuum or vacuum-like dark energy and non-baryonic clumped dark matter (presumably, non-relativistic, weakly interacting particles). This poses serious problems for both fundamental physics and cosmology:

- What are the particles of non-baryonic dark matter?

This appears to be a less difficult problem, as compared to some others listed below. Currently, a popular option is the lightest supersymmetric particle, which is stable in many supersymmetric extensions of the Standard Model. Indeed, we shall estimate in what follows that the present mass density of such particles is naturally predicted to be in the right ballpark. Of course there are
many other options, such as axions, gravitinos, Q-balls, to name a few. In any case, experimental discovery of the dark matter particle would be a great achievement of both particle physics and cosmology. This discovery may come either from experiments attempting to detect dark matter or from collider searches, or both.

- Why are there baryons, and no anti-baryons in our Universe?
  In other words, what is the origin of the matter–antimatter asymmetry of the Universe? This also appears to be a less difficult problem; we shall discuss this issue later in these lectures. Here we note only that any solution to this problem requires an extension of the Standard Model.

- Why is the mass density of the non-baryonic dark matter so similar (within less than an order of magnitude) to the mass density of baryons?
  Both these densities scale as $a^{-3}(t)$, so their ratio stays constant during most of the evolution of the Universe. It is not inconceivable that mechanisms which created baryons and dark matter particles in the early Universe are related to each other, so that the approximate equality of the mass densities is not a mere coincidence. It is, however, difficult to construct a corresponding particle physics model, and it is fair to say that existing attempts are far from being compelling.

- What is the origin of dark energy? If this is vacuum, why does the vacuum have non-zero energy density, which, however, is very small by particle physics standards?
  This is a very fundamental problem of microscopic physics. In natural units, the vacuum energy density is about $\rho_c \sim 10^{-46}$ GeV$^4$ while on dimensional grounds one would expect values like $1$ GeV$^4$ (QCD scale) or $10^8$ GeV$^4$ (electroweak scale). This enormous discrepancy cries out for an explanation, but despite numerous attempts it remains essentially an open problem. It may very well be that the solution of this ‘cosmological constant problem’ will lead to an entirely new concept of physics at ultra-large distances, or an entirely new concept of the history of the Universe, or both. On the other hand, a very small value of the cosmological constant is crucial for our existence, so the ‘solution’ of the cosmological constant problem may well be anthropic, see below.

- Why now?
  The energy densities of non-relativistic dark matter and dark energy scale differently: the for-
mer scales like $a^{-3}(t)$ while the latter stays approximately constant. Hence, at small $a(t)$ (early Universe) the energy density of non-relativistic matter exceeded by far the dark energy density. Conversely, future expansion of the Universe will be dominated by dark energy. Yet these energy densities are of the same order of magnitude today. Why is this the case? What is special about the present epoch of the evolution of the Universe?

### 5 Dark energy

It appears that by far the most difficult problem is the origin of dark energy. The most disappointing possibility would be that the carrier of dark energy is vacuum; in that case we shall hardly ever be confident of a mechanism responsible for tiny, but non-zero vacuum energy density. As a last resort, we shall possibly have to rely upon anthropic considerations [28, 29], which are based on the observation that if the vacuum energy density were substantially larger (in absolute value) than observed, the Universe would not be suitable for the existence of observers like us. Indeed, a large negative cosmological constant would give rise to early recollapse of the Universe, while with a large positive value, the accelerated expansion would start much earlier. In either case there would not be enough time for stars and galaxies to form, hence life would not develop. These considerations in fact provide rather strong bounds on the vacuum energy density, within two orders of magnitude of the observed value. The idea is then that there may be infinitely many regions in the Universe (or even infinitely many universes) where fundamental parameters like vacuum energy density are different, and span an entire range $(-\infty, +\infty)$. The observers like us can only find themselves in a region where the values of these parameters are suitable for their existence. There have been various suggestions for how such a picture can occur, ranging from wormholes/branching universes to ‘eternal inflation’ and, most recently, to string theory landscape [30]. The problem is that it will hardly be possible to check this picture experimentally even in the distant future.

Another option, more promising from an observational viewpoint, is that dark energy is due to a light field. In many models, this is a scalar field [31–33], dubbed quintessence. This scalar field must have a very flat scalar potential. We shall see towards the end of these lectures that under some circumstances, the energy density of a scalar field can indeed evolve very slowly, thus giving rise to the accelerated expansion of the Universe. In this set of models the energy density decreases in time, so that the effective $w$ parameter is negative,

$$\text{quintessence} : \quad p > -\rho \ , \ w > -1 \ .$$

The opposite case is often called ‘phantom energy’; for phantom, the energy density increases in time, and

$$\text{phantom} : \quad p < -\rho \ , \ w < -1 \ .$$

Phantom energy is way more theoretically exotic than quintessence; it is most likely that it exists only in theories with spontaneously broken Lorentz invariance. In either case, the small value of the energy density today requires extremely small values of the parameters of the quintessence or phantom models (masses and coupling constants), thus making these models quite unnatural.

Yet another option for explaining the accelerated expansion of our Universe is that gravity deviates from General Relativity at cosmological distances and time scales, so that the Friedmann equation (22) merely is not valid at the present epoch. This option would probably have a similar naturalness problem as the quintessence proposal, but even before that one meets a serious problem of constructing a theoretically consistent and phenomenologically acceptable theory which would reduce to General Relativity at distances from a fraction of a millimetre (down to which gravity is experimentally known to obey Newton’s law) to at least tens of megaparsecs, and deviate from General Relativity at cosmological scales. In known Lorentz-invariant examples of such a theory\(^\text{13}\), there either exist ghosts (fields with

\(^{13}\)See, however, Ref. [34].

219
negative energy unbounded from below) or gravity becomes strongly coupled at the quantum level (and
due to Lorentz invariance, but even this, rather exotic idea, has not yet led to a consistent
model capable of explaining the accelerated expansion of the Universe.

Quintessence/phantom models (and most likely models with infrared-modiﬁed gravity, if the latter
exist) imply that the effective dark energy density is not constant in time. Needless to say, an observa-
tional evidence for the time-dependence of $\rho_\Lambda$ would have enormous impact on fundamental physics. On
the other hand, it is hard to foresee any way of probing dark energy in a laboratory, so we have to rely
solely on cosmological observations when trying to reveal the origin of dark energy.

6 Hot Big Bang

The usual matter and dark matter have their origin in the early Universe. Before discussing plausible
scenarios of their generation, let us give some details of the evolution of the Universe in its hot stage.

One assumes that it makes sense to extrapolate the evolution of the Universe back in time by
making use of known laws of microscopic physics (nuclear physics, quantum ﬁeld theory, including
electrodynamics, QCD and electroweak physics) and General Relativity. This theory is called the ‘Hot
Big Bang’ theory. As we have seen, the Universe was hotter at earlier stages (i.e., at smaller values of
the scale factor $a(t)$), as the temperature scales as $a^{-1}(t)$. Also, the Universe was denser: the particle
number densities scale as $a^{-3}(t)$ both for relativistic and non-relativistic particles. At high enough
temperature the Universe was quite different from what we observe today: instead of almost empty space
with galaxies here and there, there was hot, dense and almost homogeneous plasma ﬁlling the Universe.
This is why microscopic physics played a role in the early Universe: at temperatures of the order of the
nuclear physics scale, roughly a few MeV, one has to deal with nuclear reactions; at temperatures of
the order of the strong interaction scale, (a few) · 100 MeV, the relevant microscopic theory was largely
QCD, etc.

6.1 Expansion of the Universe

We begin with the analysis of the evolution of the Universe, i.e., the behaviour of the scale factor $a(t)$
as a function of time. As we already discussed in Section 2.6, at early times the Universe was radiation
dominated, then matter dominated, and currently dark-energy dominated, while the curvature term $\kappa/a^2$
was never important.

6.1.1 Deceleration to acceleration

Owing to the dominance of dark energy, the Universe accelerates today. When matter was dominating,
the Universe was decelerating. To work out when the change in the regime occurred, we write down the
Friedmann equation in the following form (assuming dark matter equation of state $p = -\rho$, neglecting
spatial curvature and also neglecting ultra-relativistic matter for the moment; we shall see that ultra-
relativistic matter was important for the expansion at a much earlier stage),

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3} \rho_c \left(\Omega_M \frac{a_0^3}{a^3} + \Omega_\Lambda\right)$$

where $a_0$, $\Omega_M$, $\Omega_\Lambda$ and $\rho_c$ are the present values, thus time-independent constants. Therefore,

$$\ddot{a}^2 = \frac{8\pi G}{3} \rho_c \left(\Omega_M \frac{a_0^3}{a} + \Omega_\Lambda \dot{a}^2\right)$$

and $\ddot{a}$ is equal to zero when

$$\frac{a_0^3}{a^3} \equiv (1 + z)^3 = \frac{2\Omega_\Lambda}{\Omega_M}$$
With $\Omega_\Lambda = 0.7$ and $\Omega_M = 0.3$, we have
\[
\text{deceleration } \rightarrow \text{ acceleration : } z \approx 0.7 .
\]
The Universe was decelerating until fairly recently. Before $z \approx 0.7$, the expansion was dominated by non-relativistic matter.

### 6.1.2 Radiation domination to matter domination

Since energy densities of non-relativistic and ultra-relativistic matter (‘radiation’) scale as $a^{-3}$ and $a^{-4}$, respectively, the dominant contribution to the energy density of the Universe at very small $a$, i.e., at a very early epoch, came from ultra-relativistic matter. To estimate $z_{\text{eq}}$ at which the equilibrium between matter and radiation occurred, i.e., at which the expansion regime changed from the dominance of ultra-relativistic particles to the dominance of non-relativistic matter, we write
\[
\frac{\rho_M(t)}{\rho_{\text{rad}}(t)} = \left( \frac{\rho_M}{\rho_{\text{rad}}} \right)_0 \frac{a(t)}{a_0}
\]
where the subscript 0 still refers to the present values. Equilibrium occurs at
\[
\frac{\rho_M(t_{\text{eq}})}{\rho_{\text{rad}}(t_{\text{eq}})} \approx 1
\]
which gives
\[
\frac{a_0}{a(t_{\text{eq}})} = 1 + z_{\text{eq}} \approx \left( \frac{\rho_M}{\rho_{\text{rad}}} \right)_0 = \frac{\Omega_M}{\Omega_{\text{rad}}}
\]
We already know the energy density of relic photons; massless neutrinos\(^{14}\) of all three types have $\rho_{\nu,0} \approx 0.7 \rho_{\gamma,0}$. Thus, $\Omega_{\text{rad}} \approx 10^{-4}$, see Eq. (46). With $\Omega_M = 0.3$ we have
\[
\text{radiation domination } \rightarrow \text{ matter domination : } z_{\text{eq}} \approx 3000 .
\]
The corresponding temperature is
\[
T_{\text{eq}} = T_0 (1 + z_{\text{eq}}) \approx 10^4 \text{ K} \approx 1 \text{ eV} .
\]
(49)

At higher temperatures, the expansion of the Universe was dominated by ultra-relativistic matter.

It is important for the theory of structure formation that during much of its lifetime, the Universe was dominated by non-relativistic matter. The expansion rate at both radiation-dominated and vacuum-dominated stages is such that gravitational perturbations grow slowly, and only during the matter-dominated stage is their growth fast enough to account for the existing structures in the Universe. The bottom line is that the present composition of the Universe plus simple extrapolations back to the past are consistent with the theory of structure formation. Various ingredients of standard cosmology fit nicely together.

### 6.2 Epochs in the early Universe

We have already mentioned two important epochs in the evolution of the Universe: the **recombination epoch** (transition from plasma to neutral gas that occurred at $T \sim 3000$ K, $t \sim 3 \cdot 10^5$ yrs and lasted much less than Hubble time) and the **nucleosynthesis epoch** ($T = 1$ MeV to (a few) $\cdot 10$ keV). Another ‘event’ is **neutrino decoupling**. At high temperatures, weakly interacting particles, including neutrinos, were in thermal equilibrium with the rest of cosmic plasma. The plasma became effectively transparent

\(^{14}\)Whether or not neutrinos are exactly massless is inessential for the estimate of $z_{\text{eq}}$: the estimate remains valid if every neutrino has mass smaller than $1 \text{ eV}$.
for neutrinos at a temperature of about 1 MeV. The temperature of decoupling of neutrinos from cosmic plasma is of importance for nucleosynthesis, as it affects the neutron–proton ratio just before nucleosynthesis (and hence the abundances of light elements, which need neutrons for their formation), and also the expansion of the Universe at the nucleosynthesis epoch. The fact that neutrinos decoupled much earlier than photons implies that the present neutrino-to-photon ratio is less than one\textsuperscript{15}.

As we move further back in time, the cosmic plasma has more and more components. At temperatures above roughly 0.5 MeV (set by the mass of the electron), there are lots of electrons and positrons which frequently are pair created and annihilate; at $T > 100$ MeV the plasma contains muons and pions, etc. Simple estimates given in the next subsection show that the plasma remains in thermal equilibrium except possibly for phase transitions.

- QCD phase transition.
  At temperatures well above 100 MeV (QCD scale), strongly interacting particles are dissolved into quarks and gluons. This quark–gluon matter converts into hadronic matter (mostly pions) during the quark–hadron phase transition. Theoretical estimates and lattice simulations in QCD suggest that the temperature of this phase transition is about 170 MeV.

- Electroweak transition.
  Loosely speaking, at temperatures well above 100 GeV, electroweak symmetry is unbroken, the Higgs expectation value vanishes, and $W$ and $Z$ bosons are massless. At $T \sim 100$ GeV, the phase transition of the electroweak symmetry breaking takes place. This picture needs qualification, however: there is no local, gauge-invariant order parameter in the standard electroweak theory, and the electroweak transition is similar to vapour–liquid transition. For some values of parameters, there is the first-order phase transition, whereas for other values, a smooth cross-over takes place instead. With existing constraints on the Higgs boson mass, the \textit{Standard Model} predicts cross-over; what actually happened in the early Universe depends on what exactly is the extension of the Standard Model. Uncovering physics in the 100 GeV–1 TeV energy range will thus allow cosmologists to develop a quantitative theory of a quite early epoch of the evolution of the Universe.

- GUT transition.
  Extrapolating further back is dangerous, but if we do so, we come to the Grand Unification epoch, whose temperature is set by the GUT scale, $T_{\text{GUT}} \sim 10^{16}$ GeV. At this temperature, one expects Grand Unified phase transition to occur. However, most models of inflation suggest that the Universe never had such a high temperature after inflation.

- Even more interesting are the epoch of the generation of dark matter and the epoch of the generation of baryon asymmetry. We can only make guesses about these epochs, and some of the guesses are the subject of these lectures. Before turning to them, it is convenient to consider the expansion of the Universe at early times in a little more detail.

### 6.3 Expansion rate and lifetime at radiation domination

Before proceeding further, let us consider in a little more detail the expansion of the Universe at the radiation-dominated stage, assuming thermal equilibrium of all ultra-relativistic species\textsuperscript{16}. The energy density of all ultra-relativistic species, which enters the Friedmann equation, is

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

\textsuperscript{15}This is because photons are additionally heated, after neutrino decoupling, owing to annihilations of $e^+$ and $e^-$.\textsuperscript{16}The assumption of thermal equilibrium is in fact not valid for neutrinos at temperatures below 1 MeV. The corresponding modification of our discussion is straightforward, however.
where $g_*$ is the effective number of massless degrees of freedom at temperature $T$. The contribution of bosons into $g_*$ is equal to the number of spin states (e.g., for photons $g_\gamma = 2$, while for W bosons at temperatures above 100 GeV, $g_W = 6$ because of two charges and three projections of spin), while fermions contribute $7/8$ of the number of spin states (electrons plus positrons contribute $4 \cdot 7/8$, each type of left-handed neutrino plus its antineutrino gives $2 \cdot 7/8$, etc.). The parameter $g_*$ is the sum of contributions of all ultra-relativistic species; it slightly depends on time because at higher temperatures, more species are ultra-relativistic (say, electrons contribute at $T > 0.5$ MeV and do not contribute at lower temperatures).

