CERN–2001–003
11 September 2001

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

2000 EUROPEAN SCHOOL OF HIGH-ENERGY PHYSICS

Caramulo, Portugal
20 August–2 September 2000

PROCEEDINGS

Editors: N. Ellis, J. March-Russell

GENEVA
2001

ISSN 0531-4283
ISBN 92-9083-186-3

CERN–2001–003
11 September 2001

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

2000 EUROPEAN SCHOOL OF HIGH-ENERGY PHYSICS

Caramulo, Portugal
20 August–2 September 2000

PROCEEDINGS

Editors: N. Ellis, J. March-Russell

GENEVA
2001

CERN–Service d'information scientifique–RD/1012–2000–septembre 2001

Abstract

The European School of High-Energy Physics is intended to give young experimental physicists an introduction to the theoretical aspects of recent advances in elementary particle physics. These proceedings contain lectures on field theory and the Standard Model, quantum chromodynamics, flavour physics and physics beyond the Standard Model, as well as reports on cosmology and astrophysics, and heavy-ion physics. They also contain lecture notes on the scientific programme at JINR and the history of Portuguese music.

PREFACE

The 2000 European School of High-Energy Physics was held in Caramulo, Portugal from 20 August to 2 September. The School was attended by 101 students from 27 different countries, in addition to four Portuguese 'observer students'. The percentage of female students taking part was, at 27%, the highest yet in this series of schools. The Quality Hotel Caramulo was very comfortable, with good food, excellent sports facilities and an efficient and friendly staff. The 2000 School was the eighth in the new series organised on a yearly basis in a collaboration between CERN and JINR, Dubna.

The School was sponsored by CERN and JINR, and from Portugal by Fundação para a Ciência e Tecnologia, the EFACEC Group, Insituto do Soldadura e Qualidade, Câmara Municipal de Tondela, Câmara Municipal de Viseu and Caixa Geral de Depósitos; 35 students received financial support from UNESCO and INTAS. Local organisation and support were provided by FCTUC (School of Science and Technology, University of Coimbra), Universidade de Coimbra and LIP (Laboratório de Instrumentação e Física Experimental de Partículas). Our sincere thanks go to all the sponsors making it possible to organise the School and contributing to its success.

Our thanks are due to the lecturers and discussion leaders for their active participation in the School and for making the scientific programme so stimulating. Their personal contribution in answering questions and explaining points of theory was undoubtedly appreciated by the students who in turn manifested their good spirits during two intense weeks.

The major responsibility for organising the 2000 School was in the hands of Professor Rui Ferreira Marques together with the local organisers and support staff from LIP. Our special thanks goes to Rui, who acted as local director of the school, and whose friendly and pragmatic attitude together with his omnipresence at the premises, contributed in major ways to the success of the School. We also want to thank Fatima Loureiro for her daily assistance in sorting out all kinds of practical problems.

The indoor and outdoor sports activities were organised in the most efficient way by the professional group of 'team builders' led by Susana Guerra and Pedro Vieira. The students participated wholeheartedly, and we are convinced they will form excellent teams in the future. The organisers are particularly grateful that nobody was seriously hurt.

The necessary energy input for these activities was provided by an outstanding kitchen and its chef Joao Pereira. His variations of Bacalao and his magic desserts will long be remembered.

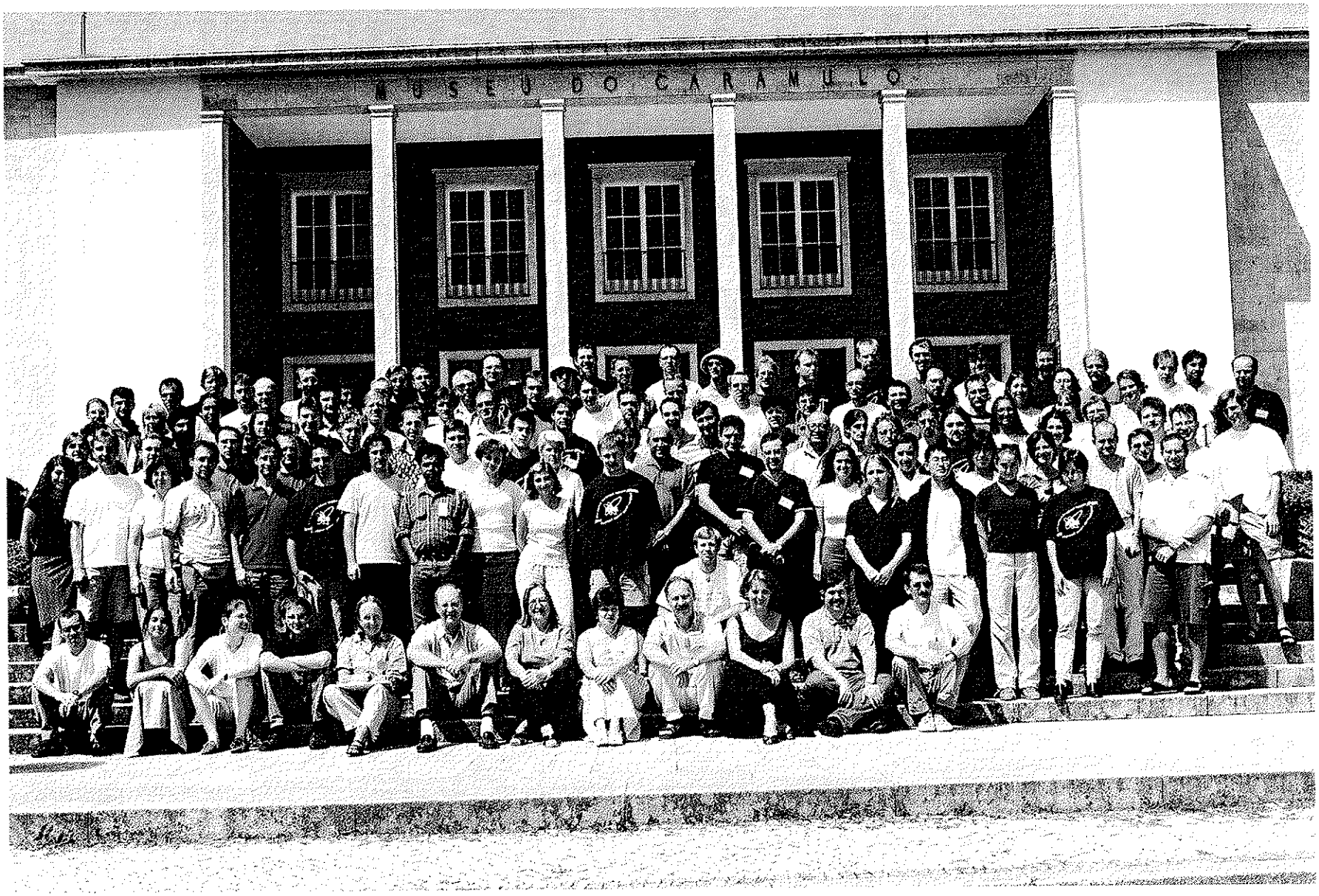
The School had wide media coverage both in the press and on television.

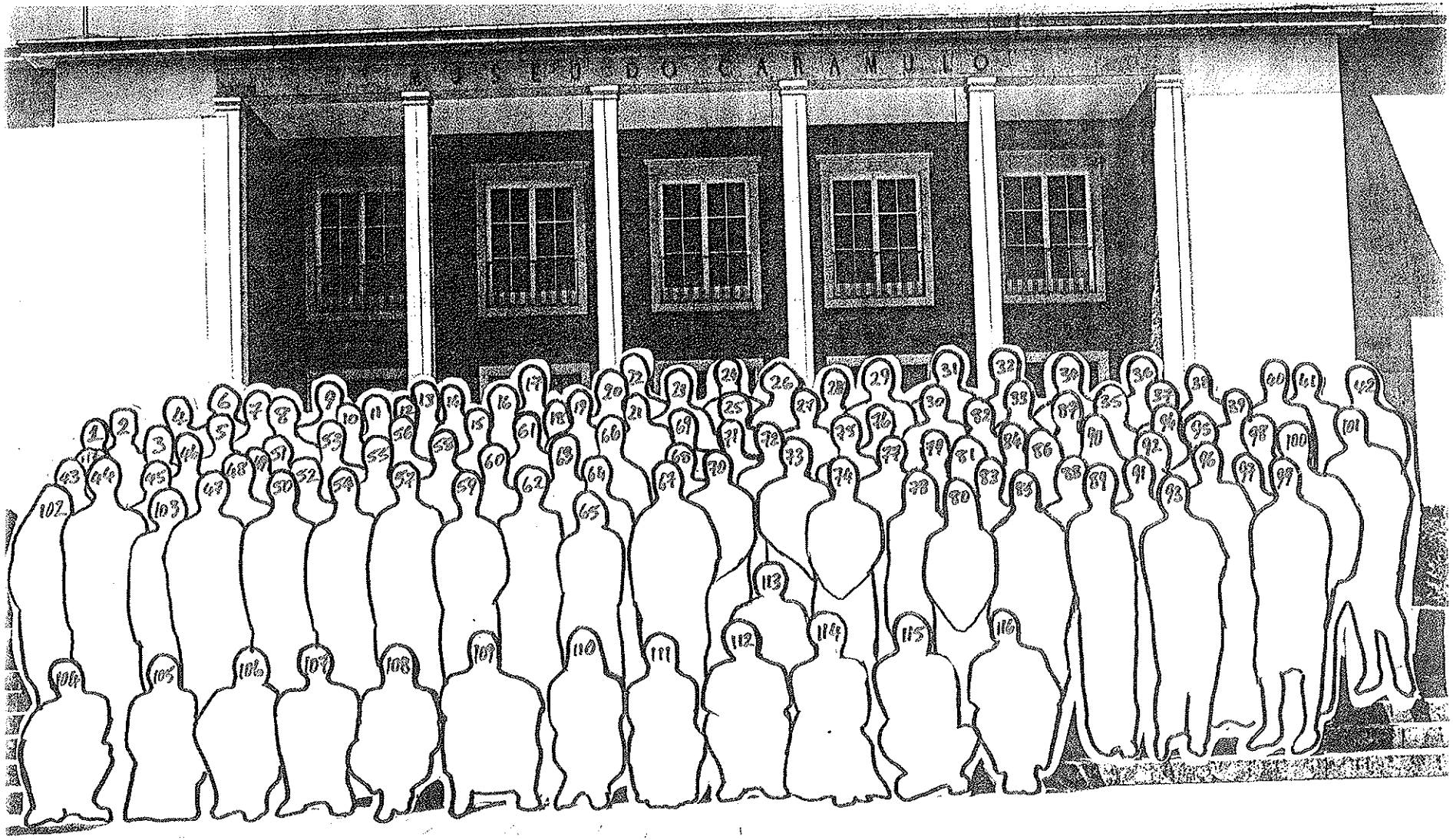
We are very grateful to Claire Earnshaw and Tatyana Donskova for their untiring efforts in the lengthy preparations for and the day-to-day care of the School. It is a great pleasure to see how much they care about the students and how well this is appreciated.

A special word of thanks goes to the General Management of the Hotel and in particular to Francisca Antunes. It is hard to imagine a better manager combining efficiency, firmness and friendliness in a most professional manner. The School certainly profited greatly from her assistance.

The students will certainly remember several interesting excursions including a visit to the Science Museum in Coimbra and its collection of old physics instruments and wine tasting at Sandeman in Porto followed by a musical lunch offered by EFACEC and hosted by Celia Adelaide Ferreira Soeiro. However, the success of the 2000 School was to a large extent due to the students themselves. Their poster session was as good as at any previous School, 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


MUSEU DO CARANGUEJO


R I S E U D O C A R A M U L O

PEOPLE IN THE PHOTOGRAPH

STUDENTS

4 Nuno Almeida
74 Nuno Anjos
10 Armen Apyan
54 Stefano Argiro
19 Sasha Artamonov
7 Daniel Barna
73 Joao Bastos
13 Alexey Berdyugin
20 Martin Blom
65 Oana Boeriu
51 Martina Bohacova
80 Bozena Boimska
102 Florencia Canelli
66 Lucio Cerrito
105 Nicole Chevalier
92 Teresa Claudino
69 Yann Coadou
88 Marialaura Colantoni
106 Caroline Collard
14 Markus Cristinziani
44 Carsten Cruse
67 Steve Dallison
21 Dmitri Dedovitch
97 Frederic Deliot
37 Camilla Di Donato
107 Julien Donini
46 Ranpal Dosanjh
28 Michael Dressel
27 Holger Fleckenstein
99 Andrea Formica
32 Paolo Gatti
63 Andrei Gheata
64 Mihaela Gheata
12 Borge Kile Gjelsten
113 David Goldstein
72 Santiago Gonzalez de la Hoz
48 Amnon Harel
40 Andreas Heiss
23 Rutger Hierck
100 Andre Holzner
93 Olya Igonkina
53 Marek Jacewicz
2 Elchin Jafarov
60 Sasha Kaoukher
43 Ben Kilminster
95 Birger Koblitz
1 Natalia Korotkova
75 Oliver Kortner
24 Carsten Krauss
3 Polina Kravtchenko
34 Stefano Lacaprara
26 Cyril Lachaud
39 Melanie Langer
87 Regis Lefevre
101 Guillaume Leibenguth
42 Axel Leins
17 Mikhail Lobanov
89 Amaya Lopez-Duran Viani
41 Ankush Mitra
90 Cristina Morone
52 Armin Nairz
50 Matteo Negrini
11 Lionel Neukermans
104 Francisco Neves
81 Andreas Nygren
5 Anna Okpara
76 Bob Olivier
86 Roman Otec
36 Riccardo Paramatti
84 Jean-Michel Pascolo
91 Francesca Pastore
79 Hugo Pereira da Costa
18 Alexander Popov
58 Yuriy Pylypchenko
68 Alina Radu
59 Azizur Rahaman
56 Alexei Raspereza
103 Melissa Ridel
6 Valera Rodionov
49 Aura Rosca
61 Roberto Sacco
94 Dorothea Samtleben
47 Giovanni Santin
71 Jose Silva
8 Karel Smolek
35 Francesca Spada
33 Antonio Stamerra
98 Jan Stark

15 Oleksandr Starodubtsev
30 Janusz Szuba
85 Hiro Tanaka
78 Valeria Tano
9 Päter Tarjan
96 Daniel Teyssier
62 Irina Titkova
55 Maksym Titov
29 Rutger Van der Eijk
31 Nick van Remortel
57 Peter Vankov
22 Roman Vasiliev
25 Jaap Velthuis
45 Andrew Washbrook
38 Dario Zurcher

LECTURERS

70 Gustavo Branco
115 Vladimir Braun
77 Jorge Dias de Deus
112 Dmitri Kazakov

DISCUSSION LEADERS

83 Beatriz de Carlos
31 Mikko Laine
117 Magda Lola (not visible)
82 Giovanni Ridolfi

ORGANIZING COMMITTEE

111 Tatyana Donskova
114 Claire Earnshaw
16 Nick Ellis
116 Rui Ferriera Marques
109 Egil Lillestol
108 Susannah Tracy

CONTENTS

Field Theory for the Standard Model

lectures given at the European School of High-Energy Physics 2000,
20 August - 2 September 2000, Caramulo, Portugal

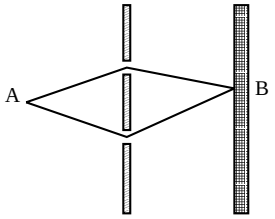
Ronald Kleiss, University of Nijmegen, the Netherlands

version of November 20, 2001

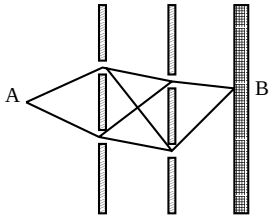
1 The Ultimate Slit Experiment: quantum mechanics and the path integral

The aim of these lectures is to show how the phenomenology of elementary particles and their interactions can be described in terms of Feynman diagrams. The treatment is not rigorous: rather, I intend to give plausibility arguments, always sticking as closely as possible to physics. In my opinion, the path integral formulation of relativistic quantum physics is one of the most directly physical ways of talking about the phenomena of the elementary particle world. It may therefore be useful to see how the path integral view naturally arises from basic quantum mechanics, using the arguments of Feynman himself.

The primordial quantum-mechanical situation is, of course, the two-slit experiment.

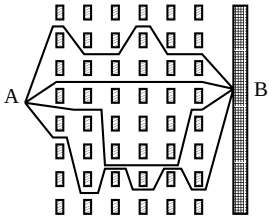


A 'particle' (electron, photon, human volunteer,. . .) is emitted at position A at time t_1, and after a while, at time t_2, it is observed to be in the 'detector', at position B out of all possible positions. Between A and B we have put a screen with 2 holes through which the particle may pass. Quantum mechanics tells us that the real, positive *probability* of finding the particle at B can be computed as the absolute value squared of the complex-valued *probability amplitude*, which in this case consists of two contributions, one from each of the two possible paths the particle 'could have taken' (indicated by straight lines). The superposition of two complex contributions gives rise to the well-known interference patterns that are observed, demonstrating the wavelike nature of matter. We now become enthousiastic, and put in yet another screen:

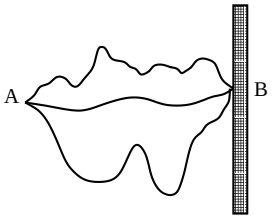


Now there are 4 different paths, each giving their contribution, and the interference pattern becomes correspondingly more complicated. Having become fanatical, we now put in many

screens, each with many holes:



There are now very many possible paths[1], and we have just drawn a few: all the possible paths give a contribution to the probability amplitude of arrriving at B. By now being rabidly frenzied, we decide on the ultimate multislit experiment: we decide to insert an infinite number of screens, each with so many holes in it that every molecule in every screen is completely drilled away. This is simpler than it sounds: the particle now travels through empty space! Still, to be fair we have to allow for all possible paths, three of which are drawn.



Our conclusion must be that *all possible paths between A and B give a contribution to the probability amplitude*, and we have to sum these over all paths: the path integral.

In many cases, we do not know that the particle was actually at A at time t_1, and the only information we have is a quantum-mechanical wave-function $\psi(x_1, t_1)$ that gives us the probability amplitude for finding the particle at a given position x_1 at time t_1: by the path integral this is then translated into the wave-function $\psi(x_2, t_2)$ at the later time t_2.

But what *is* the contribution from each path? The correct choice is to take $\exp(iS/\hbar)$, where S is the action associated with the path, that is, the time integral over the Lagrangian from t_1 to t_2. To see this, let us write the prescription as given, for this one-dimensional case:

$$\psi(x_2, t_2) = N \int \mathcal{D}x \ \exp\left[\frac{i}{\hbar}\int_{t_1}^{t_2} dt \left(\frac{m}{2}\dot{x}^2 - V(x)\right)\right] \psi(x_1, t_1) \ \ , \tag{1}$$

where $x = x(t)$ is any path in time with $x(t_1) = x_1$ and $x(t_2) = x_2$, and $\mathcal{D}x$ denotes the super-inifinitely large sum over paths. m is the particle's mass, and V any potential in which it moves. The factor N is a normalization, necessary to ensure the conservation of probability, since $|\psi|^2$ must integrate to 1 at all times. The dot denotes a time derivative. Let us now take t_2 to be infinitesimally later than t_1: $t_2 = t_1 + \tau$, with τ very very small. It is now a good approximation to replace the time integral by a multiplication by τ. Writing $x_1 = x_2 + y$, we may then replace Eq.(1) by

$$\psi(x_2, t_1 + \tau) = N \int dy \exp\left[\frac{i\tau}{\hbar}\left(\frac{my^2}{2\tau^2} - V(x)\right)\right] \psi(x_2 + y, t_1) \ \ . \tag{2}$$

[1]The number of paths is the product of the number of screens and the number of holes per screen, even if we forbid backwards motion of the particle.

It is easy to see that, if $my^2/\hbar\tau$ becomes large, the contributions will interfere utterly destructively, so that we may take $\mathcal{O}(y) = \mathcal{O}\left(\tau^{1/2}\right)$. This allows us to make an expansion in y and τ up to first order in τ:

$$\begin{aligned}
&\psi(x_2,t_1) + \tau\dot{\psi}(x_2,t_1) = \\
&N\int dy \exp\left(\frac{imy^2}{2\hbar\tau}\right) \times \\
&\left(\left(1 - \frac{i\tau}{\hbar}V(x_2)\right)\psi(x_2,t_1) + y\psi'(x_2,t_1) + \frac{1}{2}y^2\psi''(x_2,t_1)\right)
\end{aligned} \tag{3}$$

where primes denote space derivatives, and we have assumed that $\psi(x,t)$ is smooth enough. The zeroeth-order term gives us N:

$$N\int dy \exp\left(\frac{imy^2}{2\hbar\tau}\right) = N\sqrt{2i\pi\hbar\tau/m} = 1 \ , \tag{4}$$

from which

$$\begin{aligned}
N\int dy\ y\ \exp\left(\frac{imy^2}{2\hbar\tau}\right) &= 0 \ , \\
N\int dy\ y^2\ \exp\left(\frac{imy^2}{2\hbar\tau}\right) &= i\hbar\tau/m \ ,
\end{aligned} \tag{5}$$

and the term of of first order in τ then results in

$$i\hbar\dot{\psi}(x,t) = -\frac{\hbar^2}{2m}\psi''(x,t) + V(x)\psi(x,t) \ , \tag{6}$$

the Schrödinger equation.

Note that we have not 'proven' the Schrödinger equation here: we merely have replaced it, as a basic postulate of quantum physics, by another one, namely the path integral. But the above indicates that the path-integral picture is a useful way of looking at quantum physics. The various paths $x(t)$ are 'random' objects, and the only information we have about them is the probability amplitude connected with them. It is this probabilistic interpretation that we want to take with us. To make the envisageing easier, we shall for now drop the i, and talk about real, positive probabilities rather than complex probability amplitudes.

2 Small beginnings: field theory in 0+0 dimensions

It is often useful to start simply. Therefore, in this section we shall take a particularly simple model of the universe: we shall take it to consist of one single point. In this universe is a single particle species without any properties such as spin, etc. The 'field' describing the particle is a function over space(time), and therefore consists, in this case, of a single value.

2.1 The path integral and Green's functions

2.1.1 Free theory and sources

In the spirit of the path integral picture, we assume the value φ of the field to be a random number, whose probability density is proportional to $\exp(-S(\varphi))$, for some function $S(\varphi)$,

called the *action* of the model. One of the simplest choices is to assume the density to be a Gaussian:

$$S(\varphi) = \frac{1}{2}\mu\varphi^2 \ , \tag{7}$$

with some constant μ. We call this the *free* theory. Since φ is random, all the information that one can possibly *compute* about it[2] is contained in the moments of the distribution, which are called the *Green's functions* of the theory:

$$G_p \equiv \langle \varphi^p \rangle = N \int d\varphi \ \varphi^p \ e^{-S(\varphi)} \ , \tag{8}$$

where the brackets denote the average under $e^{-S(\varphi)}$, and N is the normalization, chosen such that $\langle \varphi^0 \rangle = 1$. Unless specified otherwise, all integrals run from $-\infty$ to $+\infty$. Trivially, we have in this theory

$$G_2 = \frac{1}{\mu} \ , \ \ G_{2k} = \frac{(2k)!}{2^k k!} G_2^k \ , \ \ G_{2k+1} = 0 \ , \ \ k \geq 1 \ . \tag{9}$$

We can combine the information on the Green's functions G_p into one generating function, as follows:

$$Z(J) \equiv \sum_{p \geq 0} G_p \frac{J^p}{p!} = N \int d\varphi \ e^{-(S(\varphi) - J\varphi)} \quad , \quad G_p = \frac{\partial^p}{\partial J^p} Z(J) \Big|_{J=0} \ . \tag{10}$$

The 'counting' number J is called a *source*; its physical meaning will become clear later on. We can compute $Z(J)$ directly by 'completing the square':

$$S(\varphi) - J\varphi = \frac{mu}{2} \left(\varphi - \frac{J}{\mu} \right)^2 - \frac{J^2}{2\mu} \tag{11}$$

and then doing the Gaussian integral, or by realizing, from its integral representation, that it obeys the differential equation

$$\begin{aligned} \mu \frac{\partial}{\partial J} Z(J) - J Z(J) &= N \int d\varphi \ (\mu\varphi - J) e^{-S(\varphi) + J\varphi} \\ &= -N \int d\varphi \frac{\partial}{\partial \varphi} e^{-S(\varphi) + J\varphi} = 0 \ , \end{aligned} \tag{12}$$

since we have here a total derivative. Together with $Z(0) = 1$, in either case we arrive at

$$Z(J) = \exp\left(\frac{1}{2\mu} J^2 \right) \ . \tag{13}$$

2.1.2 Interacting theories and perturbation theory

In real life, we expect theories to be more complicated than just the free one, so let us add a term to $S(\varphi)$ with a higher power of φ. The simplest choice that still leads to a finite path integral is

$$S(\varphi) = \frac{1}{2}\mu\varphi^2 + \frac{1}{24}\lambda\varphi^4 \ . \tag{14}$$

[2] You could imagine instead *measuring* the value that φ actually takes, but that would be experimental rather than theoretical physics.

Extra terms added to a free theory are called *interaction terms*, and coefficients like λ are called *coupling constants*[3]. Computing the G_p is now immediately much more difficult. A possible approach is to consider the extra term as a small 'perturbation' of the Gaussian (especially if λ is small), and make an expansion in powers of λ. For odd q, G_q is again zero, while for even $q = 2p$ we have[4]

$$\begin{aligned} G_{2p} &= N \int d\varphi \; \varphi^{2p} \; e^{-\frac{1}{2}\mu\varphi^2 - \frac{1}{24}\lambda\varphi^4} \\ &= N \sum_{k\geq 0} \frac{1}{k!} \left(-\frac{\lambda}{24}\right)^k \int d\varphi \; \varphi^{2p+4k} e^{-\frac{1}{2}\mu\varphi^2} \\ &= \frac{N\sqrt{2\pi}}{\mu^p} \sum_{k\geq 0} \left(-\frac{\lambda}{24\mu^2}\right)^k \frac{(2p+4k)!}{2^{p+2k}k!(p+2k)!} \quad . \end{aligned} \tag{15}$$

The normalization N is again fixed by $G_0 \equiv 1$,

$$N = \frac{1}{\sqrt{2\pi}} \left(1 + \frac{1}{8}\frac{\lambda}{\mu^2} - \frac{29}{384}\frac{\lambda^2}{\mu^4} + \frac{107}{1024}\frac{\lambda^3}{\mu^6} + \mathcal{O}\left(\lambda^4\right)\right) \quad , \tag{16}$$

while the first few nonzero Green's functions read

$$\begin{aligned} G_2 &= \frac{1}{\mu} - \frac{1}{2}\frac{\lambda}{\mu^3} + \frac{2}{3}\frac{\lambda^2}{\mu^5} - \frac{11}{8}\frac{\lambda^3}{\mu^7} + \mathcal{O}\left(\lambda^4\right) \quad , \\ G_4 &= \frac{3}{\mu} - 4\frac{\lambda}{\mu^4} + \frac{33}{4}\frac{\lambda^2}{\mu^6} - \frac{68}{3}\frac{\lambda^3}{\mu^8} + \mathcal{O}\left(\lambda^4\right) \quad , \\ G_6 &= \frac{15}{\mu^3} - \frac{75}{2}\frac{\lambda}{\mu^5} + \frac{445}{4}\frac{\lambda^2}{\mu^7} - \frac{1585}{4}\frac{\lambda^3}{\mu^9} + \mathcal{O}\left(\lambda^4\right) \quad . \end{aligned} \tag{17}$$

This approach is called the *perturbation expansion*: the effects of interaction are implicitly considered as 'small corrections' to the free theory[5]. Clearly, similar but more complicated expressions can be obtained in the same way for other interaction terms, even if the path integral itself is not defined, for instance in a φ^3 theory (in which case also the odd Green's functions would be non-zero). The alternative approach would be to use again the integral representation, and now derive another differential equation for $Z(J)$:

$$\frac{1}{6}\lambda\frac{\partial^3}{\partial J^3}Z(J) + \mu\frac{\partial}{\partial J}Z(J) - JZ(J) = 0 \quad . \tag{18}$$

There are 3 independent solutions, none particularly simple; and then we would have to pick the correct one, that gives the perturbative expansion above.

2.1.3 Connected Green's functions

The information about the probability density encoded in the moments G_p can of course also be described in other ways. One of the more useful descriptions (also employed in statistics) is

[3]The factor 1/24 is a convention: in general, we put a factor $1/q!$ before an interaction of type φ^q. Some authors use a different convention, so beware.

[4]Here we use the fact that for a Gaussian with zero mean and unit standard deviation, the $2n^{\text{th}}$ moment is given by $(2n)!/n!2^n$.

[5]See, however, the Appendix to this section.

that in terms of the *cumulants* C_n, defined by

$$Z(J) = \sum_{p\geq 0} \frac{1}{p!} J^p G_p \equiv e^{W(J)} \quad , \quad W(J) = \sum_{n\geq 1} \frac{1}{n!} J^n C_n \quad , \tag{19}$$

so that

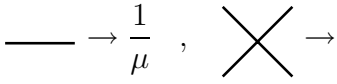
$$C_n = \left\lfloor \frac{\partial^n}{\partial J^n} W(J) \right\rfloor_{J=0} = \left\lfloor \frac{\partial^n}{\partial J^n} \log Z(J) \right\rfloor_{J=0} \quad . \tag{20}$$

In the jargon, the cumulants are called the *connected Green's functions*: the 'connected' will become clear later on. From the definition of Z and W, it can easily be checked that $C_1 = G_1$ is the mean of the probability density, while $C_2 = G_2 - G_1^2$ is its *variance*, C_3 its *skewness*. C_4 is related to the *kurtosis*, and so on. It will turn out that these correspond to the physically more interesting properties of the theory. From Eq.(13) we see that for a free theory only C_2 is nonzero, which is why it is called a free theory. For the φ^4 theory we have

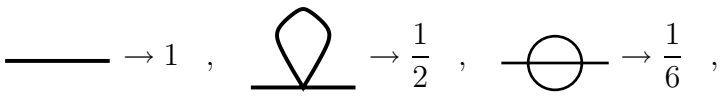
$$\begin{aligned} C_2 &= \frac{1}{\mu} - \frac{1}{2}\frac{\lambda}{\mu^3} + \frac{2}{3}\frac{\lambda^2}{\mu^5} - \frac{11}{8}\frac{\lambda^3}{\mu^7} + \mathcal{O}\left(\lambda^4\right) \quad , \\ C_4 &= -\frac{\lambda}{\mu^4} + \frac{7}{2}\frac{\lambda^2}{\mu^6} - \frac{189}{12}\frac{\lambda^3}{\mu^8} + \mathcal{O}\left(\lambda^4\right) \quad , \\ C_6 &= 10\frac{\lambda^2}{\mu^7} - 80\frac{\lambda^3}{\mu^9} + \mathcal{O}\left(\lambda^4\right) \quad . \end{aligned} \tag{21}$$

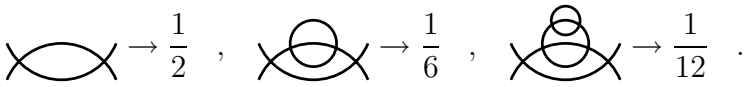
2.2 Feynman diagrams

The Feynman diagram approach is just another way to arrive at the expansions for G_p and C_n. For now, we just give a recipe to get C_n; later on we prove that it is indeed the correct one. To obtain C_n ($n \geq 1$) to order λ^k, take k vertices with 4 legs and join them by lines, in such a way that precisely n lines are sticking out: do this in all possible ways such that the graph is *connected*, so that you can walk from any external line to any other one over the graph. Closed loops are perfectly allowed, and in fact as k increases become unavoidable. Every such *Feynman graph* corresponds to a number, as follows: every line carries a factor $1/\mu$, and every vertex a factor $-\lambda$. There is also a numerical factor $1/(n!k!(4!)^k)$, the coefficient of $(-\lambda)^k J^n$ in the expansion of $e^{-S+J\varphi}$, since that is the integral we are actually trying to do. Now add the results of all Feynman graphs. The assignment of numbers to elements of Feynman graphs are called the *Feynman rules*, and are an important issue in these lectures: **the choice of Feynman rules defines the theory**. In this case, they are:

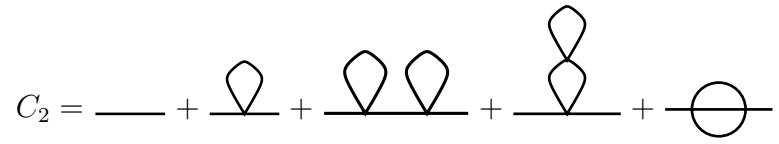


Of course, there are many diagrams that look exactly the same, differing only in the use of one of the k vertices rather than another, or one of its legs rather than another, while the diagram is being built up. It is customary to collect all such *topologically equivalent* diagrams into one. The numerical factor in front is then $1/f$, where f is the number of ways in which legs and vertices can be interchanged without changing the diagram. This is called the order of the symmetry group of the diagram, and $1/f$ is called its *symmetry factor*. Some examples of diagrams with their symmetry factors are:





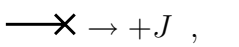
Note that the external legs are counted as distinct. The determination of symmetry factors, although in principle straightforward, can be very cumbersome[6]; fortunately in many cases they are simple or even (as in QED, as we shall see) always unity. As an excercise, you may check that



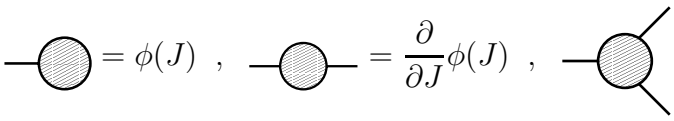
It should be kept in mind that the Feynman diagrams are just funny ways of writing definite mathematical objects: it makes perfect sense to talk about the numerical value of a diagrams.

2.2.1 The Schwinger-Dyson equation

The drawing of Feynman diagrams can be systematized. To this end, we introduce the Feynman rule for the source J, which we now picture by another vertex (denoted by a small cross):



the $+$ sign coming from the fact that we added $-J\varphi$ to the action in Eq.(10). The physical meaning of the source is now apparent: a source acts as an object that gives rise to a line entering the diagram from the 'outside'. In real life, a source may be also described by Feynman diagrams (as in QED, where an electron line may emit a photon, and acts as a souce for this photon), or we may assume the source to be 'infinitely far away' so that the line comes in from 'infinity', as it should for, say, electrons and positron coming out of an accelerator into the interaction region (we shall come back to this point later on). The symmetry factor will now also contain $1/k!$ in diagrams where k such source vertices are present[7]. Each derivative with respect to J corresponds to deleting one such vertex, so that an extra external line is generated. Let us now consider the set of all possible connected diagrams with precisely one external leg. Since (the value of) this set depends on J, we denote it by $\phi(J)$ (beware: φ stands for the 'dummy' integration variable in te path integral, and ϕ denotes a definite number, namely the value of the set of diagrams. There is a good reason for this confusing notation, as we shall see below). Denoting the set of all connected diagrams by a shaded blob, we have



(note that this works because we put in a factor $1/n!$ with J^n, as dictated by our considerations on the symmetry factors), and also

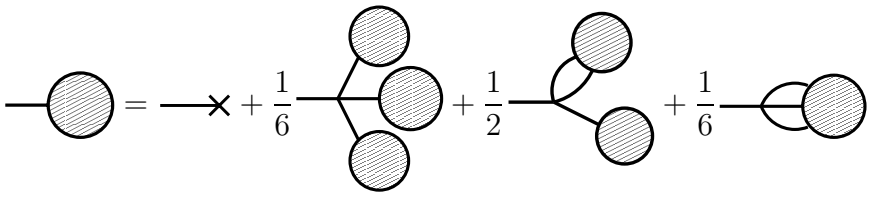
$$C_{n+1} = \left[\frac{\partial^n}{\partial J^n} \phi(J) \right]_{J=0} \quad . \tag{24}$$

Let us now follow the external line of $\phi(J)$ into the diagram. There are several possibilities. The line may immediately encounter a J vertex, and stop there, or a λ vertex, and split into

[6] It is also one of the things that is very difficult to program in a computer.

[7] Lines attached to a J vertex count as *internal* lines.

three. These lines may not come together again, or maybe two of them will, or all three. Diagrammatically, this reads



(25)

where we have to put in the symmetry factors corresponding to the interchange of equivalent internal lines[8]. This is the *Schwinger-Dyson equation* for this theory: in terms of $\phi(J)$, it reads

$$\phi(J) = \frac{J}{\mu} - \frac{\lambda}{6\mu}\phi(J)^3 - \frac{\lambda}{2\mu}\phi(J)\frac{\partial}{\partial J}\phi(J) - \frac{\lambda}{6\mu}\frac{\partial^2}{\partial J^2}\phi(J) \quad . \tag{26}$$

The Schwinger-Dyson (SD) equation allows us to construct the whole set of connected diagrams in a recursive manner[9]. That we get the right diagrams can simply be proven: if we are correct, then we should have

$$\phi(J) = \frac{\partial}{\partial J}W(J) = \frac{1}{Z(J)}\frac{\partial}{\partial J}Z(J) \quad . \tag{27}$$

By making this substitution for $\phi(J)$ in Eq.(18) we indeed obtain Eq.(26); and since the simplest term in $\phi(J)$, namely J/μ, is correct, all the other ones are correct as well. Other theories have their own SD equation: if the action is

$$S(\varphi) = \frac{1}{2}\mu\varphi^2 + \sum_{r\geq 3}\frac{\lambda_r}{r!}\varphi^r \quad , \tag{28}$$

then the SD equation in terms of $Z(J)$ reads

$$\mu\frac{\partial}{\partial J}Z(J) + \sum_{r\geq 3}\frac{\lambda_r}{(r-1)!}\frac{\partial^{r-1}}{\partial J^{r-1}}Z(J) = S'\left(\frac{\partial}{\partial J}\right)Z(J) = JZ(J) \quad . \tag{29}$$

It is left as an excercise to work out the diagrammatic form for a theory with φ^3 as well as φ^4 interactions.

It must be remarked that, once the external leg that serves as the 'starting point' has been chosen, we can for every diagram unambiguously determine to which term in the SD equation it belongs, just by looking at what happens at the 'first vertex'. Therefore, the symmetry factor of the diagram is completely determined by the factors $1/2$ and $1/6$ in the SD equation[10]. In a theory like QED, where there is only one vertex, where three *non*-equivalent lines meet, all symmetry factors are therefore unity. The same holds for the electroweak model as long as no four-boson vertices or Higgs self-interactions are involved.

Finally, it may be realized that $\phi(J)$ has a simple interpretation: in the presence of the source J, the random variable φ has probability density $N\exp(-S(\varphi) + J\varphi)$ rather than $N\exp(-S(\varphi))$. Denoting by $\langle\cdots\rangle_J$ averages with respect to *this* density, we see that

$$\phi(J) = \langle\varphi\rangle_J \quad , \tag{30}$$

which explains our use of two such similar symbols.

[8]Note that this is only unambiguous if the blobs denote *connected* diagrams.

[9]Indeed, it can easily be implemented in computer algebra, starting with $\phi(J) = J/\mu$ and iterating the right-hand side of Eq.(26) .

[10]It is *not*, however, a simple product of such factors, since for individual diagrams the lines entering the loops may be inequivalent: the equivalence holds only for the set of all diagrams.

2.2.2 *Connected, disconnected and vacuum diagrams*

Having seen that the set of all connected diagrams without external legs but with legs ending in J vertices represents $W(J)$, we now understand where the name 'connected Green's function' comes from. The set of Green's functions in $Z(J)$ is then $e^{W(J)} = 1+W(J)+W(J)^2/2+cdots$, and we see that it corresponds to the set of all diagrams, both connected ones and disconnected ones (that consist of two or more separate connected pieces), with a extra symmetry factor $1/m!$ for each connected diagram that occurs precisely m times as a factor in a given term in $Z(J)$. A disconnected diagram consisting of p connected pieces comes from the term $W(J)^p/p!$ in $e^{W(J)}$. Only one type seems to be missing, namely those diagrams that do not contain any J vertex at all: this set is given by

$$\exp\left(\quad + \quad + \quad + \cdots \right) = \exp\left(-\frac{1}{8}\frac{\lambda}{\mu^2} + \frac{1}{12}\frac{\lambda^2}{\mu^4} + \cdots \right) \quad . \tag{31}$$

That this is not really an omission becomes clear if we realize that these are just the diagrams that crop up in the computation of $Z(0)$, and hence are always precisely absorbed in the definition of the normalization N: since the same set of vacuum diagrams always occurs as a factor in any G_p, it is always divided out.

2.2.3 *The loop expansion and classical theory*

The perturbative expansion in λ is straightforward; but in a theory with more than one coupling constant ambiguities may arise, since ***a priori*** the relative orders of these constants are not known: if the action is

$$S(\varphi) = \frac{1}{2}\mu\varphi^2 + \frac{1}{6}\lambda_3\varphi^3 + \frac{1}{24}\lambda_4\varphi^4 \quad , \tag{32}$$

then are we to take λ_3 and λ_4 as being of the saem order? Or should we take, say λ_3^2 to be of order λ_4? There is a more systematic way of ordering the diagrams in a given C_n, by the number of closed loops in each diagram. Let us decide to assign a factor $\hbar$ to each loop[11]. In the SD equation, we can simply account for it by modifying Eq.(25):

$$= \quad + \frac{1}{6} \quad + \frac{\hbar}{2} \quad + \frac{\hbar^2}{6} \quad , \tag{33}$$

or, in other words,

$$\phi(J) = \frac{J}{\mu} - \frac{\lambda}{6\mu}\phi(J)^3 - \frac{\hbar\lambda}{2\mu}\phi(J)\frac{\partial}{\partial J}\phi(J) - \frac{\hbar^2\lambda}{6\mu}\frac{\partial^2}{\partial J^2}\phi(J) \quad . \tag{34}$$

This ordering of the diagrams is called the *loop expansion*. To account for the occurrence of $\hbar$, we have to modify our definitions a bit: from here on, we define

$$\phi(J) = \hbar\frac{\partial}{\partial J}\log Z(J) \quad , \quad Z(J) = N\int d\varphi \ \exp\left(-\frac{1}{\hbar}(S(\varphi) - J\varphi)\right) \quad . \tag{35}$$

[11] Here, $\hbar$ is of course just an arbitrary numerical constant: its significance as a constant of nature becomes only clear in the more-dimensional case.

and the SD equation for $Z(J)$ now becomes

$$S'\left(\hbar\frac{\partial}{\partial J}\right)Z(J) = JZ(J) \quad . \tag{36}$$

The relative order of λ_3 and λ_4, say, can now be fixed: by replacing, in the new definition of $Z(J)$, the integration variable φ by $\varphi' = \varphi\sqrt{\hbar}$, we see that μ,$\lambda_3\sqrt{\hbar}$, and $\lambda_4\hbar$ are all of the same order, and therefore λ_3^2 has the same order as λ_4 in the loop expansion.

It may be interesting to see what happens when $\hbar$ becomes vanishingly small: this is called the *classical limit*. On the one hand, we can simply put $\hbar = 0$ in Eq.(34): for a general theory, this gives

$$\begin{aligned}\phi_c(J) &= \frac{J}{\mu} - \sum_{r\geq 3}\frac{\lambda_r}{(r-1)!\mu}\phi_c(J)^{r-1} \quad \Rightarrow \\ \mu\phi_c(J) &+ \sum_{r\geq 3}\frac{\lambda_r}{(r-1)!}\phi_c(J)^{r-1} = S'(\phi_c(J)) = J \quad ,\end{aligned} \tag{37}$$

from the definition of $S(\varphi)$ as a series in φ. Since $\phi_c(J)$ contains only diagrams without closed loops, it is also called the *tree approximation.* In the path integral, on the other hand, the integrand will show extremely narrow peaks as $\hbar \to 0$, so that $\phi(J) = \langle\varphi\rangle_J$ will really only get contributions from those positions where the exponent has maxima: these correspond, of course, precisely to those values of φ where $S(\varphi) - J\varphi$ has a minimum on the integration axis. The lowest of these minima (assuming that it is unique) gives the dominant contribution, and for this value, φ_c, we of course have

$$S'(\varphi_c) = J \quad . \tag{38}$$

This is called the *classical field equation.* Note that the other minima will give contributions that are suppressed relative to the dominant one by $e^{-a/\hbar}$, with a some positive number: such contributions, that are of course totally invisible in an expansion in powers of $\hbar$, are inherently *non*-perturbative. Usually one tries to arrange things such that $\varphi_c = 0$ when J vanishes[12]; the other solutions of the classical field equation are called *instantons*.

2.2.4 The effective action

Since the classical equation looks simpler than the full quantum-mechanical one, the question arises: can we find another action (called the *effective action*) such that *its* classical approximation coincides with $\phi(J)$? If so, we would need to compute only tree diagrams, which is certainly simpler than also computing loops. Denoting the effective action by $\Gamma(\varphi)$, it must therefore be such that

$$\Gamma'(\phi(J)) = J \quad . \tag{39}$$

Let us assume that the relation between J and ϕ can be inverted, at least in some neighbourhood of $J = 0$:

$$\phi = \phi(J) \quad \leftrightarrow \quad J = y(\phi) \quad . \tag{40}$$

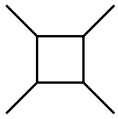
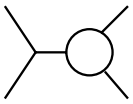
Integration by parts, and using $phi\ dJ = dW$, then gives us Γ:

$$\Gamma(\phi) = \int y(\phi)\, d\phi = \phi y(\phi) - \hbar W(y(\phi)) \quad : \tag{41}$$

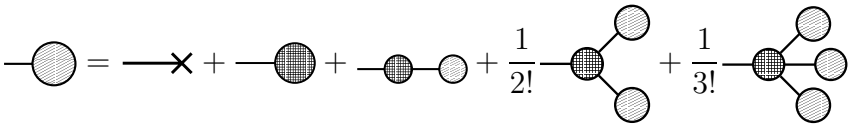
[12]This may necessitate a shift in the definition of φ, as in the case of spontaneous symmetry breaking.

The effective action is the Legendre transform of $W(J)$.

In terms of Feynman diagrams, it is useful to introduce the notion of a *one-particle irreducible* (1PI) diagram. Such a diagram can *not* be made disconnected by cutting through one internal line. For example:

 is 1PI ;  is not 1PI .

A single vertex also counts as 1PI, since there aren't any internal lines to cut. Let us now redo the reasoning that led to the SD equation. Coming into the diagram, we shall encounter a 1PI piece of it – possibly only a single vertex. Out of this piece stick $0, 1, 2, 3, \ldots$ lines that, when cut, *do* make the diagram fall apart. Denoting the 1PI pieces with dark blobs, we therefore have

 (42)

leading to

$$\mu\phi(J) + \sum_{r\geq 1} \frac{\gamma_r}{(r-1)!}\phi(J)^{r-1} = J \quad , \tag{43}$$

where $-\gamma_r$ denotes the sum of all 1PI diagrams with at least one vertex and precisely r external legs. Comparing with Eq.(37), we see that the effective action is therefore given by

$$\Gamma(\varphi) = \gamma_1\varphi + \frac{1}{2}(\mu+\gamma_2)\varphi^2 + \sum_{r\geq 3}\frac{\gamma_r}{r!}\varphi^r \quad . \tag{44}$$

Note that since it is usually easy to construct 1PI diagrams with very many external legs, the sum over r runs up to infinity. For the pure φ^4 theory, we obtain[13]

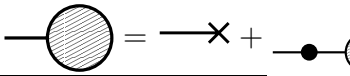
$$\begin{aligned}\Gamma(\varphi) \;&=\; \left(\frac{1}{2}\mu\varphi^2 + \frac{1}{24}\lambda\varphi^4\right) \;+\; \frac{\hbar}{2}\log\left(1+\frac{\lambda\varphi^2}{2\mu}\right) \\ &\quad -\frac{\hbar^2\lambda^2\varphi^2\left(40\mu^2+18\lambda\varphi^2+3\lambda^2\varphi^4\right)}{24\mu^2\left(2\mu+\lambda\varphi^2\right)^3} \;+\; \mathcal{O}\left(\hbar^3\right) \quad .\end{aligned} \tag{45}$$

2.2.5 Dyson summation

In the foregoing, it might seem that the φ^2 term in the action is treated differently from the interaction terms: after all, the quadratic term occurs like $1/\mu$ in the Feynman rules, while the coupling constants appear with positive power. That this is not really the case becomes clear when we consider the action

$$S(\varphi) = \frac{1}{2}\mu\varphi^2 + \frac{\lambda_2}{2}\varphi^2 + \sum_{r\geq 3}\frac{\lambda_r}{r!}\varphi^r \quad , \tag{46}$$

and decide to treat the λ_2 as a perturbation. We then have

 + (higher interaction terms) , (47)

[13] By first rewriting Eq.(34) as an equation for $y(\varphi)$ rather than for $\phi(J)$, then solving it perturbatively as a power series in $\hbar$, and then integrating over φ.

where the fat dot stands for the λ_2 vertex, or

$$\phi(J) = \frac{J}{\mu} - \frac{\lambda_2}{\mu}\phi(J) - \frac{1}{\mu}(\text{higher interaction terms}) \quad , \tag{48}$$

and we have indicated all occurrences of μ. By bringing the term linear in $\phi(J)$ to the left, we can see that this is exactly the same as

$$\phi(J) = \frac{J}{\mu+\lambda_2} - \frac{1}{\mu+\lambda_2}(\text{higher interaction terms}) \quad , \tag{49}$$

in other words, we might as well have absorbed the λ_2 into the μ from the start. This absorption of quadratic interactions into the μ term is called *Dyson summation.* The apparently different treatment of quadratic terms is related to the fact that the perturbation expansion is one around the Gaussian form in the path integral.

2.3 Renormalization

So far, we have treated all parameters in the action as if they were in some way logically prior to the physics they imply. In reality, of course, it is just the other way around: in the practice of physics, one tries to obtain the couplings etc. from some *measurement* experiment by a fit to data, and then uses them in a *prediction* for the result of some other measurement. It is only the measurements that have a claim to physical reality: strictly speaking, masses and couplings are just bookkeeping devices. This is at the basis of *renormalization*: if, by some effort, we have computed yet one higher order (in $\hbar$, say) in our prediction, this is useless unless we have the corresponding higher precision in our extraction of the parameters from their measurement. It also (and fortunately!) means that if two parameters always occur in precisely the same combination in our expressions for the measurement and prediction, we cannot, and should not try to, disentangle them. This is the way in which the infamous loop divergences in quantum field theory become 'absorbed' into the physical parameters, but we must keep in mind that, even for a perfectly finite theory, renormalization is *always* necessary as long as we are working in perturbation theory.

2.3.1 Finite renormalization in zero dimensions

To illustrate the above, let us again consider φ^4 theory. If we include the correct factors $\hbar$, the first few connected Green's functions are given by

$$\begin{aligned}
C_2 &= \frac{1}{\mu} - \frac{\hbar}{2}\frac{\lambda}{\mu^3} + \frac{2\hbar^2}{3}\frac{\lambda^2}{\mu^5} - \frac{11\hbar^3}{8}\frac{\lambda^3}{\mu^7} + \mathcal{O}\left(\hbar^4\right) \quad , \\
C_4 &= -\frac{\lambda}{\mu^4} + \frac{7\hbar}{2}\frac{\lambda^2}{\mu^6} - \frac{189\hbar^2}{12}\frac{\lambda^3}{\mu^8} + \mathcal{O}\left(\hbar^3\right) \quad , \\
C_6 &= 10\frac{\lambda^2}{\mu^7} - 80\hbar\frac{\lambda^3}{\mu^9} + \mathcal{O}\left(\hbar^2\right) \quad .
\end{aligned} \tag{50}$$

There are 2 parameters, μ and λ, so we need 2 measurements to fix them. Let us take these to be the measurements of C_2 and C_4. From the measured values of C_2 and C_4, we extract our parameters and then use them to predict C_6, say. Each new order in $\hbar$ will change our extracted value of μ and λ, and so we should write

$$\begin{aligned}
\mu &= \mu_0 + \mu_1\hbar + \mu_2\hbar^2 + \mu_3\hbar^3 + \cdots \quad , \\
\lambda &= \lambda_0 + \lambda_1\hbar + \lambda_2\hbar^2 + \lambda_3\hbar^3 + \cdots \quad .
\end{aligned} \tag{51}$$

At each fixed order in perturbation theory, we can extract the truncated values:

$$\mu^{(p)} = \sum_{k=0}^{p} \mu_k \hbar^k \quad , \quad \lambda^{(p)} = \sum_{k=0}^{p} \lambda_k \hbar^k \quad . \tag{52}$$

These lead, to the correct order in $\hbar$, to similarly truncated approximations to C_6, which we call $C_6^{(p)}$. At the lowest (tree) order, we simply have

$$\mu^{(0)} = \mu_0 = \frac{1}{C_2} \quad , \quad \lambda^{(0)} = \lambda_0 = -\frac{C_4}{C_2^4} \quad , \quad C_6^{(0)} = 10\frac{\lambda_0^2}{\mu_0^7} \quad . \tag{53}$$

To first order, we have to determine μ_1 and λ_1 such that the values of C_2 and C_4 remain unchanged, and update the (truncated) prediction for C_6:

$$\mu_1 = -\frac{\lambda_0}{2\mu_0} \quad , \quad \lambda_1 = \frac{2\lambda_0^2}{3\mu_0^2} \quad , \quad C_6^{(1)} = 10\frac{\lambda_0^2}{\mu_0^7} - 15\hbar\frac{\lambda_0^3}{\mu_0^9} \quad , \tag{54}$$

and so on. For the sake of numbers, let us assume that $C_2 = 1$ and $C_4 = -2$, and take $\hbar = 0.01$. In the table we give the updated values of μ, λ and C_6.

$\mathcal{O}(\hbar^p)$	$\mu^{(p)}$	$\lambda^{(p)}$	$C_6^{(p)}$	wrong
0	1.000000	2.000000	40.00000	40.00000
1	0.990000	2.060000	38.80000	33.60000
2	0.989767	2.060600	38.87200	34.41867
3	0.989766	2.060622	38.86600	34.31619
4	0.989766	2.060621	38.86662	34.32951
5	0.989766	2.060621	38.86654	34.32767
6	0.989766	2.060621	38.86655	34.32795
7	0.989766	2.060621	38.86655	34.32790
8	0.989766	2.060621	38.86655	34.32791
9	0.989766	2.060621	38.86655	34.32791

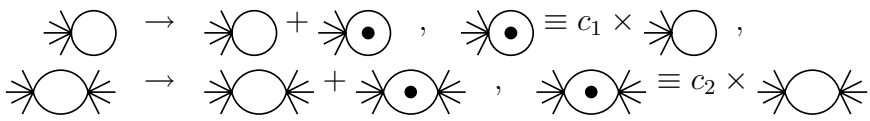
The column labelled 'wrong' is the result for C_6 where we have just included the higher orders in the prediction, while keeping μ and λ fixed to their lowest-order values 1 and 2. For this value of $\hbar$, perturbation theory appears to do nicely – but if you feel happy now, read the Appendix!

2.3.2 Divergences and renormalization: a toy model

In real, four-dimensional life, the higher-order corrections are of course both more complicated than just simple combinations of μ and λ, and more dangerous, since loop diagrams tend to give divergences. Fortunately, in many cases these are automatically taken care of by renormalization, provided the theory has the right structure. In order to see how this works, we shall study a simple toy model.

In order to include, in an admittedly crude way, the divergence structure of the loop diagrams in the four dimensionional theory into our zero-dimensional model, let us add a complication to the zero-dimensional Feynman rules, as follows. We decide that every closed loop with precisely one vertex on it shall be multiplied by a factor $1 + c_1$, and every closed loop with precisely 2 vertices on it by a factor $1 + c_2$; loops with more vertices are not modified. Here, c_1 and c_2 are numbers that eventually go to infinity. To handle this, we assume that they

both depend on some parameter Λ, and diverge as $\Lambda \to \infty$. Diagrammatically, we have the replacement



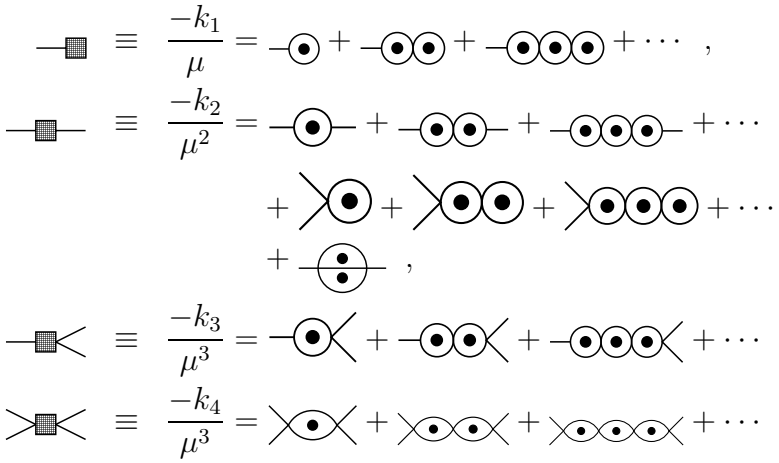
(55)

Let us consider a theory with both φ^3 and φ^4 couplings:

$$S(\varphi) = \lambda_1 \varphi + \frac{1}{2}\mu\varphi^2 + \frac{1}{6}\lambda_3\varphi^3 + \frac{1}{24}\lambda_4\varphi^4 \quad . \tag{56}$$

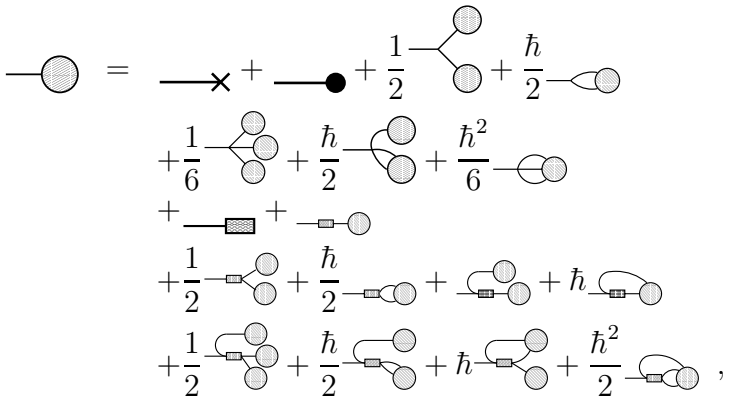
Note that we have also introduced a term linear in φ. This is because we would like to interpret particles as fluctuations in the field: for the situation with no particles, the vacuum, it is most reasonable to also require the field to be absent. Hence, we require that $\phi(0) = \langle\varphi\rangle = 0$. For pure φ^4 theory this is automatically ensured by the $\varphi \leftrightarrow -\varphi$ symmetry of the action, but when a φ^3 term is present, the coupling λ_1 has to be renormalized to achieve this.

Of course, the SD equation will have to be modified in such a way that now also the dotted diagrams are automatically generated, symmetry factors and all. This can be done by introducing new vertices $k_1, \ldots, k_4$, as follows:



(57)

Writing a dot for the λ_1 vertex, we can now give the modified SD equation that includes the dotted diagrams correctly:



(58)

where we have indicated the powers of $\hbar$ and the symmetry factors coming from equivalent lines (note that in the new 'box' vertices not all lines are equivalent). We can now see that if we define

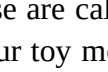
$$\lambda_1' = \lambda_1 + k_1 \quad , \quad \mu' = \mu + k_2 \quad , \quad \lambda_3' = \lambda_3 + 3k_3 \quad , \quad \lambda_4' = \lambda_4 + 3k_4 \quad , \tag{59}$$

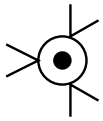
this SD equation is exactly that corresponding to the action

$$S(\varphi) = \lambda_1'\varphi + \frac{1}{2}\mu'\varphi^2 + \frac{1}{6}\lambda_3'\varphi^3 + \frac{1}{24}\lambda_4'\varphi^4 \quad . \tag{60}$$

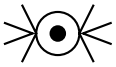
In other words, the *bare* parameters μ and λ_i are themselves invisible, and only the *renormalized* parameters μ' and λ_i' occur. The principle of renormalization then *forces* us to tune the bare parameters in such a way that they completely absorb all effects of $c_{1,2}$. Admittedly this means that the bare parameters become divergent as well – but since they show up nowhere by themselves, this is actually irrelevant.

2.3.3 Nonrenormalizable theories

The above toy model is of course very crude: in reality, there are many more divergent loop diagrams than just the ones we used. These are called overlapping divergences, occurring for instance in the diagram  , which in our toy model has no dotted partner. The significant point is that the divergencies are located in diagrams (or sub-diagrams) with just so many legs that they can be absorbed into a combination with bare vertices of the theory. That this is not always possible can be seen from some examples. Suppose we also had a divergence in loops with 3 vertices. The following diagram would then occur:



which would have to be absorbed into a 6-point vertex. In a φ^4 theory there is no such vertex, and the divergence would remain. If we try to save the day by going to a theory with additional φ^6 couplings, then again there would be diagrams like



which would necessitate a φ^8 vertex, and so on. Although a theory with an infinite number of parameters is by itself not forbidden, it would require us to make an infinite number of measurements before we could start on any prediction, and physics would be hopeless. Such theories are called *nonrenormalizable*.

Suppose the largest power of φ in the action is p. A loop with up to k vertices on it can then have up to $k(p-2)$ legs sticking out from it. If it is divergent, the theory will only be renormalizable if $k(p-2) \leq p$. If $k = 1$, all theories are renormalizable; for $k = 2$, only $p \leq 4$ is allowed, and this is the reason why φ^4 theory is used as our paradigm[14].

2.3.4 Running couplings and the β function

So far, we have not discussed the Λ dependence of the renormalization procedure. Typically, Λ contains some parameter that regularizes the loop divergences (like the famous $1/\epsilon$ in dimensional regularization) but it may also contain physical information. In actual theories like QCD, this can for instance be the energy scale at which the measurements (C_2 and C_4 in the above) are

[14]This is also the reason why the self-interactions of the Higgs boson in the Standard Model are chosen to be of type $\varphi^3 + \varphi^4$, and not higher.

performed, but in general it might be anything else, for instance the time of day or the altitude at which the measurements are performed. We shall simply assume that there is some finite parameter s, the *scale*, such that a change in s leads to a change in Λ. We can always choose s such that the dependence is linear with unit derivative: $d\Lambda = ds$ (note that what we mean here by 'scale' is not the same as what is usually implied: in QCD, for instance, our 'scale' is $\log(Q^2)$ rather than Q^2).

Now suppose that the measurements are performed at two different scales, s_1 and s_2. The bare parameters that we extract may be divergent (and we need to regularize), but their extracted values should be the same in both cases, since the action doesn't know about external experimental information like the scale. This means that the renormalized parameters at the two scales should be related, and that having performed the measurement at scale s_1 we ought to be able to predict what would have been the outcome if we had performed it at scale s_2 instead. The outcome might be different, and in that case we say that the renormalized parameters are *scale-dependent*; but the very least requirement is that at both scales the renormalized parameters are finite!

In order to illustrate all this let us suppose that we have a theory with a single dimensionless parameter v. An example is QCD with massless quarks, where $v = \alpha_s$. After absorbing the loop divergences we obtain a finite renormalized parameter, which we shall call w. This depends on the scale *via* its dependence on Λ:

$$w(s) = F(v, \Lambda) \quad . \tag{61}$$

For simplicity, we assume that $F(v, 0) = v$ (this is not really drastic: if $F(v, 0) = f(v)$ is not simply v, we just take $f^{-1}(F(v, \Lambda))$ and call *that* the renormalized parameter w, simply a finite function of the old one).

Since v can by assumption be extracted from the measurement of $w(s)$, we also have the inverse function:

$$v = G(w, \Lambda) \quad , \quad v = G(F(v, \Lambda), \Lambda) \quad . \tag{62}$$

A small change in scale $s \to s+ds$ or, equivalently, $\Lambda \to \Lambda+d\Lambda$, should lead to a corresponding change in w:

$$\frac{d}{ds} w(s) \equiv \beta(w, \Lambda) = F_2(v, \Lambda) = F_2(G(w, \Lambda), \Lambda) \quad , \tag{63}$$

where a subscript i denotes partial derivation with respect to the i^{th} argument. We have

$$F_1 G_1 = 1 \quad , \quad F_2 G_1 + G_2 = 0 \quad , \tag{64}$$

and therefore

$$\beta(w, \Lambda) = -\frac{G_2(w, \Lambda)}{G_1(w, \Lambda)} \quad . \tag{65}$$

If we insist that the renormalized parameter at the new scale is also finite, all occurrence of Λ should disappear in the β function:

$$\frac{\partial}{\partial \Lambda} \beta(w, \Lambda) = 0 \quad \Rightarrow \quad \frac{d}{ds} w(s) = \beta(w) \quad , \tag{66}$$

The last equation is known as the *renormalization group equation*. It implies that G can be written in the form[15]

$$G(w, \Lambda) = \mathcal{G}(-\Lambda + b(w)) \quad , \quad b(w) = \int \frac{1}{\beta(w)} dw \quad . \tag{67}$$

[15] By separation of variable in Eq.(65).

Since $w = v$ for $\Lambda = 0$, the functions $\mathcal{G}$ and b are each other's inverse. This also means that

$$w = \mathcal{G}(\Lambda + b(v)) = F(v,\Lambda) \quad , \quad \frac{F_2(v,\Lambda)}{F_1(v,\Lambda)} = \beta(v) \quad . \tag{68}$$

Thus, knowledge of $\beta(w)$ completely fixes the form of F: if a theory results in an F with a different form, it is only meaningful at precisely the scale at which the measurements are made, but not at any other one.

In perturbation theory, we can compute the first few terms in F by hand, given the beta function:

$$\begin{aligned} \beta(v) &= \beta_0 v^2 + \beta_1 v^3 + \beta_2 v^4 + \cdots \quad , \\ F(v,\Lambda) &= v + \beta_0 \Lambda v^2 \\ &\quad + \left(\beta_0^2 \Lambda^2 + \beta_1 \Lambda\right) v^3 + \left(\beta_0^3 \Lambda^3 + \frac{5}{2}\beta_0\beta_1\Lambda^2\beta_2\Lambda\right) v^4 + \cdots \quad . \end{aligned} \tag{69}$$

The condition $F(v,0) = v$ that we started with is of course just one of the possibilities, depending amongst other things on the precise definition of Λ. Different such *renormalization prescriptions* or *schemes* are possible, and we just used a particularly simple one.

The renormalized parameters in two different schemes must be related in a finite manner: if we adopt another scheme in which the renormalized parameter is w', then we have

$$w = h(w') = w' + h_1 w'^2 + h_2 w'^3 + h_3 w'^4 + \cdots \quad , \tag{70}$$

which leads to a new β function, β':

$$\beta'(w') = \frac{d}{ds} w' = \frac{dw'}{dw}\beta(w) = \frac{1}{h'(w')}\beta(h(w')) \quad . \tag{71}$$

where h' denotes the derivative. It is easy to check that the first two coefficients in the expansions of $\beta(w)$ and $\beta'(w')$ coincide: scheme dependence shows up in the third and higher terms.

In theories with more parameters we have a set of coupled renormalization group equations that have to be solved simultaneously.

2.4 Appendix: Convergence of perturbation theory

In setting up perturbation theory we have assumed that in φ^4 theory, say, the interaction term is a 'small disturbance' of the Gaussian shape of the free theory. Close to the peak this may be true, but for large φ values it is not: the tails of the distribution are nothing like Gaussian. As a consequence, perturbation theory does *not* converge! To see this, look again at the G_{2p} of Eq.(15), which (putting in the $\hbar$) we can also write as

$$\begin{aligned} G_{2p} &= N\sqrt{2\pi}\frac{(2p)!}{p!}\left(\frac{\hbar}{2\mu}\right)^p \sum_{k\geq 0}(-)^k T_k \quad , \\ T_0 &= 1 \quad , \quad \frac{T_k}{T_{k-1}} = \frac{\hbar\lambda}{\mu^2}\frac{(2p+4k-1)(2p+4k-3)}{24k} \quad . \end{aligned} \tag{72}$$

The radius of convergence of this series[16] is zero for any nonzero λ, since $T_k/T_{k-1} \to \infty$ as $k \to \infty$. For small λ, the terms T_k will start out decreasing, but at $k \sim 3\mu^2/2\hbar\lambda$ they will

[16] The radius of convergence of a power series in z around $z = 0$ is defined as the largest value of $|z|$ such that the power series converges, *i.e.* the subsequent terms in the series decrease fast enough.

start increasing again, and the perturbative prediction will start to oscillate wildly: perturbation theory breaks down at order $3\mu^2/2\hbar\lambda$. For $\mu = 1$, $\lambda = 2$, $\hbar = 0.01$ this means around the 75^{th} order, so it is not surprising that the table in section 2.3.1 looked reliable. This breakdown is a general feature as long as we expand the cubic or higher interactions around the quadratic part. That the problem is not cured by either dropping the vacuum diagrams, going over to connected Green's functions, or applying renormalization, becomes clear when we redo the finite renormalization procedure for C_6, this time for $\hbar = 0.3$:

$\mathcal{O}(\hbar^p)$	$\mu^{(p)}$	$\lambda^{(p)}$	$C_6^{(p)}$	wrong
0	1.000000	2.000000	40.00000	40.00000
1	.700000	3.800000	4.000000	-152.0000
2	.490000	4.340000	68.80000	584.8000
3	.463000	4.934000	-93.20000	-2182.160
4	.390100	4.20500	405.7600	8612.692
5	.517270	6.63014	-1384.664	-36118.36
6	.131629	-1.59249	5869.178	162372.5
7	1.43778	30.2795	-26669.53	-782851.6
8	-3.58970	-106.449	132819.5	404416.2
9	17.8758	535.079	-713389.2	-22335606

Mathematically, this awful behaviour is our punishment for interchanging the sum and integral in Eq.(15), and series such as these are called *asymptotic series*. Fortunately, since the terms not only increase but also oscillate we can still assign a meaning to such sums by the procedure called *Borel summation*, which in essence interchanges sum and integral back again. To see how this works, take a function

$$f(x) = \sum_{n\geq 0} (-x)^n \alpha_n \quad (x \geq 0) \ , \tag{73}$$

where α_n grows like $n!$ for large n (as in the case of the perturbative expansion). Now, $f(x)$ is not a convergent series, but

$$g(x) = \sum_{n\geq 0} (-x)^n \frac{\alpha_n}{n!} \tag{74}$$

is convergent, at least for some $x > 0$. Integrating term by term we can then see that a sensible definition of the value of $f(x)$ is

$$f(x) \equiv \int_0^\infty dy \; e^{-y} g(xy) \tag{75}$$

As an example, take, $\alpha_n = n!$; in that case

$$g(x) = \frac{1}{1+x} \quad , \quad f(x) = \int_0^\infty dy \; \frac{e^{-y}}{1+xy} \ , \tag{76}$$

and the resulting Borel-summed expression for $f(x)$ is well-defined[17] for nonnegative x, while its expansion in powers of x is not a convergent series. Note that for x real and negative, the integral is ill-defined, and the terms in the series for $f(x)$ do not oscillate any more.

[17] For the experts: it is equal to $e^{1/x} E_1(1/x)/x$, where E_1 is the exponential integral.

Perhaps of more direct computational relevance, asymptotic series like (73) have the property that if we truncate them at some order, the error made is smaller than the first neglected term, in absolute value:

$$f_{(m)}(x) \equiv \sum_{n=0}^{m-1} (-x)^n \alpha_n \quad , \quad \left| f_{(m)}(x) - f(x) \right| \leq |x^m \alpha_m| \quad . \tag{77}$$

Hence, by choosing m such that $|x^m \alpha_m|$ is minimal, we may hope to still get a good numerical approximation to the real answer. In our example, with $\alpha_n = n!$, the best m is seen to be $m \sim 1/x$, and the error made is of the order of $m! x^m \sim \exp(-1/x)$, which for small x is very small indeed! As long as we do not plan to compute the 137^{th} order, QED would appear to be safe, but in QCD the potential hazards are closer at hand, around the 10^{th} order. In any case it is good to keep in mind that perturbation theory is a crutch with cracks.

3 Enter the universe: many points make a space

3.1 One-dimensional theories

We now start to expand our model of the universe. While still keeping to a particle without intrinsic properties other than its mass, we shall allow it to live on more than just a single point. Moreover, we shall adopt the usual practice of choosing our units such that both $\hbar$ and c, the speed of light, have numerical value 1.

3.1.1 Theories with more field variables, and space

We shall start with a simple extension of our zero-dimensional model, and take an infinite set of fields, labelled by integer labels: φ_n, with $-\infty < n < \infty$. Since we are interested in expectation values of products of fields, like $\langle \varphi_2 \varphi_0 \varphi_1 \varphi_5 \rangle$, we also need to introduce an infinite number of sources J_n, one J for every φ, since we use the J's to identify and count the powers of the various φ's. In the spirit of perturbation theory, we shall for now forget interactions that contain three or more φ's. If the action only contained terms of type φ_n^2, we would end up with just an infinite number of copies of the zero-dimensional theory, and no correlations between fields would exist. Therefore, we take the action to be

$$S(\{\varphi\}) = \sum_n \left(\frac{1}{2} \mu \varphi_n^2 - \gamma \varphi_n \varphi_{n-1} \right) \quad , \tag{78}$$

and the path integral, the multi-variable generating function of the expectation values of all possible products of all possible powers of the various φ's, becomes

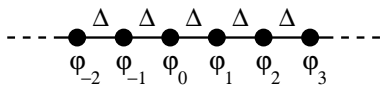
$$Z(\{J\}) = N \int \left(\prod_k d\varphi_k \right) \exp \left(-S(\{\varphi\}) + \sum_n \varphi_n J_n \right) \quad . \tag{79}$$

Now we have interactions of nearest-neighbour type, and correlations between fields further away can only go *via* the intermediate fields: we expect the correlator

$$\langle \varphi_m \varphi_n \rangle \equiv \Pi(n, m) = \Pi(|n - m|) \tag{80}$$

to be smaller when $|n - m|$ is larger, and this is just what we expect for fields at large distances. In fact, it is the 'link' γ that tells us that we might envisage the fields as living on a set of points

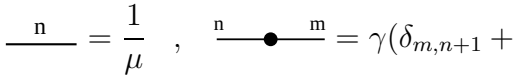
in some space, the distance between neighbouring points becoming 'larger' as γ decreases. **The neighbour-interactions tell us the structure of space**. This may at first sight appear mystical, but we should realize that we obtain our notions of space and time from ideas about how 'distances' between space-time events are defined: these are typically based on measuring correlations (for instance by sending light signals etcetera) and so they depend directly on the neighbour-interactions. In this simple case, since γ and μ do not depend on n, we can envisage the points to be equidistant, separated by some 'distance' Δ:



and the whole model is invariant under translation by one distance Δ. Of course, in the end we shall take $\Delta \to 0$, and recover normal translation invariance.

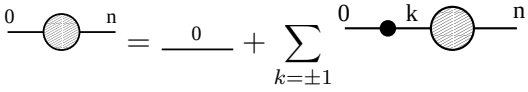
3.1.2 The one-dimensional propagator

Let us write down the Feynman rules for our model. Since there are more fields we need to label each line:



$$= \frac{1}{\mu} \quad , \quad = \gamma(\delta_{m,n+1} + \delta_{m,n-1}) \quad . \tag{81}$$

The *propagator* $\Pi(n)$ is defined as the 2-point Green's function $\langle \varphi_k \varphi_{k\pm n} \rangle$, and is given by Feynman diagrams with precisely two external legs, one corresponding to φ_k, the other one to φ_{k+n}, say. Because of translation invariance we may take $k = 0$. The SD equation for the propagator reads



$$\sum_{k=\pm 1} \quad , \tag{82}$$

or

$$\Pi(n) = \frac{1}{\mu}\delta_{n,0} + \frac{\gamma}{\mu}\left(\Pi(n+1) + \Pi(n-1)\right) \quad . \tag{83}$$

We can solve this by Fourier transform. Defining

$$\Sigma(z) = \sum_n \Pi(n) e^{-izn} \quad \leftrightarrow \quad \Pi(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} dz\, e^{izn} \Sigma(z) \quad , \tag{84}$$

the SD equation becomes purely algebraic:

$$\Sigma(z) = \frac{1}{\mu} + \frac{\gamma}{\mu}\Sigma(z)\left(e^{iz} + e^{-iz}\right) = \frac{1}{\mu - 2\gamma\cos z} \quad . \tag{85}$$

We now take the *continuum limit*, $\Delta \to 0$, and see if we can obtain a sensible limit for the propagator. We keep the 'distance' between fields fixed, so we use $x = n\Delta$ with x fixed. We also introduce the momentum as $p = z/\Delta$ such that $nz = px$. Note that then $\cos z = \cos(p\Delta) \sim 1 - p^2\Delta^2/2$ for fixed p. The only sensible and nontrivial continuum limit occurs when we choose

$$\gamma \sim \frac{1}{\Delta} \quad , \quad \mu \sim \frac{2}{\Delta} + m^2\Delta \quad , \tag{86}$$

for small Δ, with some number m^2. The correlator then becomes[18]

$$R(x) \equiv \Pi\left(\frac{x}{\Delta}\right) \sim \frac{1}{2\pi}\int dp\, \frac{e^{ipx}}{p^2 + m^2} = \frac{1}{2m} e^{-m|x|} \quad . \tag{87}$$

[18] Some care has to be taken here, since the approximation for the cosine is not valid everywhere. Fortunately, in the resulting integral only values of p of order $1/x$ contribute appreciably, and there the approximation *is* justified.