It is convenient to introduce the effective Planck mass

$$M^*_\text{Pl} = \frac{M_{\text{Pl}}}{1.66 \sqrt{g_*}}.$$  

This parameter slightly depends on temperature, and numerically is of order

$$M^*_\text{Pl} = (\text{a few}) \cdot 10^{18} \text{ GeV}.$$  

With this notation, the Friedmann equation tells us that the expansion rate is related to temperature in a simple way,

$$H(t) = \frac{T^2(t)}{M^*_\text{Pl}}.$$  

One recalls that the expansion law at the radiation-dominated stage (neglecting the dependence of $g_*$ on temperature) is

$$a(t) = \text{const} \cdot \sqrt{t}$$  

so the Hubble parameter is related to the lifetime as follows:

$$H = \frac{\dot{a}}{a} = \frac{1}{2t}.$$  

We immediately deduce the relation between lifetime and temperature,

$$t = \frac{M^*_\text{Pl}}{2T^2}.$$  

Let us make use of the latter formula to estimate the age of the Universe at different epochs:

- Nucleosynthesis.
  The temperatures relevant for BBN range from a few MeV to about 70 keV. From these we obtain that the earliest time directly probed by observations is about

$$t \sim \frac{10^{18} \text{ GeV}}{10^{-6} \text{ GeV}^2} \sim 1 \text{ s}$$  

whereas BBN ends at $t \sim 200 \text{ s} \sim 3 \text{ min}$. We do have an idea of what the Universe was like at one second old!

- Earlier epochs:

\[\begin{align*}
\text{QCD phase transition} & : T \sim 200 \text{ MeV} , \quad t \sim 3 \cdot 10^{-5} \text{ s} \\
\text{Electroweak epoch} & : T \sim 100 \text{ GeV} , \quad t \sim 10^{-10} \text{ s} .
\end{align*}\]  

(50)

One may wonder whether equilibrium thermodynamics, which we use throughout, is applicable at these early times, when the Universe expands so rapidly. To see that this is indeed the case, let us consider, as an example, electromagnetic scattering of light particles at $T > 1$ MeV. It is clear on
dimensional grounds that the mean free time of a relativistic charged particle at temperature $T$ exceeding its mass is

$$\tau \sim (\alpha^2 T)^{-1}$$

(the electromagnetic cross-section is proportional to $\alpha^2$). For thermal equilibrium with respect to the electromagnetic interactions to be established, the interaction rate $\tau^{-1}$ must be smaller than the expansion rate of the Universe $H$ which gives

$$\alpha^2 T \gg \frac{T^2}{M_{\text{Pl}}}.$$  

This inequality is indeed valid at $T \ll \alpha^2 M_{\text{Pl}}^* \sim 10^{14}$ GeV, so electromagnetic (and many other) microscopic processes are in thermal equilibrium at all temperatures of interest to us.

Thermal equilibrium is not a particularly interesting state of the Universe. What we are going to discuss in the rest of this section are in fact *inequilibrium* phenomena. It is these phenomena that may leave relics behind, and hence may have observable consequences.

### 6.4 Heavy relic: best guess for cold dark matter

As we discussed above, the observational data strongly suggest that a good part of the energy density in the present Universe is due to new stable (or practically stable) particles. At least in some cases, the Hot Big Bang theory is capable of predicting the density of such particles in terms of their interaction cross-sections and masses. Here we present the corresponding estimate in the simplest possible scenario; needless to say, this estimate can be (and has been) refined by more careful calculations.

Let us assume that there exists a heavy stable particle $Y$ (weakly interacting massive particle, WIMP) and its anti-particle $\overline{Y}$. Let us assume for definiteness that the dominant process in which these particles can be destroyed or created is their pair-annihilation or creation, with annihilation products being the particles of the Standard Model (the analysis for a neutral particle $Y$, which coincides with its own anti-particle, is very similar, provided that these particles are created and annihilate in pairs only, as is the case for the neutralino). Let us further assume that there is no asymmetry between $Y$ and $\overline{Y}$ in the early Universe, i.e., the densities of $Y$ and $\overline{Y}$ are equal to each other$^{17}$. We shall see that the overall cosmological behaviour of these particles is as follows. At high temperatures, $T \gg m_Y$, the $Y$ particles are in thermal equilibrium with the rest of the cosmic plasma; there are lots of $\overline{Y}$–$Y$ pairs in the plasma, which are copiously created and annihilate. As the temperature drops below $m_Y$, the equilibrium number density decreases. At some `freeze-out' temperature $T_f$ the number density becomes so small that $Y$ and $\overline{Y}$ can no longer meet each other during the Hubble time, and their annihilation terminates. The pair-creation of $Y$ and $\overline{Y}$ terminates too, as there are too few Standard Model particles in plasma whose energy exceeds $m_Y$. After that the number density of survived $Y$ and $\overline{Y}$ decreases like $a^{-3}$, and these relic particles contribute to the mass density in the present Universe. Our purpose is to estimate the range of properties of $Y$ particles, in which their present mass density is of the order of the critical density $\rho_c$, so that $Y$ particles serve as dark matter candidates.

Assuming thermal equilibrium, elementary considerations of the mean free path of a particle with respect to the annihilation give for the lifetime of a non-relativistic $Y$ particle in cosmic plasma, $\tau_{\text{ann}}$, the following relation,

$$\sigma_{\text{ann}} \cdot v \cdot \tau_{\text{ann}} \cdot n_{\overline{Y}} \sim 1$$

where $v$ is the velocity of the $Y$ particle, $\sigma_{\text{ann}}$ is the annihilation cross-section at this velocity, and

$^{17}$This is actually a strong assumption. It is valid in many, but not all, realistic extensions of the Standard Model. In fact, an alternative scenario with the generation of $Y$ asymmetry is appealing too, because it might relate baryon asymmetry to the density of dark matter [35].
Let us assume for definiteness that the annihilation occurs in $s$-wave (other cases give similar, but not exactly the same results), so at non-relativistic velocities

$$\sigma_{\text{ann}} = \frac{\sigma_0}{v},$$

where $\sigma_0$ is a constant about which we shall have more to say later. One should compare the lifetime with the Hubble time, or annihilation rate $\Gamma_{\text{ann}} \equiv \tau_{\text{ann}}^{-1}$ with the expansion rate $H = T^2/M_{\text{Pl}}$. At $T \sim m_Y$, the equilibrium density is of order $n_Y \sim T^3$, and $\Gamma_{\text{ann}} \gg H$ for not too small $\sigma_0$. This means that the annihilation (and, by reciprocity, creation) of $\bar{Y}Y$ pairs is indeed rapid, and $Y$ particles are indeed in thermal equilibrium with the plasma. At very low temperature, on the other hand, the number density $n_Y$ is exponentially small, and $\Gamma_{\text{ann}} \ll H$. At low temperatures we cannot, of course, make use of equilibrium formulas: $Y$ particles no longer annihilate (and are no longer created), there is no thermal equilibrium with respect to creation–annihilation processes, and the number density $n_Y$ gets diluted only because of the cosmological expansion.

The freeze-out temperature $T_i$ is determined by the relation

$$\tau_{\text{ann}}^{-1} \equiv \Gamma_{\text{ann}} \sim H$$

where we can still use the equilibrium formulas, as $Y$ particles are in thermal equilibrium (with respect to annihilation and creation) just before freeze-out. We find

$$\sigma_0 \cdot n_Y(T_i) \sim \frac{T_i^2}{M_{\text{Pl}}^s}$$

or

$$\sigma_0 \cdot g_Y \cdot \left(\frac{m_Y T_i}{2\pi}\right)^{3/2} e^{-\frac{m_Y}{T_i}} \sim \frac{T_i^2}{M_{\text{Pl}}^s}.$$

The latter equation gives the freeze-out temperature, which, up to loglog terms, is

$$T_i \approx \frac{m_Y}{\ln(M_{\text{Pl}}^s m_Y \sigma_0)}.$$

Note that this temperature is quite a bit smaller than $m_Y$, if the relevant microscopic mass scale is much below $M_{\text{Pl}}$. This means that $Y$ particles freeze out when they are indeed non-relativistic, hence the term ‘cold dark matter’. The fact that the annihilation and creation of $Y$ particles terminate at relatively low temperature has to do with a rather slow expansion of the Universe, which should be compensated for by the smallness of the number density $n_Y$.

At the freeze-out temperature, we make use of Eq. (51) and obtain

$$n_Y(T_i) = \frac{T_i^2}{M_{\text{Pl}}^s \sigma_0}.$$

Note that this density is inversely proportional to the annihilation cross-section (up to logarithms). The reason is that for a higher annihilation cross-section, the creation–annihilation processes are no longer in equilibrium, and fewer $Y$ particles survive.

To estimate the present density of $Y$ particles, it is convenient to consider the ratio $n_Y/s$ where $s$ is the entropy density,

$$s = \frac{2\pi^2}{45} g_* T^3.$$
The point is that during the adiabatic expansion after freeze-out, both entropy density and \( n_Y \) behave as \( a^{-3} \), so this ratio stays constant. Up to a factor of order 1, this ratio at freeze-out is

\[
\frac{n_Y}{s} \sim \frac{1}{g_*(T_f) M_{\text{Pl}}^2 T_f \sigma_0}.
\]

At late times, the entropy density, again up to a factor of order 1, is equal to the number density of photons, so the present number density of \( Y \) particles is of order

\[
n_Y,0 \sim n_{\gamma,0} \cdot \left( \frac{n_Y}{s} \right)_{\text{freeze-out}}
\]

and the mass density is

\[
\rho_Y,0 = m_Y n_Y,0 \sim n_{\gamma,0} \cdot \frac{\ln(M_{\text{Pl}}^2 m_Y \sigma_0)}{g_*(T_f) M_{\text{Pl}}^2 \rho_{\text{DM}}}.
\]

(52)

This formula is remarkable. The mass density depends mostly on one parameter, the annihilation cross-section \( \sigma_0 \). The dependence on the mass of the \( Y \) particle is through the logarithm and \( g_*(T_f) \), and is very mild. From this formula we immediately derive the condition ensuring that \( Y \) particles are dark matter candidates, i.e., their present mass density is of the order of \( \rho_{\text{DM}} = 0.25 \rho_c \),

\[
\sigma_0 \sim \frac{n_{\gamma,0}}{g_*(T_f) M_{\text{Pl}}^2 \rho_{\text{DM}}} \ln(M_{\text{Pl}}^2 m_Y \sigma_0).
\]

The value of the logarithm here is between 20 and 40, depending on parameters (this means, in particular, that freeze-out occurs when the temperature drops 20 to 40 times below the mass of the \( Y \) particle). Plugging in other numerical values (\( \rho_{\text{DM}} \sim 10^{-6} \text{ GeV cm}^{-3} \), \( n_{\gamma,0} \sim 400 \text{ cm}^{-3} \), \( g_*(T_f) \sim 100 \), \( M_{\text{Pl}}^2 \sim 10^{18} \text{ GeV} \)), we obtain an estimate

\[
\sigma_0 \sim 10^{-10} \text{ GeV}^{-2}
\]

which already crudely tells us what the relevant range of mass scales is. In fact, the annihilation cross-section may be parametrized as

\[
\sigma_0 = \frac{\alpha^2}{\mathcal{M}^2}
\]

where \( \alpha \) is some coupling constant, and \( \mathcal{M} \) is the mass scale. This parametrization is suggested by the picture of \( \bar{Y} - Y \) annihilation via exchange of another particle of mass \( \mathcal{M} \), which may be somewhat higher than \( m_Y \). With \( \alpha \sim 1/30 \) (SU(2)_W gauge coupling) the estimate for the mass scale is roughly

\[
\mathcal{M} \sim 1 \text{ TeV}.
\]

Thus with very mild assumptions, we find that the non-baryonic dark matter may naturally originate from the TeV-scale physics!

In supersymmetric extensions of the Standard Model, the lightest supersymmetric particle (LSP, most likely, neutralino—a mixture of superpartners of the photon, \( Z \) boson, and neutral Higgs bosons) is often stable, and its annihilation cross-section is automatically in the right ballpark. This is illustrated in Fig. 11.

Naturally, the search for both direct and indirect signals from neutralino dark matter (and more generally, WIMPs) is an active area of experimental research. For discussions of the potential of existing and future experiments, see, for example, Refs. [36, 37].

If dark matter particles are indeed WIMPs, and the relevant energy scale is of order 1 TeV, then the Hot Big Bang theory will be probed experimentally up to temperatures of (a few) \( \cdot 10 \text{ GeV} \) and down to

226
The mechanism discussed here is by no means the only mechanism capable of producing cold dark matter, and WIMPs are by no means the only candidates for dark-matter particles. Other dark-matter candidates include very heavy relics produced towards the end of inflation, axions, gravitinos, Q-balls, massive gravitons, etc. Unlike WIMPs, however, the right dark-matter density in other scenarios comes out by tuning either microscopic parameters (masses, coupling constants of new particles) or the parameters of the early Universe (e.g., maximum temperature reached after inflation), or both. Because of this, the WIMP scenario may be considered as the best guess for the origin of dark matter.

6.5 Baryon asymmetry of the Universe

In the present Universe, there are baryons and almost no anti-baryons. The number density of baryons today is characterized by the ratio $\eta$, see Eq. (18). In the early Universe, the appropriate quantity is

$$\Delta_B = \frac{n_B - n_{\bar{B}}}{s}$$
where $n_{\bar{B}}$ is the number density of anti-baryons, and $s$ is the entropy density. If the baryon number is conserved, and the Universe expands adiabatically, $\Delta B$ is constant, and its value is, up to a numerical factor, equal to $\eta$, so that

$$\Delta B \approx 10^{-10}.$$ 

Back at early times, at temperatures well above 100 MeV, cosmic plasma contained many quark–antiquark pairs, whose number density was of the order of the entropy density,

$$n_q + n_{\bar{q}} \sim s$$

while baryon number density was related to densities of quarks and antiquarks as follows (baryon number of a quark equals $1/3$),

$$n_B = \frac{1}{3}(n_q - n_{\bar{q}}).$$

Hence in terms of quantities characterizing the very early epoch, the baryon asymmetry may be expressed as

$$\Delta B \sim \frac{n_q - n_{\bar{q}}}{n_q + n_{\bar{q}}}.$$ 

We see that there was one extra quark per about 10 billion quark–antiquark pairs! It is this tiny excess that is responsible for the entire baryonic matter in the present Universe.

There is no logical contradiction to suppose that the tiny excess of quarks over antiquarks was built in as an initial condition. This is not at all satisfactory for a physicist, however. Furthermore, the inflationary scenario does not provide such an initial condition for the Hot Big Bang; rather, inflation theory predicts that the Universe was baryon-symmetric just after inflation. Hence, one would like to explain the baryon asymmetry dynamically.

The baryon asymmetry may be generated from an initially symmetric state only if three necessary conditions, dubbed Sakharov’s conditions, are satisfied. These are

(i) baryon number non-conservation;
(ii) C- and CP-violation;
(iii) deviation from thermal equilibrium.

All three conditions are easily understood. (i) If the baryon number were conserved, and assuming the initial net baryon number in the Universe equal to zero, the Universe today would be symmetric rather than asymmetric. (ii) If C or CP were conserved, then the rates of reactions with particles would be the same as the rates of reactions with antiparticles. In other words, if the initial state of the Universe was C- and CP-symmetric, then the asymmetry between particles and antiparticles may develop only if C and CP are violated. (iii) Thermal equilibrium means that the system is stationary (no time dependence at all). Hence, if the initial baryon number is zero, it is zero forever, unless there are deviations from thermal equilibrium.