Note that, whereas the fields φ_n are not independent due to the γ interaction, the *momentum modes* are in fact independent: the description of particles in terms of states with given momentum is simpler than its configuration-space description, and we shall use momentum-space Feynman rules from now on.

Finally, it should be remarked that nearest-neighbour interaction gives the standard propagator, but there are suitably perverse choices including next-to-nearest neighbour or higher interactions that result in propagators with different behaviour.

3.1.3 The scalar action in one dimension

Let us now consider the action in the continuum limit. We can collect the whole configuration of values $\{\varphi\} = (\cdots, \varphi_{-1}, \varphi_0, \varphi_1, \varphi_2, \cdots)$ into a *function* $\varphi(x)$, where $\varphi(x) = \varphi_{x/\Delta}$. The action depends on this function, and is called a *functional* of the field $\varphi(x)$. Carefully collecting terms, and rewriting $\sum_n \ \Delta \ \cdots$ as $\int \ dx \ \cdots$, we see that the limit becomes (*cf.* Eq.(78))

$$\begin{aligned} S(\{\varphi\}) &= \sum_n \left((\frac{1}{2}\mu - \gamma)\varphi_n^2 + \frac{1}{2}\gamma(\varphi_{n+1} - \varphi_n)^2 \right) \\ &\sim \sum_n \left(\frac{m^2\Delta}{2}\varphi(x)^2 + \frac{1}{2\Delta}(\varphi(x+\Delta) - \varphi(x))^2 \right) \ \rightarrow \\ \rightarrow \ S[\varphi(x)] &= \int dx \ \left(\frac{1}{2}m^2\varphi(x)^2 + \frac{1}{2}\varphi'(x)^2 \right) \ \equiv \ \int dx \ \mathcal{L} \ , \end{aligned} \tag{88}$$

where we have defined the Lagrangian density $\mathcal{L}$. A sensible limit for the source terms can only be obtained if we put

$$J_n \rightarrow \Delta J(n\Delta) \quad \text{so that} \quad \sum_n \varphi_n J_n \rightarrow \int dx \ \varphi(x) J(x) \ , \tag{89}$$

and the path integral reads

$$Z[J(x)] = N \int \ \mathcal{D}\varphi \ \exp\left(- \int dx \ (\mathcal{L} - \varphi(x)J(x)) \right) \ . \tag{90}$$

Here, $\mathcal{D}\varphi$ is the 'infinitesimal path integration element', the limiting case[19] of the earlier (enumerable) $\prod_n \varphi_n$. A word on the behaviour of $\varphi(x)$ is in order. From our interpretation of the action as describing a combined probability density for the field values φ_n, it follows that the quantity $(\varphi_{n+1} - \varphi_n)^2/\Delta$ must be finite if a particular configuration of φ's is to contribute to the path integral. So, steps in x of length Δ typically make $\varphi(x)$ jump with steps of order $\sqrt{\Delta}$, in other words: in the limit $\Delta \rightarrow 0$ this $\varphi(x)$ must be *everywhere continuous but nowehere differentiable*. Pictorially, a typical $\varphi(x)$ has a fractal structure, with zigs and zags on ever smaller lengths scales, like the path of a particle undergoing Brownian motion[20]. The term $\varphi'(x)^2$ in the action has, therefore, a rather symbolic meaning.

3.1.4 The classical field equation in one dimension

For a single field in zero dimensions, the classical field equation reads $S'(\phi) = J$, with solution $\phi_c(J)$. This does *not* mean that this solution is the only one contributing to the path integral, but

[19] This limit is actually rather more subtle than presented here, since the number of continuous functions of x is not enumerable. For our purposes, however, we may skip such fine points.

[20] In fact, for $m = 0$ it *is* Brownian motion in one dimension, where $\varphi(x)$ is the position at time x.

rather that a small interval (whose size diminishes with $\hbar$) of all φ values *around* $\varphi_c = \phi_c(J)$ gives the dominant contribution. In the discrete one-dimensional case, the analogue of the classical field equation is

$$\frac{\partial}{\partial \varphi_k} S(\{\varphi\}) = J_k \quad \forall k \ . \tag{91}$$

For the action (78), this results in

$$\mu \varphi_k - \gamma(\varphi_{k-1} + \varphi_{k+1}) = J_k \ , \tag{92}$$

so the 'equation of motion' reads

$$\begin{aligned} &\Delta m^2 \varphi(x) + \frac{1}{\Delta}(2\varphi(x) - \varphi(x-\Delta) - \varphi(x+\Delta)) = \Delta J(x) \\ &\Rightarrow m^2 \varphi(x) - \frac{1}{\Delta}\left[\left(\frac{\varphi(x+\Delta) - \varphi(x)}{\Delta}\right) - \left(\frac{\varphi(x) - \varphi(x-\Delta)}{\Delta}\right)\right] = 0 \ . \end{aligned} \tag{93}$$

The second term is seen to be the discrete variant of a derivative. Denoting this discrete derivative by $\mathcal{E}$, we see that the action in the discrete case can be written as

$$S(\{\varphi\}) = \sum_n \Delta \left(\frac{m^2}{2}\varphi^2 + \frac{1}{2}(\mathcal{E}\varphi)^2\right) \ , \tag{94}$$

and the equation of motion as

$$\frac{\partial}{\partial \varphi} S(\{\varphi\}) - \mathcal{E}\left(\frac{\partial}{\partial \mathcal{E}\varphi} S(\{\varphi\})\right) = 0 \ . \tag{95}$$

In the continuum limit, the equation of motion becomes the Euler-Lagrange equation for the scalar action:

$$m^2 \varphi(x) - \varphi''(x) = J(x) \ . \tag{96}$$

The classical solution ϕ_c is continuously differentiable for smooth source functions $J(x)$ and hence by itself is not even in the path integral measure. Again it is the set of $\varphi(x)$ functions *close to* the classical one that give the dominant contributions.

Derivatives such as the above, where the action functional is varied with respect to a single value of φ, are called *functional derivatives*, and usually denoted by curly δ's. In the text-book jargon, we may write the Euler-Lagrange equation as follows:

$$\left(\frac{\delta}{\delta\varphi(x)} - \frac{d}{dx}\frac{\delta}{\delta\varphi'(x)}\right)\left[S[\varphi(x)] - \int dx\ \varphi(x) J(x)\right] = 0 \ . \tag{97}$$

3.2 More dimensions

Rather than bothering with the Feynman rules for the one-dimensional theory, we now move on to more dimensions. We shall always aim for the simplest case, and therefore stick to nearest-neighbour interactions. We shall start with a discretized action, where the field values are labelled by a set of several integers rather than a single one. Note that a labelling with several integers can always be re-encoded in terms of a single integer label: however, simple nearest-neighbour interactions in four dimensions would look very complicated in the one-dimensional encoding, leading to non-local interactions. Our notion of space — yes, the 'everyday kind' of space – is, after all, only a mental construction, developed over the course of our development up from foetuses to adults, that allows us to comprehend, and react to, the world in an efficient manner. **Spacetime is defined in such a way interactions look simple**.

3.2.1 Euclidean multidimensional theory

Let us consider the discrete version of a D-dimensional theory. We choose vector labels $\vec{n} = (n^1, n^2, n^3, \ldots, n^D)$ for the fields. Also we define $\vec{e}_k$ to be the vector with k^{th} component equal to 1, and the other ones 0. The action (again without self-interactions) is then given by

$$S(\{\varphi\}) = \sum_{\vec{n}} \left(\frac{1}{2}\mu\varphi_{\vec{n}}^2 - \gamma \sum_{k=1}^{D} \varphi_{\vec{n}}\varphi_{\vec{n}+\vec{e}_k} \right) \quad . \tag{98}$$

There is now a translation invariance in each of the D directions. The SD equation can be solved in the same manner as above, and we find for the correlator:

$$\Pi(\vec{n}) = \frac{1}{(2\pi)^D} \int_{-\pi}^{\pi} d^D\vec{z}\, \frac{\exp(i\vec{n}\cdot\vec{z})}{\mu - 2\gamma \sum\limits_{k=1}^{D} \cos(z_k)} \quad , \tag{99}$$

where we have used the D-dimensional generalization of z. The continuum limit is now obtained by defining

$$\vec{x} = \Delta\vec{n} \;\; , \;\; \vec{z} = \Delta\vec{p} \;\; , \;\; \gamma \sim \Delta^{2-D} \;\; , \;\; \mu \sim 2D\gamma + m^2\Delta^D \;\; . \tag{100}$$

Using some algebra and standard integrals, we find for the continuum correlator

$$R(\vec{x}) = \frac{1}{(2\pi)^D} \int d^D\vec{p}\, \frac{\exp(i\vec{x}\cdot\vec{p})}{\vec{p}^2 + m^2} = \frac{1}{2\pi} K_{1-D/2}\left(m|\vec{x}|\right) \left(\frac{2\pi|\vec{x}|}{m} \right)^{1-D/2} \quad , \tag{101}$$

where K is the modified Bessel function of the second kind. Note that we have obtained not only translational but also rotational invariance: the rectangular structure of the underlying discrete grid has become invisible in the continuum limit[21]. The continuum Lagragian density is now just the more-dimensional generalization of Eq.(88):

$$\mathcal{L} = \frac{1}{2}m^2\varphi(\vec{x})^2 + \frac{1}{2}\left(\vec{\nabla}\varphi(\vec{x})\right)^2 \quad . \tag{102}$$

In the same manner, the continuum source function is defined by $J_{\vec{n}} = \Delta^D J(\vec{x})$, and the classical field equation is, in the continuum limit:

$$m^2\varphi(\vec{x}) + \vec{\nabla}^2\varphi(\vec{x}) = J(\vec{x}) \quad . \tag{103}$$

Finally, let us reintroduce the φ^4 self-interaction again. Cavalierly ignoring the fact that the propagator was derived from a SD equation without self-interaction vertices[22], we simply add to the discrete action a term $\Delta^D\lambda\varphi_n^4$ for every n. The action including sources then becomes

$$S[\varphi(x)] = \int d^D x \, \left(\frac{1}{2}m^2\varphi(\vec{x})^2 + \frac{1}{2}\left(\vec{\nabla}\varphi(\vec{x})\right)^2 + \lambda\varphi(\vec{x})^4 - J(x)\varphi(x) \right) \quad , \tag{104}$$

Note that the λ vertex also 'occurs' at all space points.

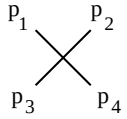
[21] Again, the rotational invariance can be corrupted by choosing a more perverse action in the discrete case: fortunately, it is the simplest action that leads to the best continuum limit.

[22] In fact, the SD equation would be very different if we include the self-interactions, and so would the continuum limit. This is precisely the sin for which the loop divergences are the punishment.

The Feynman rules for such a theory can be obtained in precisely the same manner as before, as the more-dimensional generalizations of the one-dimensional case. They are simplest in the momentum representation. For $D = 4$, we have the following Feynman rule for the propagator:

$$\overset{\text{p}}{\longrightarrow} \rightarrow \frac{1}{\vec{p}^2 + m^2} \quad , \tag{105}$$

where we recognize the propagator of the one-dimensional theory in the momentum representation (with the replacement $p^2 \rightarrow \vec{p}^2$. For the single vertex in this theory, we have



$$\rightarrow -\lambda(2\pi)^4\delta^4\left(p_1 + p_2 + p_3 + p_4\right) \quad , \tag{106}$$

where the $(2\pi)^4\delta^4()$ comes from the fact that we have to integrate this four-vertex over the whole space. Finally, the source is, in the same way, given by

$$\overset{\text{p}}{\longrightarrow\!\!\times}\ \text{q} \rightarrow J(q)(2\pi)^4\delta^4\left(p + q\right) \quad , \tag{107}$$

In the vertices all momenta are counted incoming (or all outgoing). In addition, every momentum in a line has to be integrated over, with a factor $(2\pi)^{-4}$. Usually, we will consider sources that correspond to particles of fixed momentum being absorbed or produced (that is, sources $J(\vec{x})$ consisting of a single Fourier mode), and in that case the external momenta are effectively fixed.

3.2.2 Towards Minkowski space

The construction of our field theory looks almost realistic: the only thing not yet built in is the special rôle of time. Indeed, instead of space points labelled by $\vec{x} = (x^1, x^2, x^3, x^4)$, we know it is much better to describe things on the basis of *space-time events* with coordinates $x^\mu = (x^0, x^1, x^2, x^3)$, and a metric given by $g_{\mu\nu}$, with $g_{00} = 1, g_{11} = g_{22} = g_{33} = -1$, and the other components zero. We therefore make the following substitution:

$$x^4 = ix^0 \quad . \tag{108}$$

The exponent in the path integral, Eq.(104), then becomes

$$\begin{aligned} -S[\varphi(x)] &\rightarrow iS[\varphi(x)] \ , \\ S[\varphi(x)] &= \int d^4x \ \left(\mathcal{L} + J(x)\varphi(x)\right) \ , \\ \mathcal{L} &= \frac{1}{2}(\partial^\mu\varphi(x))(\partial_\mu\varphi(x)) - \frac{1}{2}m^2\varphi(x)^2 - \frac{1}{24}\lambda\varphi(x)^4 \ . \end{aligned} \tag{109}$$

The factor $1/24 = 1/4!$ is still there by convention. Note that we have taken a factor $-i$ out of the action integral. Again moving to the momentum representation:

$$\varphi(x) = \frac{1}{(2\pi)^4}\int d^4p\ e^{-ipx}\varphi(p) \quad , J(x) = \frac{1}{(2\pi)^4}\int d^4p\ e^{-ipx}J(p) \ , \tag{110}$$

where $\varphi(p) = \varphi(-p)$ and $J(p) = J(-p)$ since $\varphi(x)$ and $J(x)$ are real, the exponent can then be written as

$$\begin{aligned} iS &= \sum\nolimits_p \Big(\frac{i}{2}(p^2 - m^2)\varphi(p)^2 + iJ(p)\varphi(p) \\ &\quad -\frac{i}{24}\lambda(2\pi)^4\delta^4\left(p_1 + p_2 + p_3 + p_4\right)\varphi(p_1)\varphi(p_2)\varphi(p_3)\varphi(p_4)\Big) \ , \end{aligned} \tag{111}$$

where the symbol $\sum_p$ means that every momentum has to be integrated over, with a factor $(2\pi)^{-4}$.

In the above, we have tacitly assumed that the 90^o *Wick rotation* implied by Eq.(108) is actually allowed: since the integrand in the path integral now is oscillatory rather than damped, this is not at all obvious. A signal of potential problems is the fact that the quadratic term in the action contains $p^2 - m^2 = (p^0)^2 - |\vec{p}|^2 - m^2$ which may become zero, and in fact does so precisely for on-shell particles! For the moment we shall blithely ignore this problem, and assume that all particles are off their mass shell; later on we will repair this.

3.3 Choosing the action

Until now we have mainly been studying φ^4 theory as a useful vehicle to develop the various techniques and concepts. Interactions such as φ^3 or interactions containing derivatives can easily be incorporated, the Feynman rules being read off in each case directly from the Lagrangian density: but which density ought we to take? A possible answer comes from the classical field equation, which as we have seen is the classical Euler-Lagrange equation for the action. In the φ^4 case, it reads

$$\left(\frac{\delta}{\delta\varphi(x)} - \frac{\partial}{\partial x_\mu}\frac{\delta}{\delta\partial^\mu\varphi(x)}\right)\mathcal{L} = 0 \quad \Rightarrow \quad \partial^\mu\partial_\mu\varphi(x) + m^2\varphi(x) + \frac{\lambda}{6}\varphi(x)^3 = J(x) \ . \tag{112}$$

If we consider the particles 'on their own', that is, without self-interaction or sources, this becomes

$$\left(\partial^\mu\partial_\mu + m^2\right)\varphi(x) = 0 \ , \tag{113}$$

which is the Klein-Gordon equation. This gives us the necessary hint: to see what the propagator of a particle type ought to be, look at its 'classical' field equation. The particle we have been studying so far was assumed to have no internal structure: its only characteristic can be its momentum, and the Klein-Gordon equation just says that for a free particle this momentum must be on the mass shell. For the vertices we may *choose* anything, but the burden of proof is of course ours: apart from the obvious conformity with experiment, there are some general considerations, for instance the resulting connected Green's functions had better be Lorentz invariant. There is, however, another fundamental requirement, which we shall come to in the next section.

4 Quantum Terrorism : of scattering, cutting, amputating and exploding

4.1 The scattering and probability

What we have done so far would be doomed to be no more than a mathematical recreation unless we attempt to make contact with physics. In quantum field theory, the connected Green's functions are postulated to relate to the quantum-mechanical transition amplitudes, and in that way enter into predictions for cross sections and lifetimes. The precise relation will follow below: first we need to make some general remarks.

4.1.1 The S matrix and the T matrix

In particle physics experiments, we typically start by preparing an initial state at some time 'far in the past' where particles are 'far away from each other': so far away, in fact, that we assume

the particles to be free of any interaction. At that moment, the initial state $|\text{in}, t = -\infty\rangle$ is then made up (ideally!) as a combination of single-particle states corresponding to, say, the incoming e^+ and e^- each with its own momentum. As time takes its course, the particles will approach and start interacting. The incoming state evolves in some complicated manner, and at a late time it will be $|\text{in}, t = +\infty\rangle$. It is important to realize that this is still *the same state*: it is rather, the experimentalist observer at $t = +\infty$ who has undergone a time translation with respect to the accelerator operator who prepared the state at $t = -\infty$. Then we perform the observation, and the state is observed to consist of, say, quarks moving away from the interaction region (we cheerfully ignore hadronization here). The corresponding state is $|\text{out}, t = +\infty\rangle$. The probability (amplitude) for this to happen is of course the overlap between the states: $\langle \text{out}, t = +\infty | \text{in}, t = +\infty \rangle$. Under the assumption that the prepared state is, at $t = -\infty$, essentially that of a combination of free particles without interaction, and the observed state at $t = +\infty$ likewise, the amplitude may be written as the element of a matrix connecting these free-particle states. This matrix is called the S matrix, and its elements are what we want to compute.

4.1.2 Unitarity and the optical theorem

If the free-particle states form a complete set, an important consequence follows. Let us denote by j a state that looks like a free-particle state at $t = -\infty$: j therefore labels a particle content as well as all kinds of momenta, spins, etc.; and let us similarly denote by k a state that looks like a free-particle state at $t = +\infty$. The element of the S matrix can then be written as S_{kj}. The completeness of the whole set of possible states j, and of the whole set of possible states k, then implies that S must be *unitary*:

$$SS^\dagger = S^\dagger S = 1 \quad . \tag{114}$$

This *requirement of unitarity* essentially means that the normalization of states is preserved, in other words: **unitarity = conservation of probability**. If there were no interactions whatsoever, j could of course only appear as j even at $t = +\infty$, and we would have $S_{kj} = \delta_{k,j}$. It therefore makes sense to write the S matrix as follows:

$$S = 1 + iT \quad , \tag{115}$$

where the i is taken out of the T matrix by convention. The unitarity of S in Eq.(114) then implies for T:

$$i(T - T^\dagger) + TT^\dagger = 0 \quad . \tag{116}$$

Let us now specialize to the case where $j = k$, *i.e.* the final and initial state happen to be identical, 'in spite of' all the interactions that have taken place. Eq.(116) then reduces to

$$\text{Im}(T_{jj}) = \frac{1}{2} \sum_n |T_{nj}|^2 \quad , \tag{117}$$

where on the right-hand side we must sum over all states n: the (imaginary part of) the amplitude T_{jj} is related to the total *probability* for the initial state j to go into *any* final state. Eq.(117) is called the *optical theorem* since it was first derived in the context of waves travelling through a medium.

Another consequence of unitarity is that S (and T) matrix elements cannot be arbitrarily large. Indeed, if all labels are discrete, we must have $\sum_k |S_{kj}|^2 = 1$. For continuous labellings (such as momenta) the condition is more involved, but the only thing we need here is the knowledge that unitarity will be endangered whenever we encounter matrix elements that show unlimited growth (for instance as a function of energy).

4.2 Towards predictions for experiments

The relation between Feynman diagrams and scattering processes is the postulate that **the T matrix elements of the theory are related to the Green's functions**, that themselves can be computed using the Feynman diagrams of the theory. Note that we have 'related to', and not 'equal to': we shall discuss the precise nature of the relation below. We shall not give rigorous derivations here[23]: instead, we aim to illustrate the consistency of our approach.

4.2.1 Density of states

The first thing to be done is to give a more precise meaning to the notion 'sum over states' as used for instance in Eq.(117). What we use for the sum over states is of course to some extent arbitrary, since we have not discussed the normalization of the states. An important requirement, however, would seem to be that the sum over states must be Lorentz-invariant, otherwise it becomes difficult to arrive at Lorentz-invariant predictions for cross sections and the like. Realizing that we assume the final-state particles to be free and on their mass shell, and to have positive energy[24], the phase-space integration element for a single particle with momentum p^μ and mass m is best defined to be

$$\frac{1}{(2\pi)^3} d^4p \; \delta\left(p^2 - m^2\right) \; \theta(p^0) = \frac{1}{16\pi^3} d^3\vec{p} \frac{1}{\sqrt{\vec{p}^2 + m^2}} \quad , \tag{118}$$

where the first alternative is explicitly Lorentz-invariant. The factor $(2\pi)^{-3}$ comes from the usual $(2\pi)^{-4}$ that comes with any momentum integral, and a factor 2π that comes with any Dirac delta function – but again, this is to some extent arbitrary. For a total final state consisting of n particles we of course also have the requirement of overall conservation of energy and momentum, and therefore we choose as the actual n-particle phase space integration element:

$$\begin{aligned} &dV(P; p_1, \ldots, p_n) \equiv \\ &(2\pi)^4 \delta^4 \left(P - \sum_{j=1}^{n} p_j \right) \; \prod_{k=1}^{n} \left[\frac{1}{(2\pi)^3} d^4p_k \; \delta\left(p_k^2 - m^2\right) \; \theta(p_k^0) \right], \end{aligned} \tag{119}$$

where P^μ is the total four-momentum. This is defined to be the 'sum over states' for particles without characteristics other than momentum. If additional quantum numbers such as spin or colour etc. are involved, we include discrete sums over these variables as well: usually we shall leave them to be understood.

4.2.2 Cross sections and widths

In making experimental predictions we have to obtain, in one way or the other, the quantum-mechanical transition amplitude $\mathcal{M}$ for some process. Presently, we shall see how $\mathcal{M}$ is related to the Feynman diagrams and the S matrix element. Having obtained $\mathcal{M}$, we have to turn it into an experimental prediction. The two commonly studied quantities are the *cross section* for $2 \to n$ processes, and the *decay width* (inverse lifetime) for $1 \to n$ processes. At this point, we simply give the prescriptions for them. By $\langle |\mathcal{M}|^2 \rangle$ we denote the absolute value squared of $\mathcal{M}$,

[23] After all, no-one can *prove* that the world is described by quantum mechanics. *Rigor mathesis perfectus mortis rigorem inducet*.

[24] The on-shell condition is essentially the *definition* of what it means to be a free particle with no energy other than that due to its motion (*i.e.* no potential energy); the positivity of the energy is an empirical fact.

summed over unobserved discrete quantum numbers of the final state such as spins or colours, and *averaged* over similar quantities for the initial state[25].

For the partial decay width $d\Gamma$ of an unstable incoming particle with mass M and momentum P we have

$$d\Gamma = \frac{1}{2M}\,\langle|\mathcal{M}|^2\rangle\; dV(P;p_1,\ldots,p_n)\; S_{BF} \quad , \tag{120}$$

and for the partial cross section $d\sigma$ for two incoming particles of masses m_a and m_b and momenta p_a^μ and p_b^μ, respectively, we have

$$\begin{aligned} d\sigma &= \frac{1}{2F}\,\langle|\mathcal{M}|^2\rangle\; dV(P;p_1,\ldots,p_n)\; S_{BF} \quad , \\ F &= \left((s-m_a^2-m_b^2)^2 - 4m_a^2m_b^2\right)^{1/2} \quad , \\ s &= P^2 \quad , \quad P^\mu = p_a^\mu + p_b^\mu \quad . \end{aligned} \tag{121}$$

Here, S_{BF} is the Bose-Fermi symmetry factor: if the final state contains a group of precisely p indistinguishable particles in the final state, we have to include a factor $1/p!$ for each such group, since if the particles are indeed indistinguishable, any permutation of them leads to overcounting in the cross section[26].

The 'flux factor' $1/2M$ is also related to our normalization of states, and we shall have to prove its consistency. That the relative 'flux factors' $1/2M$ and $1/2F$ are consistent with one another can be seen from the fact that, in the rest frame of particle b, we have

$$2F = 4m_b|\vec{p}_a| = (2m_b)\left(2m_a\cdot\frac{E_a}{m_a}\right)\left(\frac{|\vec{p}_a|}{E_a}\right) \quad , \tag{122}$$

where we recognize the normalization of the two states (with the wave of a having undergone the correct Lorentz contraction in its direction of motion, and hence a change in its normalization with respect to that for a particle at rest), and the relativistic velocity of a in the rest frame of b, inherent in the definition of cross section as the rate per unit flux.

Finally, dimensionality. Taking into account[27] that the Dirac delta function $\delta(A)$ has the inverse dimension of its argument A, the n-particle phase space factor is easily seen to have dimension (GeV)$^{3n-4}$. If an inverse lifetime is to have dimension (GeV)^1, then $\mathcal{M}(1\to n)$ must have dimension (GeV)$^{3-n}$. We conclude that an amplitude with a total of k external particles must have dimension (GeV)$^{4-k}$. It follows immediately that cross sections must have dimension (GeV)$^{-2}$, in accordance with their interpretation as an area. Note that we deal here with the *units* in which the amplitudes are expressed, and *not* with their actual behaviour as a function of the total energy: indeed, this is what leads into trouble later on.

4.2.3 Connected and disconnected diagrams

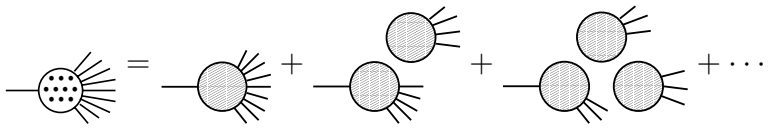
We are now in a position to take Feynman diagrams seriously as saying somthing about physics going on. At this point, the connectedness of the diagrams starts to play a significant rôle.

[25] This means, for instance, a factor 1/2 for each unpolarized incoming spin-1/2 particle, 1/3 for a massive spin-1 particle, 1/2 for an incoming massless spin-1 particle, 1/3 for an incoming quark, and 1/8 for an incoming gluon.

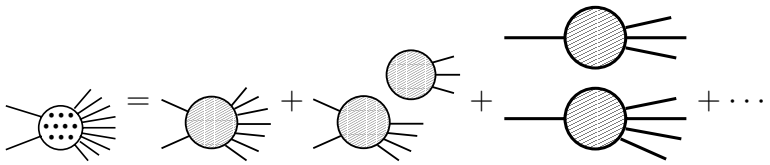
[26] On the other hand, the number of Feynman diagrams in $\mathcal{M}$ will be bigger by a factor $p!$. Some authors include a factor $\sqrt{S_{BF}}$ in the definition of $\mathcal{M}$, but this spoils some of the symmetry in case one of the *incoming* particles is indistinguishable from one of the outgoing ones.

[27] as can be easily seen from the definition $\int dA\delta(A) = 1$.

Consider the diagrams associated with the decay of a single particle: they contain one incoming, and any number of outgoing particles. Some of these diagrams may be connected, but in principle we can also write down disconnected diagrams, as indicated below, where as usual hatched blobs denote connected diagrams, and here a speckled blob denotes any combination of connected and disconnected diagrams:



If the outgoing particles all have positive energy, we see that the connected diagrams without the incoming line attached vanish because of energy conservation. Therefore, in this case only the connected Feynman diagrams can give a nonzero result. Similarly, for the two-particle cross section, the associated Feynman diagrams can be schematically given as



Again, blobs with no incoming lines vanish because of energy conservation.

The case with two separate connected contributions, each with one incoming line, is more tricky. It *can* contribute, but only if the two incoming particles are capable of decaying independently in the particular indicated final states. In most cases this will not be possible (for instance in e^+e^- collisions), and then the conclusion is that **only the connected Feynman diagrams contribute**. This is, in fact, the reason for the emphasis on connected diagrams in our zero-dimensional model. It must be kept in mind, however, that this is not really a theorem. Indeed there are cases where the disconnected diagrams do enter. For example, we may have collisions between unstable particles such as $\mu^+\mu^-$, or between, say, three or more particles. Another instance may be processes occurring at high temperature (such as the very early universe), where the surrounding 'heat bath' can actually supply energy 'out of the vacuum' such that particles may be created without explicit incoming particles. Fortunately, in typical collider situations such exotic processes can usually be neglected.

Having argued that only the fully connected diagrams give a physical contribution, we make one more step. Each such diagram contains contains an overall factor $(2\pi)^4\delta^4\left(P-\sum p\right)$, where $\sum p$ stands for the sum of all outoing momenta, and P for the total incoming momentum. The square of a Dirac δ function is not well-defined, so as usual, we enclose our system in a large space-time box, thereby turning the Dirac δ function into a simpler Kronecker δ, which can be squared. Letting the box become infinitely large again at the end, we recover precisely the overall momentum conservation that we already included in our sum over states, times volume factors that cancel against the wave function normalizations. Part of our recipe, then, is that **we should sum all the relevant Feynman diagrams, but leave out the overall momentum-conservation factor**.

4.3 Poles and residues

We are not yet able to give a precise recipe for the matrix element $\mathcal{M}$, since the propagator $i/(p^2-m^2)$ is still singular precisely for stable, on-shell particles. We are now ready to face this issue: its resolution also tells us the correct prescription for $\mathcal{M}$.

4.3.1 The propagator for an unstable particle

Let us consider the following process. We start with the vacuum, with no particles at $t = -\infty$. Then, a source emits an *unstable* particle with momentum p^μ and mass m, which propagates through spacetime, while undergoing interactions which may destroy the particle by letting it decay, but do not necessarily do so. Let us assume that the particle is not destroyed, but rather absorbed by another source. Then, at $t = +\infty$ we are left with the vacuum again. Let us denote the particle's actual propagator by $iR(p)$. At this point we do not know the precise form of $R(p)$ except that it should look like $1/(p^2 - m^2)$ if p^μ is sufficiently far off its mass shell. Putting in the source vertices as well we have

$$iT_{kk} = \quad = -iJ^2 R(p) \quad , \tag{123}$$

where the blob denotes all possible connected diagrams. The label k stands here for the vacuum: both before the source emits the particle, and after it absorbs the particle again, the states are the same, and we are indeed looking at a diagonal element of the T matrix. Now we apply the optical theorem: the imaginary part of T_{kk} is related to $\sum\limits_n |T_{nk}|^2$, where n denotes all possible final states. Pictorially, this sum reads

$$\frac{1}{2}\sum_n |T_{nl}|^2 = \frac{1}{2}\sum_n \left| \quad \text{n} \right|^2 \quad , \tag{124}$$

where the half-open blob denotes any decay process. The sum over n is nothing but our 'sum over states', summed over all possible particles configurations that could possibly result from the decay of our unstable particle. In other words, we must have a relation of the form

$$\frac{1}{2}\sum_n |T_{nl}|^2 = J^2 |R(p)|^2 \; k \; m\Gamma^{\text{tot}} \tag{125}$$

if p^μ is close to its mass shell, where Γ^{tot} is the *total decay width* of the particle. This relation simply follows from the definition of Γ^{tot}. At this point we do not know yet what k is, since that is precisely one of the factors in the transition from diagrams to $\mathcal{M}$ that we want to determine; however, we *do* know that it must be a positive dimensionless number. At any rate, we have the following condition for R from the optical theorem:

$$-\text{Im}R(p) = k \; m\Gamma^{\text{tot}} \; |R(p)|^2 \quad . \tag{126}$$

Together with the desired form for R far from the mass shell, the simplest solution is to take

$$R(p) = \frac{1}{p^2 - m^2 + ikm\Gamma^{\text{tot}}} \quad . \tag{127}$$

A more precise analysis would take into account that the invariant mass at which the decay process is evaluated is not m but rather $\sqrt{p^2}$, and the imaginary part is accordingly p^2 dependent: the above form holds for *narrow resonances*[28].

[28] The p^2 dependence of the width *does* give measurable effects in, for instance, Z^0 production at LEP.

4.3.2 *The $i\epsilon$ prescription*

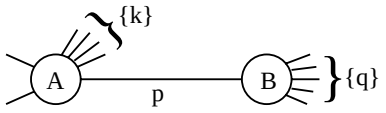
What, now, for stable particles? The simplest idea is to consider stable particles as unstable particles with an extremely long lifetime. In that case, we just let Γ^{tot} *decrease* down to almost, but not quite, zero. The precise value of Γ^{tot} is then not relevant anymore, only the fact that $km\Gamma^{\text{tot}}$ should be inifintesimally small but positve, and hence our Feynman rule for stable particles reads

$$\overline{\qquad\text{p}\qquad} \rightarrow \frac{i}{p^2 - m^2 + i\epsilon} \quad , \quad \epsilon \rightarrow 0^+ \quad . \tag{128}$$

This is the famous $i\epsilon$ prescription. Note that the sign is very important: on it hinges the fact that *unstable* particles *disappear*, or in other words the direction of time: **time is going forward in the direction in which unstable particles decay**. Incidentally, the $i\epsilon$ prescription is also sufficient to allow us to do the Wick rotation in the path integral: it could have been put in from the start, by adding a small term $i\epsilon\varphi^2/2$ to the action, thus ensuring that the rotated path integrand vanishes at infinity.

4.3.3 *Almost-stable particles and amputation*

We have seen that a good description of stable particles is to consider them as almost-stable unstable ones. In the S matrix, however, unstable particles never make it to $t = +\infty$, and we have to take a closer look at what is going on precisely. Let us assume that in some process an unstable particle is produced off-shell, with momentum p. Having propagated for a while it must decay, but if the particle is not very unstable this may be quite a distance away. Diagrammatically, we have



$+$ other diagrams , (129)

where **A** denotes the connected graphs in the 'production' process, and **B** those in the 'decay' process. The $\{q\}$ and $\{k\}$ stand for the other particle momenta involved. As indicated, there may be other diagrams leading to the same final state, but without involving this unstable particle. Let us denote by A the amplitude (a number) corresponding to **A**, and by B that corresponding to **B**. The cross section, according to the postulated rules, is given by

$$d\sigma = \frac{1}{2F}|A|^2 \frac{1}{(p^2 - m^2)^2 + k^2 m^2 {\Gamma^{\text{tot}}}^2} |B|^2 dV(P; \{q\} + \{k\}) \quad , \tag{130}$$

plus other terms that do not contain the p propagator squared. We now multiply by unity in the form $(2\pi)^{-4} d^4p (2\pi)^4 \delta^4 (p - \{q\})$, and we get

$$d\sigma = \frac{1}{2F}|A|^2 dV(P - p; \{k\}) \, \frac{1}{(2\pi)^4} \frac{d^4p}{(p^2 - m^2)^2 + k^2 m^2 {\Gamma^{\text{tot}}}^2} \, |B|^2 dV(p; \{q\}) \quad , \tag{131}$$

plus the other terms, which peak less strongly.

Let us now take the limit of almost-stability. The propagator then becomes very sharply peaked, and we may write

$$\frac{1}{(p^2 - m^2)^2 + k^2 m^2 {\Gamma^{\text{tot}}}^2} \sim \frac{\pi}{km\Gamma^{\text{tot}}} \delta\left(p^2 - m^2\right) \quad . \tag{132}$$

The normalization can be simply checked by integration over p^2. Putting this in, the expression for the cross section becomes

$$d\sigma \sim \frac{1}{2F}|A|^2 dV(P; p+\{k\})\ \frac{1}{\Gamma^{\rm tot}}\ \frac{1}{2mk}|B|^2 dV(p;\{q\})\ \ , \tag{133}$$

where we have neglected the non-resonant parts since they cannot contribute in the infinitesimally small part of phase space where $p^2 = m^2$.

The last factor in the above expression is nothing but what we are supposed to be using to compute the *partial decay width* $p \to \{q\}$, including the as yet unknown factor k. The result is exactly what we ought to expect on the basis of probability theory, namely that the probability to produce an almost-stable particle, followed by its decay into a given final state is the product of the probability of producing it, times the branching ratio for the particular decay. Since stable particles are supposed to be decaying infinitely far away, and we cannot see them decaying, we have to sum over all their possible decay modes.

The resulting rule is attractively simple: **external legs corresponding to scalar particles must be replaced by the factor 1**. This 'cutting off' of external legs goes by the grisly name of *amputation*.

Before we finish this section, the factor k has to be cleared up. Since we have decided, on the basis of the probability interpretation of the cross section, that external scalar legs should be simply amputated, the same should of course hold for the incoming legs as well, and hence we should have $k = 1$.

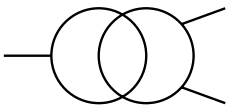
4.4 Cutting rules

Having established how external particles have to be treated, let us return to the optical theorem. Denoting the sum of all Feynman diagrams that lead from an initial state j to a final state k by $\mathcal{D}_{kj}$, the fact that the S matrix element is $S_{kj} = \delta_{kj} + \mathcal{D}_{kj}$ leads to the following expression of unitarity:

$$\mathcal{D}_{kj} + \mathcal{D}^*_{jk} + \sum_n \mathcal{D}_{nj}\mathcal{D}^*_{nk} = 0\ \ . \tag{134}$$

If perturbation theory is any good, Eq.(134) should hold term by term in an expansion in powers of all the coupling constants in the theory. Each such term may consist of many Feynman diagrams, and we shall argue that Eq.(134) in fact holds for each individual diagram.

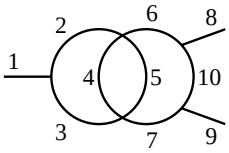
To see how this works, consider the following diagram:



This diagram comes from a theory with a single field and interaction potential

$$V(\varphi) = \frac{1}{6}\lambda_3\varphi^3 + \frac{1}{24}\lambda_4\varphi^4\ \ , \tag{135}$$

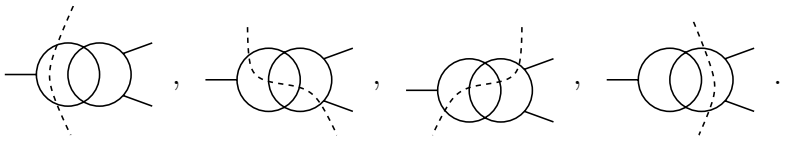
and is of order $\lambda_3^3\lambda_4^2$. There are a good many other diagrams of the same order, and Eq.(134) must hold for their sum. Now assign a different label to each line in the diagram, for instance



Let every line labelled k now correspond to a field φ_k. Consider now a theory for 10 different fields $\varphi_1, \ldots, \varphi_{10}$, with interaction potential

$$\begin{aligned} V(\varphi_1, \ldots, \varphi_1 0) &= g_1\varphi_1\varphi_2\varphi_3 + g_2\varphi_2\varphi_4\varphi_5\varphi_6 + g_3\varphi_3\varphi_4\varphi_5\varphi_7 \\ &\quad + g_4\varphi_6\varphi_8\varphi_{10} + g_5\varphi_7\varphi_9\varphi_{10} \quad . \end{aligned} \tag{136}$$

It is easy to check[29] that for *this* theory, the contribution to $1 \rightarrow 8 + 9$ of order $g_1 g_2 g_3 g_4 g_5$ contains precisely this diagram, and *only* this one. Assuming this theory to be unitary as well, we see that Eq.(134) must hold for this diagram individually. What about the summed-over state n in this case? It is easy to check that it can only be 2+3, or 2+4+5+7, or 3+4+5+6, or 6+7: every other choice would give the wrong order in the coupling constants. Diagrammatically, we can denote this by 'cutting' the diagram in half along some curve. The four possibilities in this case are given here:



A similar argument can be constructed for every Feynman diagram. The sum over n is then a sum over all possible cuts that separate j from k. The internal lines that are cut through are then 'put on the mass shell', and we have to sum/integrate over all their quantum numbers and momenta. Eq.(134), applied to individual diagrams, are called the *cutting rules*. Two remarks are in order about our treatment. In the first place, one might have worried about the symmetry factors, since diagrams with non-identical fields have symmetry factor 1, while the same diagram has another factor (in this case, 1/2) if the fields are identical. In fact there is no inconsistency, precisely owing to our prescription for the Bose-Fermi factor S_{BF} in the sum over states. Secondly, the above is not really a rigorous derivation, since we had to assume the unitarity of the 10-field theory as well: in fact one usually first proves the cutting rules from the Feynman rules, and then proves the unitarity of S from the cutting rules. Here, we merely aimed at showing the mutual consistency of unitarity and cutting rules.

4.5 Appendix: physical divergencies?

We have seen how a careful treatment of stable particles as almost-stable unstable ones led us to handle the apparent singularity in their propagators for external lines. But what about *internal* lines? That there may actually be singularities inside the physical phase space can be seen from the following simple model.

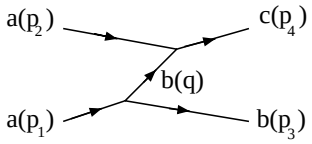
Take a model with three scalar particle species: a, b and c. Let us assume that the following decays are physically possible: $a \rightarrow bb$ and $c \rightarrow ab$. This means, of course, that some (effective-action) vertices of the type $\varphi_a\varphi_b^2$ and $\varphi_a\varphi_b\varphi_c$ exist, and also that $m_c > m_a > m_b$; let us take $m_b = 0$ for simplicity.

Now consider the process

$$a(p_1) + a(p_2) \quad \rightarrow \quad b(p_3) + c(p_4) \quad ,$$

[29] The secret is that all external lines occur in precisely one interaction term, and the internal lines in precisely two.

where we have indicated the particles' momenta. One of the diagrams contributing to this process is



where we have indicated with arrows the direction in which positive energy has to flow in this diagram. It is matter of simple two-body kinematics to check that the momentum q^μ of the internal b line can become lightlike, $q^2 = 0$, if the total invariant mass squared s is larger than $m_c^4/(m_c^2 - m_a^2)$: for each such an energy there is precisely one scattering angle (in the CM frame) for which this happens[30]. For generic couplings, the amplitude $\mathcal{M}$ and the cross section will therefore display a real, non-integrable singularity *inside* the physical phase space!

The way out of this conundrum is to realize that if the kinematics are such that the internal b can be on its mass shell, we have here the decay of the a particle inside a scattering process. If the non-stability of the a is relevant inside the scattering, then we should of course also keep track of it for the external lines: an unstable incoming a cannot come from infinitely far away, and its momentum is not fixed at $p_1^2 = m_a^2$, but rather has a distribution around this value, with a width given by the a lifetime. Instead of a pure momentum state for the incoming $a(p_1)$, we ought therefore to use a superposition of states with different invariant masses. This 'smearing' turns out to be precisely sufficient to dampen the singularity in $\mathcal{M}$, and the cross section turns out to factorize, in a similar way as above, into a part describing the production of an *on-shell* b with momentum q^μ in the decay of $a(p_1)$, followed by the collision of $b(q)$ and $a(p_2)$ into $c(p_4)$. The message, again, is: be wary of particles that claim to be stable.

5 Inner life: particles with spin

So far, we have considered only scalar particles. These are, by definition, characterized by their momentum only. In the real life of the Standard Model, almost all particles have more structure than that. Charged particles are described by complex rather than real fields, opening up the possibility of *anti*particles. The charged π is an example, although not an elementary one. Particles with additional internal spin structure are represented by fields with several components that get mixed in well-defined ways under Lorentz transformations, and coloured particles have in addition colour components that are mixed under complex rotations in colour space.

5.1 Dirac fermions

One of the most important types of spinning fields is that of spin-1/2 particles, of which the electron is the simplest type. We shall not, in these notes, go through the whole construction of Dirac spinors and their interpretation, but jump nimbly to the Feynman rules that describe them.

[30] There is, of course, also the diagram with $a(p_1)$ and $a(p_2)$ interchanged: that diagram also has the divergence, but at another scattering angle. We can therefore consider the diagrams separately as long as we only study their singularity structure.

5.1.1 *The Dirac equation and spinors*

In his search for a relativistic wave equation that, in contrast to the Klein-Gordon equation (first put forward by Schrödinger), would be of first rather than second order in time, Dirac had to construct a set of matrices γ^μ ($\mu = 0, 1, 2, 3$) that obey an anticommutation relation:

$$\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}\mathbf{1} \quad , \tag{137}$$

where **1** denotes the unit matrix. The matrix γ^0 must be hermitian, and the other three anti-hermitian. The simplest such set consists of 4×4 matrices. Together with the unit matrix, these form the basis of the Dirac algebra. Each element of this algebra can be written as a sum of products of various γ matrices. Especially the total product is important: we use

$$\gamma^5 \equiv i\gamma^0\gamma^1\gamma^2\gamma^3 \quad \Rightarrow \quad (\gamma^5)^2 = \mathbf{1} \; , \; \gamma^5\gamma^\mu = -\gamma^\mu\gamma^5 \quad . \tag{138}$$

The Dirac wave equation for a free electron reads

$$(i\partial\!\!\!/ + m)\psi(x) \equiv (\gamma^\mu\partial_\mu + m)\psi(x) = 0 \quad , \tag{139}$$

where we introduced the notation $a\!\!\!/ = \gamma^\mu a_\mu$ for any vector a, so that $a\!\!\!/a\!\!\!/ = a^2$. The wave function ψ has to have 4 components, two of which are interpreted as the two spin states of the electron e^-, and the remaining two are related to the two spin states of the *positron* e^+. Multiplying Eq.(139) on the left with $(i\partial\!\!\!/ - m)$ we recover the mass-shell condition $p^2 = m^2$ for free particles.

The full consequences of the Dirac equation, such as the proof that indeed it decribes a spin-1/2 particle with an approximate gyromagnetic ratio $g = 2$, are not important here. What we do need are the algebraic conditions for the Fourier modes corresponding to a fixed, on-shell momentum p^μ and a spin pointing in a given direction. The spin direction is described by the *spin vector* s^μ, and we require

$$p^2 = m^2 \quad , \quad s\cdot p = 0 \quad , \quad s^2 = -1 \quad , \tag{140}$$

so that in the rest frame of p^μ, s^μ is just a direction in 3-space. Having these, the *Dirac spinors* are defined to be $u(p, s)$ for a particle and $v(p, s)$ for an antiparticle, and these are defined by their *projection operators*[31]:

$$\begin{aligned} u(p,s)\bar{u}(p,s) &= \frac{1}{2}(1 + \gamma^5 s\!\!\!/)(p\!\!\!/ + m) \quad , \\ -v(p,s)\bar{v}(p,s) &= -\frac{1}{2}(1 - \gamma^5 s\!\!\!/)(p\!\!\!/ - m) \quad , \end{aligned} \tag{141}$$

where the Dirac conjugate $\bar{u}$ is defined as $\bar{u} \equiv u^\dagger\gamma^0$. The normalization chosen here is our convention: in many texts, $1/2m$ is preferred to $1/2$. The two spin states for the particle and the antiparticle are related by $s^\mu \leftrightarrow -s^\mu$. We also have the perhaps more familiar spin sums

$$\sum_s u(p,s)\bar{u}(p,s) = p\!\!\!/ + m \quad , \quad \sum_s v(p,s)\bar{v}(p,s) = p\!\!\!/ - m \quad . \tag{142}$$

[31] Strictly speaking these are not projection operators since the nonzero eigenvalue is not 1 but $2m$: the important point is that they provide a division of the Dirac space into four separate pieces.

One remark is in order here. For very energetic fermions, the spin vector may blow up. If both $\pm\vec{s}$ and $\vec{p}$ point in the same direction $\vec{e}$ in some frame, Eq.(140) gives

$$p^\mu = (E, p\vec{e}) \quad , \quad s^\mu = \pm\left(\frac{p}{m}, \frac{E}{m}\vec{e}\right) \sim \pm\frac{1}{m}p^\mu + \mathcal{O}\left(\frac{m}{E}\right) \quad , \tag{143}$$

where the last term holds in the high-energy limit. That this problem is only apparent can be seen from the projection operator in that case:

$$\begin{aligned} u(p,s)\bar{u}(p,s) &= \left(1 \pm \frac{1}{m}\gamma^5 \not{p} + \mathcal{O}\left(\frac{m}{E}\right)\right)(\not{p} + m) \\ &= \frac{1}{2}(1 \pm \gamma^5)\not{p} + \mathcal{O}(m) \quad . \end{aligned} \tag{144}$$

The resulting projection operator in the limit $m \to 0$ is that for *helicity states.* The divergence of s^μ with energy disappears owing to the Dirac algebra.

5.1.2 Fermion propagators

In the same way that the propagator of a scalar particle is related to the Klein-Gordon equation, the spin-1/2 Dirac propagator is (the momentum representation of) the Green's function for the free Dirac equation. We find the following Feynman rule:

$$\underset{\text{p}}{\longrightarrow} = i\frac{\not{p} + m}{p^2 - m^2 + i\epsilon} \quad . \tag{145}$$

Note that this propagator is *not* even in p^μ: the direction in which the momentum flows is important, and hence we orient the propagator with an arrow and count the momentum *along* the arrow. Notice the occurrence of a nontrival numerator: it tells us that the various spin states propagate together through space. This becomes even more clear when we write, somewhat symbolically,

$$i\frac{\not{p} + m}{p^2 - m^2 + i\epsilon} = \frac{i\sum\limits_s u(p,s)\bar{u}(p,s)}{p^2 - m^2 + i\epsilon} = \frac{-i\sum\limits_s v(-p,s)\bar{v}(-p,s)}{p^2 - m^2 + i\epsilon} \quad , \tag{146}$$

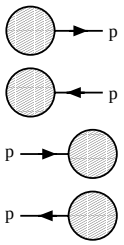
although this is strictly speaking only correct on the mass shell, to which we shall now move.

5.1.3 Amputation for fermions

The above Feynman rule holds for stable fermions (hence the $i\epsilon$); for unstable fermions, we again have to insert the $m\Gamma^{\text{tot}}$. We can now repeat the arguments that led us to the amputation formula for scalar particles, with the added complication that the propagator has a matrix structure. Since a good amplitude $\mathcal{M}$ should have no Dirac indices hanging around, a fermion line in a diagram ought to have the form $\bar{u}$(matrices)u or similar. The ordering of the components of a Feynman diagram becomes important now, and we adopt the usual convention of matrix multiplication, which means that the fermion lines are written down *against* the direction of the arrow.

Let us now assume that a fermion rather than an antifermion is produced and decays far away, as depicted for a sclar particle in Eq.(129). The amputation argument is completely analogous, since it relies only on the denominator of the propagators, and we see that the denominator of the fermion propagator is again amputated away, but the numerator is left. The

spinor part $u(p)$ must now be assigned to the decay amplitude B, while the conjugate spinor part $\bar{u}(p)$ must go with the production amplitude A. The sum over spins in Eq.(146) is now justified, and indeed consistent since the spin of the propagating fermion is by assumption not detected (otherwise it would not be an internal line). A similar argument holds for the case of an antifermion. The upshot is that we may choose the following Feynman rules for external, on-shell fermions:

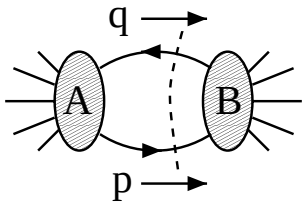


$$\begin{aligned} &\to \quad \bar{u}(p,s) \quad \text{outgoing fermion} \ , \\ &\to \quad v(p,s) \quad \text{outgoing antifermion} \ , \\ &\to \quad u(p,s) \quad \text{ingoing fermion} \ , \\ &\to \quad \bar{v}(p,s) \quad \text{ingoing antifermion} \ , \end{aligned} \tag{147}$$

where the momentum is defined to be physical, *i.e.* with positive energy[32]. In our derivation of the antifermion amputation rules we have neglected the minus sign in front of the factor $\sum v\bar{v}$: not unreasonably, since in the derivation we dealt with the cross section, hence with $|\mathcal{M}|^2$ rather than $\mathcal{M}$ itself. We might have decided to take, say, $-\bar{v}$ for an incoming antifermion in order to include this sign, but we prefer to stick to the more usual assigments. But the minus sign rears its head in other places, as we shall now discuss.

5.1.4 The Fermi minus sign

Let us take a graph in which two fermion lines appear in opposite directions, and let us consider the cut through these two fermion lines: this is depicted below.



The cut may also go through other internal lines. The momenta are p^μ and q^μ, both with positive energy in the direction going from blob A to blob B (as required by the cutting rules). The two fermions are not necessarily of the same type and mass. According to our Feynman rules, the diagram *before* cutting reads

$$\text{Tr}\left[B\frac{\not{p}+m_1}{p^2-m_1^2+i\epsilon}A\frac{-\not{q}+m_2}{q^2-m_2^2+i\epsilon}\right] \ .$$

The trace over spinorial indices is necessary because both A and B must have the form of Dirac matrices (possibly in a complicated combination). The Feynman rules we have extracted for external fermions, however, tell us to take, for the cut diagram:

$$\text{Tr}\left[Bu(p)\bar{u}(p)Av(q)\bar{v}(q)\right] = \bar{u}(p)Av(q)\ \bar{v}(q)Bu(p) \ ,$$

where we have implied summation over the spins in the spinors. We see that these results are inconsistent by an overall sign because $\sum \bar{v}(q)v(q) = \not{q} - m_2$. To repair this (and, hence, repair

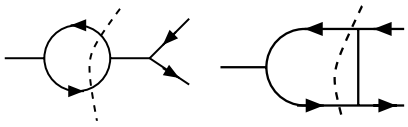
[32] This convention allows us to decide whether we have, say, an outgoing fermion or an incoming antifermion in our process.

unitarity) we have to include an extra minus sign in the Feynman rules. The most economical way to do this is by two extra prescriptions.

By the structure of the Dirac spinors, every diagram consist of zero or more *fermion lines* connecting an external 'outgoing spinor', $\bar{u}$ or $\bar{v}$, with an 'ingoing spinor', u or v, and zero or more fermion lines going in closed loops. All these are connected and dressed up by bosonic pieces. If we adopt the convention that in every diagram contributing to a certain $\mathcal{M}$ the *ingoing* external fermions, say, are put in a fixed order, the ordering of the *outgoing* is unambiguously given for each diagram, and these orderings may differ between different diagrams. The minus sign precriptions can now be given as follows:

- Every closed fermion loop gives a factor -1.
- Every interchange of two outgoing external fermion lines (in the above sense) gives a factor -1.

That the second prescription is also necessary to maintain unitarity can be seen from the following example, where the solid lines without arrows stand for any bosonic particle:



The first diagram must, by the first rule, have a relative minus sign with respect to the second one when the diagrams are uncut: the second rule is necessary to ensure the same relative minus sign when the diagrams are cut.

Finally, it is important to observe that the Fermi minus sign holds for *all* fermions, not only those that you would naively consider as identical particles, such as two electrons: the two tree-level diagrams for Moeller scattering, $e^-e^- \to e^-e^-$, have a relative minus sign. Indeed, electron and positron are easily distinguishable (and we have a relative minus sign in the two tree-level diagrams for Bhabha scattering, $e^+e^- \to e^+e^-$), but the minus sign rules also hold for exotic interchanges such as between the muon neutrino and the top quark (somewhat less exotically, the two tree-level diagrams for $e^+e^- \to \nu_e\bar{\nu}_e$ also have a relative minus sign). Really, it depends on the set of interaction vertices whether or not one fermion can replace another in a Feynman diagram. In QCD for instance, there are no vertices that allow us to interchange a single up quark and a single down quark, and their relative minus sign is invisible; but if we take the weak interactions into account, extra diagrams and interchanges becomes possible, and the sign *does* matter.

5.1.5 Aside: long-distance communication

There is a difference between scalar and non-scalar particles. As we have seen from the amputation arguments, when a scalar particle is produced and decays very much later, the production and decay are separated in spacetime to such an extent that the only information going between them is the momentum of the scalar particle, and for the rest the particle's decay can be considered to be independent from its production. Not so for spinning particles: the amputation for, say, an outgoing fermion gives a factor $\bar{u}(p)$, which is coupled to a factor $u(p)$ in its subsequent decay: and both spin states of this spinor propagate together from production to decay. Therefore, more information than just momentum is exchanged between production and decay, no matter how far they are separated. Moreover, this will hold also for two particles that are produced in one point and subsequently move very far away from each other before decaying.

5.2 Massive vector bosons

5.2.1 Polarization vectors and the Proca equation

We now turn to yet another particle type with internal structure, corresponding to spin-1. The best-known such particle is of course the photon, but it is simpler to start with *massive* spin-1 particles such as the W and Z. The reason is that for such particles we can define a rest frame. Apart from their momentum p^μ, such particles are decribed by their *polarization vector* ε^μ, which in the rest frame of the particle is nothing but the direction in which the 'electric' field points, hence we have

$$p^2 = m^2 \quad , \quad p\cdot\varepsilon = 0 \quad , \quad \varepsilon\cdot\varepsilon^* = -1 \quad , \tag{148}$$

where m is again the mass. Note the similarity between these and Eq.(140). The last equation is written such that it holds for real as well as complex polarization vectors. In the massive case, there are 3 linearly independent solutions, correspoding to the fact that a massive spin-1 system has 3 different values that the spin component in a given direction can take. Real polarization vectors, such as ε_1^μ, correspond to linear polarization, and complex ones, such as $(\varepsilon_1^\mu + i\varepsilon_2^\mu)/\sqrt{2}$, to elliptic or even (as in this case) circular polarization. The sum over spins in the rest frame must read

$$\sum_{j=1}^{3} \varepsilon_j^\mu (\varepsilon_j^*)^\nu = \begin{cases} 1 \text{ if } \mu = \nu \neq 0 \\ 0 \text{ otherwise} \end{cases}$$

and therefore the spin sum in any Lorentz frame reads[33]

$$\sum_{j=1}^{3} \varepsilon_j^\mu (\varepsilon_j^*)^\nu = -g^{\mu\nu} + \frac{p^\mu p^\nu}{m^2} \quad . \tag{149}$$

The conditions on ε can convienently be combined[34] in the so-called *Proca* equation:

$$\partial^\nu \partial_\nu W_\mu - \partial^\nu \partial_\mu W_\nu + m^2 W_\mu = 0 \quad , \tag{150}$$

where ε_μ is the coefficient of the mode of the W field with momentum p. If we contract Eq.(150) with ∂^μ we obtain $\partial^\mu W_\mu = 0$, or $p\cdot\varepsilon = 0$, and inserting that again the Eq.(150) gives the Klein-Gordon equation which takes care of the mass-shell condition for free particles. Note that this only works if $m \neq 0$.

5.2.2 Vector boson propagators and amputations

Just as in the Dirac case, the propagator is related to the Green's kernel of the field equation. The Greens kernel R must satisfy the Proca equation with a Dirac delta function on the right-hand side:

$$\partial^\nu \partial_\nu R_{\mu\alpha}(x) - \partial^\nu \partial_\mu R_{\nu\alpha}(x) + m^2 R_{\mu\alpha}(x) = \delta(x) g_{\mu\alpha} \quad , \tag{151}$$

In momentum representation, we have of course $R_{\mu\nu}(p)$, and the Proca equation reads

$$-p^2 R_{\mu\nu}(p) + p_\mu p_\beta R_{\beta\nu}(p) + m^2 R_{\mu\nu}(p) = g_{\mu\nu} \quad , \tag{152}$$

with solution

$$R_{\mu\nu}(p) = \frac{-g_{\mu\nu} + p_\mu p_\nu / m^2}{p^2 - m^2} \quad . \tag{153}$$

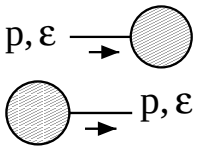
[33] The expression is correct in the rest frame, and it is Lorentz-invariant.

[34] The orthogonality of p and ε, $\partial\cdot W = 0$, follows from contracting with ∂^μ, and the mass-shell condition then follows by inserting this condition again.

Notice again that this will not wirk for $m = 0$. As usual, for unstable particles a width, and for stable particles an $i\epsilon$ has to be added for reasons of unitarity. The Feynman rule for a stable spin-1 particle propagator therefore is seen to be

$$\mu \xrightarrow[p]{} \nu \rightarrow \frac{i}{p^2 - m^2 + i\epsilon}\left(-g^{\mu\nu} + \frac{p^\mu p^\nu}{m^2}\right) = \frac{i\sum_{j=1}^{3}\varepsilon_j^\mu \varepsilon_j^{*\nu}}{p^2 - m^2 + i\epsilon} \quad . \tag{154}$$

Again, we recognize the spin-sum in the 'symbolic' last expression. The Feynman rules for external stable spin-1 particles follow again from the amputation argument:



$$\rightarrow \varepsilon^\mu \ , \qquad \rightarrow \varepsilon^{*\mu} \ , \tag{155}$$

where the arrow shows the flow of positive energy. The index on the polarization vector has to be contracted with a corresponding vector index in whatever vertex the external line is attached to, just as in the Dirac case vertices involving fermions must carry a Dirac index. Since the propagator is *even* in the momentum we do not need to orient the spin-1 lines, nor is there a minus sign to worry about.

5.3 Massless vector bosons

5.3.1 Longitudinal and transverse polarization

Since the photon, at least, is massless or almost so[35], we need to consider the $m \rightarrow 0$ limit of the spin-1 case discussed so far. Here, we run into trouble. Consider a massive spin-1 boson moving along the z axis with momentum p and energy $E = \sqrt{p^2 + m^2}$. By boosting from the rest frame, we see that the three possible polarization vectors can be chosen as follows:

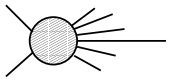
$$\begin{aligned} p^\mu &= (E, 0, 0, p) \ , \\ \varepsilon_1^\mu &= (0, 1, 0, 0) \ , \\ \varepsilon_2^\mu &= (0, 0, 1, 0) \ , \\ \varepsilon_3^\mu &= \left(\frac{p}{m}, 0, 0, \frac{E}{m}\right) = \frac{p^\mu}{m} + \mathcal{O}\left(\frac{m}{E}\right) \ . \end{aligned} \tag{156}$$

The vectors ε_1 and ε_2 correspond to the field being at right angles to the velocity, and hence of course unaffected by the boost. They (and real or complex linear combinations of them) are called the *transverse polarizatoins*. The vector ε_3, however, corresponds to the field being in the direction of motion, and hence by Lorentz contraction becoming more and more intense as the velocity increases: this is called the *longitudinal polarization*. As the velocity approaches that of light when $m \rightarrow 0$, it blows up. That this is a problem can be seen from the Feynman rule for external particles: the matrix element $\mathcal{M}$ contains ε as an overall factor, so $\mathcal{M}$ will blow up when ε does, and we are bound for unitarity violation, especially when $m = 0$. Nevertheless, the propagators and external lines are (more or less) unambiguously determined.

[35] The best limit to date is about 10^{-25} GeV.

5.3.2 Unitarity and current conservation

A possible way out of our dilemma is to realize that a theory contains more than propagators and external lines: there are also interaction vertices. If a particle behaves awfully but has no interaction with the rest of the cosmos, it is just the same as if it doesn't exist, and we might as well ignore it. Let us consider a process where a very light and fast spin-1 particle is emitted. If it is transversely polarized, we do not particularly worry about unitarity, so let us take it longitudinally polarized (denoted by L):



$$= \mathcal{M} = \mathcal{M}^\mu(\varepsilon_L)_\mu = \mathcal{M}^\mu \left(\frac{p^\mu}{m} + \mathcal{O}\left(\frac{m}{E}\right)\right) \quad . \tag{157}$$

Here, we have written $\mathcal{M}^\mu$ for the rest of the diagram (or set of diagrams). We now adopt the law that **the vertices in the theory have to be such that in all cases such as this, we have**

$$\mathcal{M}^\mu p_\mu = 0 \quad . \tag{158}$$

If we can arrange this, then the longitudinal spin-1 particle decouples when $m \to 0$: it has no interaction with anything else in the world, and is effectively absent from the theory. In fact, the above requirement is well-known. The diagram piece $\mathcal{M}^\mu$ acts, after all, like a source for the 'photon', and (since we work in the momentum representation) it is the Fourier transform of a spacetime current J^μ. The requirement of Eq.(158) is seen to be

$$\partial_\mu J^\mu = 0 \quad , \tag{159}$$

which is nothing but conservation of electromagnetic charge in the case of photons.

Another way to arrive at the same result is to consider more closely the statement that 'the electromagnetic field is transverse'. This is obviously not a Lorentz-invariant statement: even if the polarization vector happens to be transverse in one frame, a simple boost may make it non-transverse. In general, a massless spin-1 particle will have momentum and polarization given by

$$p^\mu = (|\vec{p}|, \vec{p}) \quad , \quad \varepsilon^\mu = (\varepsilon^0, \vec{\varepsilon}) \quad , \quad |\vec{p}|\varepsilon^0 = \vec{p}\cdot\vec{\varepsilon}. \tag{160}$$

Now we can split $\vec{\varepsilon}$ into a longitudinal part $\vec{\varepsilon}_L$ and a transverse part $\vec{\varepsilon}_T$:

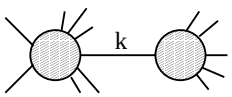
$$\vec{\varepsilon} = \vec{\varepsilon}_L + \vec{\varepsilon}_T \quad , \quad \vec{\varepsilon}_L = \frac{(\vec{p}\cdot\vec{\varepsilon})\vec{p}}{|\vec{p}|^2} = \varepsilon^0 \frac{\vec{p}}{|\vec{p}|} \quad , \tag{161}$$

so that

$$\varepsilon^\mu = (\varepsilon^0, \vec{\varepsilon}_L + \vec{\varepsilon}_T) = (0, \vec{\varepsilon}_T) + \frac{\varepsilon^0}{|\vec{p}|} p^\mu \quad , \tag{162}$$

with the same result, namely that the transversality statement is effectively Lorentz-invariant provided $\mathcal{M}^\mu p_\mu = 0$.

So far, we have been able to decouple external longitudinally polarized photons, but a danger lurks yet, since photons may of course also occur as internal lines. To solve this, we require that condition (158) also holds for off-shell momenta k^μ. Consider the diagram(s) given by



where we have indicated the momentum of the internal massless spin-1 line by k^μ. We can write this, in the almost-massless case, as

$$\mathcal{M} = \mathcal{M}_1^\mu \frac{i(-g_{\mu\nu} + k_\mu k_\nu/m^2)}{k^2 - m^2 + i\epsilon} \mathcal{M}_2^\nu \quad . \tag{163}$$

We now request that for all such diagrams,

$$\mathcal{M}_1^\mu k_\mu = 0 \quad \text{and} \quad \mathcal{M}_2^\nu k_\nu = 0 \quad . \tag{164}$$

This *gauge invariance* requirement will allow us to let m go to zero. The condition is too strict to satisfy for each individual diagram: the best we can arrange is that it holds for the complete set of diagrams that make up $\mathcal{M}_1$ and $\mathcal{M}_2$. That, in turn, means that it is extremely dangerous to leave out diagrams, since delicate cancellations are necessary to satisfy gauge invariance: forgetting one term can lead to a cross section that is off by many orders of magnitude.

Finally, a word on the connection with Dirac particles. For those, the spin vector s^μ blows up just like ε^μ for longitudinal polarization. It is one of the many small miracles of Dirac algebra that $u(p, s)$ remains finite even if s^μ explodes; our insistence on gauge invariance for spin-1 particles is necessary because no such mechanism is around for ε^μ.

APPLICATIONS OF QCD

V.M. Braun
Institut für Theoretische Physik, Universität Regensburg,
Universitätsstr. 31, D-93053 Regensburg, Germany

Abstract
In these lectures I try to explain basic ideas behind the applications of quantum chromodynamics to the description of hard processes and indicate the directions of current activities.

1 INTRODUCTION

These notes are intended as a 'crash course' for experimental particle physicists who do not intend to do QCD calculations themselves, but are curious to know about recent developments and, most importantly, want to understand why — and to which extent — the calculations done by theorists have anything to do with the observed reality. The quantum chromodynamics in its present form is essentially a "knowhow" to calculate propagation and interactions of colored objects at small distances. On the other hand in experiments one deals with colorless hadrons at large distances from the interaction point; strong interaction between hadrons is known very poorly and it is not known how to derive it from the underlying quark-gluon interaction, apart from a few exceptional cases. One should wonder, therefore, how the signatures of quark and gluon interactions survive through the hadronization stage. The main topic of this lectures is to explain the physical principles that lie behind the calculations; how accurate the theory predictions can be, and what to do if a disagreement between theory and experiment is found.

2 INFRARED SAFETY

In this section we will consider an example of hadron production in e^+e^- annihilation, and, in particular, its total cross section. In order to get rid of various kinematical factors, it has become standard to consider the ratio

$$R(q^2 = E^2_{\rm cm} = s) = \frac{\sigma(e^+e^- \to \text{ hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)}. \tag{1}$$

In the c.m. frame the interesting part of this process is initiated by the virtual photon (or Z^0) decay into a quark-antiquark pair, see Fig. 1. The quark and the antiquark are created at one point in space and fly

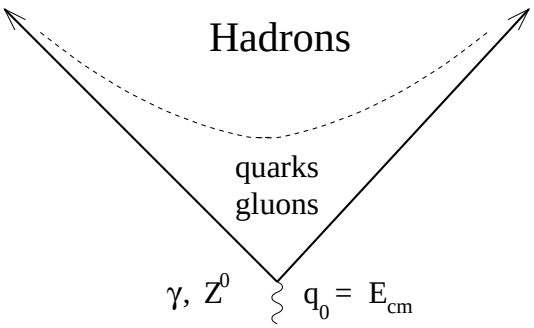


Fig. 1: Hadron production in e^+e^--annihilation

in opposite directions with the velocity of light (if quark masses are neglected); the time of the creation can be determined to the accuracy up to $\delta t \sim 1/E$ because of the uncertainty principle. Our approach in these lectures will be to take for granted that quark propagation and interactions at small distances are

described by the nonabealian gauge theory — the QCD. In this section we will try to identify a class of physical observables for which this incomplete knowledge proves to be sufficient for their description to a certain accuracy.

2.1 Unitarity+Analiticity+QCD

The basic idea behind all QCD applications is that although large-distance structure of QCD interactions remains a mystery and the dynamics of hadronization process is unknown, it must obey certain general rules in order to preserve Lorentz invariance and causality of the theory.

One of these general principles is the conservation of probability. Consider the total cross section, summed over contributions of all possible hadron states and integrated over the hadron momenta. This set of states is *complete*, which means that the sum of probabilities for *any given* initial-state configuration decaying into *some* hadron state is unity. This completeness condition is usually formulated as unitarity of the 'S-matrix' of the transition matrix elements between the initial and final states in a scattering process, and is postulated in axiomatic field theory. One familiar consequence of unitarity is the optical theorem that relates the total cross section to the imaginary part of the forward scattering amplitude. Similarly, unitarity allows to calculate the total cross section of e^+e^- annihilation as the imaginary part of the correlation function of two electromagnetic currents:

$$R(s) = \mathrm{Im}\ \Pi(q^2 = s), \tag{2}$$

where

$$i\int d^4x\, e^{iqx}\langle 0|T\{j_\mu^{\rm em}(x) j_\nu^{\rm em}(0)\}|0\rangle = (q_\mu q_\nu - q^2 g_{\mu\nu})\Pi(q^2) \tag{3}$$

and $j_\mu^{\rm em} = e_u \bar{u}\gamma_\mu u + e_d \bar{d}\gamma_\mu d + \ldots$. The premium is obvious since any reference to a particular hadronic state and the summation over states disappeared altogether! One may hope that the calculation of $\Pi(q^2)$ does not involve some of the strong interaction problems but is it simple enough, at least for large energies q^2?

The answer to this question is in fact not trivial. First, notice that although using unitarity is a great help to avoid the question how exactly hadrons are produced from quarks, we still do not get rid of long-distance interactions. Indeed, an inspection of the integral in (3) shows that the main contribution to its imaginary part comes from the integration region in coordinate space $x^2 \sim 1/q^2$ and since $x^2 = x_0^2 - \vec{x}^2$ this means that both time and distance can be large. Second, the detailed behavior of the cross section as a function of energy has to be rather elaborate in order to reproduce the kinematical hadron production thresholds. Let the photon energy be close to an exact multiple of the pion mass: $E_n = nm_\pi$. Because of the energy conservation, this value serves as a boundary for the possibility to produce at most $n-1$, or at most n pions in the final state. The annihilation cross section has, therefore, to exhibit a typical threshold behavior with a small cusp at $E = E_n$. It is rather clear that a complicated structure of thresholds corresponding to the production of the whole enormous variety of hadrons and hadron resonances cannot be reproduced in perturbative QCD calculations. One may hope that the actual contribution of such processes to the total cross section is small (at LEP, they correspond to events with very large multiplicity), but this requires further analysis.

There exists an elegant way out, however, that allows to avoid both difficulties by going over to unphysical, imaginary values of the energy. It can be proven that if the strong interactions obey special relativity and preserve causality (and we strongly hope they do!), then the correlation function $\Pi(q^2)$ must be analytic in the complex q^2 plane, with a cut at real q^2 stretching from the lowest hadron threshold to infinity, see Fig. 2. The trick is to use the Couchy theorem that allows to calculate the value of an analytic function at an arbitrary point in the complex plane provided its discontinuity (imaginary part) is known at all singularities. In our case, choose the integration contour encircling a (negative) value of q^2 of interest, and deform the counter as shown in Fig. 2. Assuming that the integral over the large circle vanishes, one is left with the difference of integrals above and below the real axis that can be

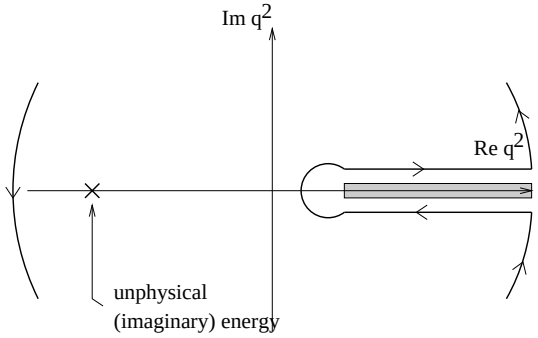


Fig. 2: Complex q^2 plane in e^+e^--annihilation

expressed in terms of the imaginary part $\mathrm{Im}\,\Pi(s=q^2)=R(s)$. In fact, the integral over the large circle does not vanish, and to repair this one has to take a derivative over q^2 (write the dispersion relation with one subtraction). The result presents a so-called duality sum rule:

$$q^2\frac{d}{dq^2}\,\Pi(q^2)=q^2\int\limits_{\rm threshold}^{\infty}ds\,\frac{R(s)}{(s-q^2)^2}. \tag{4}$$

The way to read this formula is the following. Assume q^2 is large and negative. The integral on the r.h.s. involves a physical cross section (2) and is measurable. On the other hand, the correlation function $\Pi(q^2)$ is now calculable since it only involves contributions of small distances $|x_0|\sim|\vec{x}|\sim 1/\sqrt{-q^2}$. To see this explicitly one has to examine the structure of the integral in (3). A qualitative explanation is as follows. The wave function of the quark-antiquark pair produced via the virtual photon decay has a typical oscillating time dependence familiar from quantum mechanics $\Psi\sim e^{-iEt}$. Negative values of q^2 correspond to imaginary energies $E=-i\mathcal{E}$ and, therefore, the corresponding wave function decays exponentially with time $\Psi\sim e^{-\mathcal{E}t}$ so that quarks cannot 'propagate' far from the origin.

All 'gold-plated' tests of QCD are organized by the same general scheme, relating observable cross sections integrated over large energy regions, with quantities that are calculable (dominated by contributions of small distances) at unphysical values of energy. The most important examples are Bjorken and Gross-Llewellyn sum rules in deep-inelastic lepton-nucleon scattering, we will consider them in what follows. Their derivation relies on the applicability of QCD at short distances combined with using dispersion relations that follow from the (generally assumed) unitarity and causality of strong interactions; these are the 'most exact' QCD predictions that the theory can provide us with at present.

2.2 Space-time picture

The strength — and the shortcoming — of the approach outlined above is that it allows to avoid the necessity to consider the space-time evolution of quark-gluon interactions and the strong interaction processes in general. This is a strong point, because it allows to make rigorous predictions, and it is a weak point since the number of such predictions appears to be very small. We will try, therefore, to develop a space-time picture of quark-gluon interactions in the hope that the number of applications will increase considerably.

The heuristic argument can be given as follows. We assume that uncalculable — let us say nonperturbative — effects come entirely from large distances, whereas the interactions at small distances can be systematically taken into account. This picture tacitly implies that the e^+e^- annihilation proceeds via a two-step process with the 'short distance' and 'long distance' subprocesses occuring at different time scales, see Fig. 3. The crucial observation is that to the extent that they occur at different time scales, the short- and long-distance interactions cannot have a quantum-mechanical interference. Because of this

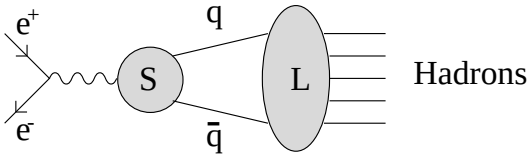


Fig. 3: Short-distance S and long-distance L interactions have no quantum interference

the probability for the production of a particular hadron state h can be written as a classical product of probabilities

$$P_{e^+e^-\to h} = P_{e^+e^-\to \bar{q}q} * P_{\bar{q}q\to h} \tag{5}$$

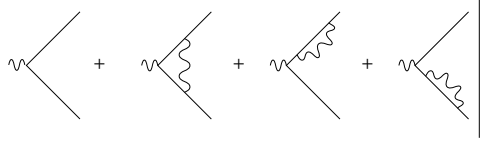
rather than amplitudes. The total cross section is then equal to

$$\sigma_{\rm tot} = \sum_h P_{e^+e^-\to h} = P_{e^+e^-\to \bar{q}q} \sum_h P_{\bar{q}q\to h} = P_{e^+e^-\to \bar{q}q}\,, \tag{6}$$

where in the last equality we have used $\sum_h P_{\bar{q}q\to h} = 1$ that is again the completeness condition (unitarity), but applied to long-distance interactions *separately*.