There are two well-understood mechanisms of baryon number non-conservation. One of them emerges in Grand Unified Theories and is due to the exchange of super-massive particles. It is very similar, say, to the mechanism of charm non-conservation in weak interactions which occurs via the exchange of heavy W bosons. The scale of these new, baryon-number-violating interactions is the Grand Unification scale, presumably of order $10^{16}$ GeV.

Another mechanism is non-perturbative and exists already in the Standard Model, and, possibly with slight modifications, operates in all of its extensions (see Ref. [38] for details). Baryon and lepton currents $j_B^\mu$ and $j_L^\mu$ have electroweak anomalies, which arise due to quantum effects,

$$\partial^\mu j_B^\mu = \partial^\mu j_L^\mu = 3\frac{\alpha_W}{8\pi} \epsilon^{\mu\nu\lambda\rho} F_{\mu\nu}^a F_{\lambda\rho}^a.$$
where $\alpha_W = g_W^2/(4\pi)$ is the SU(2)$_W$ gauge coupling, $F_{\mu\nu}^a$ is the SU(2)$_W$ gauge field ($a = 1, 2, 3$) and $\epsilon^{\mu\nu\lambda\rho}$ is the antisymmetric tensor. Integrating these relations over space–time, one finds that the changes of baryon and lepton numbers are related to a space–time integral involving the gauge field,

$$\Delta B = \Delta L = 3 \frac{\alpha_W}{8\pi} \int d^4 x \epsilon^{\mu\nu\lambda\rho} F_{\mu\nu}^a F_{\lambda\rho}^a .$$

The right-hand side of this equation may be non-zero for fluctuations of the gauge fields whose field strengths are large

$$F_{\mu\nu}^a \propto \frac{1}{g_W} .$$

Obviously, such fluctuations, being inversely proportional to the coupling constant, cannot be described within the perturbation theory; electroweak baryon number non-conservation is indeed invisible in Feynman diagrams. In vacuo, these fluctuations are instantons; the probability for them to happen is extremely small, so electroweak processes violating $B$ and $L$ practically do not take place. At high temperatures there are thermal fluctuations (a keyword here is sphaleron), and they are quite probable. By comparing the rate of electroweak baryon-number violating processes with the Hubble rate one sees that these processes are fast in the early Universe at temperatures $T$ obeying

$$T < 10^{11} \text{ GeV}$$

and

$$T > v(T)$$

where $v(T)$ is the expectation value of the Higgs field, which in turn depends on temperature. At $T \approx v(T)$ electroweak baryon number violation turns off in the cosmic plasma.

As seen from Eq. (53), electroweak processes do not conserve $B$ and $L$ separately, but their combination is not violated,

$$B - L = \text{ conserved} .$$

In fact, the complete set of conserved quantum numbers is $(B - 3L_e)$, $(B - 3L_v)$ and $(B - 3L_\tau)$, but we shall not need this qualification in what follows.

Realistic mechanisms of baryon number non-conservation are thus not numerous, yet there are several ways the baryon asymmetry could have been generated. They differ by the characteristic temperature at which the asymmetry is produced.

(i) **Grand Unification mechanisms** operate at extremely high temperatures, $T \sim M_{\text{GUT}} \sim 10^{15} - 10^{16}$ GeV. The most commonly discussed source of the baryon asymmetry in this context are $B$- and CP-violating decays of ultra-heavy particles. At later times, the baryon number is violated by anomalous electroweak processes whose effect is basically to wash out $(B + L)$. They would therefore reprocess part of the baryon asymmetry, but if non-zero $(B - L)$ is generated at GUT temperatures, then this $(B - L)$ would survive until the present epoch (provided there are no strong lepton-number-violating interactions at intermediate temperatures, $100 \text{ GeV} < T < 10^{11} \text{ GeV}$, otherwise all fermion quantum numbers would be violated at those temperatures, and no asymmetry would survive). Part of this $(B - L)$ would be carried by baryons.

It is unlikely that GUT baryogenesis operated at the hot stage of the evolution of the Universe, after post-inflationary reheating. In the vast majority of inflationary models, the reheat temperature is well below $M_{\text{GUT}} \sim 10^{15} - 10^{16}$ GeV. However, it is not unrealistic that GUT baryogenesis operated during the reheating epoch, which occurred between inflation and the hot stage. Unfortunately, there does not seem to exist any direct or indirect observational test of this scenario.\footnote{It is important here that the SU(2)$_W$ gauge theory is non-Abelian, and that only left-handed fermions participate in the SU(2)$_W$ interactions.}
(ii) *Electroweak baryogenesis* is a scenario in which the baryon asymmetry is generated entirely due to the anomalous electroweak processes (53). Its generation would occur at a temperature of order 100 GeV, at which these anomalous processes are switching off.

Since the Universe expands slowly at the electroweak epoch (as compared to the rates of microscopic interactions), considerable departure from thermal equilibrium is possible due only to the first-order phase transition.

At low temperatures, the scalar potential has the shape shown in Fig. 12(a), so the Higgs field has a non-vanishing expectation value. At high temperatures, the relevant quantity is the effective potential $V_{\text{eff}}(\phi, T)$, which is defined as the free energy density at temperature $T$ in the presence of the homogeneous scalar field $\phi$. As often happens in condensed matter systems, symmetry is restored at high temperatures, i.e., the effective potential has the shape shown in Fig. 12(b).

Symmetry breaking in principle can proceed via phase transitions of the first and second order, or may occur as a smooth cross-over\(^\text{19}\). In the case of the first-order phase transition, the value of $v(T) \equiv \langle \phi \rangle_T$ jumps at a certain critical temperature, as shown in Fig. 13(a). This distinguishes the first-order phase transition from the second-order phase transition, Fig. 13(b). The reason behind this difference is that the effective potential changes with temperature as shown in Fig. 14: in the case of the

\(^{19}\)In the latter case there is no order parameter discriminating the phases, so one cannot, in fact, talk about symmetry breaking. A well-known condensed matter example is the liquid–gas transition which is first order at low pressure and a smooth cross-over at high pressure.
first-order phase transition, in some range of temperatures there are two minima of the effective potential, the minimum with \( v \neq 0 \) being deeper below the critical temperature. In realistic situations (including the early Universe), the system supercools, i.e., the transition actually occurs when the depths of the two minima are not equal to each other, as indicated in Fig. 14(a). For the second-order phase transition the situation is different. The minimum of the effective potential continuously moves away from \( \phi = 0 \), there is nothing like supercooling, the system evolves smoothly.

The first-order phase transition cannot occur simultaneously everywhere in space: in that case the
system would have to climb the infinitely high potential barrier (recall that $V_{\text{eff}}$ is free energy density, so the free energy of the homogeneous scalar field away from the minimum of $V_{\text{eff}}$ is proportional to the total volume of the system). Instead, the transition proceeds via thermal nucleation of bubbles of the new phase, their growth and percolation, as sketched in Fig. 15. Obviously this is a highly non-equilibrium process.

To generate the baryon asymmetry during the phase transition, the following necessary requirements have to be fulfilled:

(a) The baryon numbers have to be efficiently violated before the phase transition.

(b) The baryon number violation has to be switched off after the phase transition (otherwise the baryon asymmetry generated during the transition would be washed out at later times when cosmic plasma is again close to the thermal equilibrium); in other words, the inequality (54) has to be valid after the phase transition.

(c) There should be enough CP-violation during the phase transition.

The requirement (a) is satisfied automatically (provided there is indeed the first-order phase transition). Whether requirements (b) and (c) are fulfilled depends on microscopic theory.

In the Standard Model, neither (b) nor (c) are satisfied. The strength of the phase transition in the Standard Model depends on its only unknown parameter, the Higgs boson mass. With the current limit, $m_H > 114$ GeV, there is no electroweak phase transition at all—the Standard Model exhibits smooth cross-over as the temperature decreases (see Ref. [39] and references therein). Also, CP-violation coming from the Kobayashi–Maskawa mechanism is too weak at temperatures of order 100 GeV to produce enough baryon asymmetry.

On the other hand, extensions of the Standard Model with additional relatively light scalars (including a corner of MSSM with light stop) work fine in an appropriate range of parameters. The interactions of the Higgs field with additional scalars often strengthen the phase transition, so that the requirement (a) can be satisfied. An extra source of CP-violation, needed to satisfy (c), comes from CP phases in the scalar sector. Interestingly, these phases induce electric dipole moments of the neutron and electron, which are typically 1 to 3 orders of magnitude below the current limits, i.e., are in a potentially observable range.

The electroweak generation of baryon asymmetry occurs when the extra particles are not very heavy, otherwise they would not have any effect on the Higgs effective potential at $T \sim 100$ GeV: the masses of at least some of these particles have to be in the range of 100–200 GeV. Clearly, this range is accessible at the LHC, so experiments in the near future will either confirm or reject the electroweak mechanism.

(iii) A currently popular mechanism is leptogenesis. It may occur at some intermediate temperature (the estimates range from $10^7$ GeV to $10^{11}$ GeV) due to $L$- and CP-violating decays of heavy Majorana neutrinos [40]. Of course, the generation of lepton asymmetry requires lepton number violation, i.e., extension of the Standard Model, but such an extension is favoured by neutrino oscillation data anyway. The lepton asymmetry would then be partially reprocessed into baryon asymmetry by anomalous electroweak processes. Interestingly, the range of Majorana masses of ordinary neutrinos compatible with this mechanism is indeed consistent with the range inferred from neutrino oscillations. Let us discuss the leptogenesis scenario in a little more detail.

Let us assume that ordinary neutrinos get their small masses via the see-saw mechanism. This mechanism invokes heavy ‘sterile’ (neutral with respect to the Standard Model gauge interactions) neutrinos $N_i$, where $i = 1, 2, 3$ is the generation number. They have large Majorana masses; in an appropriate basis the corresponding mass matrix is diagonal and real. They are assumed to have Yukawa interactions with conventional lepton doublets $L$ and the Standard Model Higgs field $H$. Thus, besides
Fig. 16: Diagrams for the decay $N_1 \to l_i h$. Note that no CP-asymmetry emerges at the tree level, so to the lowest non-trivial order, this asymmetry is given by the interference of tree and one-loop diagrams. In the latter, a summation over $k$ and $l$ is assumed.

In the kinetic terms, the Lagrangian involving $N_i$ contains two terms

$$L = \sum_i M_i \bar{N}_i N_i + \left( \sum_{ij} h_{ij} \bar{L}_i N_j \tilde{H} + h.c. \right)$$

where the fermions $N$ are right, the Yukawa couplings $h_{ij}$ are in general complex, and $\tilde{H}_\alpha = \epsilon_{\alpha\beta} H_\beta^*$ is a weak doublet whose vacuum expectation value is $(v, 0)$. Owing to this vacuum expectation value, there is mixing between ordinary neutrinos and $N$'s. Upon diagonalizing the mass matrix, one finds that there are heavy states, predominantly $N$'s, with masses $M_i$, and light states with mass matrix

$$m_{ij} = \sum_k h_{ik}^* h_{jk}^* \frac{v^2}{M_k}.$$  \hspace{1cm} (55)

These masses are naturally small for large $M_i$, which is the advantage of the see-saw mechanism. Obviously, the original Lagrangian does not conserve any of the lepton numbers.

Now, assume that at high temperatures, heavy neutrinos are in thermal equilibrium\(^{20}\). As the Universe cools down below $T \sim M_1$ (the smallest of $M_i$), the lightest of $N$'s (call it $N_1$) starts decaying\(^{21}\). There is also an inverse decay process which tends to keep $N_1$ in thermal equilibrium. The simplest version of the leptogenesis scenario assumes that $N_1$ is not in thermal equilibrium with respect to decay and inverse decay at $T \sim M_1$. This requires that its width be small enough,

$$\Gamma_1 < H(T \sim M_1).$$  \hspace{1cm} (56)

Now, since the Yukawa couplings contain CP-phases, there is an asymmetry in decays of $N_1$'s,

$$\delta \equiv \frac{\Gamma(N_1 \to LH) - \Gamma(N_1 \to \bar{L}\bar{H})}{\Gamma(N_1 \to LH) + \Gamma(N_1 \to \bar{L}\bar{H})} \neq 0.$$  

This asymmetry occurs because of the interference of the tree diagram and one-loop diagram in which all $N_i$ run in intermediate states, Fig. 16. This 'microscopic' asymmetry is

$$\delta = \frac{1}{8\pi} \frac{1}{(h^\dagger h)_{11}} \sum_{k=2,3} \text{Im} \left[ (h^\dagger h)_{1k}^2 \right] F \left( \frac{M_k}{M_1} \right)$$

\(^{20}\)This may be due either to some new interactions in which $N$’s participate, or to the Yukawa interactions themselves. In the latter case, thermal equilibrium is established with large enough Yukawa couplings only, i.e., the usual neutrinos must be sufficiently heavy, $m_\nu > 3 \times 10^{-3}$ eV. This is consistent with neutrino oscillation data.

\(^{21}\)Generically, the lightest of $N$’s is most relevant, since any asymmetry produced in decays of heavier $N$’s is washed out by interactions involving the lightest one.
where $F(M_1/M_k) \sim M_1/M_k$ for $M_1 \ll M_k$, up to a factor of order 1. Provided the out-of-equilibrium condition (56) is satisfied, the generated lepton asymmetry is of order

$$\frac{n_L}{s} \sim 0.01\delta$$

where the factor 0.01 is due to the large number of ultra-relativistic species at temperature $T \sim M_1$. A good part of this asymmetry gets reprocessed into the baryon asymmetry by electroweak processes, so the generated baryon asymmetry is roughly of the same order of magnitude as (57).

Let us first see what the out-of-equilibrium condition (56) means in terms of neutrino masses. Let $L_1$ be a linear combination of lepton fields to which $N_1$ is coupled (generally, this combination is neither a flavor nor a mass eigenstate),

$$L_1 \propto \sum_k h_{k1}^* L_k .$$

After rotating into the lepton basis whose first basis vector is $L_1$, the lightest $N$ produces the only term in the neutrino mass matrix,

$$m_{11} \bar{\nu}_1 \nu_1$$

where

$$m_{11} = (h_{11}^*)^2 \frac{v^2}{M_1} .$$

The width of $N_1$ is

$$\Gamma_1 = \frac{1}{16\pi} |m_{11}|^2 M_1 = \frac{|m_{11}|^2}{16\pi v^2} .$$

Let us now recall that $H(T \sim M_1) = M_1^2/M_{Pl}^2$, so the relation (56) gives

$$m_{11} < \frac{16\pi v^2}{M_{Pl}^2} .$$

Inserting numbers, one finds

$$m_{11} < 3 \cdot 10^{-3} \text{ eV} .$$

Miraculously, this number is in the right ballpark of neutrino masses, suggested by neutrino oscillations!

In fact, the situation is even better. If the out-of-equilibrium condition (56) is not satisfied, the resulting lepton (and hence baryon) asymmetry is suppressed roughly by a factor $H(T = M_1)/\Gamma_1$ (up to logarithm), so that

$$\Delta_B \sim \Delta_L \sim 0.01 \cdot \delta \cdot \frac{3 \cdot 10^{-3} \text{ eV}}{|m_{11}|} .$$

This means that the asymmetry is in the right ballpark even for $m_{11} \sim 0.05 \text{ eV}$, the value corresponding to $\Delta m^2$ determined by atmospheric neutrino oscillations.

The masses of heavy neutrinos must be quite large. For a crude estimate, let us assume that all these masses are of the same order of magnitude; let us also assume that all Yukawa couplings are of the same order. Then, assuming that CP phases are large, we have

$$\delta \sim \frac{1}{8\pi} h^2 .$$

We need $\delta > 10^{-8}$, so

$$\frac{1}{8\pi} h^2 \sim \frac{m_\nu M}{8\pi v^2} > 10^{-8} .$$

This gives for $m_\nu \sim 0.1 \text{ eV}$ the following estimate,

$$M > 10^8 \text{ GeV}$$
which is not unreasonable for the see-saw mechanism. The conclusion is that the mechanism described can indeed produce the right amount of the baryon asymmetry. Unfortunately, the CP phases responsible for the asymmetry are generally different from CP phases in the mass matrix of light neutrinos. So, future observation of CP violation in neutrino oscillations, although it will add weight to the leptogenesis scenario, it will not be proof of it.