The quantum-mechanical expressions (5) and (6) are still oversimplified in one important aspect. A brute-force calculation of the cross section $\sigma_{e^+e^-\to \bar{q}q}$ to first order in perturbation theory brings an embarassing result: infinity. To repair this, one has to remember that the idea behind writing (5) and (6) was to separate contributions of large and small distances. This separation implies that we identify some scale — for example time t_0 that passed from the moment of annihilation — beyond which we are going to treat gluon emission as nonperturbative and include in the block labeled 'L' in Fig. 3. Conversely, only the part of the gluon interactions that occur at times less than t_0 have to be included in block 'S' and in what we call $\sigma_{e^+e^-\to \bar{q}q}$. In practice, using a time cutoff is inconvenient because it leads to ugly integrals and the theorist's trick (one of) is to achieve the same goal of cutting away the contribution of large distances by giving a small mass $m_g \sim 1/t_0$ to a gluon. This helps indeed, and the straightforward calculation yields a finite result

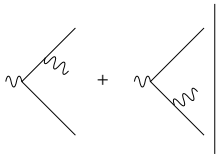
$$\sigma_{e^+e^-\to \bar{q}q} = \Big|$$



$$\Big|^2 = \sigma_0 \left\{ 1 + \frac{4}{3}\frac{\alpha_s}{\pi}\left[-2\ln^2\frac{Q}{m_g} + 3\ln\frac{Q}{m_g} - \frac{7}{4} + \frac{\pi^2}{6}\right]\right\}. \tag{7}$$

Having solved one problem, we run into another one: the cross section for the quark-antiquark production depends on the scale at which we measure it. However, this is not the whole story. Calculation of the total cross section implies that we have to take into account contributions of all possible states of the system at all times and at $t = t_0$ in particular. The difference with quantum mechanics is that in a field theory the number of particles is not conserved and in addition to the quark-antiquark states one also has to include multiparticle contributions involving extra gluons. To first order in perturbation theory only one gluon can be produced. We get

$$\sigma_{e^+e^-\to \bar{q}qg} = \Big|$$



$$\Big|^2 = \sigma_0 \frac{4}{3}\frac{\alpha_s}{\pi}\left[+2\ln^2\frac{Q}{m_g} - 3\ln\frac{Q}{m_g} + \frac{5}{2} - \frac{\pi^2}{6}\right] \tag{8}$$

and in the sum of both contributions the dependence on m_g luckily goes away:

$$\sigma_{\rm tot}^{e^+e^-} = \sigma_{e^+e^-\to\bar{q}q} + \sigma_{e^+e^-\to\bar{q}qg} = \sigma_0\left[1+\frac{\alpha_s}{\pi}+\mathcal{O}(\alpha_s^2)\right]. \tag{9}$$

As m_g decreases (time t_0 increases) the probability to have a gluon in the final state increases as $\sim \ln^2\frac{Q}{m_g}$ and the probability to have a 'pure' quark-antiquark state decreases by the same amount. The cancelation in (9) means that the total cross section of e^+e^--annihilation appears to be not sensitive to this redistribution and, as a consequence, does not depend on the (unknown) details of quark-gluon interactions at large distances. Another way to state this: using a gluon mass corresponds to an *ad hoc* modification of the theory at large distances; since this modification apparently does not influence the total cross section, we conclude that the latter is not sensitive to the exact structure of large distance interactions.

Note that the argument is general. For a given observable, check whether it is sensitive to the gluon mass if calculated in perturbation theory (theorists speak of an 'infrared cutoff'). If there is no dependence (better to say, the dependence is mild), one may hope that this observable will not be modified by *any* long-distance interactions, both perturbative and nonperturbative. The property of being not sensitive to the infrared cutoff is called infrared safety and the corresponding observables are usually referred to as infrared-safe.

Best about this is that the explicit calculation is often not needed since it is rather easy to formulate intuitive criteria for an infrared safe quantity. Indeed, long-distance interactions require long time, and, therefore, can only proceed between the two partons (quark and gluons) with small relative momenta. In a given frame, this means that the two partons either have to have small momenta — partons are called 'soft' in this case — or they have both to fly in the same direction — so they are called 'collinear'. Long-distance interactions are possible only between soft and collinear partons. It follows that observables that do not change when a) one particle splits in two collinear and b) a very soft particle is emitted, are infrared-safe and have chance to be insensitive to nonperturbative effects. This suggests that summing over the states with 'jets' of nearly collinear particles and possibly adding contributions of very soft particles can be a good idea. A famous example of an infrared safe cross section is provided by the energy flow into constrained angular regions, see Fig. 4. The two-jet cross section in e^+e^--annihilation

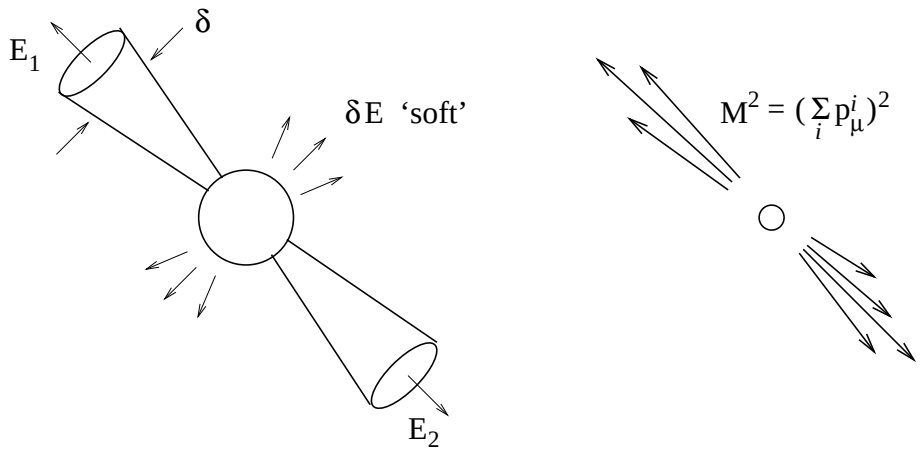


Fig. 4: Calorimetric jet measurements

can be calculated as

$$\sigma_{2J} = \frac{3}{8}(1+\cos^2\theta)\left[1+c_1\frac{\alpha_s}{\pi}+c_2\left(\frac{\alpha_s}{\pi}\right)^2+\dots\right] \tag{10}$$

where the coefficients c_k depend on the jet definition. For example, if a jet is defined as a collection of particles within a solid angle δ (c.f. Fig. 4), then

$$\sigma_{2J}(\delta) = \frac{3}{8}(1+\cos^2\theta)\left[1 - \frac{4}{3}\frac{\alpha_s}{\pi}\left(4\ln^2\delta + 3\ln\delta + \frac{\pi^2}{3} + \frac{5}{3}\right) + \ldots\right]. \tag{11}$$

At this order only one gluon can be emitted so that the total cross section is equal to the sum of the two-jet and the three-jet cross sections, $\sigma_{\rm tot} = \sigma_{2J} + \sigma_{3J}$. As $\delta \to 0$, more events are identified as three jets. One can study properties of individual jets as well, for example the jet invariant mass distribution is an infrared-safe quantity.

Note that the expressions for $\sigma_{e^+e^-\to\bar{q}q}$ in (7) and $\sigma_{e^+e^-\to 2\text{ jets}}$ in (11) are mathematically very similar. The contributing Feynman diagrams are precisely the same and the only difference (apart from using a different infrared cutoff) is that quark production is *interpreted* in the latter case as the *quark jet* production. The possibility of such an interpretation is established by the infrared safety of the jet.

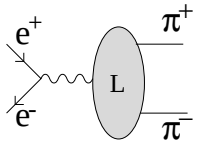
In fact there is a theorem by Kinoshita-Lee-Nauenberg (KLN) that applies to theories with massless vector particles (e.g. QED or QCD) and formulates precise conditions under which a given observable is not affected by long-distance interactions. The requirement is, roughly speaking, that an observable has to be sufficiently 'inclusive' and does not distinguish between the states that have the same energy. Summing over the states with arbitrary many quark and gluons (or hadrons) within a jet presents an example of the KLN sum. Using the KLN theorem (in fact, a weaker version of it) one can argue that e.g. the cancelation of the gluon mass dependence in (9) is valid to all orders of perturbation theory.

2.3 How things might go wrong

Notice that the concept of infrared safety apparently allows to make more strong predictions compared to the dispersion approach outlined in Sect. 2.1. For example, using the infrared safety argument we are seemingly allowed to calculate the total cross section of e^+e^- annihilation *for each value of energy*, compared to the complicated integral (4) in the dispersion approach. This comes at a price, however, that the prediction is less rigorous.

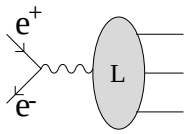
Technically, any perturbative calculation is done by evaluation of Feynman diagrams which contain four-dimensional integrations over the momenta (or coordinates) of virtual quanta. These integrals are tricky because of the poles in the propagators (so that the integrand is singular) and the whole idea of the Feynman method is, essentially, to define these integrals by analytic continuation from the Euclidian region $x_0 \to i\tau$. Referring to Fig. 2, this means that any perturbative calculation is in fact done at imaginary energies, and then continued analytically to the physical cut at $q^2 > 0$, consistent with our intuition that any perturbative calculation is in fact done at small distances. The difference between the dispersion sum rules and the perturbative calculations is that the analytic continuation is performed in the opposite directions: In the first case, the physical observable is continued from the cut to the unphysical (Euclidian) region, while in the second case the perturbative calculation is continued from the Euclidian to the physical region. The problem with the latter is that any (perturbative) calculation is only an approximation to the exact result and small errors can be amplified by the analytic continuation. A famous example is the following: Imagine the perturbative calculation is missing a correction of order $\exp\{-\sqrt{-q^2}\}$ that is very small at large $-q^2$ and can safely be neglected in the dispersion sum rule (4). The analytic continuation of this tiny correction to the physical region gives, however, $\exp\{-\sqrt{-q^2}\} \to \sin\{\sqrt{q^2}\}$ that is oscillating and not suppressed at all! (From this consideration it also becomes clearer how the dispersion sum rule actually works: The oscillations are smoothened by the integration and average to zero.) It follows that within a perturbative calculation of the total cross section we are not guaranteed against the existence of corrections that are out of control. In the literature such corrections are referred to as violation of duality; they came to the spotlight recently in connection with inclusive B-decays. Up to now, very little is known about them.

To conclude, the concept of infrared safety is very important, but less rigorous compared to the dispersion sum rules considered in the Sect. 2.1. Apart from the above technical argument, it is worthwhile to have in mind another example of the effect that could be missing in this framework. Imagine that there exists a purely nonperturbative vertex for the transition of the virtual photon directly into hadrons, without any quarks/gluons in the intermediate stage. Since no short-distance subprocess exists, the calculation outlined in Eqs. (6)–(9) and all our further argumentation do not apply. One should expect that such an exotic nonperturbative contribution to, say, pion pair production is very small, of order



$$\sim e^{-c/\alpha_s} \tag{12}$$

where c is some constant. (This is an almost generic way of writing of a contribution that cannot be expanded in powers of the coupling.) In perturbation theory emission of extra particles would always yield extra powers of the coupling g_s, $\alpha_s = g_s^2/(4\pi)$. Beyond perturbation theory this is not so, and it is possible to argue that each extra particle rather brings a factor $1/g_s$, e.g.



$$\sim \frac{1}{g_s} \cdot e^{-c/\alpha_s}\,, \tag{13}$$

etc. Summing up contributions with n hadrons in the final state, for the full nonperturbative contribution to the total cross section one gets

$$\sigma_{\rm tot} = \sum_n \sigma_n \sim e^{-2c/\alpha_s} \sum_n \frac{1}{\alpha_s^n}. \tag{14}$$

Under favourable circumstances, the sum over inverse powers of the coupling may produce an exponential enhancement $\sum_n 1/\alpha_s^n \sim e^{+2c'/\alpha_s}$ and if $c' \simeq c$ then the nonperturbative contribution to the cross section may become of order unity! (Note that the sum is dominated by final states with very large multiplicity; cf. the discussion of particle thresholds effects in Sect. 2.1.)

A possibility of large nonperturbative corrections to total cross sections at large energies has triggered intensive discussions in 1989–1994, mainly in connection with possible baryon number violation in the electroweak theory. The outcome of this discussion is that such processes still remain to be strongly suppressed (i.e. $c' < c$) although the exact suppression rate is difficult to calculate; in fact it remains an open problem. The situation in QCD is probably similar and purely nonperturbative contributions are very small; their very existence is, however, interesting as it shows limitations of the standard approach.

3 GLUON BREMSSTRAHLUNG

We have discussed general ideas how to convert results of calculations in terms of quarks and gluons into predictions for hadron observables. As the next step, let us have a more close look at the gluon radiation in QCD perturbation theory and learn some of its basic features.

3.1 Soft and collinear emission

We start with a simple calculation. Imagine a very fast quark exits the interaction region and emits a soft gluon at a small angle, see Fig. 5. Let p be the quark momentum after the emission, $p^2 = m_q^2$ where m_q is the quark mass. And let q be the gluon momentum. We will assume that the gluon has a small mass m_g as well, so that $q^2 = m_g^2$. The probability amplitude of the emission is given by the product of

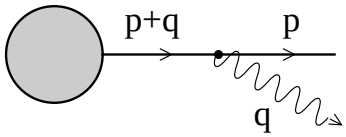


Fig. 5: Gluon bremsstrahlung

several factors: the quark propagator to the emission point $i/(\not p + \not q - m_q)$, the QCD vertex $ig_s\gamma_\mu$, the gluon polarization vector $\varepsilon_\mu^{(\lambda)}$ and the quark spinor $u(p)$:

$$A = -g_s \frac{\not p + \not q + m_q}{(p+q)^2 - m_q^2} \cdot \not\varepsilon^{(\lambda)} \cdot u(p) \sim -g_s \frac{2\varepsilon \cdot p}{2pq + m_g^2}, \tag{15}$$

where to arrive at the second expression I used that $(\not p + m_q)\, \not\varepsilon u(p) = 2\varepsilon \cdot p u(p)$ because of the Dirac equation $(\not p - m_q)u(p) = 0$ and assumed that $\not q \ll \not p$ so it can be neglected. Next, let θ be the emission angle of the gluon relative to the quark and take the frame of reference such that $p_\mu = (E, 0, 0, p)$ and $q_\mu = (\omega, 0, q\sin\theta, q\cos\theta)$. Further assuming that $\theta \ll 1$, $E \gg m_q$, $E \gg \omega \gg m_g$ we can expand

$$\begin{aligned} &\cos\theta \simeq 1 - \theta^2/2\,, \qquad p \simeq E[1 - m_q^2/(2E^2)]\,, \qquad q \simeq \omega[1 - m_g^2/(2\omega^2)]\,, \\ &2pq = 2E\omega - 2|\vec p||\vec q|\cos\theta \simeq E\omega\left(\theta^2 + \frac{m_q^2}{E^2} + \frac{m_g^2}{\omega^2}\right). \end{aligned} \tag{16}$$

Finally, for a physical (transverse) gluon polarization

$$\varepsilon \cdot p \simeq E\sin\theta \simeq E\theta\,. \tag{17}$$

Collecting everything, we obtain a simplified expression for the probability amplitude

$$A \sim \frac{1}{\omega} \cdot \frac{\theta}{\theta^2 + m_q^2/E^2 + m_g^2/\omega^2}. \tag{18}$$

Here and below I neglect numerical factors. The emission probability (differential cross section) is given by the amplitude squared, integrated over the gluon phase space:

$$d\sigma \sim |A|^2 \frac{d^3q}{2\omega} \sim |A|^2 \omega d\omega\, \theta d\theta \sim \frac{d\omega}{\omega} \cdot \frac{\theta^2 d\theta^2}{\left[\theta^2 + m_q^2/E^2 + m_g^2/\omega^2\right]^2}\,, \tag{19}$$

which is the desired result. Note that the cross section is strongly peaked at small angles $\theta \sim m_q/E$ or $\theta \sim m_g/\omega$ depending on which of the two ratios is larger; at the same time the forward cross section $\theta = 0$ vanishes. The total bremsstrahlung cross section is obtained by integrating (19) over the emission angle and over the gluon energy $m_g < \omega < E$. If $m_g \ll m_q$, one obtains

$$\sigma \sim \ln\frac{E}{m_g} \ln\frac{E}{m_q}. \tag{20}$$

This regime is realized for heavy quarks, with masses larger than the QCD scale Λ. For light quarks one rather has to assume that $m_g \gg m_q$. In this case the dependence on m_q drops out and the cross section becomes

$$\sigma \sim \ln^2\frac{E}{m_g}. \tag{21}$$

This is precisely the behavior that we observed on several examples in Sect. 2. The two logarithms can easily be traced to the behavior

$$d\sigma \sim \frac{d\omega}{\omega} \cdot \frac{d\theta}{\theta} \tag{22}$$

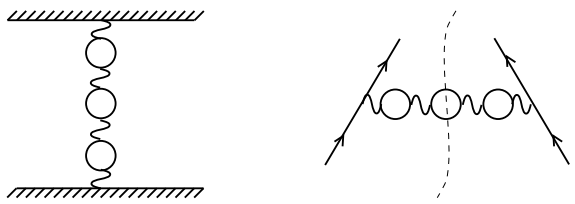
in the formal large-energy limit $E \to \infty$, $m_g \to 0$. In this limit, the integrated cross section formally diverges, with the divergences coming from the region $\theta \to 0$, that is emitting a gluon *collinear* with the parent quark, and $\omega \to 0$, that is emitting a *soft* gluon. This proves to be a general statement: all (infrared) divergencies of amplitudes in perturbation theory are due to emission of soft or collinear particles. Since the necessity to introduce an infrared regulator (here, the gluon mass) has to be interpreted as the intervention of long-distance interactions, this result is in agreement with our intuition that such interactions are only possible between soft or collinear particles.

In many practical QCD applications a logarithm of energy can be considered as a large parameter. In such cases the calculation can often be simplified by observing that such logarithms, as we just found, come from either soft or collinear gluon emission and the approximations in (16)–(19) can be applied.

3.2 Scale of the coupling

Apart from the two logarithms of energy, the probabilty of the gluon emission in Fig. 5 is proportional to the strong coupling $\alpha_s = g_s^2/(4\pi)$, cf. Eq. (15). As we know, the QCD coupling is scale-dependent, that is it depends on the distance at which it is measured. So, what is the relevant scale for the bremstrahlung process?

Let us consider a similar process for the photon bremsstrahlung in quantum electrodynamics which is simpler. The effective charge (scale-dependent coupling) in QED can be defined through the interaction between two very heavy charged plates. Going over from the 'bare' to the scale-dependent coupling corresponds to taking into account the vacuum polarization by virtual electron-positron pairs that is given by multiple insertions of 'bubbles' in the photon propagator see Fig. 6a. The corresponding potential in



a b

Fig. 6: The QED running coupling

momentum space is the usual Coulombic interaction with the effective coupling

$$\widetilde{V}(\vec{q}) = -\frac{4\pi\alpha(\vec{q})}{\vec{q}^2}. \tag{23}$$

Using $\alpha(\vec{q})$ in QED calculations is, therefore, a compact way to take into account a certain class of Feynman diagrams — multiple insertions of fermion bubbles in a single photon line.

For the gluon (well, so far photon) bremsstrahlung, the question about the scale of the coupling is, similarly, the question of taking into account contributions of more complicated diagrams corresponding to vacuum polarization and, what is crucial, photon splitting into the electron-positron pair in the final state as shown in Fig. 6b: the two parts of the diagram to the left and to the right of the dashed line symbolize a product of the initial and the final amplitudes and the phase space integration over the momenta of the electron and the positron in the final state is tacitly assumed. Notice that taking into account the contributions with photon splitting into the electron-positron pair in the final state we are, in effect, changing the question: instead of photon emission, we are, in effect, considering the emission of a *photon jet*!

The calculation of the amplitude with the e^+e^- pair in the final state is very similar to the calculation carried out in the previous section, with the only difference that the gluon mass is substituted by the invariant mass of the e^+e^- pair. Since electron-positron pair with different invariant mass can be produced, one has to integrate over all massses (or photon virtuality q^2, which is the same). Finally, using dispersion relations one can show that the sum of all such diagrams can be written as a simple one-photon emission with the effective coupling taken at the scale that is the *maximum possible* invariant mass of the e^+e^--pair (or photon jet mass, in general). For the double logarithmic regime corresponding to the soft and collinear photon emission the corresponding restriction is

$$q^2 < \omega^2\theta^2 = q_\perp^2 \tag{24}$$

(to see this, use (19) and put $m_q = 0$) where $q_\perp$ is the photon transverse momentum with respect to the parent quark. We conclude that *if* the photon emission is interpreted as photon jet emission, and *if* the energy is large so that the approximation by soft and collinear emission can be applied, then the relevant coupling has to be taken at the scale of the photon transverse momentum. The result in QCD is the same, although the derivation is more subtle. The difficulty is that the running coupling in QCD is not reduced to the effects of vacuum polarization insertions in the gluon propagator, but also receives contributions of vertex corrections and the quark self-energy (in QED these effects cancel thanks to the Ward identity).

For more sophisticated readers, I want to add some explanation of the above statement that the scale of the coupling is given by the *maximum possible* invariant mass of the jet. Because of unitarity, the product of the two photon lines (with bubble insertions) in the direct and final amplitudes in Fig. 6b is equal to the imaginary part of the polarization operator $\Pi(q^2)$ (3), integrated from $q^2 = 0$ to $q^2 = q^2_{\max} \simeq q^2_\perp$. This integral can be written as a contour integral in the complex q^2 plane as shown in Fig. 7a (cf. Fig. 2) and the integration contour can be moved away from the origin as in Fig. 7b without changing

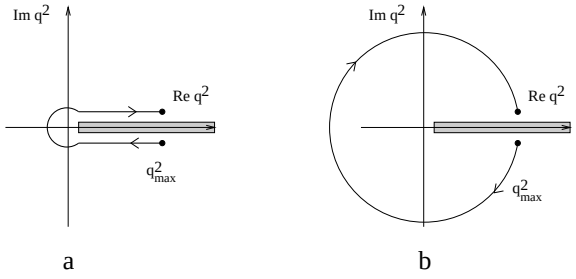


Fig. 7: From the imaginary part to the contour integral, see text

the value of the integral. As the result, the polarization operator is now being integrated over a cirle of the radius $q^2_{\max} \simeq q^2_\perp$ in the complex plane. Since $\Pi(q^2)$ is calculated in terms of $\alpha(|q^2|)$ (we have not discussed this, but it is a common knowledge), it follows that the bremsstrahlung cross section can be expressed in terms of the coupling at the scale $q^2_\perp$, as anticipated.

3.3 Angular ordering

As known from classical electrodynamics, radiation from a charged particle is proportional to the acceleration — we need to struck the quark in order that it starts to radiate gluons. Consider the simplest process shown in Fig. 8. A free quark (on mass-shell) gets struck by a virtual photon with large momentum q so it (abruptly) changes the direction of motion. Gluons can be radiated both from the initial and the final state; we have learnt that this emission is mostly soft and collinear, so we would expect to see two narrow radiation cones along the initial and the final quark directions. However, what about the interference? Each amplitude for the gluon emission has the form (15) and it is easy to show that the contributions of the initial and the final state radiation have opposite signs. Neglecting all masses

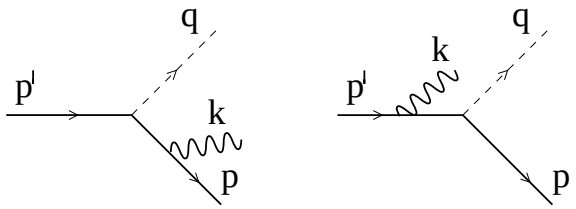


Fig. 8: Gluon emission from the initial vs. final state

and summing up contributions of all gluon polarizations $\sum_\lambda e_\mu^{(\lambda)} e_\nu^{*(\lambda)} \sim -g_{\mu\nu}$, we get a very simple expression

$$\sum_\lambda \left| \frac{e^{(\lambda)} p}{pk} - \frac{e^{(\lambda)} p'}{p'k} \right|^2 = \frac{2pp'}{(pk)(p'k)} \tag{25}$$

so that it is *only* the interference term that survives! A useful observation that has become known as 'angular ordering' is that the classical picture of the emission can, nevertheless, be restored for the emission cross section, integrated in some special way over the azimuthal angle. It proves convenient to introduce unit vectors in the directions of the participating momenta:

$$n_\mu = \frac{p_\mu}{p_0}, \qquad n'_\mu = \frac{p'_\mu}{p'_0}, \qquad (n_k)_\mu = \frac{k_\mu}{k_0}, \tag{26}$$

so that $n_\mu = (1, \vec{n}),\ |\vec{n}|^2 = 1$, etc. The differential cross section is then given by

$$\frac{2pp'}{(pk)(p'k)} \frac{d^3k}{2k_0} = \frac{2nn'}{(nn_k)(n'n_k)} \frac{d\omega}{2\omega}\, d\Omega_k, \tag{27}$$

where $d\Omega_k$ is the infinitesimal solid angle for the gluon emission. Write

$$\begin{aligned} \frac{2nn'}{(nn_k)(n'n_k)} &= \frac{2}{nn_k} + \frac{2}{n'n_k} - 2\,\frac{nn_k + n'n_k - nn'}{(nn_k)(n'n_k)} \\ &= \left(\frac{2}{nn_k} - \frac{nn_k + n'n_k - nn'}{(nn_k)(n'n_k)} \right) + \left(\frac{2}{n'n_k} - \frac{nn_k + n'n_k - nn'}{(nn_k)(n'n_k)} \right). \end{aligned} \tag{28}$$

The three terms in the first line of this equation can be interpreted as, loosely speaking, the would-be emission from the final and initial quarks and the interference, respectively. In the second line we have regrouped them in two contributions, with half of the interference term subtracted in each term. Now comes the basic idea: consider emission probability integrated over the azimuthal angle φ measured with respect to the parent quark momentum, that is *different* for the two contributions in (28). The calculation is simple, so I reproduce it in full detail in order that the result does not look as a miracle.

For the first term in the second line in (28) we choose the geometry as shown in Fig. 9 and expand the scalar products in terms of the corresponding angles

$$n'n_k = 1 - (\cos\theta_{n'n} \cos\theta_{nn_k} - \sin\theta_{n'n} \sin\theta_{nn_k} \cos\varphi), \qquad nn_k = 1 - \cos\theta_{nn_k}\,. \tag{29}$$

We will need the following integral

$$\int_0^{2\pi} \frac{d\varphi}{2\pi}\, \frac{1}{a - b\cos\varphi} = \frac{1}{\sqrt{a^2 - b^2}} \tag{30}$$

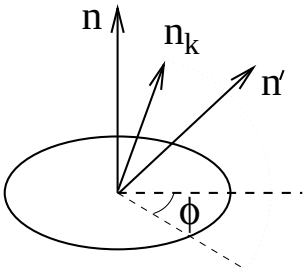


Fig. 9: The azimuthal integration

with $a = 1 - \cos\theta_{n'n}\cos\theta_{nn_k}$ and $b = \sin\theta_{n'n}\sin\theta_{nn_k}$ so that $\sqrt{a^2-b^2} = |\cos\theta_{n'n} - \cos\theta_{nn_k}| = |n'n - nn_k|$. Using these expressions, we easily obtain

$$\begin{aligned}\int_0^{2\pi}\frac{d\varphi}{2\pi}\left[\frac{2}{nn_k} - \frac{nn_k + n'n_k - nn'}{(nn_k)(n'n_k)}\right] &= \frac{2}{nn_k} - \frac{nn_k - nn' + |nn' - nn_k|}{(nn_k)|nn' - nn_k|} \\ &= \frac{2}{nn_k} - \frac{2}{nn_k}\theta(nn_k - nn') \\ &= \frac{2}{nn_k}\theta(nn' - nn_k), \end{aligned} \tag{31}$$

which is the desired result. The effect of the interference appears to be, after azimuthal integration, to forbid the gluon radiation at the angles larger than the angle between the directions of the initial and the final quark. The situation with the second term in (28) is exactly the same, but the azimuthal integration has to be done with respect to the n'-axis. We conclude that if, by some reason, the azimuthal dependence of the radiation is not important, the correct quantum-mechanical answer can be reproduced by the independent (classical) radiation from the initial and the final quark, with the restriction on the emission angle.

This result is very interesting because it suggests that under certain conditions the emission of gluons is independent of the 'history' of the process: when several gluons are emitted successfully, the only way how the latter ones 'know' about the previous emissions is through the kinematical condition of the angular ordering. A detailed analysis shows that the calculation of mutiple gluon emission using angular ordering indeed reproduces the major part of corrections that are beyond the double-logarithmic approximation.

This is the basis of parton shower models:

- Probabilistic picture of the sequence of gluon emissions
- Emission angles are strictly ordered
- Used $\alpha(k_\perp)$ for each emission
- Plus a lot of machinery!

Parton shower models proved to be very successful in the description of gluon radiation. They are usually combined with simple hadronization models and reproduce a rather detailed structure of the final hadronic state. Because of this, they are indispensable for planning the experiments, estimating the backgrounds etc.

3.4 String effect

Up to this point, our discussion of gluon radiation was in fact not specific for QCD and the results apply to any theory with massless vector particles, and QED in particular. The major difference between QED and QCD is that a gluon, in difference to a photon, has a (color) charge. The gluon itself is a source of

bremstrahlung (and in fact stronger than that of a quark because a gluon has a larger charge) while in QED the photon radiation from a photon has to be mediated by the production of an electron-positron pair and is small. Experimentally, this difference is seen most directly through the so-called string effect. Consider the three-jet events in e^+e^- annihilation and compare the particle flow in between the jet directions in the two cases when the gluon jet is substituted by a photon jet (in the same kinematics), see Fig. 10. In the case of a gluon jet there will obviously be additional gluon radiation from the gluon, and

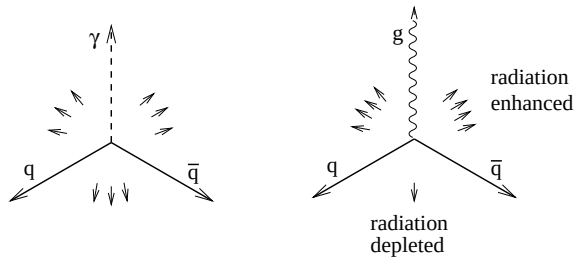


Fig. 10: The gluon emission in three-jet events: the quark, antiquark and photon jets (left) compared to the quark, antiquark and gluon jets (right)

this radiation will interfere quantum-mechanically with the radiation from quarks. The corresponding calculation shows that the interference is constructive in the angular region between the quark and the gluon jet, and the interference is destructive between the two quark jets. As the result, one expects to have more soft gluons (and therefore pions) in between the quark and the gluon jet compared to the quark and the photon jet, and less gluons (pions) in the direction opposite to the gluon jet (compared to the photon).

In order to get an idea how this happens, consider the toy-model example illustrated in Fig. 11. In

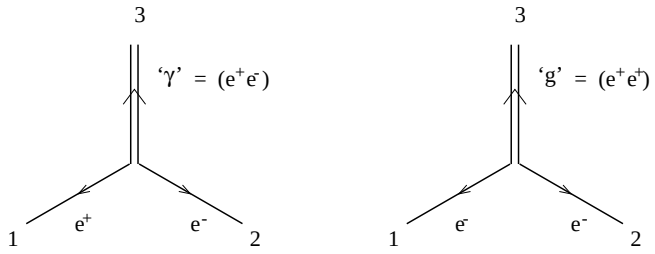


Fig. 11: The string effect in a toy-model

the left figure we have an electron jet, a positron jet and a 'photon' jet made of a e^+e^- pair. In the right figure we see two electron jets and a 'gluon' jet made of two positrons. The differential cross section is given in the first case by, schematically

$$d\sigma_{q\bar{q}\gamma} \sim \frac{d\omega}{\omega}\, d\Omega_k\, (\widehat{1\,2}) \tag{32}$$

where $(\widehat{1\,2})$ stands for the emission amplitude from the jets 1 and 2 (the numeration is according to Fig. 11), cf. Eq. (27). In the second case, the radiation cross section can be calculated as

$$d\sigma_{q\bar{q}g} \sim \frac{d\omega}{\omega}\, d\Omega_k \left[2(\widehat{1\,3}) + 2(\widehat{2\,3}) - (\widehat{1\,2})\right]. \tag{33}$$

The factors 2 in the first two terms reflect the larger 'gluon' charge and the negative sign in the third term is there because the interaction between the two electron jets is repulsive. Working out the amplitudes

and using the angular ordering condition (31), it is easy to show that the probability of emission in the direction opposite to the (e^+e^+)-pair vanishes. I leave this exercise to the reader.

The string effect is indeed observed, see Fig. 12. Note that for the gluon jet there are *less* particles.

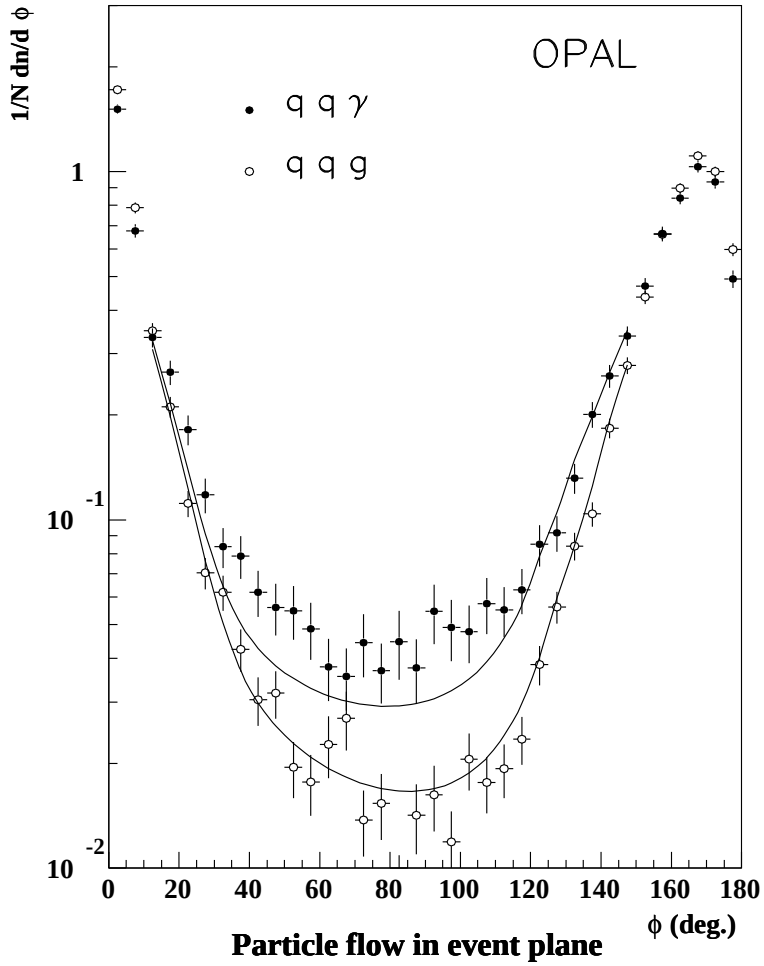


Fig. 12: The interjet hadron flow projected into the plane of a $q\bar{q}g$ or $q\bar{q}\gamma$ event as a function of the angle between the quark jet and the hadron [7].

Historically, the observation of the string effect was very important. It has given a crushing blow to old models of independent jet fragmentation (suggested and worked out by Feynman and Field), and gave strong evidence supporting a more modern picture of string fragmentation by the Lund group. The original idea behind the string effect was to explain nonperturbative hadron production by breaking of the colored string connecting the quark and the antiquark (in this picture emission of a gluon corresponds to a cusp in the string). The explanation of the string effect as due to quantum interference was invented only later.

4 FACTORIZATION

All our discussion up to this point assumed that hadron production (and all long-distance effects) can be dispensed off using a completeness condition (unitarity). Roughly speaking, the idea is that although we do not know how the quarks become hadrons, we do know that this always happens. There exists a more strong version of this statement which is sometimes called a 'local parton-hadron duality' (LPHD): the momentum distributions of hadrons more or less follow that of the quarks and gluons. The theoretical

status of LPHD is unclear.

In this section we will discuss an even broader class of reactions for which the long-distance effects do not cancel — and our previous argumentation does not apply — but one is able to isolate (factorize) their effect in certain phenomenological functions, of which the celebrated quark and gluon parton distributions in the nucleon provide the prime example.

The driving principle of factorization is the separation of scales: for a given observable that involves a 'hard' scale Q one tries to invent a representation as a product (convolution, in general) of two (or more) functions that take into account the effects of small and large distances, respectively. Schematically one can write

$$\begin{aligned}&\text{Physical quantity}\,(Q^2, \text{hadron momenta}) = \\ &= \mathcal{C}(Q^2, \text{parton momenta}\,|k|>\mu) \otimes \mathcal{P}(\text{parton momenta}\,|k|<\mu, \text{hadron momenta}).\end{aligned} \tag{34}$$

The 'hard' function $\mathcal{C}$ describes the evolution of the system at small distances and does not 'know' about hadrons, while the 'soft' function $\mathcal{P}$ stands for the parton transition to hadrons. By construction, $\mathcal{P}$ only involves partons with momenta less than the factorization scale μ and does not depend on Q. 'Factorization' means exactly this: the dependence on the hard scale Q is isolated (factorized) from the dependence on soft (hadronic) momenta at the cost of introducing an auxilary separation scale.

Before going over to QCD, let us illustrate this idea on a simple example. Define

$$I(Q^2, m^2) = Q^2 \int_0^1 dx\, \frac{x(1-x)}{x^2Q^2 + (1-x)m^2}. \tag{35}$$

I am going to use the technique of factorization to calculate this integral in the limit $Q^2 \gg m^2$. To this end, define the separation scale

$$m \ll \mu \ll Q \tag{36}$$

and split the integral in two parts

$$I(Q^2, m^2) = Q^2 \int_{\mu/Q}^1 dx\, \frac{x(1-x)}{x^2Q^2 + (1-x)m^2} + Q^2 \int_0^{\mu/Q} dx\, \frac{x(1-x)}{x^2Q^2 + (1-x)m^2}. \tag{37}$$

In the first integral we can safely put $m^2 \to 0$ after which the integral becomes trivial and we obtain

$$I_1(Q^2, \mu^2) = \int_{\mu/Q}^1 dx\, \frac{(1-x)}{x} = \ln\frac{Q}{\mu} - 1\,. \tag{38}$$

In the second integral in order to avoid the singularity at $x \to 0$ one has to keep m^2 finite, but still can simplify the integral replacing $1 - x \to 1$. The result is

$$I_2(\mu^2, m^2) = Q^2 \int_0^{\mu/Q} dx\, \frac{x}{x^2Q^2 + m^2} = \ln\frac{\mu}{m} + \mathcal{O}(m/\mu)\,. \tag{39}$$

Collecting the both pieces, we obtain the final result

$$I(Q^2, m^2) = I_1(Q^2, \mu^2) + I_2(\mu^2, m^2) = \ln\frac{Q}{m} - 1 \tag{40}$$

up to corrections that fall down as powers of m/Q.

With some imagination one can interpret x as a gluon momentum fraction (in terms of Q) so that $I_1(Q^2, \mu^2)$ is the hard 'coefficient function' in the sense of Eq. (34) and $I_2(\mu^2, m^2)$ is the soft 'parton distribution'. In our example both functions are, of course, calculable. In QCD, we would say that $I_1 \sim \mathcal{C}$ is calculable since it comes from large momenta, while $I_2 \sim \mathcal{P}$ is not accessible in perturbation

theory and has to be considered as a phenomenological input (or calculated using some nonperturbative method). Notice that since the dependence on the scale separation parameter μ must cancel in the sum, the scale dependence of $I_2 \sim \mathcal{P}$ *is calculable* as it has to coincide (with the opposite sign) with that of the coefficient function.

What do we achieve if factorization is possible?

- The Q^2 dependence of the observable is calculable since it comes entirely from the hard coefficient function, cf. the above example.
- It usually happens so that the same soft functions $\mathcal{P}$ enter different physical processes; in this case one can use one of them in order to extract the function $\mathcal{P}$ from experiment and later use it in other processes. This property is sometimes referred to as 'universality'. E.g. quark and gluon parton distributions are universal in the sense that they appear in the QCD description of many hard reactions.
- As we have seen on our simplistic example already, factorization brings in considerable technical advantages. A very important observation is that in order to conserve dimensions, the coefficient function $\mathcal{C}(Q^2, \mu^2)$ must depend, in fact, on the single dimensionless ratio Q^2/μ^2, i.e. $\mathcal{C}(Q^2, \mu^2) \equiv \mathcal{C}(Q^2/\mu^2)$. This means that in order to learn about the Q^2 dependence it is enough to study the μ dependence. This is in many cases simpler because the scale separation can be done in a way that is convenient for the calculation (e.g. use dimensional regularization). Last but not least, one can use the powerful machinery of the renormalization group.

4.1 Deep inelastic scattering and parton sum rules

A canonical example of factorization in QCD is provided by the deep inelastic lepton-nucleon scattering (DIS), see Fig. 13. The quantity of interest is the total cross section, measured as a function of the

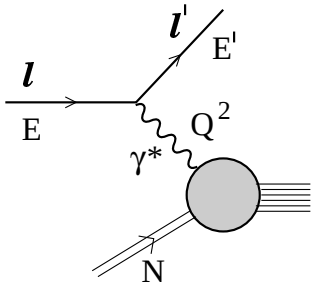
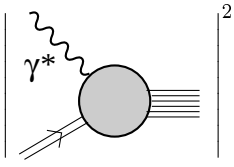
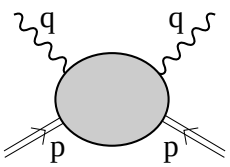


Fig. 13: The deep inelastic lepton-nucleon scattering

lepton energy and the scattering angle. The part of the amplitude involving the leptons can easily be calculated and dispensed off. One is left with a cross section of virtual photon scattering from a nucleon which is parametrized in terms of the two structure functions (for unpolarized targets) $F_1(x_B, Q^2)$ and $F_2(x_B, Q^2)$, where $Q^2 = -q^2$ is the photon virtuality and $x_B = Q^2/(2pq)$ is the Bjorken variable; p is the nucleon momentum.

As the first step, we make use of the completeness condition (unitarity) that allows to rewrite the total cross section as imaginary part of the forward Compton scattering amplitude, schematically

$$\sum_{\text{final states}}$$  $$= \mathbf{Im}$$  , (41)

where the formal definition of the Compton amplitude in the r.h.s. is

$$T_{\mu\nu}(p,q) = i \int d^4x\, e^{iqx} \langle N(p)|T\{j^{\rm em}_\mu(x) j^{\rm em}_\nu(0)\}|N(p)\rangle, \tag{42}$$

cf. (2) and our discussion of the e^+e^- annihilation. In difference to the latter case, the virtuality of the photon is now negative $q^2 = -Q^2 < 0$ but this does not make life easier since the invariant energy $s = (p+q)^2 \simeq 2pq - Q^2 > 0$ is positive. As the result, the forward Compton amplitude (42) at large Q^2 still contains contributions of large distances and cannot be evaluated directly.

The second step is, therefore, to make use of the analytic properties of the Compton amplitude in the complex energy plane. It proves to be convenient to introduce the variable

$$\omega = 2pq/Q^2 = 1 + s/Q^2\,, \qquad s = (p+q)^2\,, \tag{43}$$

which is nothing but the invariant energy in units of Q^2. The analytic structure of the Compton amplitude in the complex ω plane is shown in Fig. 14. It has two cuts along the real axis stretching from $\omega = \pm 1$

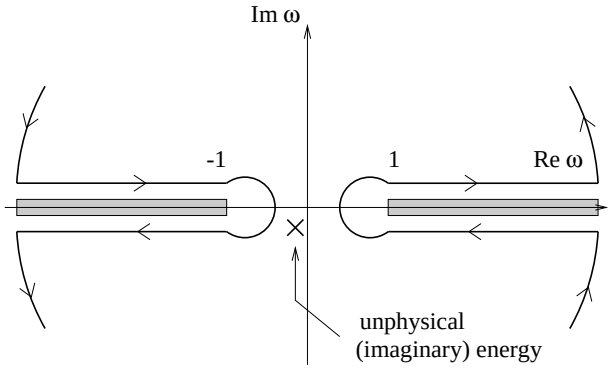


Fig. 14: Analytic structure of the Compton amplitude

to infinity. The cut for positive ω corresponds to the physical region of the deep inelastic scattering and the corresponding discontinuity determines the DIS cross section. The cut for negative ω corresponds to the physical region of the so-called 'u-channel' process with the ingoing and the outgoing photon lines in the r.h.s. of Eq. (41) interchanged. Using the integration contour shown in Fig. 14 we can calculate the Compton amplitude $T(\omega, Q^2)$ at $|\omega| < 1$ as

$$T(\omega, Q^2) = \int\limits_1^\infty \frac{d\omega'}{\omega' - \omega}\, F(\omega, Q^2) + \text{ contribution of the u-channel cut} \tag{44}$$

where $F(\omega) = \mathrm{Im}\, T(\omega)$ is the DIS cross section (structure function), cf. Eq. (4). (In the present case we do not need to take the derivative since it can be argued that the integral over the large circle vanishes.) The contribution of the 'u-channel' process is a nuisance since it corresponds to a different physical process that is not measured in DIS experiments. Luckily enough, it can be dispensed off using the so-called crossing symmetry of the Compton amplitude: $T(\omega + i\epsilon, Q^2) = T(-\omega - i\epsilon)$; as the result of this symmetry the effect of the crossing channel in the dispersion relation is, roughly speaking, to guarantee that $T(\omega)$ is an even function and to give a factor of 2.

Note that the choice $\omega \ll 1$ corresponds according to (43) to a large and negative c.m. energy squared $s \sim -Q^2$ and, therefore, to an unphysical large imaginary energy in the photon-nucleon collision. A particularly useful form of Eq. (44) is obtained by the Tailor expansion of both sides in powers of ω:

$$T(\omega, Q^2) = \sum_{n=0}^\infty \omega^n T_n(Q^2)\,, \qquad \frac{1}{\omega' - \omega} = \frac{1}{\omega'} \sum_{n=0}^\infty \left(\frac{\omega}{\omega'}\right)^n. \tag{45}$$

To be precise, only even values of n are allowed in the first expansion because of the crossing symmetry. Comparing the coefficients with the same powers of ω on both sides and introducing the Bjorken variable $x_B = 1/\omega$ we obtain (for even n)

$$T_n(Q^2) = \int\limits_1^\infty \frac{d\omega'}{\omega'^{n+1}} F(\omega', Q^2) \equiv \int\limits_0^1 dx_B\, x_B^{n-1}\, F(x_B, Q^2)\,. \tag{46}$$

Eq. (46) states that (even) moments of the DIS structure function are given by the coefficients T_n in the expansion of the forward Compton amplitude in the unphysical region $\omega \to 0$.

Similar to what happens for the polarization operator (3), large imaginary energy translates to the requirement that the space-time points where the initial and the final state photon gets coupled to the hadronic (quark and gluon) system are close to each other: The integral in (42) is dominated by the region $|x| \sim 1/Q$. The difference with the e^+e^- annihilation is, however, that the process happens to occur inside a nucleon. For the polarization operator, all quarks and gluons produced by the initial photon have to be annihilated at a small distance. Therefore, they cannot get out of the small space-time region of the size of the order of $1/|q|$. In deep inelastic scattering, the virtual photon collides with the quarks that already exist in the proton and their wave function had a long time to develop. As a consequence, the process is not 'confined' (kinematically) to small distances.

A typical contribution to the forward Compton amplitude looks as shown in Fig. 15a. While the

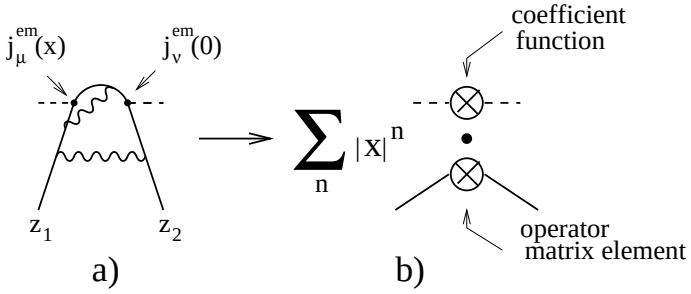


Fig. 15: The short-distance (operator product) expansion

distance between the two electromagnetic current attachments is small, $|x| \sim 1/Q$, the quark coordinates z_1, z_2 are not restricted and can be arbitrarily large. It can be shown (this is easy but requires some writing) that the expansion at small ω in Eq. (45), (46) corresponds to the expansion of the amplitude in coordinate space in powers of x, the separation between the currents. The result is presented schematically in Fig. 15b. The difficulty is that each loop correction typically brings in a factor $\ln x^2p^2$, where p^2 is the quark virtuality so that for real quarks the result diverges, signalling intervention of large distances. The amplitude is, therefore, not infrared-safe.

The way out is given by factorization. We introduce the scale separation parameter $|p| \ll \mu \ll Q$ and divide the contribution of any given Feynman diagram in two parts, corresponding to short and long distances. The short distance part (the coefficient function) contains, by construction, contributions of energetic gluons with momenta $|k| > \mu$[1]. It can be interpreted as the n-th moment of the cross section of a virtual photon scattering off a quark with virtuality $p^2 = \mu^2$. The contribution of long distances corresponding to small loop momenta $|k| < \mu$ can formally be written as the nucleon matrix element of a certain local operator built of quark fields and their derivatives; it contains the information about the quark/gluon wave function of the nucleon that cannot be obtained in QCD perturbation theory. In technical terms, the idea is to write each logarithm $\ln x^2p^2$ as a sum of two: $\ln x^2p^2 = \ln x^2\mu^2 + \ln p^2/\mu^2$

[1] To be precise, one has first to rotate the integration contour over the gluon momentum to the Euclidian space and only then impose the cutoff.

and then replace the second logarithm by a nonperturbative parameter, in full analogy with the simplistic example at the beginning of this Section.

As already discussed, the main premium of the factorization is that one is able to calculate the dependence of the cross section on Q^2. In a few cases it happens, however, that the relevant nonperturbative parameters are known exactly as they are given by matrix elements of conserved currents. For example, the number of quarks minus the number of antiquarks in the nucleon is a constant that cannot be changed by any long-distance interactions (e.g. because quarks and antiquarks are produced in pairs). These exceptional cases give rize to the so-called parton sum rules that are among the gold-plated QCD tests. Because of the solid theoretical background, DIS sum rules are ideally suited (in principle, at least) to extract the single fundamental parameter of QCD — the strong coupling constant.

One example is provided by the Bjorken sum rule for the polarized deep inelastic scattering that relates the integral of the difference between the structure functions $g_1(x_B, Q^2)$ of the proton and the neutron to the axial nucleon charge g_A that is known from the neutron β-decay:

$$\int\limits_0^1 dx_B\, g_1^{p-n}(x_B, Q^2) = \frac{1}{6}|g_A| \left[1 - \frac{\alpha_s}{\pi} - 3.58\left(\frac{\alpha_s}{\pi}\right)^2 - 20.22\left(\frac{\alpha_s}{\pi}\right)^3 + \ldots\right], \tag{47}$$

where $\alpha_s = \alpha_s(Q)$. The perturbative series corresponds to the coefficient function and $|g_A|$ is the nucleon matrix element of the axial current. The experimental value for the Bjorken integral in the l.h.s. is $0.176 \pm 0.003 \pm 0.007$ [8] at $Q^2 = 5$ GeV^2. It compares very well with the theoretical calculation using the world average for α_s, cf. the compilation in Fig. 16.

Another important example is provided by the Gross–Llewellyn Smith sum rule that relates the integral of the sum of the structure functions $F_3(x_B, Q^2)$ for neutrino-proton and antineutrino-proton scattering to the number of valence quarks (three) in the proton:

$$\int\limits_0^1 dx_B\, F_3^{\nu p+\bar\nu p}(x_B, Q^2) = 3\left[1 - \frac{\alpha_s}{\pi} - 3.58\left(\frac{\alpha_s}{\pi}\right)^2 - 18.97\left(\frac{\alpha_s}{\pi}\right)^3 + \ldots\right]. \tag{48}$$

Measurement of the GLS integral can be used to extract the value of α_s. Rather accurate data exist at the relative low $Q^2 \sim 3$ GeV^2 [9]. Splitting them in six Q^2-bins and making the global fit, CCFR/NuTev extract the value of the strong coupling at the scale of the Z-boson mass: $\alpha_s(M_Z) = 0.114 \pm 0.005 \pm 0.008 \pm 0.005$ where the first error is statistical, the second is due to the systematics and the third corresponds to the theoretical uncertainty of the higher twist effects.

A recent compilation of the α_s measurements [10] is shown in Fig. 16, where the second and the third entry from above correspond to the Bjorken and the GLS sum rules, respectively. It attracts attention that the accuracy of these determinations is not better than the others, in particular the determination from the τ-lepton hadronic decays (in the fourth line) seemingly has a much smaller error.[2]

The reason for this is that in practice one has to keep balance between the theoretical rigour and the experimental (and calculational) possibilities. One notorious problem with the DIS sum rules is that they require to take the integral over the whole x_B range, and in practice this means that experimental data have to be extrapolated both to $x_B \to 0$ and $x_B \to 1$. The other problem is that although the structure of nonperturbative power suppressed corrections $1/Q^2$ to the DIS sum rules is well *understood* (we will discuss such corrections in the last Section), these corrections appear to be rather large and up to now there exist only crude estimates of them. In contrast to this, the *known* nonperturbative corrections to τ-decays are small and the experimental situation is easier, hence the smaller error.

[2] I have no time to discuss the τ-decays in these Lectures. The following remarks are general and apply to any α_s extraction that is using perturbative factorization (a space-time picture discussed in Sec. 2.2) as opposed to the operator product expansion and dispersion relations.

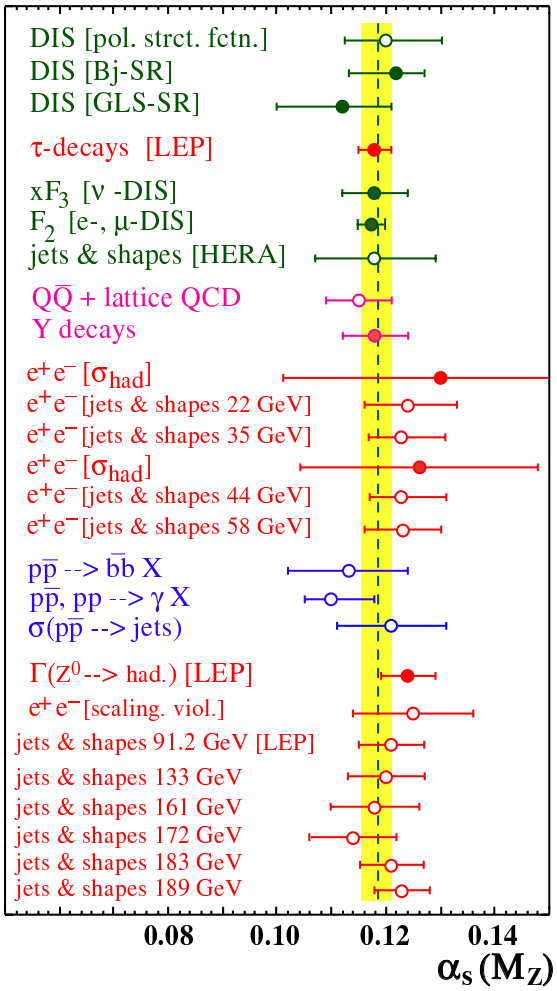


Fig. 16: Summary of $\alpha(M_Z)$ [10]. Filled symbols represent results based on complete NNLO QCD.

It has to be stressed, however, that the α_s determinations from τ-decays could be affected by the effects of duality violation in a way that the present theory cannot predict. A very good agreement with the world average has to be interpreted as an experimental verification of the high accuracy of duality. In other words, if the α_s extracted from τ decays appeared to be very much off compared to other measurements, this would be interpreted as violation of duality (and no one would be surprised). At the same time, if the α_s extracted from DIS sum rules appeared to be in appaling contradiction with the other data, this would create a serious problem. By saying that, I do not want to discredit most of the α_s measurements in Fig. 16, but rather emphasize that one has to be aware of the existing problems and keep an open eye.

4.2 The parton model

The derivation of parton sum rules uses the analytic properties of the Compton scattering amplitude in an essential way: going over to imaginary energies ($\omega \to 0$) is crucial to ensure that the integral in (42) is dominated by short distances $|x| \to 0$. The price to pay is that the information about the space-time development of the DIS process is lost: in Euclidian space, particles do not 'propagate', they just 'occur'. The actual space-time picture of the deep-inelastic scattering depends on the frame of reference. This picture becomes particularly intuitive if the frame of reference is chosen such that the nucleon has a large energy, the reason being that in such a case it can be described as a superposition of Fock states in terms of quarks and gluons. We cannot calculate the structure of these states but assume that interactions of 'free' partons are given by the QCD perturbation theory.

Consider the ep scattering in the c.m. frame, Fig. 17. At large energies, the nucleon gets contracted

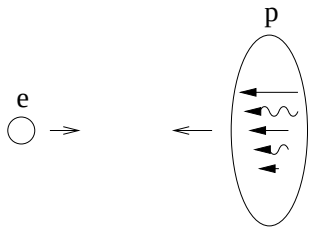


Fig. 17: The ep scattering in the c.m. frame

in the longitudinal direction and the interactions between the partons inside the nucleon get slowed down. As the result, the lifetime of any virtual parton state is lengthened and the time it takes the electron to traverse the hadron is shortened. These simple consequences of the special relativity have profound implications:

- Since the partons have no time to interact, each of them can be considered as 'free' and, in particular, having a fixed momentum fraction 'x' of the proton.
- If the momentum transfer is high, the virtual photon is 'heavy' and cannot travel far. Thus the electron will be able to interact with a single parton only, unless the parton density is very large.
- Final state interactions occur at time scales too long to interfere with the hard photon scattering. The scattering becomes classical and incoherent and may be computed by combining probabilities rather than amplitides. For the total cross section, the final state interactions can be dispensed off using the completeness condition.

The overall picture that arises in this way has become known as the QCD parton model. In the parton model, the deep inelastic cross section is given by the convolution

$$\sigma_{eN}(x,Q^2) = \sum_p \int_x^1 \frac{dy}{y}\, f_{p/N}(y)\, \sigma_{ep}(x/y,Q^2)\,, \tag{49}$$

where $x = x_B$, the summation goes over all sorts of partons (quarks, antiquarks and gluons), $\sigma_{ep}(\xi, Q^2)$ is the deep inelastic cross section of the electron scattering from a (free) parton p that carries the momentum fraction ξ, and $f_{p/N}(y)$ is the probability distribution to find the parton p in the nucleon with the momentum fraction y. In QCD, the cross section $\sigma_{ep}(\xi, Q^2)$ is infrared divergent because of the emission of collinear gluons. Interpretation of this divergence is rather obvious since collinear gluons require a long time for their emission and have to be included, therefore, in the wave function of the initial parton state alias the parton distribution function $f_{p/N}(y)$. In order to decide which gluon is hard and which is soft, one is forced to introduce the separation scale μ so that both the hard cross section and the parton distribution are scale-dependent. The scale-dependence can be calculated in perturbation theory and is governed by the celebrated Dokshitzer-Gribov-Lipatov-Altarelli-Parisi evolution equations that can be found in any QCD textbook.

In order to make the connection with the operator product expansion in the previous Section one has to take moments. Eq. (49) becomes

$$\int_0^1 dx\, x^{n-1} \sigma_{eN}(x, Q^2) = \sum_p \left[\int_0^1 d\xi\, \xi^{n-1}\, \sigma_{ep}(\xi, Q^2) \right] \cdot \left[\int_0^1 dy\, y^{n-1}\, f_{p/N}(y) \right] \tag{50}$$

so that moments of the hard cross section and the parton distribution can be identified with the coefficient functions and the matrix elements of local operators. Notice that the dispersion relation technique combined with the operator product expansion predicts moments of the deep inelastic structure function, while the parton model allows one to calculate the structure functions themselves, in precise analogy to the analysis of the total cross section in e^+e^- annihilation. Although the two procedures are equivalent from the practical (calculational) point of view, the parton model predictions are not protected against possible duality violation effects, see Sect. 2.3. Technically speaking, the difference is again in the direction in which the analytic continuation is performed: The predictions for the moments of structure functions are based on the analytical continuation of the physical cross section to the unphysical Euclidian region, where they are matched to the operator product expansion of the T-product of the two electromagnetic currents at small distances. The parton model predictions for the structure functions themselves are obtained by the analytic continuation of the calculations at small distances (in the unphysical domain $\omega \to 0$) to the physical cut. An interesting detail: whereas the crossing symmetry was crucial in order that the dispersion relations work, it plays no rôle in the parton model argumentation.

The power of the parton model is that it extends beyond the deep inelastic scattering and can be applied to a broad variety of reactions. In particular, we can replace the electron in Fig. 17 by another nucleon, and consider the process in which two partons belonging to the colliding nucleons annihilate and produce a $\mu^+\mu^-$ pair with a large invariant mass Q^2, see Fig. 18. This is called the Drell-Yan

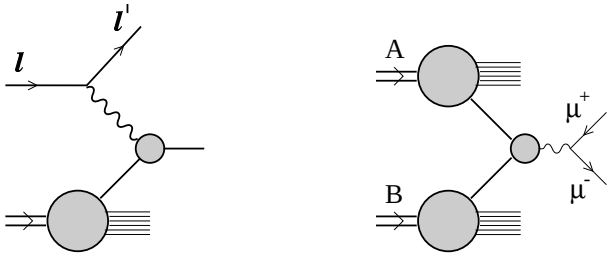


Fig. 18: The deep inelastic scattering and Drell-Yan pair production in the parton model

process, and the corresponding cross section can be calculated as

$$\frac{d\sigma}{dQ^2 dy} = \sum_p \int_{x_A}^1 \frac{dy_A}{y_A} \int_{x_B}^1 \frac{dy_B}{y_B} f_{p/A}(y_A) f_{\bar{p}/B}(y_B)\, \sigma_{p\bar{p}\to\mu^+\mu^- + X}(x_A/y_A, x_B/y_B, Q^2)\,, \tag{51}$$

where y is the rapidity of the muon pair, $x_A = e^y\sqrt{Q^2/s}$, $x_A = e^{-y}\sqrt{Q^2/s}$ and s is the total energy.

The crucial statement of the parton model is that the parton distributions $f_{p/N}(x)$ are universal and are the same in DIS and the Drell-Yan production. This universality is a direct consequence of the Lorentz contraction. Indeed, without the Lorentz contraction, partons belonging to the colliding hadrons would overlap a finite time before the hard scattering, and the initial-state interactions would modify their distributions. What I want to demonstrate now is that the Lorentz contraction is not at all 'selfobvious' in a gauge field theory![3]

To start with, consider a classical scalar field produced by a heavy charged particle located at the origin:

$$\phi_{\rm cl}(\vec{x}) = \frac{e}{|\vec{x}|}. \tag{52}$$

The same field in a frame where the charged particle is moving at the velocity βc along the third axis is given by

$$\phi'_{\rm cl}(x') = \frac{e}{[x_\perp^2 + \gamma^2(\beta ct' - x'_3)^2]^{1/2}}, \tag{53}$$

where $\gamma = (1-\beta^2)^{-1/2} \gg 1$. For an observer in the primed system, the field is concentrated in a narrow strip in the direction of motion: $|\beta ct' - x'_3| \sim 1/\gamma$ and has a finite tranverse extension. In other words, the field is indeed Lorentz-contracted. This means that if such a scalar particle is approached by another one at nearly the speed of light, they affect each other during a very short period of time, just as supposed in the parton model.

Let us make the same exercise with the vector field (e.g. like in classical electrodynamics). The field potential in the rest frame of the charge looks very similar

$$A^\mu_{\rm cl}(\vec{x}) = \frac{e}{|\vec{x}|}\delta_{\mu 0}, \tag{54}$$

apart from the fact that it presents a time-like component of a four-vector rather than a scalar. Because of this, the field of the moving charge looks, however, very differently:

$$\begin{aligned} A'^{\,0}_{\rm cl}(x') &= \frac{e\gamma}{[x_\perp^2 + \gamma^2(\beta ct' - x'_3)^2]^{1/2}}, \\ A'^{\,3}_{\rm cl}(x') &= \frac{-e\beta\gamma}{[x_\perp^2 + \gamma^2(\beta ct' - x'_3)^2]^{1/2}}, \\ A'^{\,\perp}_{\rm cl}(x') &= 0. \end{aligned} \tag{55}$$

For large γ, the field components in the zero and the third direction have finite extension and are not contracted at all! The situation is saved by the observation that the force experienced by the test charge in the primed frame is in fact proportional to the field strength rather than the vector potential itself. If we look at the field strength transformations rather than those of the vector potential, we find a much better situation. The electric field in the direction of motion, for instance, is given for the moving charge by

$$E'^{\,3}_{\rm cl}(x') = \frac{-e\gamma(\beta ct' - x'_3)}{[x_\perp^2 + \gamma^2(\beta ct' - x'_3)^2]^{3/2}}, \tag{56}$$

which shows a γ^{-2} falloff. Another way to understand what actually happens is to notice that, as $\gamma \to \infty$, the vector potential approaches the total derivative $A^\mu(x') \sim e\partial^\mu \ln(\beta ct' - x'_3)$ so that it can be eliminated by a suitable gauge transformation.

We conclude that factorization in a gauge theory like QCD is a nontrivial issue. The parton model *picture* of the interaction is valid, at best, in a certain gauge (if we suceed to find one such that the

[3] The following example is taken from the book [1].

potential is Lorentz contracted). Since the gauge invariance is involved in an essential way, one has to suspect (and this is true) that the factorization is only valid in the sum of Feynman diagrams to given order, but not for a given diagram separately. Full-scale factorization proofs are normally rather lengthy and have only been completed in a few cases. A few cases are known where the 'naive' factorization proves to be broken, and there exist many situations where the answer is not known. In fact, this is an active field of research.

4.3 New developments

Now that we have learnt the basic principles, I want to spend some time to discuss the directions of current activities.

Higher orders

One main-stream theoretical activity has always been trying to extend the QCD perturbative calculations to higher orders. Technically, this task is reduced to the evaluation of multidimensional integrals of a special type and the most efficient (known) technique for solving such integrals uses a version of the familiar trick called 'integration by parts' that has been reinvented and adapted to the present context about 20 years ago [11]. This technique allows to calculate three-loop diagrams of two sorts: either self-energy type (with two external lines) with massless quarks, or vacuum bubbles with massive quarks. Physical quantities for which the 3-loop (or 4-loop) result is known are usually those that can be reduced to the above basic integrals (the reduction may be quite nontrivial). Analytic problems are traded in this way for algebraically extensive tasks done by computers. These 'record' calculations include the 4-loop QCD β-function and the quark mass anomalous dimension, α_s^3 corrections to the total cross section of the e^+e^- annihilation and DIS sum rules (47), (48).

Quantities that involve several scales (say, quark mass m and external momentum q) or on-shell lines require new methods. In particular, the asymptotic expansion in m^2/q^2 or q^2/m^2 or around other special kinematic limits, followed by the (approximate or exact) summation has turned out to be very useful. Consider an example: the three-loop α_s^3 correction to the heavy quark pole mass

$$m_{\rm pole} = m_{\overline{\rm MS}}(m_{\overline{\rm MS}})\Big[1 + r_1\alpha_s + r_2\alpha_s^2 + r_3\alpha_s^3 + \dots\Big]. \tag{57}$$

A typical contribution to the coefficient r_3 is given by the Feynman diagram in Fig. 19 where the external quark lines have to be taken on-shell: $q^2 = m^2$. Instead, it was suggested to expand around $q^2 = 0$. In this way the diagram that originally was a self-energy type becomes a vacuum bubble, albeit with a complicated integrand, and the standard technique can be applied. In the original paper [12] the

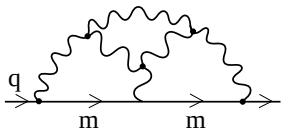


Fig. 19: A three-loop contribution to the heavy-quark pole mass

expansion was performed to order $(q^2/m^2)^{14}$ and the series extrapolated to $q^2 = m^2$ using the Padé summation. For this particular case, an exact analytic result $r_3 = 3.0451\dots$ [13] was later obtained using a different technique. The similar method was applied to the calculation of the α_s^2 corrections to the heavy quark production cross section in e^+e^--annihilation $e^+e^- \to Q\bar{Q}X$ and semileptonic decay rates of heavy quarks $b \to c\ell\nu$, $b \to u\ell\nu$, $t \to bW$. The method is rather general and more applications will probably follow. For example, one may try to calculate $2 \to 2$ parton scattering amplitudes by expanding in the ratios of the Mandelstam variables s/t or t/s, or etc.

A large amount of work is traditionally being invested in multiloop jet calculations. These pose a different sort of challenge because the kinematics becomes complicated as the number of jets increases, and because the calculation is done on the amplitude level. Infrared singularities either cancel in an intricate way or are factorized into parton distributions after the no less intricate cancelations. The next-to-leading order (NLO) calculations have become standard and many new results have appeared in recent years: $e^+e^- \to 4$ jets, $e^+e^- \to 3$ jets with heavy quarks, $p\bar{p} \to 3$ jets is to my knowledge almost completed, etc. In contrast to this, the NNLO jet calculations are still at the exploratory stage and present a new and interesting frontier. Up to now the NNLO QCD corrections are only known for Drell-Yan production. One would like to calculate, e.g. $pp \to 2$ jets (or 1 jet inclusive), $pp \to \gamma\gamma X$ (Higgs background), or $e^+e^- \to 3$ jets with the same high accuracy. The task is formiddable, but progress is being made in all directions. In particular, an analytic expression has been obtained recently [14] for the 'double-box' contribution to the gluon-gluon scattering and the techniques of phase-space integration have been advanced considerably to enable taking into account four partons in the final state. (Situations where two partons simultaneously become soft, or three become collinear, turn out to be very nontrivial.) Many algebraic and numerical tasks are yet to be done, however, so that the completion of this project requires concentrated effort.

For an overall consistency one also needs NNLO corrections to the DGLAP splitting functions. The corresponding work is in progress and the approximate kernels exist already, incorporating the known NNLO results for the evolution of a few lowest moments. It is likely, however, that the numerical impact of these corrections for $x > 0.01 - 0.05$ will be minimal.

Novel factorization 'theorems'

As already mentioned, although the concept of factorization is very general, the particular structure depends on the process in question and has to be analysed (and proved) case by case. This is a developing field of research, and new applications appear all the time. I will discuss (very briefly) a few recent examples that are important for the modern phenomenology.

One interesting development is related to the HERA physics of deep inelastic scattering at small Bjorken x and is usually referred to as hard diffractive scattering: DIS events with a large rapidity gap between the hadron system (with invariant mass M_X^2) and the recoiling proton, see Fig. 20. The process

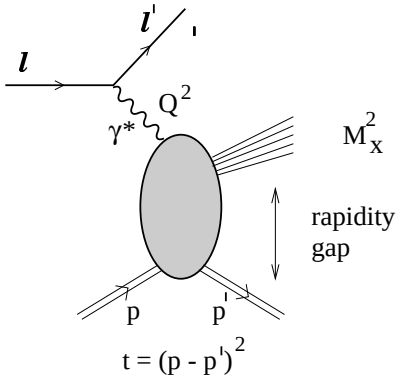


Fig. 20: Hard diffractive scattering

is usually described in terms of the Bjorken x variable, $x = Q^2/(Q^2 + W^2)$ where W^2 is the total invariant energy of the hadronic system, and $\xi = (Q^2 + M_X^2)/(Q^2 + W^2)$ which is the momentum fraction that the initial photon looses to the color neutral system X. Another common variable β is just

$\beta = x/\xi$. The statement of factorization [16] reads:

$$\frac{d\sigma_{\gamma^* p \to p' X}(x, Q^2, \xi, t)}{d\xi dt} = \sum_{i=q,\bar{q},g} \int_x^\xi dy\, \sigma_{\gamma^* i}(Q, x, y; \mu) \frac{df_i^D(y, \xi, t; \mu)}{d\xi dt}, \tag{58}$$

where $\sigma_{\gamma^* i}$ is the calculable cross section of the virtual photon scattering from the parton i (the same as in usual DIS) and $df/(d\xi dt)$ is called a 'diffractive parton distribution'. It represents the probability to find the parton i with momentum fraction y under the condition that the proton stays intact and looses the momentum fraction ξ.

The physical picture of diffraction is clearest in the proton rest frame [17]. At small Bjorken x meaning large energy, the virtual photon splits into a $\bar{q}q$ pair and develops a complicated partonic wave function long before it hits the proton. The slowest, 'wee' partons have a possibility to fluctuate to large transverse sizes, of hadronic scale. The diffraction occurs, because components in the wave function that have large transverse size simultaneously have a larger cross section and get absorbed. The smaller-size components can escape the collision but only present a part of the full (virtual) photon wave function: Re-expanding these components over the complete set of wave functions of possible hadronic states, one ends up with the inelastic diffraction. The diffraction cross section constitutes a finite (and large) fraction of the total deep inelastic cross section because the short distance processes are the same in both cases. The probability to have a gap is determined by the soft physics and decouples from the hard scale of the process.

Hard diffraction in hadron-hadron collisions appears to be much more complicated and factorization is not expected to hold, at least in the general case.

A particularly nice example of factorisation is provided by the deeply virtual Compton scattering (DVCS) [18] and hard diffractive meson production [19], shown schematically in Fig. 21. In constrast to

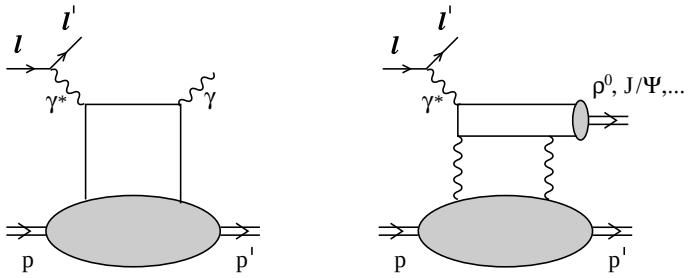


Fig. 21: Deeply virtual Compton scattering and hard diffractive meson production

the previous discussion both processes are fully *exclusive*, with only two particles in the final state. The challenge of the factorization proof is in both cases to demonstrate that the final-state interactions between the outgoing particles can be neglected. In the case of the hard diffractive vector meson production it only happens to be true (at large Q^2) for the longitudinal polarization.

The description of both processes requires new, generalized parton distributions because the proton is scattered with a non-zero momentum transfer and, therefore, the parton emitted with a certain momentum fraction x is returned with a different momentum fraction x'. These objects, dubbed 'skewed' (also non-forward, off-forward etc.) parton distributions have hybrid properties, representative of that of usual parton distributions and of light-cone distribution amplitudes in different kinematical regions. Observation of hard exclusive deep inelastic processes of this type would give many new insights in the hadron structure and is very interesting. From the experimental point of view, such a program is very demanding because of the necessity to have a decent value of Q^2 (i.e. large energy) and at the same time make sure that there are no soft pions in the target proton region.

As the last example, consider exclusive B-meson decays in two (light) mesons M_1 and M_2. A general two-body decay amplidude can be written as

$$A(B \to M_1 M_2) = A_1 e^{i\delta_1} e^{i\delta_1^W} + A_2 e^{i\delta_2} e^{i\delta_2^W} + \dots \tag{59}$$

where real $A_{1,2}$ denote the absolute values of the amplitudes, $\delta_{1,2}$ are strong interaction phases and $\delta_{1,2}^W$ weak phases. In a typical situation there are several terms in the r.h.s. (59) due to different reaction mechanisms. The weak phases are CP-violating and of primary interest. Yet to determine them the strong phases and the amplitudes must be known, unless we are so lucky that there exists a single term only, or the experimental information is so abundant that it allows to separate the contributions of separate terms.