The bottom line is that the observed baryon asymmetry may be explained by a number of mechanisms, all of which, however, exist in extensions of the Standard Model only. The problem is that, with the notable exception of electroweak baryogenesis, direct proof that any given mechanism is indeed responsible for the baryon asymmetry does not seem possible in the near future. For reviews of baryogenesis see, for example, Refs. [39, 41–43].

6.6 Remarks

We have seen in this section that extensions of the Standard Model can explain the dark matter and the baryon asymmetry of the Universe. Let us stress again, however, that the mechanisms outlined here do not explain the approximate equality between the baryon mass density and the dark-matter mass density in the present Universe,

$$\rho_{DM} \sim \rho_B , \quad \text{within a factor of } 5 .$$

Indeed, these mechanisms of the generation of baryon asymmetry and dark matter are completely unrelated to each other. The approximate relation (58) may well be a coincidence; still, a compelling model that would explain this coincidence is more than welcome.

Another remark is that some of the most appealing mechanisms for the generation of dark energy and baryon asymmetry operate at relatively low temperatures, $T \sim 10$–$100$ GeV. The corresponding extensions of the Standard Model contain new particles in the $100$ GeV–$1$ TeV mass range. Hence at least some of these ideas will be probed at the LHC. Phrased differently, there is a good chance that the LHC will open a window to the Universe at temperature $10$–$100$ GeV and age $10^{-8}$–$10^{-10}$ s.

The final remark is that the mechanisms that we discussed here operate at the hot stage of the cosmological evolution. Therefore, the resulting composition of the cosmic plasma is the same everywhere in the (visible part of the) Universe, i.e.,

$$\frac{\rho_B}{s} = \text{independent of } x \quad (59)$$

and

$$\frac{\rho_{DM}}{s} = \text{independent of } x . \quad (60)$$

The temperature itself (and hence entropy density), however, varies in space—these are density perturbations in the early Universe. Equations (59) and (60) set the initial data for the evolution of the cosmological perturbations at the hot stage,

$$\frac{\delta \rho_B(x)}{\rho_B} = \frac{\delta \rho_{DM}(x)}{\rho_{DM}} = \frac{\delta s(x)}{s} .$$

Perturbations with this property are called ‘adiabatic’; once baryon asymmetry and dark matter are generated at the hot stage, there is only the adiabatic mode in the density perturbations.

Off hand, there may be other modes, for example, with

$$\delta \rho_B(x) \neq 0 , \quad \delta T(x) = 0$$

or

$$\delta \rho_{DM}(x) \neq 0 , \quad \delta T(x) = 0 .$$
These are called baryon and dark matter isocurvature modes, respectively; they correspond to perturbations in the composition of cosmic plasma without perturbations in the energy density of the relativistic component. Cosmological observations, notably CMB measurements and galaxy surveys, can discriminate between different types of modes. Currently, the observations are consistent with the adiabatic mode only, while the limits on isocurvature modes are not very strong: crudely speaking,

$$\delta \rho_{\text{isocurvature}} < 0.2 \delta \rho_{\text{adiabatic}}.$$  

We stress that the detection of any admixture of baryon and/or dark matter isocurvature modes would have a profound effect on cosmology: this detection would mean that the baryon asymmetry and/or dark matter were generated prior to the hot stage by a mechanism quite different from the family of mechanisms discussed in this section. A good candidate would then be the Affleck–Dine mechanism [44].

7 Inflation

The Hot Big Bang theory, although very successful in many aspects, is not free of problems. These have to do with the initial conditions for the cosmological evolution: the initial data required are very special, and in several respects very unnatural. This situation is improved dramatically if the Universe had undergone an inflationary epoch before the Hot Big Bang stage. In this section we first discuss the motivation for inflation, and then briefly study mechanisms of inflation and observational predictions of the inflationary theory. A major success of inflation, from the observational point of view, is that it provides a mechanism of the generation of primordial density perturbations in the early Universe, whose spectrum is almost flat. The approximate flatness of the spectrum has been confirmed by the measurements of the angular anisotropy of cosmic microwave background radiation, while many alternative mechanisms of the generation of density perturbations, like topological defects, are ruled out by the CMB data. Inflation also predicts a certain spectrum of primordial gravitational waves in our Universe, which, in principle, is observable through CMB. In these lectures we shall illustrate the basic mechanism of the generation of density perturbations and gravitational waves at inflation. Of course, the corresponding theory is rather involved, so our discussion here will be fairly qualitative. For reviews on inflationary cosmology, see Refs. [45–47].

7.1 Problems of Hot Big Bang theory

Within Hot Big Bang theory, the Universe started its expansion from the singularity. Of course, the singularity is a property of classical General Relativity, and it may be replaced by something else in full quantum theory of gravity and matter. In other words, it is not legitimate to extrapolate the evolution, by making use of the classical Einstein equations, to curvatures and energy densities of the order of the Planck values or higher. Still, the classical theory is applicable soon after the Planck epoch, and we may ask what were the properties of the Universe at that time. This is what we mean by initial conditions in the Hot Big Bang theory.

Off hand, one would think that these initial conditions are more or less random: different parts of the Universe would have different properties after the Planck epoch, both spatial curvature and energy density would be close to the Planck values, etc. It is straightforward to see, however, that with such random initial conditions, Hot Big Bang theory alone would not be able to explain the high homogeneity and isotropy of the Universe that we see today. When quantifying this statement, one arrives at several ‘problems of Hot Big Bang theory’.

– Horizon problem.  
As we know, relic photons were emitted/last scattered when the Universe was rather young, $t_{\text{rec}} \approx 3 \cdot 10^5$ yrs. In the Hot Big Bang theory, the horizon size at that time was about $l_{\text{hor, rec}} = 3t_{\text{rec}} \approx 10^6$ light yrs. In the Hot Big Bang theory, there was no cross-talk, by the time of recombination,
between regions separated by a distance larger than $l_{\text{hor, rec}}$, i.e., these regions were not in causal contact with each other. Hence, there is no reason for such regions to have the same properties, e.g., the same temperature; CMB photons coming from different regions should a priori have quite different temperatures.

The present size of the region whose size at recombination was $l_{\text{hor, rec}}$ is

$$l_{\text{hor, rec}}^{\text{today}} = l_{\text{hor, rec}} \cdot (1 + z_{\text{rec}})$$

$$\approx 10^6 \cdot 10^3 \text{ light yrs}$$

$$\approx 300 \text{ Mpc} (61)$$

while the present horizon size is $l_{H,0} \approx 10^4 \text{ Mpc}$. This means that the present angular size of the horizon at recombination is

$$\theta_{\text{rec}} = \frac{l_{\text{hor, rec}}^{\text{today}}}{l_{H,0}} \approx 0.03 \approx 2^\circ . (62)$$

We conclude that there is no reason for angular isotropy of CMB temperature at angular scales greater than $2^\circ$. Yet CMB is isotropic to better than $10^{-4}$! Why is this so? Why are the initial conditions for the Hot Big Bang so homogeneous and isotropic even over causally disconnected regions of space? This is the horizon problem, which cannot be solved in the context of the Hot Big Bang theory.

- Flatness problem.

The Universe today is almost spatially flat. Quantitatively,

$$|\Omega_{\text{curv}}| \equiv \frac{1}{a^2} \frac{1}{\frac{8\pi}{3} G \rho_c} < 0.02 .$$

Thus, the curvature term in the Friedmann equation scales as $1/a^2$, while matter and radiation contributions scale as $1/a^3$ and $1/a^4$, respectively. Thus, the curvature term in the Friedmann equation was even less important at earlier epochs, for example

- nucleosynthesis epoch : $|\Omega_{\text{curv}}| = \frac{|\rho_{\text{curv}}|}{|\rho_{\text{tot}}|} < 10^{-16}$,

- electroweak epoch : $|\Omega_{\text{curv}}| < 10^{-26}$.

Thus the spatial curvature of the Universe was tiny at the beginning. Why was space so flat initially? In other words, one can compare the radius of spatial curvature $a$ with the radius of space-time curvature, the latter being of the order of the inverse Hubble parameter. One obtains, for example, at the electroweak epoch,

$$a > 10^{13} H^{-1} .$$

Why are the initial conditions such that the radius of the Universe is so large? This ‘flatness problem’ again cannot be solved within Hot Big Bang theory.

- Entropy problem.

Let us estimate the entropy of the visible part of the Universe, i.e., entropy inside a sphere of size $l_{H,0}$. This entropy is of the order of the number of photons inside this sphere,

$$S \sim N_{\gamma} \sim n_{\gamma} l_{H,0}^3$$
which gives
\[ S \sim 10^{88}. \]

In the Hot Big Bang theory, the expansion of the Universe is almost adiabatic, so this huge entropy should be built in as an initial condition. Certainly, this initial condition is very special: off hand, one would rather expect that all dimensionless quantities are roughly of order 1 at the beginning of the Universe.

Besides these problems which basically mean that the Hot Big Bang theory does not explain why our Universe is so large, hot and homogeneous, there is another problem of a different kind. This is the problem of primordial perturbations. At early times (e.g., at the recombination epoch), the Universe was not exactly homogeneous: there were density perturbations at the level \( \delta \rho / \rho \sim 10^{-5} \).

These density perturbations grew up, and finally gave rise to structures in our Universe (galaxies, galactic clusters, etc.). The problem is that in the Hot Big Bang theory, the density perturbations are to be built in as initial conditions, and there is no way to explain their origin.

Inflation is a dynamical mechanism that makes the Universe large, homogeneous, flat, and hot. As a bonus, it provides a mechanism for the generation of primordial density perturbations (and also gravitational waves). These perturbations originate from vacuum fluctuations of quantum fields, which get enhanced during the inflationary epoch.

### 7.2 Basic picture

The idea of inflation is that before the Hot Big Bang (but after the Planck era), the Universe was in a vacuum-like state and underwent the exponential expansion\(^{22}\),

\[
a(t) = \text{const} \cdot e^{\int H_{\text{inf}} \, dt}
\]

where \( H_{\text{inf}} \) is almost constant in time. Owing to the exponential expansion, a small patch of the Universe expands to great size. Say, if the duration of inflation \( t_{\text{infl}} \) exceeds 140 Hubble times\(^{23}\),

\[
t_{\text{infl}} > \frac{140}{H_{\text{inf}}}
\]

then a patch of initial Planck size \( l_{\text{Pl}} = 1/M_{\text{Pl}} \sim 10^{-33} \) cm expands to the size exceeding the present horizon size \( l_{H,0} \sim 10^{28} \) cm. Obviously, the Universe flattens out, any initial inhomogeneities get diluted and, by the end of inflation, the Universe becomes spatially flat, homogeneous and isotropic at exponentially large spatial scales. This solves the horizon and flatness problems.

A natural and most popular way to ensure that the Universe expands exponentially is to assume that the matter at the inflationary stage is in the vacuum-like state characterized by energy density \( \rho_{\text{inf}} \) which is almost constant in time. At some point, however, this energy density should transform into the conventional energy density of hot plasma. This transformation is called reheating, and after reheating the Hot Big Bang era begins. During reheating, huge entropy is released, and this solves the entropy problem.

This scenario automatically solves three problems mentioned above, which have to do with horizon, flatness, and entropy. It is not at all obvious that inflation solves the fourth problem of primordial perturbations, but it does!

\(^{22}\)It is not absolutely necessary for the expansion to be exponential. What is needed is that by the end of inflation, the size of the cosmological horizon be very large. As an example, power-law behaviour of the scale factor (34) with \( \alpha > 1 \) would also correspond to inflation.

\(^{23}\)In fact, the number of e-foldings required is smaller than 140, since the Universe expanded considerably after the end of inflation. The minimum number of e-foldings is 50–70, depending on the Hubble parameter at inflation, duration of the reheating stage, etc.
7.3 Slowly rolling inflaton

Currently, the most popular models of inflation invoke a new scalar field—inflaton, which drives inflation and then automatically ensures an exit from the inflationary stage. Depending on the scalar potential, several versions of the inflationary scenario have been designed. Their common feature is inflation, the slow-roll regime, in which the inflaton field slowly changes in time.

Let us assume that ‘at the beginning’ there is a sufficiently large patch in the Universe where the field \( \phi(x) \) is reasonably homogeneous. We should stress that these assumptions are rather mild: if ‘the beginning’ is just after the Planck era, then ‘sufficiently large’ means ‘somewhat larger than the Planck size’, and ‘reasonably homogeneous’ means that the gradient term in energy density is somewhat smaller than the potential term. Under these assumptions, one may consider both the metric and scalar field as homogeneous and isotropic, which means that the metric has the FRW form, and the scalar field does not depend on spatial coordinates. With non-trivial scalar potential, the action (3) for the scalar field in the expanding Universe is modified as follows:

\[
S = \int dt \, d^3x \, a^3(t) \left[ \frac{1}{2} \dot{\phi}^2 - \frac{1}{2a^2(t)} (\partial_i \phi)^2 - V(\phi) \right].
\]

Neglecting the gradient term and then varying the action, one obtains the field equation

\[
\ddot{\phi} + \frac{3}{a} \dot{\phi} = -\frac{\partial V}{\partial \phi}.
\]  

(63)

It is worth noting that this equation formally coincides with the equation of classical mechanics for a ‘particle’ with coordinate \( \phi \) moving with friction in the potential \( V(\phi) \), the friction coefficient being \( 3H(t) \equiv 3\dot{a}/a \).

To complete the system of equations, we have to write the Friedmann equation. For the homogeneous scalar field, the energy density is

\[
\rho = \frac{1}{2} \dot{\phi}^2 + V(\phi).
\]

Hence the Friedmann equation has the following form:

\[
H^2 \equiv \left( \frac{\dot{a}}{a} \right)^2 = \frac{8\pi}{3M_{\text{Pl}}^2} \left( \frac{1}{2} \dot{\phi}^2 + V(\phi) \right).
\]  

(64)

The vacuum-like state occurs when the energy density is almost constant in time. This is possible when the friction in Eq. (63) is so strong that the scalar field barely evolves. Its kinetic energy \( (1/2)\dot{\phi}^2 \) is small compared to the potential energy \( V(\phi) \), the latter stays almost constant in time. This is called the slow-roll regime. The conditions for slow roll are

\[
H \dot{\phi} \gg \dot{\phi}
\]  

(65)

and

\[
V(\phi) \gg \dot{\phi}^2.
\]  

(66)

If these conditions are satisfied, then the system of equations (63) and (64) simplifies; one has instead

\[
3H \dot{\phi} = -\frac{\partial V}{\partial \phi}
\]  

(67)

and

\[
H^2 = \frac{8\pi}{3M_{\text{Pl}}^2} V(\phi).
\]  

(68)
It follows from Eq. (68) that at large enough $V(\phi)$, the Hubble parameter $H \equiv \dot{a}/a$ is large, and then it follows from Eq. (67) that the field $\phi$ indeed rolls down slowly. The potential $V(\phi)$ indeed remains almost constant, and the Universe expands exponentially. Thus, once the slow-roll regime is ensured, inflation occurs automatically.

The slow-roll regime terminates when the Hubble parameter $H$ becomes relatively small, so that the friction term in (63) no longer prevents the scalar field from evolving quickly. Then the inflaton field quickly rolls down to the minimum of its potential, and possibly oscillates about this minimum. This produces a scalar field background rapidly varying in time, in which particles (either inflatons themselves, or particles of other fields) are copiously produced. The latter is the reheating process, which eventually leads to thermal equilibrium. Of course, one needs to introduce interactions between the inflaton and conventional matter to have the standard hot era in the end.

### 7.4 Simple model: one-field inflation

Let us study whether the slow-roll conditions can be satisfied. We consider the simplest potential of the form shown in Fig. 17. Let the potential at large $\phi$ have the power-law behaviour, e.g.,

$$V(\phi) = \frac{m^2}{2} \phi^2 \quad \text{or} \quad V(\phi) = \lambda \phi^4.$$ 

Making use of Eq. (67), and then Eq. (68), and dropping numerical factors of order one, we write

$$\dot{\phi}^2 \sim \left( \frac{\partial V}{\partial \phi} \right)^2 \cdot \frac{1}{H^2} 
\sim \left( \frac{\partial V}{\partial \phi} \right)^2 \frac{M_{Pl}^2}{V}.$$

Therefore, the slow-roll condition (66) takes the form

$$\frac{\partial V}{\partial \phi} \ll \frac{V}{M_{Pl}}.$$

For potentials which have power-law behaviour at large $\phi$ one has

$$\frac{\partial V}{\partial \phi} \sim \frac{V}{\phi}$$

so the slow-roll condition (66) is satisfied at

$$\phi \gg M_{Pl}.$$

It is straightforward to see that the first slow-roll condition (65) is also satisfied for power-law potentials at $\phi \gg M_{Pl}$. Inflation occurs whenever the value of the scalar field is larger than the Planck mass.

It is worth noting that when considering the field values of order of and larger than the Planck mass one makes quite an extrapolation. One may suspect that there may be contributions to the scalar potential which are generated by gravitational effects and have the form

$$\sum_N C_N \phi^{N+4} \frac{M_{Pl}^N}{M_{Pl}^4} \quad \text{(70)}$$

with coefficients $C_N$ of the order of one. Such a behaviour would destroy the slow roll; in fact, one can place strong bounds on $C_N$ from the analysis of density perturbations (see below): we shall see that the correct amplitude of density perturbations is obtained for very flat scalar potentials. This brings up the issue of the nature of the inflaton field and a mechanism forbidding contributions like (70).
Also, one may worry that even with sufficiently flat scalar potentials, the classical analysis of the evolution of the Universe is not applicable at $\phi \gg M_{Pl}$. This is not the case, since, at such large values of $\phi$, energy density may still be well below the Planck value $M_{Pl}^4$. For example, consider the quartic potential, $V(\phi) = \lambda \phi^4$, where $\lambda$ is a dimensionless coupling constant. The energy density at the Planck value of the scalar field is smaller than the Planck energy density, provided $\lambda \ll 1$,

$$V(\phi \sim M_{Pl}) \sim \lambda M_{Pl}^4 \ll M_{Pl}^4, \quad \text{for } \lambda \ll 1.$$  

We shall see that the correct amplitude of density perturbations is obtained when $\lambda$ is very small indeed, $\lambda < 10^{-10}$. Taking this value, we see that inflation occurs well below the Planck energies, and our classical analysis makes sense. Furthermore, we shall point out that there is direct observational evidence that towards the end of inflation, its energy scale is well below the Planck scale,

$$H < 10^{-5} M_{Pl}, \quad \text{end of inflation}.$$  

This bound has to do with the generation of gravitational waves at inflation and their effect on CMB anisotropy.

One’s best guess about the initial value of the scalar field $\phi_b$ is that the energy density at the beginning of inflation is of the order of the Planck energy density,

$$V(\phi_b) \sim M_{Pl}^4.$$  

For the quartic potential this means

$$\phi_b \sim \lambda^{-1/4} M_{Pl} \gg M_{Pl}. \quad (72)$$

Starting from this value, the field slowly rolls down the potential, until it reaches the value $\phi_e \sim M_{Pl}$. Inflation occurs during all this period of time. Once the field becomes of order $\phi_e \sim M_{Pl}$, the slow-roll regime terminates, the field quickly rolls down to $\phi \approx 0$, and then oscillates around its minimum $\phi = 0$. Field oscillations generally lead to particle creation from vacuum, so the oscillations get damped, and the Universe begins to be filled with particles. This is a reheating process, which ends when the coherent oscillations of the scalar field terminate, the classical scalar field settles down to the minimum of the scalar potential, and particles created by the oscillations get in thermal equilibrium. After the system thermalizes, it is described by the Hot Big Bang theory.

The reheating process is quite complex, and may occur in several stages. We shall not discuss it here; on general grounds it is clear that the outcome of this process is a thermal state anyway. The
end of reheating is at the same time the beginning of the Hot Big Bang. A most naive (and, in fact, unrealistic) estimate of the temperature of the Universe after reheating is obtained if one assumes that reheating takes of order one Hubble time. Under this assumption, all energy density of the inflaton field $V(\phi_e)$ transforms into heat, and the energy density is not substantially diluted, because of the expansion of the Universe, during reheating. This picture implies that the Hubble parameters at the end of inflation and in the beginning of the Hot Big Bang are of the same order,

$$H_{\text{end of inflation}} \sim \frac{T^2}{M_{\text{Pl}}}$$

where we made use of the standard formula for the Hubble parameter at the radiation-dominated stage. According to this estimate, the temperature of the Universe at the beginning of the Hot Big Bang may be quite high: a model-independent bound comes from Eq. (71), which gives

$$T < 3 \cdot 10^{-3} \sqrt{M_{\text{Pl}}/M^2} \sim 10^{16} \text{ GeV}.$$  

More realistic estimates give the reheat temperature which is several orders of magnitude lower, because reheating takes more than one Hubble time. In fact, one can design inflationary models with arbitrarily low reheat temperature.

Let us see that the inflationary stage naturally lasts long, so that the scale factor increases a lot during inflation. In this way we make sure that the three problems of the Hot Big Bang theory (horizon, flatness, and entropy) are naturally solved. During inflation, the scale factor increases by

$$\frac{a_e}{a_b} = e^{N_{\text{e-folds}}},$$

$$N_{\text{e-folds}} = \int_{t_b}^{t_e} H dt$$  \hspace{1cm} (73)

where subscripts $b$ and $e$ refer to the beginning and end of inflation. We obtain from Eqs. (67) and (68) that the number of e-foldings may be written as follows (again omitting factors of order one),

$$N_{\text{e-folds}} \sim \int_{\phi_e}^{\phi_b} \frac{d\phi}{H^2 \partial V/\partial \phi}$$

$$\sim \int_{\phi_e}^{\phi_b} d\phi \frac{V(\phi)}{M^2_{\text{Pl}}(\partial V/\partial \phi)}. \hspace{1cm} (74)$$

For power-law potentials, and $\phi_b \gg \phi_e$ one estimates

$$N_{\text{e-folds}} \sim \int_{\phi_e}^{\phi_b} \frac{\phi d\phi}{M^2_{\text{Pl}}}$$

$$\sim \frac{\phi_e^2}{M^2_{\text{Pl}}}. \hspace{1cm} (75)$$

We have seen that at the beginning of inflation, the scalar field is naturally very large, $\phi_b \gg M_{\text{Pl}}$. Therefore, the number of e-foldings is indeed naturally very large. As an example, for the quartic potential we have an estimate (72), and taking $\lambda \sim 10^{-10}$ we estimate

$$N_{\text{e-folds}} \sim 10^5.$$  

The size of the Universe after inflation in this case is of order

$$a \sim e^{N_{\text{e-folds}}} \sim 10^{100,000} \hspace{1cm} (76)$$
(it does not matter in which units!); this is more than enough to solve the Hot Big Bang problems.

One remark is in order. When discussing the Hot Big Bang theory, we saw that the horizon size was of the order of the Hubble distance, $H^{-1}$. If there was inflation before the Hot Big Bang, this estimate is no longer valid: the actual horizon size is much larger, as the size of the causally connected region of space increased during inflation by a factor $e^{N_{\text{e-folds}}}$. In particular, the present Hubble volume of the size of $10^4$ Mpc makes only a small fraction of the true horizon volume. Yet it is the distance of order $10^4$ Mpc from which the earliest electromagnetic radiation—relic photons—reaches us. In this sense the distance $10^4$ Mpc is still the size of the visible part of the Universe even in theories with an inflationary stage.

One more point is that the estimates like (76) show that inflation naturally predicts that the spatial curvature of the Universe is extremely small today, i.e., $\rho = \rho_c$ with extremely high precision. At some point many cosmologists believed that the Universe was open, with $\rho \approx 0.3\rho_c$. This was a problem for most inflationary models. The situation was cleared up with the data on CMB anisotropy, which tell us that $\rho = \rho_c$ within 2%, and show no evidence for spatial curvature.

Finally, we stress that one-field inflation studied in this subsection is not at all the only model of inflation (and it was not the first historically). Gross features of an inflationary scenario are quite similar in various models. Their detailed predictions for density perturbations and gravitational waves differ, however, so detailed measurements of the properties of CMB and large-scale structure make it possible to discriminate between various inflationary models.

8 Generation of primordial perturbations at inflation

8.1 Fluctuations of the inflaton field

In the inflationary scenario, density perturbations are generated from vacuum fluctuations of the inflaton field, for reviews see Refs. [48–50]. In this subsection, we illustrate the main idea and obtain the basic formulas for the scalar field perturbations. In effect, we shall be dealing with quantum field theory of the free scalar field in time-dependent background, and study pair creation. It is amazing that such a simple mechanism is likely to be the origin of all the structure of our Universe.

To warm up, let us estimate the amplitudes of vacuum fluctuations of the massless scalar field $\varphi$ in Minkowski space–time. The energy functional reads

$$ E = \int d^3x \left[ \frac{1}{2} \dot{\varphi}^2 + \frac{1}{2} (\nabla \varphi)^2 \right]. $$

(77)

Let us denote by $\delta \varphi_{\text{vac}}$ the vacuum amplitude at the momentum scale $p$ (length scale $\lambda \sim p^{-1}$). To estimate this amplitude, we note that the relevant frequency is $\omega_p = p$, and the typical volume is $\lambda^3 \sim p^{-3}$. We require that the energy (77) of a fluctuation in this volume be of the order of the zero-point energy of a field oscillator,

$$ E \sim \frac{1}{2} \omega_p. $$

This gives the relation

$$ \lambda^3 \cdot \omega_p^2 \cdot (\delta \varphi_{\text{vac}})^2 \sim \omega_p. $$

In this way we obtain the estimate

$$ \delta \varphi_{\text{vac}} \sim \omega_p = p. $$

(78)

Small-frequency, long-wavelength vacuum fluctuations have small amplitudes, and vice versa.

To substantiate this estimate, let us make use of the standard expression for the free-field operator in Minkowski space–time,

$$ \varphi(t, x) = \int \frac{d^3p}{(2\pi)^{3/2} \sqrt{2\omega_p}} \left( e^{i\omega_p t - i p_i x_i} A_p^\dagger + e^{-i\omega_p t + i p_i x_i} A_p \right). $$

(79)
where $A_p^\dagger$ and $A_p$ are the creation and annihilation operators obeying
\[ [A_p, A_p^\dagger] = \delta(p - p') . \tag{80} \]
It is straightforward to calculate the variance of the field in vacuo,
\[
\langle \varphi^2(x) \rangle = \int \frac{d^3p}{(2\pi)^3 2\omega_p} = \frac{1}{(2\pi)^2} \int_0^\infty p^2 \frac{dp}{p} . \tag{81}
\]
We see that the contribution of a decimal place in momenta into the variance of $\varphi$ is indeed estimated as $(\delta\varphi)^2$ where $\delta\varphi$ is given by (78), more precisely,
\[
\delta\varphi_{\text{vac}} = \frac{p}{2\pi} . \tag{82}
\]
Needless to say, these amplitudes are tiny at cosmologically interesting scales, $p^{-1} \sim \text{Mpc}$ or so.

One remark is in order. From (79) it follows that the vacuum fluctuations make up a Gaussian field: the multi-point correlation functions factorize into products of two-point functions; as an example, for equal-time correlators one has
\[
\langle \varphi(x_1)\varphi(x_2)\varphi(x_3)\varphi(x_4) \rangle = \langle \varphi(x_1)\varphi(x_2) \rangle \langle \varphi(x_3)\varphi(x_4) \rangle + \text{permutations} .
\]
In quantum field theory this property is known as Wick’s theorem.

Let us now turn to perturbations of the inflaton field at the inflationary epoch. We begin with dividing the quantum inflaton field $\hat{\varphi}$ into its classical part and quantum fluctuation operator,
\[
\hat{\varphi}(x,t) = \phi(t) + \varphi(x,t)
\]
where the classical part $\phi(t)$ is precisely the object we discussed in the previous subsection, and $\varphi$ is the operator describing small perturbations. We are going to develop linear theory in $\varphi$. Furthermore, we are going to neglect the curvature of the scalar potential, as it is small at the inflationary stage. Finally, for the time being we approximate the Hubble parameter $H$ at inflation by a constant, i.e., neglect its dependence on time. Introducing the weak dependence of $H$ on time is not difficult; we shall comment in appropriate places on the effects coming from the time dependence of $H$.

Under these assumptions, the action for perturbations coincides with the action (3), and the field equation is precisely (4). Let us write those formulas here again,
\[
S = \int dt \ d^3x \ a^3(t) \left[ \frac{1}{2} (\dot{\varphi})^2 - \frac{1}{2a^2(t)} (\partial_i \varphi)^2 \right] \tag{83}
\]
and, in momentum representation, the field equation is
\[
\ddot{\varphi}_k + 3H \dot{\varphi}_k - \frac{k^2}{a^2(t)} \varphi_k = 0 . \tag{84}
\]
Recall that $k$ is the time-independent coordinate momentum, the physical momentum being
\[
p(t) = \frac{k}{a(t)} .
\]
Properties of solutions to this equation depend on whether the physical momentum is greater or smaller than $H$. Modes with
\[
p(t) \gg H
\]
are called subhorizon, as their wavelengths $\lambda = 2\pi/p$ are shorter than the Hubble length (horizon size) $H^{-1}$. Modes with $p(t) \ll H$

are superhorizon. The WKB solution (6),

$$\varphi_k(t) = \frac{\text{const}}{a(t)} e^{i \int \omega(t) dt}$$

(85)

applies to subhorizon modes only. For superhorizon modes, the friction is so strong that they stay constant in time \(^{25}\),

$$\varphi_k = \text{const}, \quad p \ll H.$$  

The Hubble parameter stays (almost) constant at inflation, while the physical momentum gets red-shifted. Thus a mode of given $k$ is first subhorizon and then superhorizon. This is what inflation is about: short scales get stretched beyond the Hubble radius.

It is worth noting that this sequence of events for a mode with fixed $k$ is specific to inflation. In a radiation-dominated and a matter-dominated Universe, the scale factor behaves as $a \propto t^{1/2}$ and $a \propto t^{2/3}$, respectively. Thus, the physical momentum red-shifts as

$$p \propto \frac{1}{t^{1/2}} \quad \text{and} \quad p \propto \frac{1}{t^{2/3}}$$

while the Hubble parameter behaves as

$$H \propto \frac{1}{t}.$$  

Thus at the radiation- or matter-dominated stage, a mode with given $k$ is first a superhorizon, and then becomes a subhorizon. This is just the opposite to the situation at inflation.

Coming back to the inflationary stage, we see that for modes of any given $k$, the expansion of the Universe is effectively adiabatic at early times, so these modes experience vacuum fluctuations like in Minkowski space–time, and their amplitudes are of order $\varphi \sim p(t)$. When a mode of a given $k$ is still a superhorizon, its amplitude decreases in time because of the redshift. These fluctuations get frozen in at the time when they exit the horizon. At the radiation- or matter-dominated epoch these modes would re-enter the horizon and would start to oscillate again; at this moment their amplitudes would be much greater than the amplitudes of vacuum fluctuations at the same frequency \(^{26}\).

As the modes cross out the horizon when $\omega(t) = p(t) \sim H$, the field amplitude freezes in at the inflationary stage when $\varphi \sim p \sim H$, and no longer decreases, even though the physical wavelength continues to increase. Relative to vacuum, the amplitude increases like $a(t)$, and is thus enhanced by a huge factor. We immediately obtain an estimate for the amplitudes of perturbations of superhorizon modes, created from vacuum,

$$\varphi \sim H, \quad p \ll H$$

and infer that these amplitudes are (almost) independent of wavelengths (flat, Harrison–Zeldovich spectrum).