In this case the large mass of the b-quark sets the hard scale, but it turns out that the standard methods do not apply because the soft (light) quark spectator in the B-meson may go into the final state meson without participating in a hard scattering. Because of this, the required nonperturbative input is not reduced to quark distributions in the three participating hadrons. It was suggested [20] that the factorized expression can still be written, at the cost of adding an additional contribution written in terms of the $B \to M_1$ semileptonic form factor. Schematically

$$A(B \to M_1 M_2) = F_{B\to M_1}(q^2=0) \int_0^1 dx\, T_1(x)\, \Phi_{M_2}(x) + \int_0^1 d\xi dx dy\, T_2(\xi,x,y)\, \Phi_B(\xi) \Phi_{M_1}(y) \Phi_{M_2}(x)\,, \tag{60}$$

where T_1 and T_2 are calculable short-distance functions, $\Phi_{B,M_1,M_2}(x)$ are the (nonperturbative) distribution amplitudes of quarks in the mesons and $F_{B\to M_1}(q^2=0)$ is the form factor which is not supposed to be further factorized. The two contributions to the factorization formula (60) are shown schematically in Fig. 22. To my opinion, this work still has to be cleaned up. In particular, the separation of the two

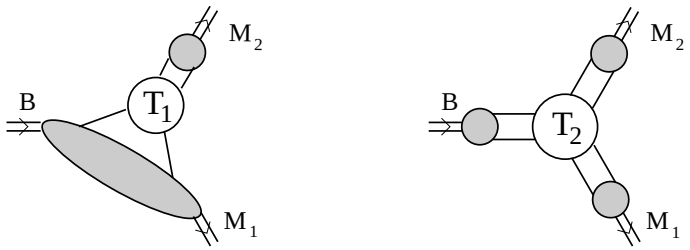


Fig. 22: QCD factorization in exclusive B-decays

contributions has to be scale-dependent itself and the corresponding renormalization-group treatment is so far missing. The potential of this approach is high; since the participating nonperturbative quantities can (at least in principle) be measured or calculated on the lattice, the strong phases and the amplitudes turn out to be completely predicted, at least in the formal $m_b \to \infty$ limit. Whether the predictions are accurate enough at the realistic m_b scale, remains to be seen.

5 RESUMMATIONS

It often happens that perturbative contributions of a certain type have to be summed to all orders in perturbation theory because they are enhanced by kinematical factors. The clearest example — on which the problem was initially formulated by Gell-Mann and Low and solved by Landau and collaborators — is provided by the photon polarization operator (3) in QED. To leading order

$$\Pi^{\rm QED}(q^2) = \frac{\alpha}{3\pi} \ln \frac{q^2}{m_e^2}\,, \tag{61}$$

where $\alpha = 1/137\ldots$ is the fine structure constant and m_e is the electron mass. The photon propagator, therefore, receives a correction

$$D_{\mu\nu}(q^2) = \frac{g_{\mu\nu}}{q^2[1-\Pi^{\rm QED}(q^2)]} = \frac{g_{\mu\nu}}{q^2}\left[1+\frac{\alpha}{3\pi}\ln\frac{q^2}{m_e^2}+\ldots\right] \tag{62}$$

that grows logarithmically with q^2. It is clear that the second-order contribution contains $\alpha^2\ln^2 q^2/m_e^2$ and in general higher-order corrections are accompanied by the ever increasing powers of the logarithm. As the result, for the sufficienly large q^2 the perturbative expansion does not work. The solution to this problem is well known nowadays: In order to avoid the accumulation of logarithms in high orders of perturbation theory, it is sufficient to reexpand the photon propagator in terms of the effective charge

$$\alpha_{\rm QED}(q^2) = \frac{\alpha}{1-\frac{\alpha}{3\pi}\ln\frac{q^2}{m_e^2}}. \tag{63}$$

The photon polarization operator $\Pi(q^2)$ is expressed in terms of $\alpha_{\rm QED}(q^2)$ as a regular perturbative series with numerical coefficients and no logarithms (to be precise, one either has to take an imaginary part of $\Pi(q^2)$ or a derivative in q^2). In QCD, such a procedure is standard and is considered as 'self-obvious': E.g. the total cross section of e^+e^- annihilation at the energy $s=q^2$ is expressed in terms of $\alpha_s(q^2)$ 'because there exists no other scale'.

In general, one speaks of a resummation, when perturbative contributions of some type are enhanced by a certain kinematical factor (typically a logarithm or a double-logarithm of the ratio of the contributing disparate scales) and have to be summed to all orders in perturbation theory. Technically speaking, the resummation amounts to the rearrangement of the series as, schematically

$$\sum_n \alpha_s^n\left[r_n^n\ln^n x + r_n^{n-1}\ln^{n-1}x+\ldots+r_n^0\right] = \sum_n r_n^n(\alpha_s\ln x)^n + \sum_n r_n^{n-1}\alpha_s(\alpha_s\ln x)^{n-1}+\ldots, \tag{64}$$

where it is assumed that $\ln x \gg 1$, $\alpha_s\ln x\sim\mathcal{O}(1)$. The task of the resummation is to calculate the first few (infinite) sums in the r.h.s exactly. The leading-order resummation amounts to taking into account contributions with the maximum number of logarithms for each power of the coupling (there is often more than one logarithm for each power of α_s), the next-to-leading-order resummation corresponds to taking into account the terms with one logarithm less, etc.

Although the idea of the resummation is very general, there exists no general solution which reflects the fact that the physics of kinematical enhancements can be very different. In fact, there exists a large variety of applications and we will consider several of them in what follows. Working out specific resummation formulas is often technical and belongs to the most challenging problems that the theorists have to deal with.

5.1 Coulomb corrections

As the first example, consider top-quark production in the e^+e^- annihilation near the kinematical threshold: $e^+e^-\to\bar{t}tX$. The explicit calculation of the first-order QCD correction corresponding to the gluon exchange between the t and $\bar{t}$ reveals a large enhancement factor $1/v$ where $v=\sqrt{E/m_t}$ is the classical velocity of the outgoing t-quark; $E=\sqrt{q^2}-2m_t$ is the energy measured relative to the $t\bar{t}$ threshold and we assume that $E\ll m_t$. The reason for this enhancement is easy to understand. The top quark and the antiquark are produced at one point and move very slowly. In fact, they decay before the distance between them becomes comparable to the strong interaction scale Λ. This is a good news since it indicates that perturbative QCD may be applicable. At the same time, however, since the quarks spend a long time in the vicinity of each other, contribution of the gluon exchange is enhanced as it can occur any time. For the same reason, taking into account a single gluon exchange is not enough and one has to sum up the series of ladder-type diagrams (in the Coulomb gauge) shown in Fig. fig:4-23. Luckily enough, the solu-

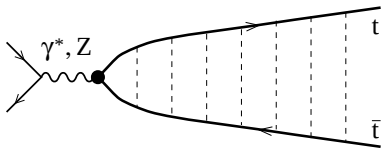


Fig. 23: Top quark production in e^+e^-annihilation near the energy threshold. The dashed lines correspond to instantaneous Coulomb-like gluon exchanges.

tion is known from quantum mechanics (to leading order) since it corresponds to solving the Schrödinger equation for the top quark moving in the static Coulomb-like potential $V(\vec{r}) = -4\alpha_s/(3|r|)$ created by the (anti)quark. Neglecting all relativistic corrections, the cross section is given by the imaginary part of the nonrelativistic Green function corresponding to the propagation of a charged particle with the reduced mass $m_t/2$ in the static Coulomb field, starting from and ending at the origin of the field in the coordinate space, calculated in the energy representation [21]:

$$\sigma_{e^+e^-\to t\bar{t}}(E) = \frac{8\pi^2\alpha_{\rm QED}^2}{3m_t^4}\,{\rm Im}\,G(r_1=0,r_2=0;E+i\Gamma_t); \tag{65}$$

Γ_t is the total top quark width. The explicit expression for the Green function in dimensional regularization with $\overline{\rm MS}$ subtraction is [22]

$$G(0,0,E) = -\frac{m_t^2\alpha_s}{3\pi}\left[\frac{1}{2\lambda}+\frac{1}{2}\ln\frac{-4m_tE}{\mu^2}-\frac{1}{2}+\gamma_E+\psi(1-\lambda)\right], \tag{66}$$

where $\lambda = 2\alpha_s/(3\sqrt{-E/m_t}) = -2i\alpha_s/(3v)$ is the so-called Coulomb parameter, $\psi(x)$ is the logarithmic derivative of the Γ-function and γ_E is Euler constant. The cross section requires only the discontinuity of G which is scheme-independent. Note that the Green function contains a continuum at $E>0$ and poles at $E<0$ from which the energies and wave functions at the origin of toponium resonances can be extracted. It follows that the expression in (66) effectively organizes all corrections in the 'naive' perturbative expansion that are enhanced by inverse powers of the quark velocity.

A generalization of this result to the subleading powers of $1/v$ has proven to be much more complicated. The phenomenologically interesting interval of energies is the interval in which the $t\bar{t}$ cross section deviates considerably from the background as the consequence of the color attraction. This happens when E is of the order of the binding energy in the ground state of the nonrelativistic $t\bar{t}$ system $E_0\sim m_t\alpha_s^2$ or, which is the same, $v\sim\alpha_s$. Disregarding possible logarithms, we can envisage a generic expansion of the type

$$\sigma_{e^+e^-\to t\bar{t}} = v\sum_{n=1}^{\infty}\left(\frac{\alpha_s}{v}\right)^n\left\{1;[\alpha_s,v];[\alpha_s^2,\alpha_sv,v^2];\ldots\right\} \tag{67}$$

in which the first term corresponds to the leading-order resummation and is given by the solution of the nonrelativistic Schrödinger equation in (66); the terms suppressed by either one power of α_s or one power of velocity both have to be resummed in NLO, etc. This expansion has recently been made systematic by finding essential integration regions and formulating a non-relativistic effective field theory for each of them. The difficulty of the problem is that the physics appears to be different at scales of order m_t, m_tv and m_tv^2 corresponding to the top quark mass, momentum and energy, respectively. Without going into details I want to mention that, as it turns out, integrations over the quark and gluon momenta have to be separated in contributions of different regions, called 'hard' (h), 'soft' (s), 'potential' (p) and 'ultrasoft'(u) [gluons only]. The idea of an effective field theory is nothing but a yet another language to introduce factorization. The so-called non-relativistic QCD Lagrangian (NRQCD) is designed to take

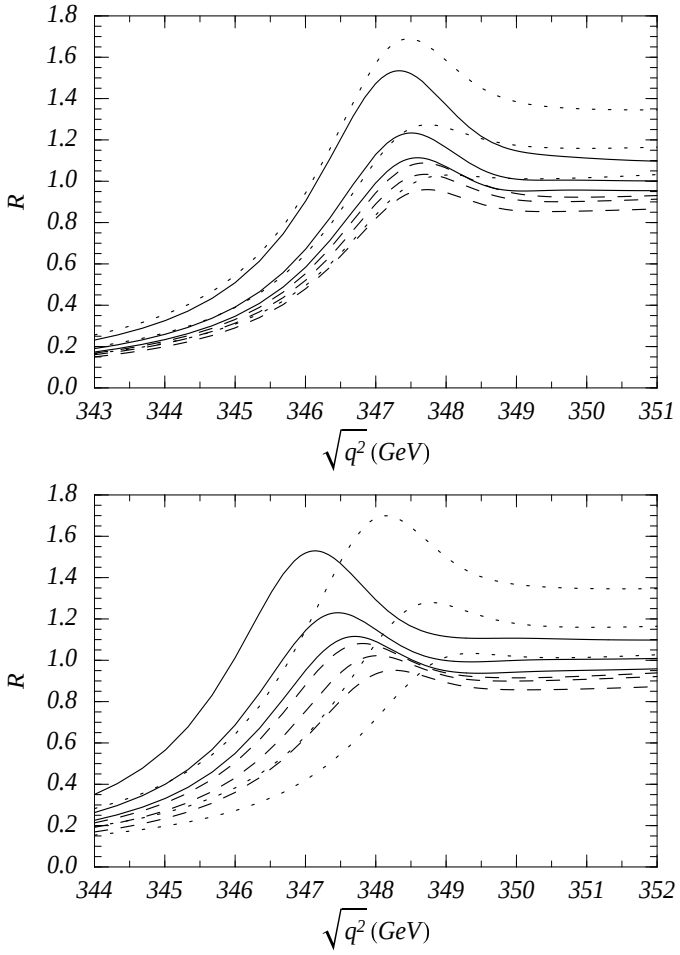


Fig. 24: The total normalized $t\bar{t}$ cross section (virtual photon contribution only) in the threshold region at leading order (dotted), next-to-leading order (dashed) and next-to-next-to-leading order (solid) versus the center-of-mass energy [23]. Both plots use $\bar{m}_t(\bar{m}_t) = 165\,\text{GeV}$, $\Gamma_t = 1.43\,\text{GeV}$ and $\alpha_s(m_Z) = 0.119$ as input. The three curves for each case refer to the renormalization scale $\mu = \{15(\text{upper}); 30(\text{central}); 60(\text{lower})\}\,\text{GeV}$. The PS mass corresponding to $\bar{m}_t(\bar{m}_t) = 165\,\text{GeV}$ is $m_{t,\text{PS}}(20\,\text{GeV}) = 173.30\,\text{GeV}$, the pole mass $m_{t,\text{pole}} = 175.05\,\text{GeV}$. The upper panel shows the successive approximations in the PS mass scheme, the lower panel in the pole mass scheme.

into account the effects at the scales less than the factorization scale μ that is supposed to be chosen in the interval $m_t v \ll \mu \ll m_t$, schematically

$$\mathcal{L}_{\text{QCD}}[Q(h,s,p); g(h,s,p,us)] \overset{m_t v \ll \mu \ll m_t}{\longrightarrow} \mathcal{L}_{\text{NRQCD}}[Q(s,p); g(s,p,us)]. \tag{68}$$

The difference between the QCD and the NRQCD Lagrangian is that contributions of hard quarks and gluons are omitted from the latter; they have to be included in the (calculable) coefficient functions that relate the observables calculated using NRQCD with those in the real world (QCD). It turns out that using NRQCD is not enough for the systematic expansion of the top production cross section and one more step is necessary, introducing the so-called 'Potential NRQCD' Lagrangian in order to separate the contributions of soft and potential (ultrasoft) integration regions:

$$\mathcal{L}_{\text{NRQCD}}[Q(s,p); g(s,p,us)] \overset{m_t v \ll \mu \ll m_t}{\longrightarrow} \mathcal{L}_{\text{PNRQCD}}[Q(p); g(p,us)]. \tag{69}$$

Using this framework, it has become possible to resum the Coulomb corrections to the top quark production cross section to the NNLO accuracy. The results are shown in Fig. 24. The two calculations in the upper and in the lower panel differ by the definition of the top quark mass. We will come back to this issue later, in the discussion of nonperturbative corrections. The main outcome of these calculations is

that although the normalization of the cross section receives large corrections, the position of the peak can be predicted rather precisely — within ~ 200 MeV. Hence the mass of the top quark can be measured with a comparable accuracy.

A one more application of the same formalism has been the determination of the b-quark mass from the data on $e^+e^- \to b\bar{b}X$ in the threshold region and $e^+e^- \to \Upsilon(1s)$. The result is $m_b^{\overline{\rm MS}}(m_b^{\overline{\rm MS}}) = 4.23 \pm 0.08$ GeV and has a much better accuracy compared to all other existing determinations.

5.2 Sudakov effects

In a typical event in e^+e^--annihilation two narrow jets are produced, with the invariant mass much smaller compared to the total energy $\rho = M_{\rm jet}^2/s \ll 1$, see Fig. 25. Broadening of the jets is mainly due

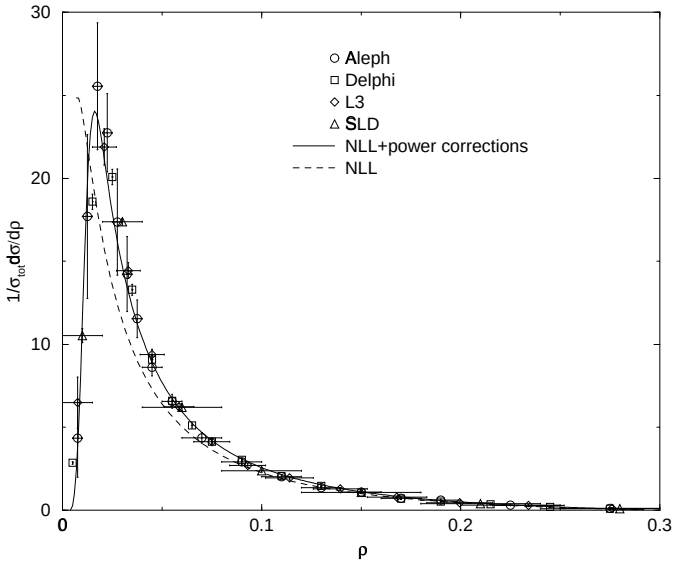


Fig. 25: Heavy jet mass distribution $\rho = \max(M_R^2/s, M_L^2/s)$ at $\sqrt{s} = M_Z$. M_R^2 and M_L^2 are defined as total invariant masses flowing into the right and the left hemispheres with respect to the thrust axis. The solid and the dashed curves correspond to theoretical calculation with and without taking into account nonperturbative effects, respectively [24].

to gluon emission and a large invariant mass requires emission of a sufficiently energetic gluon at a large angle. Since such processes are rare — we remember that the gluon bremsstrahlung mainly involves soft and collinear emission that does not alter the jet mass considerably — the rise of the jet mass distribution at small (but not very small) values of ρ in Fig. 25 is understandable. This rise is interrupted, however, at $\rho \sim 0.02 - 003$ and at yet smaller masses is replaced by a sharp drop. This is the region that I want to discuss in this section. The jets, apparently, do not want to be *too* narrow. Why?

In order to enforce a small value of the jet mass one has to constrain gluon radiation. Classically, the probability to have a pencil-like jet is equal to the probability that not a single gluon is emitted. The change of the 'jet survival probability' with time is given by

$$dP = -w\,P\,dt\,, \tag{70}$$

where w is the probability of gluon emission (in unit time). Solving this equation, one ends up with a familar exponential decay law

$$P(t) \sim e^{-wt}\,. \tag{71}$$

The crucial condition for the exponential decay (in general) is that the decay probability w has to be independent on the 'history' of the process — in our case we would have to require that the probability

to emit a gluon at time t_0 is independent on how many gluons (and with which momenta) have already been emitted at times $t < t_0$. In fact, this is exactly the statement of the so-called Low theorem, that is applicable to emission of soft gluons (and photons). The reason for the Low theorem to hold is that soft gluons cannot resolve the internal structure of the jet; soft gluons only 'see' the total color charge of the jet as a whole (and the jet direction) but not the charge distribution among the jet parton constituents. It follows that the condition for the exponential pencil-like jet decay is indeed satisfied in QCD in so far as we can restrict ourselves to the soft emission (not necessarily collinear). The explicit calculation gives, to leading order

$$\frac{1}{\sigma_{\rm tot}}\frac{d\sigma}{d\rho} = \frac{8\alpha_s}{3\pi}\frac{\ln(1/\rho)}{\rho}\exp\left[-\frac{4\alpha_s}{3\pi}\ln^2(1/\rho)\right]. \tag{72}$$

Note the exponentiation which is the consequence of independent gluon emission, and the double logarithm that is the same as in (21) with the jet mass replacing the gluon mass as the infrared regulator. Expanding the exponential, one obtains a 'regular' perturbative series in which each power of the QCD coupling is accompanied by two powers of the logarithm $\ln 1/\rho = \ln s/M_{\rm jet}^2$. The expression in (72) corresponds to the leading-order resummation with the maximum amount of logarithms kept to each order of perturbation theory.

The resummation of this type was invented long time ago by Sudakov [25] who studied the electron form factor in QED. The latter case is extremely beautiful, since it turns out that the leading order resummation is exact — there exist no further corrections! In QCD, the situation is much more complicated and a lot of work was invested (and is being invested) to make the corresponding calculations systematic.

The Sudakov-type resummations of threshold corrections in QCD are well known to the next-to-leading order accuracy for Drell-Yan pair production and event shapes in e^+e^--annihilation. In the last years, the similar NLO resummations have been worked out for more complicated reactions involving $2 \to 2$ scattering processes at the parton level (heavy quark production ($b\bar{b}$, $t\bar{t}$), prompt photons, dijets, ...). These cases are more complicated, because soft gluons transfer color to the final state so that one has to deal with the anomalous dimension matrix. Also, there exist more kinematical variables: The scattering angle or relative rapidity of the two final state particles.

The practical outcome of these (complicated) calculations so far has been relatively modest. The resummation effects in integrated quantities (like the total top quark production cross section at the Tevatron or $d\sigma/dE_T$ for γX) are typically small, well within the NLO QCD correction. The resummation does lead to a significant reduction of the scale-dependence, but whether this can be interpreted as the reduction of the theoretical uncertainity, is not obvious. A general problem seems to be that at present energies the effects of resummation become significant too close to the kinematic thresholds, where nonperturbative corrections are already important. This can already be seen from Fig. 25: the perturbative Sudakov suppression 'switches on' much too late, and the description of the jet mass distribution at small values of ρ requires a large nonperturbative (hadronization) correction.

5.3 BFKL equation and the Regge limit

The most important result of HERA was the finding that the deep inelastic structure functions strongly increase in the region of small Bjorken x, see Fig. 26. As seen from the data, the x-dependence apparently becomes more steep at larger values of Q^2. Let us see whether we can understand this. In the parton model picture, a large structure function has to reflect a large parton distribution and it is easy to convince oneself that at small x the gluon distribution is the one that is most important. Moreover, the Q^2 dependence of the structure function can largely be traced to the scale dependence of the parton distribution. This is because, after the factorization is established, one can push the factorization scale to large values $\mu^2 \sim Q^2$. In this way, the cross section of virtual photon scattering (cf. (49)) is expanded in powers of $\alpha_s(Q)$ and does not contain any logarithms so that it is a very slow function. Hence, it is worthwhile to have a closer look on the scale dependence of the gluon parton density that is governed

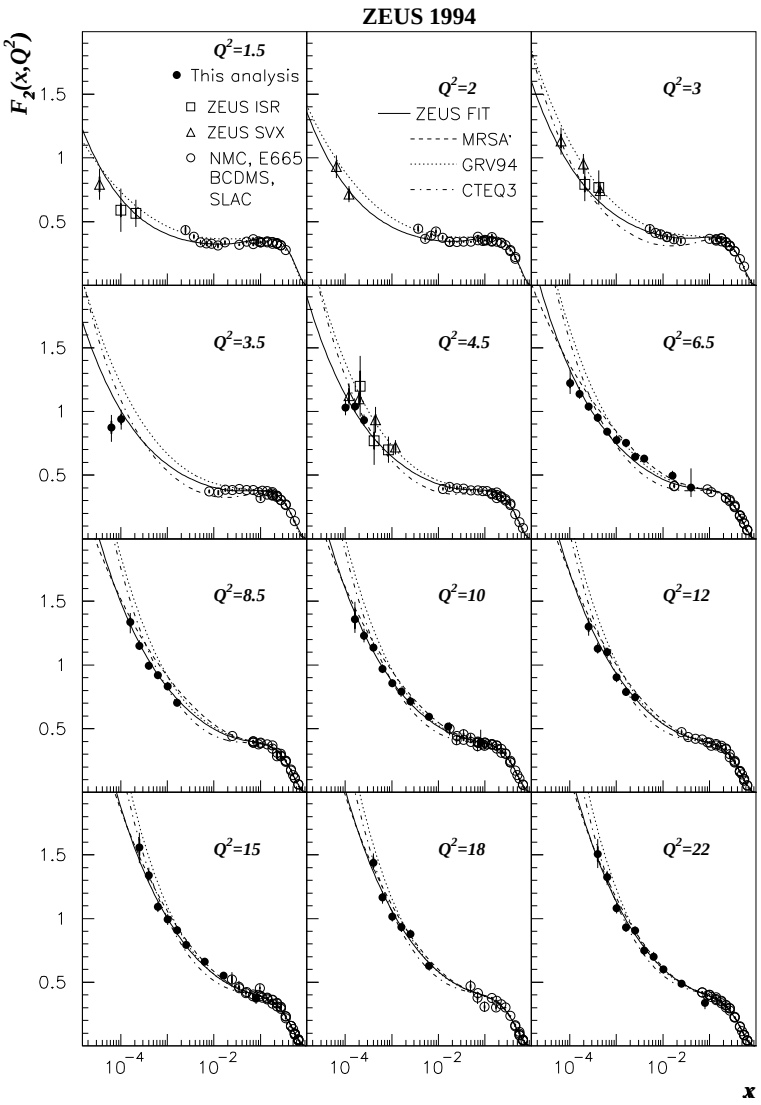


Fig. 26: Structure function $F_2(x, Q^2)$ as a function of x for fixed values of Q^2 [26]. The solid lines indicate the fit to the data. Also shown are some popular parametrizations. The Q^2 values are given in GeV^2.

by the famous DGLAP equation which I have already mentioned in Sect. 4.1 but so far never written explicitly. For simplicity, let us neglect quark contributions altogether. In this case the DGLAP evolution equation takes the form

$$\frac{d}{d\ln Q^2} f_g(x,Q^2) = \int\limits_x^1 \frac{dy}{y}\, P_{gg}(y,\alpha_s)\, f_g(x/y,Q^2)\,. \tag{73}$$

In this expression $P_{gg}(y,\alpha_s)$ is the probability of gluon splitting into a pair of gluons, with y being the longitudinal momentum fraction of the initial gluon carried on to the gluon in the final state. The huge popularity of the DGLAP equation is due to the simplicity of its interpretation. The equation (73) describes the dependence of the gluon density distribution in the nucleon on the resolution (scale) with which it is measured. Going over to higher values of Q^2 corresponds to using a better microscope so that finer details of the nucleon structure can be seen. In particular, if we see a gluon at $Q^2 = Q_0^2$ and increase the resolution to $Q^2 = Q_1^2 > Q_0^2$, we may find that what originally appeared to be a gluon with momentum fraction x, turns out to be a pair of two collinear gluons with momentum fractions x_1 and $x_2 = x - x_1$. Obviously, $x_1, x_2 < x$ so that increasing the resolution we see more gluons with smaller momentum fractions and less gluons with larger momentum fractions — the gluon density at large Q^2 is shifted towards small x values.

The crucial property of $P_{gg}(y,\alpha_s)$ is that it is singular in the $y \to 0$ limit:

$$P_{gg}(y,\alpha_s) \stackrel{y\to 0}{=} \frac{\alpha_s}{2\pi}\frac{6}{y} + \mathcal{O}(\alpha_s^2) \tag{74}$$

where the appearance of $1/y$ can easily be traced to the $1/\omega$ factor in the expression for the gluon bremstrahlung in Sect. 3.1, see Eq. (18). Let us use this expression and rewrite Eq. (73) as a finite-difference equation

$$f_g(x,Q_1^2) - f_g(x,Q_0^2) = \frac{3\alpha_s}{\pi}\ln\frac{Q_1^2}{Q_0^2}\int\limits_x^1 \frac{dy}{y}\,\frac{1}{y}\, f_g(x/y,Q_0^2)\,. \tag{75}$$

This is legitimate if the difference between $\ln Q_1^2$ and $\ln Q_0^2$ is not large. As a first guess, assume that the gluon density f_g at the lower scale Q_0^2 is completely flat: $f_g(x,Q_0^2) = \text{const}$. In this case the integration in (75) becomes trivial and the result reads

$$f_g(x,Q_1^2) = \text{const}\ \cdot\ \left[1 + \frac{3\alpha_s}{\pi}\frac{1}{x}\ \ln\frac{Q_1^2}{Q_0^2}\right]. \tag{76}$$

Note the $1/x$ factor. It means that once gluon radiation is allowed, the gluon density *cannot* stay flat and starts to grow rapidly at small x. In particular, the gluon multiplicity

$$N_g = \int\limits_0^1 dx\, f_g(x,Q_1^2) \tag{77}$$

is formally infinite. In fact, it also means that our flat ansatz for the gluon density at $Q^2 = Q_0^2$ is not a clever one: If Q_0^2 is not very small so that we already 'see' a few gluons, their distribution, most probably, already contains the $1/x$ factor. Hence, we make a second try: Assume $f_g(x,Q_0^2) = \text{const}/x$, or, equivalently, $x f_g(x,Q_0^2) = \text{const}$. With this new ansatz, performing the integration in (75) and multiplying the result by x, we obtain

$$x f_g(x,Q_1^2) = \text{const}\ \cdot\ \left[1 + \frac{3\alpha_s}{\pi}\ln\frac{1}{x}\ \ln\frac{Q_1^2}{Q_0^2}\right] = x f_g(x,Q_0^2)\ \left[1 + \frac{3\alpha_s}{\pi}\ln\frac{1}{x}\ \ln\frac{Q_1^2}{Q_0^2}\right] \tag{78}$$

which looks sensible.

Solution of the differential equation (73) is nothing but the repetition of the finite-difference equation (75) many times, for the set of scales $Q_0^2 < Q_1^2 < Q_2^2 < \ldots < Q^2$. On each step one gets a QCD coupling accompanied by the logarithms as in Eq. (78), so that the final result presents the resummation of $(\alpha_s \ln Q^2 \ln 1/x)^k$ corrections to the so-called double-logarithmic accuracy. In fact, one can do somewhat better, taking into account that the QCD coupling always enters at the relevant scale and is changing with the evolution. The explicit solution was obtained in an early work [27]:

$$x f_g(x, Q^2) \sim \exp \sqrt{\frac{48}{b} \ln \frac{\alpha_s(Q_0^2)}{\alpha_s(Q^2)} \ln \frac{1}{x}}\,, \tag{79}$$

where $b = 11 - 2/3 n_f$ is the first coefficient of the QCD β-function. Note that this behavior corresponds to the approximation when, first, the splitting function $P_{gg}(y)$ is calculated to the leading order in α_s and, second, all contributions to $P_{gg}(y)$ that are not singular at $y \to 0$ are neglected.

The trouble is that higher-order contributions to the DGLAP splitting function appear to be enhanced at small x:

$$P_{gg}(x) = \frac{3\alpha_s}{\pi x} \left[1 + 9\zeta(3) \left(\frac{\alpha_s}{\pi} \ln \frac{1}{x} \right)^3 + \frac{81}{20} \zeta(5) \left(\frac{\alpha_s}{\pi} \ln \frac{1}{x} \right)^5 + \mathcal{O}\left(\left(\frac{\alpha_s}{\pi} \ln \frac{1}{x} \right)^6 \right) \right], \tag{80}$$

where $\zeta(3) = 1.20206$ and $\zeta(5) = 1.03693$. Note that there remains to be a single overall factor $1/x$, but each power of the coupling is multiplied by a logarithm of $1/x$. (Absence of the terms $\sim (\alpha_s \ln x)^{1,2,4}$ is essentially accidental). The expansion written in (80) corresponds to the leading logarithmic approximation in $(\alpha_s \ln x)^k$ meaning that the corrections with less logarithms are neglected.

The resummation of leading logarithmic corrections $(\alpha_s \ln x)^k$ to the splitting function is accomplished using the so-called BFKL equation [28, 29]. In both cases (DGLAP and BFKL), one deals with the same ladder-type Feynman diagrams for the gluon emission as the one shown schematically in Fig. 27, and the challenge is to calculate the contributions with arbitrary many emitted gluons (including the necessary virtual corrections, of course). In both cases, this is achieved using the same

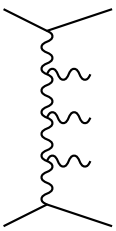


Fig. 27: Multiple gluon emission

general scheme: summation over the contributions of many gluons is reformulated as a certain integro-differential equation, and the kernel in this equation takes into account a single gluon emission to some accuracy. Solving the DGLAP equation (73) we have to split the whole evolution interval in smaller pieces $Q_0^2 < Q_1^2 < Q_2^2 < \ldots < Q^2$ so that for each piece a finite-difference approximation for the derivative can be used. This, of course, exactly corresponds to taking into account gluon emission one by one. As we discussed in Sect. 3.2, the scale parameter Q^2 roughly corresponds to the transverse momentum of the emitted gluon so that the procedure incorporated in the DGLAP equation is to enforce the condition that gluon transverse momenta are *strictly ordered* along the ladder. This does *not* mean that 'wrongly' ordered configurations are discarded, but rather that they are taken into account as higher-order corrections to the splitting functions. Hence, in principle, the DGLAP equation incorporates all corrections, but if by some reason 'wrongly ordered' configurations give a large contribution, this equation is not adequate in a sense that one is required to calculate the splitting functions to all orders.

This is exactly what happens at small x. If $\ln Q^2 \ll \ln 1/x$, the dominant kinematic region corresponds to the emission of gluons that are strictly ordered in rapidity ($\sim \ln 1/x$) but not ordered in transverse momenta. It is therefore almost natural to look for an equation in which $\ln 1/x$ would serve for an evolution parameter and higher-order corrections to the (new) evolution kernels take into account corrections related to the 'wrong' ordering in rapidity. This equation was invented by Kuraev, Fadin and Lipatov [28] and further elaborated in a somewhat later work by Balitsky and Lipatov [29] where it was shown that giving up the ordering in the transverse momenta does not produce any nasty infrared singularities.

The solution of the BFKL equation with the kernel calculated to lowest order corresponds to the resummation of $(\alpha_s \ln x)^k$ corrections to the leading-logarithmic accuracy. The result is

$$x f_g(x) \sim x^{-12 \ln 2 \alpha_s/\pi} \tag{81}$$

indicating an even steeper increase at small x compared to the DGLAP double-logaritmic limit (79).

The next-to-leading order corrections to the BFKL kernel have been calculated only recently [30]. They correspond to the NLO resummation — corrections to (80) with one power of the logarithm less — and turned out to be very large. The corresponding modification of the low-x behavior is to replace

$$12 \ln 2 \frac{\alpha_s}{\pi} \left[1 - 20 \frac{\alpha_s}{\pi} + \ldots\right] \tag{82}$$

There have been a lot of recent works trying to 'tame' this huge correction by some well (or not so well) motivated partial resummations beyond NLO. The overall impression is that this is indeed possible, by summing collinear double logarithms to all orders, taking into account the running coupling effects etc. After such massaging, one can achieve that the energy conservation is strictly maintained, the resummed anomalous dimension appears to be close to the exact DGLAP one for not small x, and reasonable bounds on the power in (81) can be established. One probably still needs time to get a coherent picture, however.

In these lectures, I do not want to enter a discussion of whether the DGLAP or the BFKL evolution is better suited for the description of the HERA data, since to my (heretic) opinion this question is not that interesting. Far more appealing about the BFKL equation is that it offers the first insight on a general problem of the high-energy behavior of cross sections in QCD.

The basic formalism for the description of high energy behavior of amplitudes was invented by Regge and developed in works of Gribov, Mandelstam, Pomeranchuk etc. The rough idea is the following. Since amplitudes at high energies involve summation over contributions over many partial waves, the summation can be approximated by the integration. Under favourable circumstances, the integration over real angular momenta (alternatively, one may speak of the summation of contributions of hadron resonnaces in the s-channel) can be shifted to the complex plane. In this way, it can be shown that high-energy behavior of cross sections is determined by singularities of the scattering amplitude in the complex angular momentum plane. These are called Regge singularities, or Regge poles (if they are poles). In the early days of the Regge theory, there was a hope that the structure of singularities in the complex angular momentum plane is relatively simple — compared to the incredible variety of hadron resonances that were found in experiments just around this time. In the Regge picture, the sum over *all* s-channel resonances can be substituted (in the sense of analytic continuation) by an exchange of a *single* funny object with complex spin in the t-channel, called the reggeon. Reggeons have quantum numbers and in many respects behave like particles. The pomeron (named after I.Pomeranchuk) is nothing but the particular reggeon that mediates high-energy scattering of two particles without the change of quantum numbers in the t-channel so that it has itself the quantum numbers of the vacuum.

It has been shown that many properties of high-energy amplitudes and of reggeons in particular do not depend on the underlying field theory and can be derived using unitarity and analyticity only — the two basic properties that any 'good' theory has to obey. Most importantly, it is possible to prove that

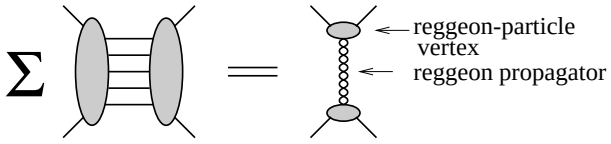


Fig. 28: Substitution of the sum over s-channel intermediate states by the reggeon exchange. The 'bubble-line' denotes the reggeon

the picture of the Reggeon exchange is by itself not complete: one has to take into account contributions of several reggeon exchanges (so-called Reggeon cuts) and reggeon interactions. It follows that any field theory in the Regge limit will effectively be reduced to the reggeon field theory (rather than quantum mechanics, e.g.). As the result, chances for the simplicity of the Regge description appear to be slim.

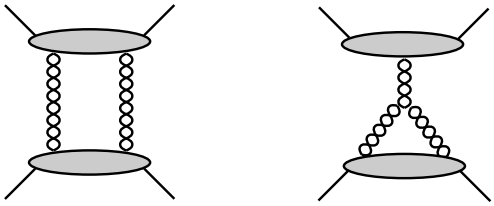


Fig. 29: Reggeon diagrams

In the Regge theory language the BFKL result is interpreted (in fact this was the original interpretation) as the explicit construction of the pomeron in QCD perturbation theory. The next step: the theory of interacting pomerons (interacting gluon ladders) still presents an open problem (and we do not in fact know whether the BFKL construction presents a good starting approximation). This problem is very challenging and at the same time one of the most interesting, corresponding to 'new physics' that arises in QCD at small coupling and large parton densities.

To explain this, remember that one of the assumptions of the parton model was that the parton density in the nucleon is not 'too high' so that the virtual 'heavy' photon can only interact with one parton. This condition is broken at very large energies. Indeed, the transverse 'size' of the partons at the scale Q^2 is just $r_\perp \sim 1/Q$. If Q^2 is kept fixed and the energy increases, so does the number of partons which the virtual photon can interact with. At some energy this number becomes so large that the partons occupy the whole proton disc area and start to overlap. Once it happens, the picture of the interaction changes since the virtual photon is able to interact coherently with the overlapping partons. This is precisely what happens in a deep inelastic scattering on a large nucleus and is responsible for the deviation of the structure function of a nucleus compared to that of a free nucleon. In the Regge theory language, effects of parton screening correspond to the contributions of the type shown in Fig. 29. They are expected to be very important in high energy nucleus-nucleus collisions that will be observed and investigated in much detail in the coming decade. As a consequence, it is easy to predict that the high energy behavior of cross sections in QCD is going to remain to be a hot frontier.

Several theoretical approaches are currently being developed. It was shown [31] that the 'quantum mechanics' of pomerons that includes interactions between gluon ladders but not their splitting (i.e. the first diagram in Fig. 29 but not the second one) has a highly nontrivial hidden symmetry and a large number of conservation laws. It is mathematically equivalent to the quantum mechanics of the so-called completely integrable Heisenberg spin chains so that the same methods can be used for the solution. The approach involves very elegant (and powerful) mathematics but so far the practical outcome was rather limited. The second idea [32] that has been around for some time, is to formulate the theory

of interacting gluon ladders as an effective field theory in which the BFKL solution would correspond to the Born approximation. This is indeed possible to do, but the corresponding theory appears to be prohibitively complicated. In a more pragmatic approach, several models of high-energy behavior have been suggested that incorporate the BFKL solution (e.g. the color-dipole model [33]). Finally, there exists an interesting suggestion by McLerran and Venugopalan [34] that high energy scattering of two large nuclei may be dominated by production of a strong classical color field and may be easier to describe compared to the nucleon-nucleon collisions, see [6] for a review and the comparison with the more conventional approaches.