To obtain quantitative estimates, let us write the WKB solution for the operator of the quantum field $\varphi$ at the early stage of inflation as follows,

$$\varphi(t, x) = \int \frac{d^3k}{(2\pi)^{3/2}} \frac{1}{a^{3/2}} \left( \frac{1}{\sqrt{2\omega(t)}} e^{i \int \omega(t) dt - ik \cdot x} A_k + \text{h.c.} \right)$$  

(86)

\(^{24}\)The term ‘horizon’ here refers to the de Sitter horizon, not to be confused with the cosmological horizon. In fact, what matters here is the length scale $H^{-1}$.

\(^{25}\)Another superhorizon solution rapidly decays in time, and is therefore irrelevant.

\(^{26}\)In fact, our analysis may not apply to the period after inflation, as the classical part $\phi$ is small at that time, and the curvature of the scalar potential may not be negligible. However, we will need the behaviour of the fluctuations at the inflationary stage only.
where \( \omega(t) = |p(t)| = k/a(t) \). To write this solution, we made use of two points:

(i) interesting modes are subhorizon at early stages of inflation, so they behave as given in (85): the correspondence follows from the fact that \( a^{3/2} \sqrt{\omega} \propto a \);

(ii) prefactor \( a^{-3/2} \) compensates for the volume factor in the action (83); with this prefactor, the action coincides with the usual action for the scalar field in Minkowski background, modulo slow dependence of the physical momentum (and hence frequency) on time.

Now, the solution (86) is valid only for subhorizon modes. At the moment \( t_k \) of horizon crossing,

\[
\frac{k}{a(t_k)} = H(t_k)
\]

(87)

the mode of the coordinate momentum \( k \) gets frozen in, so after that it stays constant. This constant is obtained by matching the solution (86) to a constant solution at \( t = t_k \). In this way we find that after the mode becomes superhorizon, it is equal to (up to a phase)

\[
\varphi_k = \frac{1}{(2\pi)^{3/2} a^{5/2}(t_k)} \frac{1}{\sqrt{2H(t_k)}} A_k^\dagger + \text{h.c.}
\]

Eliminating \( a(t_k) \) via (87), we end up with the following expression for the contribution of superhorizon modes into the field operator (again up to an unimportant phase in the integrand),

\[
\varphi(x) = \int \frac{d^3k}{(2\pi)^{3/2}} \frac{H(t_k)}{\sqrt{2}} (A_k^\dagger + \text{h.c.})
\]

(88)

Making use of this expression, one can calculate all correlation functions involving superhorizon modes. One finds that these are the Gaussian fluctuations with an almost flat (i.e., scale-independent) spectrum. For example, for time-independent \( H \) one finds

\[
\langle \varphi^2 \rangle = \frac{H^2}{2} \int \frac{d^3k}{(2\pi)^3 k^3}
= \frac{H^2}{2} \int \frac{dk}{k}.
\]

(89)

This corresponds to fluctuations of scale-independent amplitude \( H/(2\pi) \) in a decimal interval of wavelengths.

Two remarks are in order. First, the Hubble parameter entering the fluctuation spectrum (89) is the Hubble parameter at inflation. In fact, this is the Hubble parameter towards the end of inflation, as relevant fluctuations of the scalar field cross out the horizon towards the end of inflation. The spectrum is exactly flat only if the Hubble parameter is constant in time; if this is not so, the spectrum is slightly tilted (there is extra dependence on \( k \)). The reason is that fluctuations of different momenta \( k \) cross out the horizon at different times, and the relevant value of the Hubble parameter is the value at the time when a fluctuation of a given wavenumber crosses out the horizon, see (87). This tilt is different in different models of inflation; we shall come to this point later.

Second, when writing (86) we made an implicit assumption that the modes are described by the usual quantum field theory at very early times. This is certainly questionable, if 'very early times' means, say, the beginning of inflation, \( t_b \). Indeed, at that time the physical wavelength of every interesting mode was extremely short,

\[
\lambda(t_b) = \lambda_0 \frac{a(t_b)}{a_0}
\]

(superscript 0 still means 'present'). In view of estimates like (76), the wavelength \( \lambda(t_b) \) is naturally many orders of magnitude smaller than the Planck length, for example. It is certainly of interest to understand how robust the predictions of inflation are with respect to possible new effects at such short distances. For a review of the work in this direction see, for example, Ref. [49].
8.2 Density perturbations

How does the above discussion relate to density perturbations? An intuitive way to understand the creation of density perturbations by fluctuations of the inflaton field is as follows. Let us consider a region of the Hubble size towards the end of inflation. In this region, the actual value of the scalar field is

$$\phi_{\text{act}} = \phi + \varphi$$

where $\phi$ is the unperturbed value determined by the classical field equations; $\phi$ is homogeneous over the whole inflating Universe. Now, $\varphi$ is a linear combination of scalar field fluctuations of superhorizon size. Because of the second term, the scalar field is different in different regions of the Hubble size; if $\varphi > 0$, the field is larger than the average over the whole Universe, and vice versa. The Hubble regions evolve in time independently, so a region where $\varphi > 0$ exits from inflation later than on average. The delay time is

$$\delta t = \frac{\dot{\varphi}}{\phi}.$$ 

This delay leads to higher energy density, after inflation, in the region we look at: in other regions there is more time for the density to get diluted due to the expansion. We obtain the density perturbation

$$\delta \rho = \dot{\rho} \delta t$$

where

$$\dot{\rho} \sim H \rho.$$ 

Combining all factors, we arrive at scales which at the end of inflation exceed the inflationary Hubble size

$$\frac{\delta \rho}{\rho}(x) = \frac{H}{\phi} \varphi(x).$$

This means that primordial density perturbations are proportional to fluctuations of the inflaton field; they form a random (Gaussian) field whose amplitude in a decimal interval of wavelengths is [see Eq. (89)]

$$\frac{\delta \rho}{\rho} \sim \frac{H^2}{\dot{\phi}}.$$ 

(90)

If $H$ and $\dot{\phi}$ were constant in time, this would be a flat, Harrison–Zeldovich spectrum with no preferred length scale; as we discussed at the end of the previous subsection, the spectrum is in fact slightly tilted, the tilt being model dependent.

To obtain the correct magnitude for primordial density perturbations,

$$\frac{\delta \rho}{\rho} \sim 10^{-5},$$ 

(91)

one has to tune the parameters of the inflaton potential. As an example, let us consider one-field inflation with quartic potential, discussed in the previous section. In the slow-roll approximation, one has

$$H^2 = \frac{8\pi}{3M_{Pl}^2} V(\phi)$$

and

$$\dot{\phi} = -\frac{1}{3H} \frac{\partial V}{\partial \phi}.$$ 

This gives

$$\frac{\delta \rho}{\rho} \sim \left[ \frac{V^{3/2}}{M_{Pl}^2 |\partial V/\partial \phi|} \right]_{\text{end of inflation}}.$$
Inflation ends when $\phi$ is about $M_{Pl}$, so for the quartic potential $V = \lambda \phi^4$ one has a crude estimate
\[
\frac{\delta \rho}{\rho} \sim \sqrt{\lambda}.
\]
The right magnitude of the density perturbations, Eq. (91), is, roughly speaking, obtained for
\[
\lambda \sim 10^{-10}.
\]
In fact, one should take into account the fact that the modes which currently have the cosmologically interesting wavelengths crossed out the horizon some time (50 to 70 e-foldings) before the end of inflation. This gives an improved estimate,
\[
\lambda \sim 10^{-12}.
\]
This tiny value of the self-coupling is not a peculiarity of the quartic potential: generally speaking, small density perturbations require a flat inflaton potential.

Coming back to the tilt, the traditional way to parametrize the spectrum is\(^{27}\)
\[
\frac{\delta \rho}{\rho}(k) = A_s k^{n_s - 1/2},
\]
where $A_s$ and $n_s$ are the amplitude and spectral index. The flat spectrum corresponds to $n_s = 1$, the blue and red tilted spectra have $n_s > 1$ and $n_s < 1$, respectively. Let us again consider an example of one-field inflation. Using the slow-roll equations, one finds from Eq. (90)
\[
\frac{\delta \rho}{\rho} \propto \frac{V^{3/2}}{V'}.
\] (92)
For a mode of momentum $k$ the quantities entering the right-hand side of (92) are to be evaluated at the time it exits the horizon, which is determined by (87). Modes with higher momenta exit the horizon later, at later times the value of $\phi$ is smaller, hence for power-law potentials the right-hand side of (92) is smaller. Thus, one-field models predict less power at high momenta, i.e., a slightly red spectrum of primordial density perturbations. Other models may give rise to a slightly blue spectrum; in any case, a non-vanishing tilt is an interesting and quite generic prediction of inflationary models.

The primordial density perturbations stay constant until they re-enter the horizon at the radiation-dominated or matter-dominated stage. After that they make sound waves. This picture, together with the flatness of the spectrum of density perturbations, immediately implies that small structures (galaxies) form earlier than larger structures (clusters): shorter wavelengths re-enter the horizon earlier, and hence smaller structures start to develop earlier. This general prediction is in accord with observational data on the structure in the Universe.

### 8.3 Gravity waves: the scale of inflation

Inflation creates not only scalar field perturbations, but also gravitational waves (cosmologists often call them tensor perturbations, while density perturbations are scalar in the three-dimensional sense). The mechanism is precisely the same as that outlined in Section 8.1, and, in fact, the equation for gravitational waves (tensor perturbations) is literally the same as the equation for perturbations of the massless scalar field. The action for metric perturbations, however, contains an extra factor $1/G = M_{Pl}^2$, so the canonically normalized field differs from metric perturbation by the factor $M_{Pl}$. With this qualification, we make use of Eq. (89) to obtain the amplitude of gravitational waves in a decimal interval of wavelengths,
\[
h = \frac{H}{2\pi M_{Pl}}\]
(93)
\(^{27}\)In the sense that $\left\langle \left( \frac{\delta \omega}{\rho} \right)^2 \right\rangle = \int \frac{dk}{k} A_s^2 k^{(n_s - 1)}$. 

V. A. RUBAKOV

248
where $H$ is the Hubble parameter towards the end of inflation. The primordial spectrum of the gravitational waves is almost flat, like the spectrum of the scalar field fluctuations, see Eq. (89).

Primordial gravitational waves make a contribution to the CMB anisotropy. This contribution is potentially important at large angular scales, $\Delta \theta > (a \text{ few})^\circ$. At these scales,

$$\left[ \frac{\delta T}{T} \right]_{\text{grav. waves}} \sim h.$$ 

The very fact that the CMB anisotropy $\delta T/T$ does not exceed $10^{-5}$ at large angular scales implies the bound on the Hubble parameter towards the end of inflation,

$$H < 10^{-5} M_{\text{Pl}}, \quad \text{end of inflation}.$$ 

We have already discussed the significance of this bound in Section 7.

---

**Fig. 18:** Contour plots (68% and 95%) allowed by WMAP (left upper panel) and WMAP combined with other measurements, in the plane $(n_s, r)$ where $r$ is the tensor-to-scalar ratio. Solid and dashed lines are predictions of one-field inflationary models for different durations of the preheating stage. Open and filled circles are predictions of $m^2 \phi^2$ and $\lambda \phi^4$ models, respectively. The rectangle denotes the pure Harrison–Zeldovich spectrum, $n_s = 1$, $r = 0$ (flat spectrum of primordial density perturbations, no gravity waves).

In most inflationary models, the contribution of gravitational waves into the CMB anisotropy is smaller (sometimes much smaller) than the contribution of density perturbations, even at large angular scales. Indeed, let us introduce the tensor-to-scalar ratio

$$r = \left[ \frac{h}{\delta \rho/\rho} \right]^2.$$ 

Making use of (90) and (93) we obtain

$$r \sim \left[ \frac{\dot{\phi}}{H M_{\text{Pl}}} \right]^2.$$
where we used the slow-roll equations to obtain the last estimate. At slow roll, $\dot{\phi}^2 \ll V$, so a generic prediction of inflation is that

$$r < 1.$$ 

Often, however, $r$ is not extremely small (it can easily be at the 10% level), as the modes of interest cross out the horizon towards the end of inflation, when the slow-roll conditions are about to be violated.

There is definitely a chance to detect the contribution of gravitational waves, and discriminate it from the contribution of density perturbations, in future measurements of the CMB polarization. In this way one would be able to find the scale of inflation, and possibly even reconstruct part of the inflaton potential [51].

Recent WMAP data [19] seem to provide the first evidence for inflation. The point is that these data appear to require either a tilt in the spectrum of scalar perturbations, or a contribution of gravity waves, or both, see Fig. 18. If confirmed, this result would be extremely exciting for cosmology, and for physics in general.

9 Conclusions

At first sight, our Universe appears infinitely complex. Yet, with not so many parameters, a coherent picture of the present and past Universe emerges, which has already passed precision tests of CMB anisotropy, Big Bang Nucleosynthesis, structure formation, etc. Even more precise measurements are due to come, which makes the whole field lively and fascinating.

Even the gross features of cosmology are ‘orthogonal’ to the Standard Model of particle physics:

– Most of the energy in the present Universe is in unknown forms. Furthermore, cosmology requires the existence of both new stable particles (clumped non-baryonic dark matter) and dark energy. The latter is the most fundamental and mysterious of all aspects of cosmology, as we know it today.

– Inflation needs an inflaton, a new scalar field with a very flat potential. No such field exists in the Standard Model, nor does it emerge naturally in the simplest extensions of the Standard Model.

– Baryogenesis needs new sources of CP-violation and mechanisms of baryon and/or lepton number violation.

Cosmology certainly has its own intrinsic problems, some of which have been mentioned in these lectures. We all know that the Standard Model has its intrinsic problems too. Experiments and theory in particle physics and cosmology still have a lot to tell us about the micro and the macro world, as well as about the interconnections between them.

References

Cosmology


High-energy neutrino astronomy with IceCube

O. Botner
Uppsala University, Uppsala, Sweden

Abstract
The prospect of extending our knowledge of the astrophysical processes in the deepest recesses of the Universe by using neutrinos as astronomical messengers has been a dream of scientists since the 1960s. The vision is finally becoming a reality: the first-generation AMANDA neutrino telescope at the South Pole designed to search for high-energy neutrinos is being upgraded to a kilometre-scale array, IceCube, with a much improved sensitivity. A summary of the results from AMANDA, and the perspectives for IceCube are presented.

1 Introduction
Astronomical observations traditionally exploit a wide range of the wavelengths of the electromagnetic spectrum. Photons are abundantly produced in astrophysical processes, and relatively easy to detect. However, they are prone to interact with matter and radiation and therefore do not reach us from the interior of stars or from the far-away regions of the Universe since they are absorbed by the interstellar matter and in pair-production processes with infrared radiation or the cosmic microwave background, Fig. 1. At TeV energies, γ-rays have a range of only about 100 Mpc, and at PeV energies they barely reach us from the edge of our galaxy.

Fig. 1: Logarithm of energy vs. distance for photons and protons. From P. Gorham and D. Saltzberg, www7.nationalacademies.org/bpa/EPP2010_Presentation_Gorham.pdf. The shaded areas represent regions which cannot be observed because of interactions with diffuse radiation.
High-energy cosmic rays constitute another probe of the energetic processes throughout the Universe. Cosmic rays are mostly protons with a small admixture of alpha-particles and heavier nuclei. They were discovered by Victor F. Hess in 1912 [1] and have since been studied extensively with space-borne and ground-based detectors. As seen in Fig. 1, at a given energy the reach of protons is larger than that of photons. However, protons are deflected by the intergalactic and interstellar magnetic fields and hence the information on the direction back to their sources is lost. Figure 2 shows the measured all-particle cosmic-ray energy spectrum, extending over more than 25 decades in flux out to about $10^{20}$ eV and following approximately a power law. The feature at $3-4 \times 10^{15}$ eV is called the ‘knee’ and indicates a change in slope from $\sim E^{-2.7}$ to $\sim E^{-3.1}$ [3]. A ‘second knee’ is seen at $\sim 4-6 \times 10^{17}$ eV, and a flattening called the ‘ankle’ is observed at $\sim 10^{19}$ eV. A cut-off is finally expected owing to interactions with the cosmic microwave background radiation (CMBR). For protons this occurs through the resonant reaction $p + \gamma \rightarrow \Delta^+ \rightarrow N + \pi$. The cut-off, known as the Greisen–Zatsepin–Kuzmin (GZK) limit [4], for nucleons is at $\sim 5 \times 10^{19}$ eV. The detection by AGASA [5] of events with energies apparently beyond the GZK cut-off is to date still under debate.

Considering the broken power-law shape of the cosmic-ray energy spectrum, it is commonly believed that the mechanisms causing the acceleration of these charged particles might be connected to diffusive shock phenomena, with different types of sources contributing fluxes with different spectral indices. Enrico Fermi was the first to realize that charged particles crossing back and forth over a moving shock-front can be accelerated to very high energies by stochastic scattering processes on moving magnetized plasma-clouds, and to show that the resulting energy spectrum follows a power law [6]. It is generally assumed that cosmic rays in the energy range below the knee are of galactic origin, accelerated in shocks originating in supernova explosions into the interstellar medium [7]. This assumption has been claimed to be supported by X-ray observations of the supernova remnant SN1006 [8a], indicative of electron acceleration to very high energies. This claim has not been corroborated by other observations, however. Today the best evidence for acceleration of charged particles in young supernova remnant (SNR) shocks comes from observations of the SNR RX J1713.7-3946 [8b] where the energy spectrum of the observed $\gamma$-rays indicates acceleration of charged particles to $\sim 100$ TeV. For energies above the knee, galactic magnetic fields are not strong enough to confine the cosmic rays. Therefore it is likely that the sources of these particles are to be found beyond the Milky Way. Actually, at the very highest energies, proton astronomy might be possible since the...
magnetic deflection decreases with energy—unfortunately those protons are extremely rare (fewer than 1 per km² per year) and their range is limited by the GZK process. So the location of the sources of high-energy cosmic rays remains a puzzle.

In this context, high-energy neutrinos emerge as extremely interesting cosmic messengers. Unaffected by intergalactic magnetic fields and weakly interacting, neutrinos should be able to penetrate regions opaque to protons and photons and point straight back to their sources.

In ‘bottom-up’ schemes, ultra-high energy (UHE) neutrinos originate in decays of pions and kaons produced when protons, accelerated to ultra-high energies by ‘cosmic accelerators’ driven by gravitational energy release, interact with radiation or matter in the vicinity of the sources. Models of such processes typically predict a neutrino energy spectrum of the form \( \frac{dN}{dE} \sim E^{-2} \) and make a connection between the expected UHE neutrino flux, the flux of the UHE cosmic rays, and the flux of coincidently produced TeV photons. The possible accelerators include extra-galactic sources like Active Galactic Nuclei (AGN) or massive collapsing objects giving rise to Gamma Ray Bursts (GRB), and galactic sources like micro-quasars and SNR.

A second category of models exists, where UHE neutrinos arise from decays of heavy cosmological remnants \( (M \sim 10^{23} \text{ eV}) \). These so-called top-down models predict higher neutrino fluxes than the ‘bottom-up’ acceleration scenarios, certainly within reach of kilometre-scale neutrino observatories [9]. Both schemes make predictions of the shape of the neutrino spectrum and correlate the neutrino flux to that of photons and protons.

2 Neutrino telescopes

The main objective of neutrino telescopes such as AMANDA and its successor IceCube is the search for very high energy neutrinos from cosmic sources. Predictions of the rates of such neutrinos are what drives the telescope design. It turns out, however, that the design also allows additional goals to be pursued. These include searches for neutrinos from annihilations of dark matter candidate particles, Weakly Interacting Massive Particles (WIMPs), at the centre of the Earth or the Sun, searches for magnetic monopoles and Q-balls [10], and investigations of the limits of validity of Lorentz invariance using huge samples of atmospheric neutrinos [11]. It is also possible to monitor our galaxy for low-energy neutrinos from supernova bursts.

The predicted UHE neutrino production rates from cosmic sources [12] together with the small neutrino interaction cross-section result in low expected rates at Earth, of the order a few (tens) per year per km². This limits the choice of detector medium to naturally abundant materials. The neutrinos can only be observed through their interactions with the atoms of the medium which produce long charged tracks \( (\nu_e \rightarrow e^-) \) and/or ‘cascades’ i.e., fairly localized energy depositions (from the hadron showers in both charged current (CC) and neutral current (NC) interactions, and the electromagnetic showers when \( \nu_e \rightarrow e^- \)). In both cases charged particles are produced moving with speeds exceeding the speed of light in the medium—giving rise to Cherenkov radiation, which can be observed provided the detector material is transparent in the blue/UV wavelength range where most of the light is emitted. This calls for the choice of water or ice as detector medium.

The AMANDA neutrino telescope has been constructed in the clear Antarctic glacier close to the Amundsen–Scott South Pole station [13]. During the period 1996–2000, 677 optical sensors were deployed on 19 strings, at depths between 1500 and 2000 metres. The strings were lowered into holes drilled with hot water at 80°C and allowed to freeze in. The optical properties of the ice at depths down to 2300 m, essential for understanding and optimizing the telescope performance, were studied using steady and pulsed light sources deployed with the strings [14]. The long absorption length \( (\sim 110 \text{ m}) \) and the relatively long scattering length \( (\sim 20 \text{ m}) \) at 400 nm allow observation of neutrino interactions exterior to the array. This, together with the long range of high-energy muons in ice.
(about 1 km at 300 GeV), results in an effective volume for observation of muons from $\nu_\mu$ interactions widely exceeding the instrumented detector volume. Figure 3 shows schematically the AMANDA array with an inset optical sensor, the 8$^\circ$ Hamamatsu R5912-2 photomultiplier coupled via a UV-transparent optical gel to the glass pressure housing designed to withstand a pressure of at least 600 bar. Figure 4 indicates how the neutrino direction and energy can be reconstructed by registering the Cherenkov light pattern of the individual photons. The arrival times at the optical sensors allow reconstruction of the direction, whereas the total number of registered photons is related to the deposited energy.

![Fig. 3: A sketch of the completed AMANDA array](image-url)
Fig. 4: Left panel: muons travelling through the ice give rise to a conical wave front of Cherenkov light. Right panel: cascades produce a fairly spherical wave front. By registering the arrival times of single photons, the direction of the neutrino and the energy are reconstructed.

3 Results from the AMANDA telescope

The results obtained with the AMANDA array were recently summarized in Ref. [15]. To avoid unintentional bias, data is analysed using blind techniques. The exact approach varies depending on the topic but may imply randomization of event times or optimization of cuts on randomly selected data sub-samples, on low-energy data, or on data outside a certain time range.

3.1 Atmospheric neutrinos

Atmospheric neutrinos arise in the interactions of cosmic rays in the Earth’s atmosphere. Since their energy spectrum is well determined theoretically, atmospheric neutrinos can be used as a calibration beam to verify the understanding of the detector response. AMANDA has measured the atmospheric neutrino energy spectrum out to $\sim$300 TeV [16]. The spectrum ties in well to lower energy measurements with other detectors and is consistent with theoretical expectations. The overwhelmingly important background for this analysis is the atmospheric muons abundantly produced in cosmic-ray interactions with the atmosphere. When searching for neutrinos, this background can be largely removed by using the timing information to distinguish between down-going and up-going charged tracks. Whereas the down-going flux is dominated by muons, only neutrinos can penetrate the Earth giving rise to up-going tracks on interaction. The residual background—mainly down-going tracks mistakenly reconstructed as up-going—is removed by strict requirements on the track quality parameters.

3.2 Diffuse neutrino flux

The summed flux of extra-terrestrial neutrinos, not ascribable to individual sources, is called ‘diffuse’. The observations integrate over the full Northern sky and over long time periods. Atmospheric neutrinos constitute the most important background, which is discriminated against using AMANDA’s ability to correlate the registered signal to neutrino energy. The atmospheric neutrino spectrum varies like $E^{-3.7}$ whereas the spectrum of extra-terrestrial $\nu$’s is expected to be harder ($\sim E^{-2}$). Hence, at sufficiently high energies the diffuse neutrino energy spectrum should contain an excess compared to the atmospheric expectation. Diffuse searches in AMANDA have been pursued with a three-pronged approach:

- TeV to PeV muon-neutrinos have been searched for using up-going well-reconstructed tracks [17],
- neutrinos of all flavours have been searched for by estimating the energy of NC and CC cascades [18,19],
– at very high energies (>10^{14} \text{eV}) where the neutrino cross-sections are larger and the Earth becomes opaque to up-going neutrinos, close-to-horizontal muon or cascade events were looked for [20]. In this case, energetic muon bundles created by cosmic-ray interactions in the atmosphere are the most important background.

In all cases, the results are handled statistically comparing the number of observed events with the expected background. The results are shown in Fig. 5 with the most stringent (preliminary) limit on $E^2dN/dE$ of $8.8 \times 10^{-8}$ GeV cm$^{-2}$ s$^{-1}$ sr$^{-1}$ for the neutrino energy range $16-2.5 \times 10^3$ TeV [15]. No statistically significant excess above the background expectation can be claimed, which leads to stringent limits on models of neutrino production.

![90% c.l. limits and sensitivities on $\nu_e$ E$^2$ diffuse fluxes](image)

**Fig. 5:** Flux upper limits on diffuse $\nu_\mu$ flux, assuming $E^2$ spectrum. All-flavour limits (divided by 3 based on the assumption of 1:1:1 flavour ratio at Earth) are shown for comparison. References: Fréjus [21], MACRO [22], Baikal [23], AMANDA B-10 1997 $\nu_\mu$ [17a], AMANDA B-10 1997 UHE [20], AMANDA-II 2000 cascades [19], AMANDA-II 2000 unfolding [16], AMANDA-II 2000-2003 $\nu_\mu$ [17b], Bartol [24], Honda [25].

### 3.3 Point sources

The AMANDA telescope has been used to search for steady or transient sources of extra-terrestrial neutrinos. Again, different strategies have been pursued. In the case of steady sources we conducted:

– searches over the full Northern sky for clusters of events [26],
– searches for $\nu$’s from individual promising astrophysical sources within a predefined set of 32 objects [26],
– searches for $\nu$’s from ‘stacked’ sources, grouping AGNs according to their photon emission characteristics in classes of about 10 and summing events in each class [27].
In all these cases, pointing accuracy is important, so only well-reconstructed muon tracks from $\nu_\mu$’s were looked for. The expected background is determined from off-source data. Figure 6(a) is a map of the Northern sky showing the 4282 neutrino candidates found in the 2000–2004 data set, corresponding to 1001 days of live time. Figure 6(b) shows the significance obtained by scanning the sky for clusters of events. The significance is positive for excess and negative for deficit of events compared to the expected background of atmospheric neutrinos. The highest excess is $3.7\sigma$ consistent with expectation for a random distribution of background events. Hence, as yet there is no evidence for hadronic acceleration in astrophysical sources.

The most spectacular transient, potential neutrino sources are the events giving rise to the GRBs detected at a rate of a few per day by satellite-borne instruments. AMANDA uses the spatial and temporal information provided by the Burst And Transient Source Experiment (BATSE) aboard the CGRO satellite (de-commissioned in 2000) and the satellites of the Third Interplanetary Network (IPN [28]) to perform an essentially background-free search for coincident $\nu$'s from GRBs. The background is determined from on-source/off-time data and amounts in the muon channel to of the order 1 event for the whole data set corresponding to about 400 bursts. A statistical analysis assuming average burst characteristics [29] allows putting a limit on the $\nu$ flux from GRBs, which, however, is still a factor of about 4 above prediction [30]. Specific GRBs can be addressed taking into account individual burst characteristics [31]. Further such analyses are in progress.
3.4 Indirect dark matter searches

It is important to realize that AMANDA is also a particle physics detector capable of detecting neutrinos with energies far above those within reach of Earth-based accelerators, and hence can address issues of interest to particle physics. One such topic is the search for cold dark matter particle candidates (WIMPs), notably the neutralino predicted to be the lightest supersymmetric particle in many extensions of the Standard Model. WIMPs have been proposed as a solution to the dark matter puzzle. Over time they would be gravitationally trapped by astronomical objects like the Earth or the Sun and eventually accrete in their centres. With sufficiently large densities, annihilations would start—the end-product being ‘ordinary’ Standard Model particles producing high-energy neutrinos in their decays. The ν’s are expected to have energies of the order 30% of the WIMP mass and hence could not be confused with low-energy ν’s from radioactive processes. AMANDA searches for an excess of ν’s from the Earth or the Sun. Examples are shown in Fig. 7 displaying the muon flux from the centre of the Earth and the Sun expected from predictions of the Minimal Supersymmetric extension of the Standard Model (MSSM) with varying parameters and subject to cosmological constraints, as a function of the neutralino mass. The AMANDA limit is competitive compared to other indirect dark matter searches, and complementary to the direct dark matter searches, which, moreover, explore a different epoch of the WIMP population and a different part of its velocity distribution [32].

Fig. 7(a): Muon flux from neutralino annihilations at the centre of the Earth as a function of the neutralino mass. Each dot corresponds to a prediction from a supersymmetric model with a given set of parameters [32c].

Fig. 7(b): Muon flux from neutralino annihilations at the centre of the Sun as a function of the neutralino mass. Each dot corresponds to a prediction from a supersymmetric model with a given set of parameters [32b].

3.5 Supernova watch

Since the interior of the Antarctic ice-cap is essentially noise-free, AMANDA can explore its sensitivity to coherent rate increases in excess of the dark noise to search our galaxy for supernova events. A supernova produces a burst of neutrinos, which on interaction in the detector give rise to isotropically distributed (mostly) positron-induced cascades with energies in the MeV range. The result is a simultaneous rate increase for many optical modules. AMANDA is participating in the SuperNova Early Warning System (SNEWS) [33].
4 IceCube

Building on the experience from AMANDA, a kilometre-scale, gigatonne, neutrino observatory, IceCube, is currently under construction at the South Pole. Figure 8 shows the IceCube array consisting of 80 strings with a total of 4800 optical modules deployed at depths between 1450 m and 2450 m in the Antarctic glacier. The strings are arranged in a hexagonal pattern with inter-string distances of 125 m, surrounding the AMANDA array. The geometry is optimized for the detection of muons at TeV energies and above. An extended air shower detector on the surface, IceTop, complements the deep-ice array with optical modules frozen inside water tanks, in total 320 optical sensors. The combination of IceCube and IceTop provides an opportunity for cosmic-ray composition studies close to the ‘knee’ of the cosmic-ray spectrum and above. Moreover, IceTop tagged muons can be used for calibration and survey of the deep-ice array. Also important is the possibility of using IceTop information for rejecting cosmic-ray backgrounds to physics studies with the deep detector.

Fig. 8: A sketch of the IceCube observatory with 80 in-ice strings and, on the surface, 80 IceTop stations, one above each string location. The grey cylinder represents AMANDA. The red dots indicate the positions of the nine strings deployed by spring 2006.

The fundamental component of IceCube is the Digital Optical Module (DOM), a 30 cm diameter pressure sphere enclosing a $10''$ Hamamatsu PMT similar to AMANDA’s $8''$ PMT but, unlike AMANDA, also including a system for time-stamping, digitizing, and storing the registered waveforms locally, transmitting to the surface on request. The local time-stamp is correlated to a global clock by a time calibration procedure allowing one to achieve a relative timing accuracy between modules of the order of 3 ns, which is small compared to the effects of light scattering in ice. Each DOM also contains 12 LEDs which provide light-sources for geometrical calibration of the array.
4.1 Current status

IceCube installation started during Austral summer 2004–2005 with the deployment of one string and four IceTop stations. In 2006, 9 strings and 16 IceTop stations are being commissioned, the detector having reached an instrumented volume comparable to that of AMANDA. Performance measurements so far show behaviour according to design specifications [34]. Figure 9 shows a typical waveform captured by a DOM, which is well described by a fit representing a superposition of terms, each corresponding to a single photo-electron response.

![Figure 9](image.png)

**Fig. 9:** A typical waveform captured by an IceCube DOM. The waveform features can be used in event reconstruction [34].

Simulations lead us to expect an increasing sensitive aperture with increasing energy, and for muon-neutrinos in the PeV range, full-sky observations become possible. The effective area for upward-moving muons above 10 TeV exceeds the geometric instrumented area of 1 km², increasing with energy. The median angular resolution for muons improves with energy, approaching 0.8° for TeV muons. This result does not take into account the DOM waveform information which when fully used should give a further improvement. The sensitivity of IceCube to point sources and diffuse flux of neutrinos has been estimated [35] leading us to believe that a flux at the level of 10% of the present four-year limit would be detectable after three years of observation with complete IceCube at 5σ significance.

Assuming independently available spatial and temporal information, for instance from the Swift or GLAST satellites, fully instrumented IceCube expects to achieve within three years a sensitivity at the level of 20% of the flux of neutrinos from GRBs predicted by Waxman and Bahcall [30].

5 Outlook

The implementation of the IceCube observatory is proceeding according to plan and will be completed in four more years. At that time IceCube will be the world’s largest neutrino telescope with an effective volume exceeding 1 km³. To further extend the energy reach of IceCube into the domain of the very highest energies, like those expected for neutrinos produced in the interactions of the cosmic rays at GZK energies with the cosmic microwave background, the IceCube Collaboration is looking into ways of expanding the sensitive volume by a factor 100 by deploying acoustic and/or radio-sensitive devices. Owing to the long absorption length of acoustic (radio) waves in ice, of the order 1 km, such an extension does not need to be extremely sensor-intensive and would provide not only additional sensitivity but also redundancy for the observation of rare signals.

The IceCube Collaboration includes about 200 researchers from 30 institutions in Belgium, Germany, Japan, The Netherlands, New Zealand, Sweden, the U.K., and the U.S.A. [36].
Acknowledgements

I wish to thank the organizers of the 2006 European School of High-Energy Physics for providing the opportunity to present this talk. I also wish to thank Carlos P. de los Heros, Christian Spiering and Thomas K. Gaisser of the IceCube Collaboration for reading and commenting on the manuscript.

References


   L. Gerhardt (for the IceCube Collaboration), Search for ultra high energy neutrinos with
   AMANDA in Proc. 14th International Conference on Supersymmetry and the Unification of
   Fundamental Interactions SUSY06, Irvine, CA, USA, 2006.


[29] K. Kuehn (for the IceCube Collaboration and the IPN Collaboration), The search for neutrinos
   from gamma-ray bursts with AMANDA in Proc. 29th ICRC, Pune, India, 2005 (also in


[31] M. Stamatikos, J. Kurtzweil and M.J. Clarke (for the IceCube Collaboration), Probing for
   leptonic signatures from GRB030329 with AMANDA-II in Proc. 29th ICRC, Pune, India, 2005


Organizing Committee

Egil Lillestøl (CERN and University of Bergen, Norway), Schools Director
Tord Ekoslöf (Uppsala University, Sweden), Local Director
Johan Rathsman (Uppsala University, Sweden)
Tatyana Donskova (JINR, Dubna, Russia), Secretary
Nick Ellis (CERN, Geneva, Switzerland)
Robert Fleischer (CERN, Geneva, Switzerland)
Danielle Métral (CERN, Geneva, Switzerland), Administrator
Barbara Strasser (CERN, Geneva, Switzerland), Secretary

Lecturers

Ikaros Bigi (University of Notre Dame, USA)
Olga Botner (Uppsala University, Sweden)
Tord Ekoslöf (Uppsala University, Sweden)
Gunnar Ingelman (Uppsala University, Sweden)
Boris Kayser (FermiLab, USA)
Mariette Mankelew (Uppsala University, Sweden)
Larry McLerran (Brookhaven National Laboratory, USA)
Antonio Pich (IFIC, Valencia, Spain)
Riccardo Rattazzi (CERN, Geneva, Switzerland)
Valery Rubakov (INR, Moscow, Russia)
Torbjörn Sjöstrand (Lund University, Sweden)

Discussion Leaders

Andrey Arbuzov (JINR, Dubna, Russia)
Gennaro Corcella (Univ. Roma, Italy)
Leif Lönnblad (Lund University, Sweden)
Silvia Pascoli (CERN, Geneva, Switzerland)
Alexei Pivovarov (INR, Moscow, Russia)
Johan Rathsman (Uppsala University, Sweden)

Other Attendees

Jos Engelen, Representing CERN Director-General
Rupert Lettner (Co-chair 2007 European School)
Alexei Sissakian, Vice-Director, JINR

Swedish Support Students (Uppsala University)

Elias Coniavitis
Martin Flechfl
Bjarte Mohn

Sophie Ohlsson
Henrik Pettersson
Oscar Stål

265
Students

Mohamed Aharrouche
Akmaral Ashimova
Angei Asryan
Konstantinos Bachas
Spasimir Balev
Danilo Banfi
Andrea Bangert
Baatar Batgerel
Paolo Bellan
Christoph Bloch
Carlo Bombonati
Jérémie Borel
Grzegorz Brona
Elena Bruna
Giuseppe Cerati
Alan Cervi
Nathalie Chon-Sen
Eric Conte
Giacomo Contin
Zofia Czyczula
Carlos Filipe Da Silva Costa
Umberto De Sanctis
Georgios Dedes
Julien Descamps
Benedetto Di Ruzza
Rafael Diaz Valdes
Anton Dimitrov
Peder Eliasson
Maria Luísa Ferreira da Gama Velho Arruda
Salvatore Fiore
Alexander Floessdorf
Thomas Gadfort
Henning Gast
Gabriele Giacomini
Marc Goulette
Christian Helebrant
Franziska Hofmann
Pieter Houwen
Natalia Ilina
Ben Jeffery
Andreas Werner Jung
Darius Jurciukonis
Ioannis Katsanos
Gordon Kaussen
Matthias Kirsch
Stefan Koerig
Vladimir Kolesnikov
Iro Koletsou
Natalia Kondrateva
Katherine Korcsak-Gorzo
Gernot Krobath
Nils Lävesson
Alexander Linn
Inga Ludwig
Vladimir Lyubushkin
Philipp Mack
Luis March
Camillo Mariani
Paolo Mastrandrea
Joao Gentil Mendes Saraiva
Volker Michels
Corinne Mills
Krzysztof Nowak
Mark Owen
Anastasios Papageorgiou
Bakur Parsanyan
Francesco Perfetto
Tatevik Poghosyan
Bernardo Resende
Hélène Ricaud
Svenja Richter
Catalin Ristea
Gael Rospabe
Leonid Rumyantsev
Renat Sadykov
Joergen Samson
Georgy Sharkov
Andrzej Siodmok
Giovanni Siragusa
Federica Sozzi
Martin Spousta
Teresa Spreitzer
Thorsten Stahl
Andrzej Szcel
Claire Timlin
Konstantin Toms
Ingo Torchiani
Selma Uhlig
Martin Vala
Stefano Venditti
Pedro Manuel Vieira de Castro Ferreira da Silva
Philip von Doetsch
Daniel Walker
Lotte Wilke
Mark Williams
Jan-Christopher Winter
Adam Wren
Nicolas Zwahlen
## Posters

<table>
<thead>
<tr>
<th>Author</th>
<th>Poster title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohamed Aharrouche</td>
<td>Atlas CTEB 2004 and measurement of the weak mixing angle in the LHC</td>
</tr>
<tr>
<td>Akmeral Ashimova</td>
<td>The constraint on FCNC interaction of the top quark with a gluon from e-p collisions</td>
</tr>
<tr>
<td>Angei Asryan</td>
<td>PSM and PYTHIA simulations for p-p and heavy-ion collisions</td>
</tr>
<tr>
<td>Spasimir Baiev</td>
<td>Search for direct CP-violation in $K^\pm \rightarrow \pi^0 \mu^0 \nu^0$ decays by the NA14/2 experiment</td>
</tr>
<tr>
<td>Baatar Batgerel</td>
<td>Antideuteron and deuteron production in 158 A GeV Pb–Pb collisions</td>
</tr>
<tr>
<td>Grzegorz Brona</td>
<td>Gluon polarization in the nucleon measured directly by the COMPASS experiment</td>
</tr>
<tr>
<td>Elena Bruna</td>
<td>Open charm reconstruction in ALICE: $D^+ \rightarrow K^- \pi^+ \pi^+$</td>
</tr>
<tr>
<td>Alan Cervi</td>
<td>MWPC at LHCb</td>
</tr>
<tr>
<td>Nathalie Chon-Sen</td>
<td>Oscillation of $\nu_{\mu} \rightarrow \nu_{\tau}$ at the OPERA experiment</td>
</tr>
<tr>
<td>Eric Conte</td>
<td>Study of the violation of CP and T symmetries in $\Lambda_B \rightarrow \Lambda V$ with V a vector meson (LHCb experiment)</td>
</tr>
<tr>
<td>Carlos Filipe Da Silva Costa</td>
<td>The Alpha Magnetic Spectrometer on the International Space Station: The silicon tracker of AMS-02</td>
</tr>
<tr>
<td>Umberto De Sanctis</td>
<td>Discovery potential for supersymmetric particles in the focus point region of SUGRA with the ATLAS detector at the LHC</td>
</tr>
<tr>
<td>Georgios Dedes</td>
<td>Search for the MSSM Higgs boson decay $A \rightarrow \pi^- \pi^+$ in the ATLAS detector</td>
</tr>
<tr>
<td>Julien Descamps</td>
<td>Periodic position measurement in the CMS electromagnetic calorimeter</td>
</tr>
<tr>
<td>Benedetto Di Ruzza</td>
<td>The silicon vertex tracker in the CDF detector at the Tevatron</td>
</tr>
<tr>
<td>Rafael Diaz Valdes</td>
<td>Jet finder algorithms and background subtraction in the ALICE experiment</td>
</tr>
<tr>
<td>Anton Dimitrov</td>
<td>CMS barrel resistive plate chambers</td>
</tr>
<tr>
<td>Maria Luiza Ferreira da Gama</td>
<td>The RICH detector of the AMS experiment: test beam results with the RICH prototypes</td>
</tr>
<tr>
<td>Velho Arruda</td>
<td>Positrons with AMS (with Philip von Doetinchem)</td>
</tr>
<tr>
<td>Henning Gast</td>
<td>Radiation detector with internal amplification based on the BJT effect</td>
</tr>
<tr>
<td>Gabriele Giacomini</td>
<td>Higgs decay into two photons at NLO</td>
</tr>
<tr>
<td>Franziska Hofmann</td>
<td>Association of jets with the signal vertex</td>
</tr>
<tr>
<td>Natalia Ilina</td>
<td>Flavour and charge ID at the linear collider</td>
</tr>
<tr>
<td>Ben Jeffery</td>
<td>The H1 fast track trigger</td>
</tr>
<tr>
<td>Andreas Werner Jung</td>
<td>The CMS silicon tracker endcaps: building and testing of substructures (Petals) (with Alexander Linn)</td>
</tr>
<tr>
<td>Gordon Kaussen</td>
<td>Ageing studies for the ATLAS muon MDT gas system</td>
</tr>
<tr>
<td>Stefan Koenig</td>
<td>SANC, application to LHC: QCD sector</td>
</tr>
<tr>
<td>Vladimir Kolesnikov</td>
<td>The start T0 detector for the heavy-ion experiment ALICE at CERN</td>
</tr>
</tbody>
</table>

267
<table>
<thead>
<tr>
<th>Author</th>
<th>Poster title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gernot Krobath</td>
<td>Search for 2nd generation leptoquarks with ATLAS at the LHC</td>
</tr>
<tr>
<td>Alexander Linn</td>
<td>Building and testing of CMS substructures (Pascal) 1/2 (with Gordon Kaussen)</td>
</tr>
<tr>
<td>Inga Ludwig</td>
<td>Isolated electrons in ATLAS</td>
</tr>
<tr>
<td>Vladimir Lyubushkin</td>
<td>Measurement of the quasi-elastic neutrino cross-section section at the NOMAD experiment (CERN)</td>
</tr>
<tr>
<td>Philipp Mack</td>
<td>Exclusive reconstruction of the decay $B_s \rightarrow D_s \pi \rightarrow KK$</td>
</tr>
<tr>
<td>Luis March</td>
<td>Search for hadronic decays of $ZH$ and $WH$ in the little Higgs model using the ATLAS detector</td>
</tr>
<tr>
<td>Camillo Mariani</td>
<td>K2K - SciBoone - T2K</td>
</tr>
<tr>
<td>Corrinne Mills</td>
<td>Top dilepton cross-section using lepton plus track selection at CDF (with Teresa Sprietzer)</td>
</tr>
<tr>
<td>Mark Owea</td>
<td>Search for $Higgs \rightarrow \tau \tau$ with the D0 detector</td>
</tr>
<tr>
<td>Bakur Parsanyan</td>
<td>Predictions for double spin asymmetry ALT in semi-inclusive DIS</td>
</tr>
<tr>
<td>Francesco Perfetto</td>
<td>Dalitz plot analysis of $n - 3$ final state</td>
</tr>
<tr>
<td>Tatevik Poghosyan</td>
<td>Search for the multiplicity fluctuations over azimuthal angle in NA60 data on charged-particle production in indium–indium central collisions at beam energy of 158 GeV per nucleon.</td>
</tr>
<tr>
<td>Bernardo Resende</td>
<td>Top quark polarization studies in ATLAS</td>
</tr>
<tr>
<td>Hélène Ricaud</td>
<td>Strangeness at intermediate p–p collisions in the ALICE experiment at the LHC, Pythia simulations</td>
</tr>
<tr>
<td>Svenja Richter</td>
<td>A neural network b tagger for top-physics at CDF</td>
</tr>
<tr>
<td>Leonid Rumyantsev</td>
<td>Processes of the H-boson production and decay in SANC</td>
</tr>
<tr>
<td>Renat Sadykov</td>
<td>SANC, application to LHC: Drell–Yan-like single W and Z boson production</td>
</tr>
<tr>
<td>Georgy Sharkov</td>
<td>Experimental study of cumulative effects</td>
</tr>
<tr>
<td>Federica Sozzi</td>
<td>Transversity signals at COMPASS</td>
</tr>
<tr>
<td>Teresa Sprietzer</td>
<td>Top dilepton cross-section using lepton plus track selection at CDF (with Corrinne Mills)</td>
</tr>
<tr>
<td>Thorsten Stahl</td>
<td>$B_s$ mixing at ATLAS</td>
</tr>
<tr>
<td>Konstantin Toms</td>
<td>Rare B decays at ATLAS</td>
</tr>
<tr>
<td>Ingo Torchiani</td>
<td>Search for new physics in $\tau$ lepton final states at D0</td>
</tr>
<tr>
<td>Philip von Doetinchem</td>
<td>Positrons with AMS (with Henning Gast)</td>
</tr>
<tr>
<td>Lotte Wilke</td>
<td>Study of the decay $B_s \rightarrow J/\psi \phi \rightarrow \mu^+\mu^-K^+K^-$ at the CMS experiment</td>
</tr>
<tr>
<td>Jan-Christopher Winter</td>
<td>SHERPA's CKKW merging in practice: the case of $W^+W^-$ production</td>
</tr>
</tbody>
</table>
Photos