5.4 Particle spectra at low x in e^+e^--annihilation

As we have just seen, the total gluon multiplicity in the nucleon is infinite, as reflected by the rapid increase of the gluon parton density at small x due to multiple emission of soft gluons. This implies, in particular, that the total multiplicity of gluons (hadrons) produced in a deep inelastic scattering is not infrared safe — it is formally infinite in perturbation theory, which means in practice that it cannot be calculated. At first sight, the situation with particle multiplicity in e^+e^- annihilation has to be the same, since both processes are similarly affected by the soft gluon radiation. We will find, however, that the structure of the corrections enhanced by extra powers of $\ln 1/x$ turns out to be very different and the situation is changed drastically compared to DIS. The presentation in this section essentially follows a good review by B. Webber [3] where further details can be found.

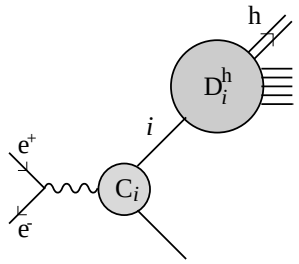


Fig. 30: Inclusive particle production in e^+e^- annihilation

The total inclusive cross section for the production of hadrons of type h in e^+e^- annihilation at c.m. energy $q = \sqrt{s}$ is defined as

$$F^h(x,s) = \frac{1}{\sigma_{\rm tot}} \frac{d\sigma}{dx}(e^+e^- \to hX) \tag{83}$$

where $x = 2E_h/\sqrt{s}$ is the fraction of energy of the virtual photon (in c.m. frame) transferred to the hadron, see Fig. 30. It can be represented as a sum of contributions from the different partons i

$$F^h(x,s) = \sum_i \int_x^1 \frac{dy}{y}\, C_i(y,\mu^2)\, D_i^h(x/y,\mu^2). \tag{84}$$

In the lowest order in QCD the coefficient function C_g for gluons is zero, while for quarks it is proportional to $\delta(1-y)$ with the coefficient given by the appropriate electromagnetic or electroweak coupling.

The $D_i^h(x,\mu^2)$ are called fragmentation functions and their scale dependence is given by the DGLAP equations that have the same form as in deep inelastic scattering:

$$\frac{d}{d\ln Q^2} D_i^h(x,Q^2) = \sum_j \int_x^1 \frac{dy}{y}\, P_{ij}(y,\alpha_s)\, D_j^h(x/y,Q^2)\,. \tag{85}$$

The similarity is further emphasized by the fact that the lowest order splitting functions P_{ij} are the same in DIS and fragmentation, although the higher-order terms are different. The effect of the splitting is in both cases to shift (at higher scales) the x distributions towards lower values. In particular, the $g \to g$ splitting function at small x values is given by Eq. (74). As a consequence, repeating the same simple calculation (75), (76) which yielded the conclusion that the average number of gluons in a proton is infinite (in QCD perturbation theory) (77), we conclude that the multiplicity of hadrons of type h in a jet initiated by a gluon at scale Q^2

$$N_g^h(Q^2) = \int_0^1 dx\, D_g^h(x, Q^2) \tag{86}$$

is infinite as well.

Similar to the DIS case, the higher-order contributions to the DGLAP splitting function in e^+e^- annihilation appear to be enhanced by powers of $\ln 1/x$, but here the similarity ends because the structure of such corrections turns out to be totally different:

$$P_{gg}(x) = \frac{3\alpha_s}{\pi x}\left[1 - 3\frac{\alpha_s}{\pi}\ln^2 x + 3\left(\frac{\alpha_s}{\pi}\ln^2 x\right)^2 + \mathcal{O}\left(\left(\frac{\alpha_s}{\pi}\ln^2 x\right)^3\right)\right]. \tag{87}$$

This expansion is known to all orders [35] and the result can be presented in a compact form as the answer (to leading logarithmic accuracy) for the resummed anomalous dimension

$$\gamma_{gg}(N) = \int_0^1 dx\, x^{N-1} P_{gg}(x) = \sqrt{\frac{(N-1)^2}{16} + \frac{3\alpha_s}{2\pi}} - \frac{1}{4}(N-1). \tag{88}$$

For fixed $N \neq 1$ and small α_s one can expand the square root in powers of $\alpha_s/(N-1)^2$, reproducing the series in (87).

Comparing (87) and (80) we see two differences: First, each power of α_s in (87) is accompanied by two powers of $\ln x$ instead of one, and, second, the series has become sign-alternating. The latter turns out to be crucial. As we have seen, the resummation of $\ln x$ corrections in deep inelastic scattering enhances the rise of the gluon density which acquires the factor (79) or (81) in addition to an overall $1/x$ behavior. In contrast to this, the resummation of $\ln x$ corrections in e^+e^- annihilation leads to a suppression of soft gluon emission so that their distribution becomes less singular.

To see this, one has to expand the resummed anomalous dimension (88) in a different way, assuming that $N \to 1$ and α_s is a constant:

$$\gamma_{gg}(N) = \sqrt{\frac{3\alpha_s}{2\pi}} - \frac{1}{4}(N-1) + \frac{1}{32}\sqrt{\frac{2\pi}{3\alpha_s}}\,(N-1)^2 + \mathcal{O}((1-N)^3)\,. \tag{89}$$

Absence of a pole in the anomalous dimension at $N = 1$ means that the integral of the splitting function is finite:

$$\int_0^1 dx\, P_{gg}(x) = \gamma_{gg}(1) = \sqrt{\frac{3\alpha_s}{2\pi}}. \tag{90}$$

Integrating the DGLAP equation (85) over x and using this result, one obtains the differential equation

$$Q^2\frac{d}{dQ^2}N_g^h(Q^2) = \sqrt{\frac{3\alpha_s(Q^2)}{2\pi}}N_g^h(Q^2). \tag{91}$$

which is easily solved (for running coupling $\alpha_s(Q^2) = 4\pi/(b\ln Q^2/\Lambda^2)$) to give for the average multiplicity

$$N^h(Q^2) \sim \exp\frac{4\pi}{b}\sqrt{\frac{6}{\pi\alpha_s(Q^2)}} = \exp\sqrt{\frac{24}{\pi b}\ln(Q^2/\Lambda^2)}. \tag{92}$$

In the same approximation one obtains for the small-x behavior

$$xF^h(x,Q^2) \sim \exp\left[-\frac{1}{2\sigma^2}\ln^2(x/x_0)\right], \tag{93}$$

where $\ln 1/x_0 = \pi/(b\alpha_s(Q^2)) \sim \frac{1}{4}\ln Q^2$ and σ is a certain constant $\sim \ln^{3/4} Q^2$, see [3]. Both predictions appear to be in excellent agreement with the data, see Fig.31.

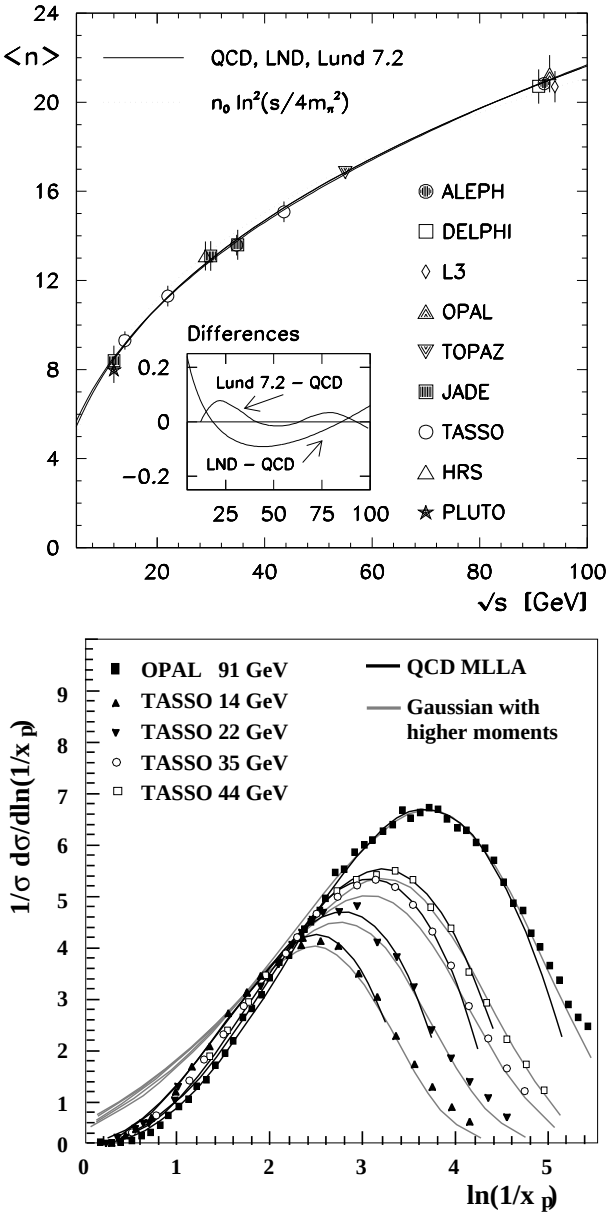


Fig. 31: Average multiplicity of charged hadrons in e^+e^- annihilation (upper panel) and their distribution over $\xi = \ln 1/x$ (lower panel) [3]

6 NONPERTURBATIVE EFFECTS IN HARD PROCESSES

Nonperturbative effects in QCD have two aspects. In soft processes they are dominating. There exist many theoretical phenomena (confinement, chiral symmetry breaking etc.) that cannot be explained

in perturbation theory and these phenomena are responsible for the hadron spectrum, parton distributions and fragmentation functions at low scales etc. On the other hand, nonperturbative effects in hard processes correspond to relatively small corrections to the perturbative predictions. A nonperturbative contribution to a short-distance observable can be thought of as a correction of order

$$\exp\left[-\frac{\text{const}}{\alpha_s(Q)}\right] \leftrightarrow \left(\frac{\Lambda}{Q}\right)^p. \tag{94}$$

The two representations in (94) are equivalent since $\alpha_s(Q) = 4\pi/(b\ln Q^2/\Lambda^2)$. The first way of writing emphasizes the fact that such correction cannot be obtained by perturbative expansion in the QCD coupling at the hard scale, while the second expression illustrates that such corrections are suppressed by powers of the large momentum.

One rationale for the study of nonperturbative power-suppressed (higher-twist) corrections to hard processes is, therefore, to quantify the accuracy of perturbative predictions. This is needed for precision tests of QCD and, more interestingly, making sure that the signals of electroweak or 'new' physics that is happening at very small distances are not lost in the strong interaction background. In this context, one is looking for nonperturbative corrections in order to minimize their influence.

Another motivation is that the hard momentum typically provides one with a 'handle' that allows to separate nonperturbative effects of different origin and make the problem simpler. From the celebrated successes of the perturbative description of hard processes we have learnt that hadrons at small distances are built of quarks and gluons, while studies of higher-twist corrections allow one to ask the next question: *How* are they built?

6.1 Short-distance expansion

Our task in this Section is, essentially, to find out the precise meaning of the statement 'calculable' that was used throughout the text. It is natural to start with the simplest example which is the duality sum rule for the photon polarization operator in Eq. (4). I have claimed that the l.h.s. of the sum rule is dominated by contributions of small distances and is calculable. But what does it really mean 'calculable'? The simplest interpretation might be that the corresponding perturbative calculation gives a finite result, free of infrared divergences. In this sense, a 'calculable' quantity is the one which is infrared safe. This is a bad definition, however, since it is not clear why a finite perturbative result is necessarily correct. In order to clarify this point we have to look 'inside' a perturbative calculation in some more detail.

The leading-order QCD expression for the polarization operator is well known to be

$$q^2\frac{d}{dq^2}\Pi(q^2) = \Pi_0\left[1+\frac{\alpha_s(Q)}{\pi}\right], \tag{95}$$

where $Q^2 = -q^2$ and Π_0 stands for an overall constant which is not important for what follows. The α_s correction in (95) corresponds to the contribution of the three Feynman diagrams shown in Fig. 32. The calculation involves the gluon propagator which is defined as the vacuum expectation value of the

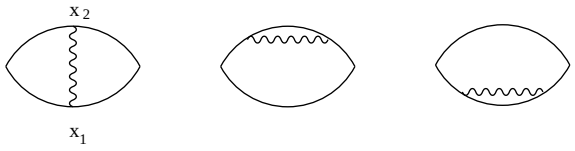


Fig. 32: The leading QCD correction to the correlation function of two electromagnetic currents

product of two gluon fields at a certain distance, and is given in perturbation theory by the standard

expression

$$G^{ab}_{\mu\nu}(x_1-x_2)=\langle 0|T\{A^a_\mu(x_1)A^b_\nu(x_2)\}|0\rangle=\delta^{ab}g_{\mu\nu}\int\frac{d^4k}{(2\pi)^4 i}\,e^{-ik(x_1-x_2)}\frac{1}{k^2+i\epsilon} \tag{96}$$

(in Feynman gauge). Each Feynman diagram in Fig. 32 (and their sum) can be written, schematically, as

$$\int d^4x\, f^{\mu\nu}(x,q)\, G_{\mu\nu}(x) \tag{97}$$

where $x=x_1-x_2$ (see the first Figure in Fig. 32) and it is assumed that all other propagators, vertices and other factors are included in the function $f(x,q)$. The color indices are not shown for simplicity of notation.

We suspect that the 'true' gluon Green function can differ from the expression in (96) by some nonperturbative corrections so that

$$G^{\rm QCD}_{\mu\nu}(x)=G^{\rm pert}_{\mu\nu}(x)+G^{\rm nonpert}_{\mu\nu}(x)\,, \tag{98}$$

where $G^{\rm pert}_{\mu\nu}(x)$ refers to the expression in (96) (and includes further α_s corrections) while $G^{\rm nonpert}_{\mu\nu}(x)$ stands for 'the rest'. We assume (at least this is the standard wisdom) that nonperturbative effects occur at large distances. In fact, the words 'large distance' are rather confusing for the following reason. A proton is certainly a large object but it can contain a small-distance substructure as well. A very naive example: let the proton be a ball with radius ~ 1 fm and a sharp boundary. Then its radius, mass, magnetic moment etc. are 'long distance' and nonperturbative, but e.g. the width of the boundary (if it is small) is 'short' and should be calculable! If you wish, this is a way to argue that the real proton does not have a sharp boundary — we do not find anything that could justify this in perturbation theory. To say this a bit more accurately — what matters is not the bare size but the structure: perturbative (predictable) phenomena correspond to small-scale substructures that can be created by the exchange of gluons with a short wave-length (high frequency) and nonperturbative (complicated) phenomena are due to gluons with a large wave-length (low frequency).

If we accept this, it follows that the nonperturbative contribution to the Green function of a gluon in (98) is related to the contribution of the slowly varying gluon fields and, therefore, by itself is a slowly varying function of the distance. Since, on the other hand, the integral in (97) is dominated by the integration region of small $|x|\sim 1/|q|$, we should not make a big error by replacing $G(x)$ under the integral by $G(x=0)$:

$$\int d^4x\, f^{\mu\nu}(x,q)\, G^{\rm nonpert}_{\mu\nu}(x)\simeq G^{\rm nonpert}_{\mu\nu}(0)\int d^4x\, f^{\mu\nu}(x,q)\,. \tag{99}$$

The remaining integral can be done while $G(x=0)$, by definition, presents a vacuum expectation value of the product of the two nonperturbative gluon fields at one point:

$$\langle 0|A_\mu A_\nu|0\rangle=\frac{1}{4}g_{\mu\nu}\langle 0|A^2|0\rangle. \tag{100}$$

Since $\Pi(q^2)$ carries no dimension, and since $\langle 0|A^2|0\rangle$ has dimension GeV^2 (because any vector field has dimension of a mass), it follows that the integral in (99) must have the dimension GeV^{-2} and, therefore, must be proportional to $1/q^2$ since there is no other dimensionful parameter. Without making any calculation, we can conclude that a potential nonperturbative correction to the polarization operator coming from 'softly varying' color fields has to be

- suppresssed as $1/q^2$ at large energies;
- proportional to the vacuum average of the gluon field squared.

and, therefore, ... it cannot exist! Indeed, a nonzero vacuum expectation value of A_μ^2 would break the gauge invariance (spontaneously) so that, roughly speaking, the vacuum would become 'colored'. We know that this does not happen (from experiment) and, therefore, $\langle 0|A^2|0\rangle = 0$. As a consequence, there cannot be any nonperturbative correction of order $1/q^2$ to the polarization operator.

In order to do better, we need to improve the approximation made in (99). Write

$$\int d^4x\, f^{\mu\nu}(x,q)\, G_{\mu\nu}^{\rm nonpert}(x) \;=\; \int d^4x\, f^{\mu\nu}(x,q) \left\{ G^{\rm nonpert}(0) + x_\alpha \partial^\alpha G^{\rm nonpert}(0) + \right. \\ \left. + \frac{1}{2} x_\alpha x_\beta \partial^\alpha \partial^\beta G^{\rm nonpert}(0) + \ldots \right\}_{\mu\nu}, \tag{101}$$

where $\partial^\alpha = d/dx^\alpha$ etc. Note that each power of x translates to the inverse power of energy $1/q$ by dimension counting so that the series is well converging (at large energies). On the other hand, derivatives of the Green function at $x = 0$ correspond to vacuum matrix elements of local operators built of two gluon fields and their derivatives. All what we have to do is to find out the operator of the lowest dimension that is gauge invariant and, therefore, can have a nonzero vacuum expectation value. This operator is $F_{\mu\nu}^2$ — the square of the gluon field-strength tensor — the one that appears in the QCD Lagrangian. It has dimension GeV^4 and, therefore, the corresponding nonperturbative correction to the polarization operator has to be suppressed as $1/q^4$. The only thing that remains to be calculated is the numerical coefficient. This requires some work [36] and the result reads

$$q^2 \frac{d}{dq^2} \Pi(q^2) = \Pi_0 \left[1 + \frac{\alpha_s(Q)}{\pi} + \ldots + \frac{1}{12q^4} \langle 0| \frac{\alpha_s}{\pi} F_{\mu\nu}^2 |0\rangle \right] + \mathcal{O}(1/q^6)\,. \tag{102}$$

To summarize, we made two assumptions: a) nonperturbative corrections are related to slowly varying color fields and b) QCD gauge invariance is not broken. We then used the fact that the polarization operator at large and negative q^2 is dominated by contributions of small distances and arrived to the conclusion that possible nonperturbative corrections to the polarisation operator (alias to the sum rule (4)) are given by a power series in $1/q^2$ in terms of vacuum matrix elements of gauge-invariant local operators (vacuum condensates). The coefficients in front of these operators can be calculated (in perturbation theory) while the vacuum condensates have to be taken as nonperturbative parameters. The premium is *power counting* ($1/q^4$) and *universality*: one can expect that the gluon condensate $\langle 0| \frac{\alpha_s}{\pi} F_{\mu\nu}^2 |0\rangle$ appears in the description of many physical processes; we can therefore sacrifice one measurement in order to extract the value of the condensate and use this value in order to refine the QCD predictions in other processes. The accepted value is $\langle 0| \frac{\alpha_s}{\pi} F_{\mu\nu}^2 |0\rangle \simeq 0.012$ GeV^4 with, probably, 50% error.

The construction is general and works in precisely the same way in all cases where one can reduce the problem to a short-distance expansion (using dispersion relations). The only thing that varies is the set of contributing operators, their dimension (and, hence, the power of the power correction) and, of course, the coefficients. To give another example, the leading nonperturbative correction to the Gross-Llewellyn Smith sum rule described in Sect. 4.1 equals to

$$\int_0^1 dx_B\, F_3^{\nu p + \bar\nu p}(x_B, Q^2) = 3 \left[1 - \frac{\alpha_s}{\pi} - \ldots - \frac{8}{27Q^2} \langle\langle O \rangle\rangle \right], \tag{103}$$

where $\langle\langle O \rangle\rangle$ is the reduced nucleon matrix element of the quark-antiquark-gluon operator

$$\langle N(p)| \bar\psi \widetilde{F}_{\mu\nu} \gamma_\nu \gamma_5 \psi |N(p)\rangle = 2 p_\mu \langle\langle O \rangle\rangle\,. \tag{104}$$

The value of $\langle\langle O \rangle\rangle$ was estimated using quark models and QCD sum rules [37] and turns out to be of the order of ~ 0.3 GeV^2.

Unfortunately, there is a subtlety that makes improvement of the QCD predictions by adding power corrections difficult in practice. The difficulty is related to the fact that the distinction between perturbative and nonperturbative contributions in (98) cannot be made rigorous. The reason for this will be explained in the next Section. For now, note that the derivation that I have sketched above never uses a 'nonperturbative' nature of a power correction, but rather is based on the observation that they are related to low frequency fluctuations of the color field. The contribution of low frequency fluctuations (perturbative or nonperturbative alike) to a short-distance dominated observable corresponds to a power correction and can be encoded in the matrix element of a certain local operator. This implies that the only possible way to understand the separation made in (98) is to *define* nonperturbative contributions to the propagator as those corresponding to fields of low frequency $|k| < \mu$, and therefore introduce the separation scale. The perturbative contribution is then defined as the contribution of high frequencies $|k| > \mu$ and, in order to avoid the double counting, one has to make this cutoff in the gluon propagator (96) explicitly.

As the result, the propagator becomes scale-dependent and so does the perturbative QCD prediction. On the other hand, the 'nonperturbative' contribution now depends on μ as well, and the dependence on μ has to cancel in the final result. Take the Gross-Llewellyn sum rule as an example. Since the nonperturbative contribution is proportional to $1/Q^2$, the only way such a cancelation can take place is if the modification of the perturbative contribution is of the order of $\sim 1/Q^2$ as well, and in such a case it has to be proportional to μ^2 in order to conserve dimensions:

$$\int\limits_0^1 dx_B\, F_3^{\nu p+\bar{\nu}p}(x_B,Q^2) = 3\left\{1-\frac{\alpha_s}{\pi}\left[1-c\cdot\frac{\mu^2}{Q^2}\right]-\ldots-\frac{8}{27Q^2}\langle\!\langle O(\mu)\rangle\!\rangle\right\}. \tag{105}$$

Moreover, the scale dependence of $\langle\!\langle O(\mu)\rangle\!\rangle$ has to be

$$\langle\!\langle O(\mu)\rangle\!\rangle = c\cdot\frac{27}{8}\alpha_s\,\mu^2 + \text{terms that depend on } \mu \text{ at most logarithmically}, \tag{106}$$

with the same constant c as in (105).

The trouble is that making calculations with an explicit momentum cutoff is very difficult in practice (if one goes beyond the leading order). The worst of all is that the cutoff procedure used to 'correct' the perturbation theory has to be exactly the same as the one used in the nonperturbative calculation of the matrix element (e.g. on the lattice) and since the techniques of perturbative and nonperturbative calculations are in many respect 'orthogonal' to each other, using the same cutoff (regularization) can be very inconvenient.

6.2 ... and beyond

In more complicated situations, one can turn things around and use the scale separation dependence of power corrections to estimate their magnitude. The approach has become known as the renormalon model of power corrections for historical reasons, and also because the necessary calculations are most easily done using the same (renormalon-related) techniques. The idea is, however, very simple and presents a straightforward extension of the standard practice to estimate the uncertainty of perturbative QCD predictions by their scale- and scheme-dependence.

Let us remind ourselves how it actually works. Any 'honest' perturbative calculation involves the 'bare' coupling normalized at a certain scale μ that serves as an ultraviolet cutoff. The leading-order correction to, say, the total cross section of e^+e^- annihilation is, therefore, equal to

$$R^{(1)}(q^2) = R^{(0)}\left[1+\frac{\alpha_s(\mu)}{\pi}\right]. \tag{107}$$

It depends on the scale μ inasmuch as $\alpha_s(\mu)$ does. We know that this dependence is spurious and is canceled by contributions of higher orders. In particular, if the calculation is extended to the second order, one obtains [4]

$$R^{(2)}(q^2) = R^{(0)} \left[1 + \frac{\alpha_s(\mu)}{\pi} + \left(1.917 \ln \frac{\mu^2}{q^2} + 1.411 \right) \left(\frac{\alpha_s(\mu)}{\pi} \right)^2 \right]. \tag{108}$$

(for $n_f = 5$ light flavors). If the scale is shifted from μ to μ_1, the coefficient in front of $(\alpha_s/\pi)^2$ is changed by the amount $1.917 \ln \mu_1^2/\mu^2$ and this correction is exactly compensated by the change of the coupling in the leading term

$$\frac{\alpha_s(\mu_1)}{\pi} = \frac{\alpha_s(\mu)}{\pi} - \frac{b}{4} \ln \frac{\mu_1^2}{\mu^2} \left(\frac{\alpha_s(\mu)}{\pi} \right)^2 + \dots \tag{109}$$

where $b/4 = (11 - 2/3 n_f)/4 \simeq 1.917$. Because of this, the logarithmic contribution $\sim \alpha_s^2 \ln \mu^2/q^2$ a) does not need to be calculated — it is easily guessed from the known $\sim \alpha_s$ term and b) is in fact spurious since its only rôle is to cancel the scale dependence of the coupling to the lowest order correction; it disappears if the scale it chosen to be $\mu = Q$.

The only new information in (108) compared to (107) is the scale-independent constant $r_2 = 1.411$. This constant cannot be guessed, it has to be calculated (and this is a hard work!). The question is whether one can, nevertheless, get a crude idea of how big this constant might be, and thus estimate an accuracy of the first-order result in (107) without doing the full two-loop calculation?

The standard procedure to estimate the accuracy of the existing perturbative prediction is by the uncertainty that is generated when the scale is taken within the interval of factor several times the hard scale Q. As we see from the above example, this dependence is in fact spurious and is canceled by the parts of higher-order contributions that are known without calculation. It has therefore no meaning by itself. Rather, an implicit assumption is being made (but kept secret, mostly) that the uncalculated contributions of higher orders of perturbation theory are of the same order of magnitude as the coeffcients in front of the spurious logarithms which are there to cancel the scale dependence. This is indeed true for our example: 1.917 in front of the logarithm is certainly of the same order of magnitude as 1.411.

It is easy to see that the same logic can be extended to obtain semiquantitative estimates of power corrections, the only difference being that the scale-dependence is power-like in this case. The scale-dependence of the subtracted perturbative contribution in (105) is a pure artefact of the scale separation and it has to cancel in the end. However, following the same logic, it is natural to assume that the true 'physical' contribution to the matrix element in the nonperturbative correction has to be of the same order of magnitude as the coefficient in front of the spurious $\sim \mu^2$ term (106). Since this coefficient refers to the low-momentum contributions in Feynman diagrams, it can be calculated order by order in perturbation theory (in $\alpha_s(\mu)$) and in this way a crude estimate of the expected nonperturbative effects can be done.

In the case of the GLS sum rule the corresponding estimates indicate that the nonperturbative correction is expected to be as big as the $\sim \alpha_s^3$ perturbative contribution, in agreement with model calculations. Much more interestingly, the same idea can be applied to estimate nonperturbative corrections to any infrared safe observable since it only relies on Feynman diagrams.

One useful application has been to the structure functions of deep inelastic scattering. In this case one looks for a parametrization of the form

$$F_2(x,Q) = F_2^{\mathrm{tw}-2}(x,Q) \left[1 + \frac{D(x)}{Q^2} + \mathcal{O}(1/Q^4) \right], \tag{110}$$

where $F^{\mathrm{tw}-2}(x,Q)$ is the leading-twist (perturbative) prediction. The aim is to find a motivated approximation for $D(x)$. To this end, remember that the perturbative result for $F_2(x,Q)$ is given by the

convolution of the coefficient function and the parton distribution, see (49). Perturbative corrections to the coefficient function can be calculated and the contributions of small momenta subtracted, resulting in the following structure, schematically

$$F_2(x,Q) = 2x\sum_q e_q^2\left[\delta(1-x) + \alpha_s\left(r(x) - c(x)\frac{\mu^2}{Q^2}\right)\right] \otimes f_{q/N}(x,Q) + \frac{C(x,\mu)}{Q^2}, \tag{111}$$

where $\otimes$ stands for the convolution and I have only shown the quark contribution, for simplicity. The function $c(x)$ was calculated [38] to be

$$c(x) = -\frac{4}{[1-x]_+} + 4 + 2x + 12x^2 - 9\delta(1-x) - \delta'(1-x)\,, \tag{112}$$

where the '+' prescription is defined as $\int_0^1 dx[f(x)]_+ t(x) = \int_0^1 dx f(x)[t(x) - t(1)]$ for any test function $t(x)$. We see that in order to cancel the quadratic scale dependence the function $C(x,\mu)$ has to behave as $C(x,\mu) \sim c(x) \otimes f_{q/N}(x)$, which translates to

$$D(x) = \frac{\Lambda^2}{F_2^{\rm tw-2}(x,Q)}\int_x^1 \frac{dy}{y} c(y) f_{q/N}(x/y,\mu)\,, \tag{113}$$

where Λ is a dimensionful constant of order $\Lambda_{\rm QCD}$ that cannot be fixed without additional assumptions. As shown in Fig. 33 the shape of the power correction indeed reproduces the data very well.

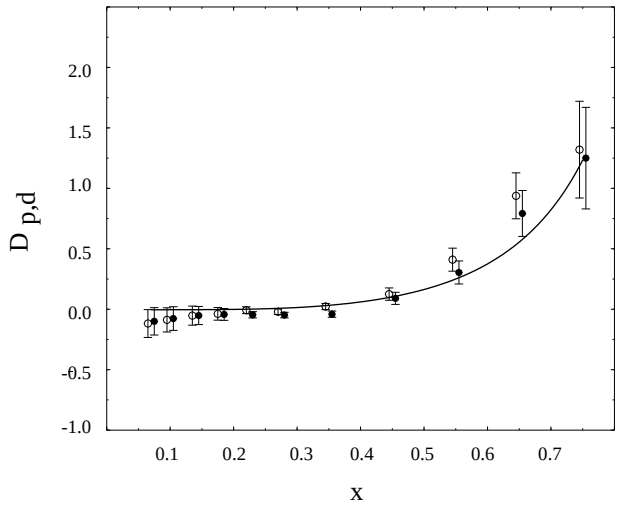


Fig. 33: Relative twist-4 contribution $D(x)$ defined in Eq. (113) to the proton (deuteron) structure function $F_2(x)$ (normalization is adjusted) compared to the proton (filled circles) and deuteron (empty circles) data [39].

In more complicated situations not only the shape but also the power of the power correction is unknown and can be guessed using the same method. As an example, consider the so-called event shape observables in e^+e^- annihilation. They are usually defined to be zero at tree level and therefore related to the matrix element for gluon emission $\gamma^* \to q\bar{q}g$ to leading order. The average value of a given event shape observable S can be calculated as

$$\langle S\rangle = \int d[p_i]|\mathcal{M}_{q\bar{q}g}|^2 S(p_i)\,, \tag{114}$$

where $d[p_i]$ stands for the Lorentz-invariant phase-space integration over all partons in the final state and $S(p_i)$ is the particular event shape function. It can be argued that the leading power correction — in this

case $\sim 1/Q$ — arises neither from emission of collinear partons, nor from soft quarks, but only from soft gluons. Introducing the energy fractions $x_i = 2(p_i q)/q^2$ and reserving x_3 for the gluon, this implies that the only relevant integration region is $x_3 \to 0$. Note that $x_1 + x_2 + x_3 = 2$. Therefore, the gluon emission can be calculated in the 'soft' approximation and the corresponding result is (see e.g. [4])

$$|\mathcal{M}_{q\bar{q}g}|^2 = 32 g_s^2 \frac{2}{(1-x_1)(1-x_2)}. \tag{115}$$

The phase space is, on the other hand

$$\int d[p_i] = \int_0^1 dx_1 dx_2 \, \theta(x_1 + x_2 - 1) \tag{116}$$

and since the matrix element (115) is singular as $x_{1,2} \to 1$ there is a potential logarithmic singularity. This is in agreement with what we already know: the cross section $\sigma_{e^+e^- \to q\bar{q}g}$ (8) is given by the expression in (114) with $S(p_i) \to 1$. This cross section is infrared divergent so that we had to introduce a gluon mass to make it finite. In the total cross section this divergence is canceled against a similar divergence in the $\sim \alpha_s$ virtual correction to $\sigma_{e^+e^- \to q\bar{q}}$. In event shapes, a finite result is achieved differently, by requiring that the event shape function $S(p_i)$ is constructed such that it vanishes when the gluon momentum goes to zero. The generic situation with the commonly used infrared-safe event shapes is that the suppression of soft gluons is *linear*: $S(x_i) \propto x_3$ as $x_3 \to 0$[4]. It is easy to see that this property implies a contribution of order μ/Q to the integral in (114) from gluons with energy less than μ unless there is some cancelation. By implication, one has to expect nonperturbative power corrections of order $1/Q$ to the event shapes since a perturbative calculation for small gluon momenta is illegal.

Linear nonperturbative power corrections are indeed observed for most of the event shape observables, see Fig. 34 for an example.

The last example that I want to consider concerns the quark masses. Since quarks are not observed in nature, their masses generally have to be considered as parameters in the QCD Lagrangian, on par with the strong coupling. Because the coupling is conventionally defined using dimensional regularization, it is most natural to also employ the $\overline{\rm MS}$ scheme for the mass definition, introducing running quark masses $\bar{m}_q(\mu_{\overline{\rm MS}})$ and fixing their values at a certain reference scale. This is indeed the procedure used to deal with light quarks $q = u, d, s$ and also heavy quarks provided the hard scale in the process is larger than or of order of the quark mass. On the other hand, using $\overline{\rm MS}$ heavy quark masses is not convenient in processes where the hard scale is significantly smaller than the mass of the quark itself. The reason for this is that the usual renormalization group expression

$$m_b(\mu) \simeq \left(\frac{\alpha_s(\mu)}{\alpha_s(m_b)}\right)^{4/b} m_b(m_b) \tag{117}$$

is physically irrelevant at $\mu \ll m_b$ because it is derived by assuming that μ is the ultraviolet cutoff and thus the largest scale. (There is, *formally*, nothing wrong with taking $\mu \ll m_b$ in Eq. (117). However, in calculations of physical observables, e.g. heavy quark decay rates, we obtain large perturbative corrections in higher orders in this case, because unphysical, large logarithms are generated.)

On the other hand, at small scales $\mu \ll m_b$ the heavy quark interacts with gluons through the color Coulomb potential $V(\vec{r}) = -(4/3)\alpha_s(1/r)/|\vec{r}|$. A restriction on the gluon momentum $|k| < \mu$ corresponds to the cutoff at large distances $r > 1/\mu$ so that the dependence of the mass parameter on μ at $\Lambda \ll \mu \ll m_b$ is in fact linear.

$$m_b(0) - m_b(\mu) \sim \alpha_s \mu \,, \tag{118}$$

see Fig. 35 for an illustration. The quantity $m(0)$ would correspond to a 'physical' quark mass if it existed and in this case would be defined as the location of the pole in the 'exact' quark propagator. In

[4] e.g. for the thrust $1 - T = 1 - \max(x_1, x_2)$ with $x_1 + x_2 + x_3 = 2$

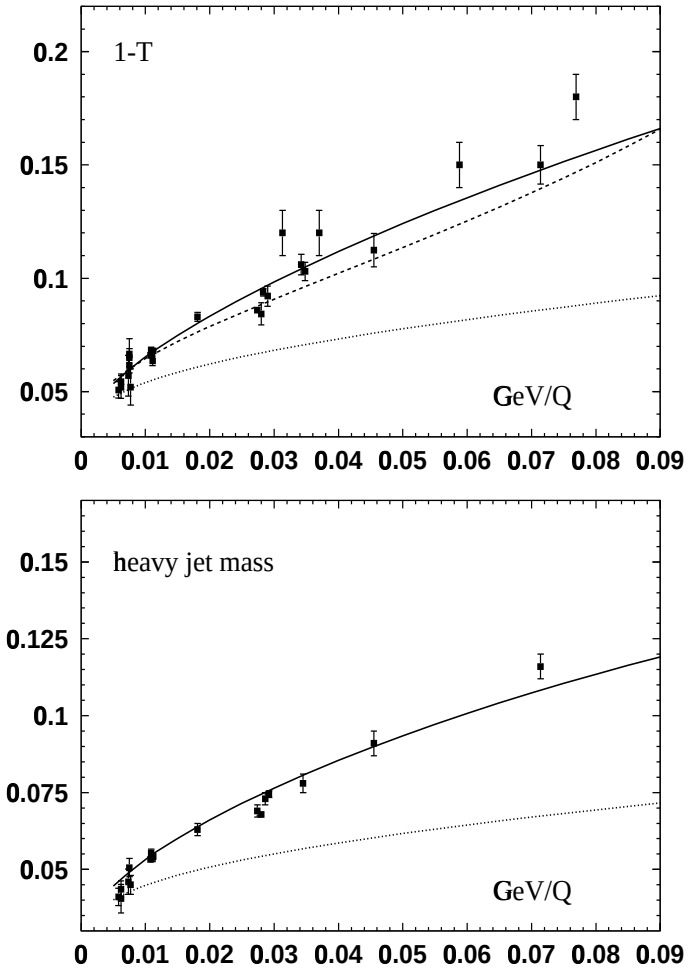


Fig. 34: Energy dependence of $\langle 1-T\rangle$ (upper panel) and the heavy jet mass $\langle M_H^2/Q^2\rangle$ (lower panel) plotted as a function of $1/Q$. [40] Dotted line: second-order perturbation theory with scale $\mu = Q$. Solid line: second-order perturbation theory with power correction $1/Q$ added. The dashed line shows second order perturbation theory at the very low scale $0.07Q$ with no power correction added. For both observables $\alpha_s(M_Z)$ has been fixed to 0.12.

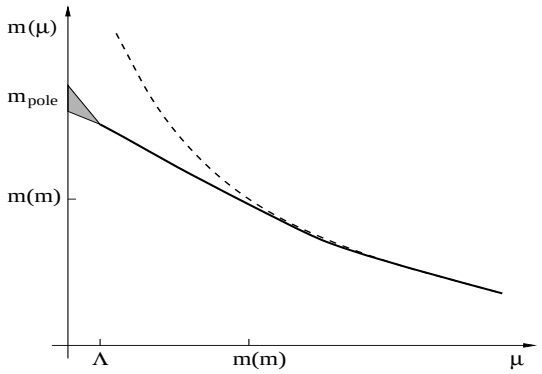


Fig. 35: A schematic representation of the scale dependence of a heavy quark mass defined with a 'physical' infrared cutoff (solid line) and in the dimensional regularization (dashed line). The shaded area illustrates uncertainties of the mass definition when the scale is driven to values of order $\Lambda_{\rm QCD}$

QCD, the best we can do is to calculate the perturbative quark propagator to a certain order, find position of its pole, and define the 'pole mass' by the series:

$$m_{\rm pole} = m_{\overline{\rm MS}}(m_{\overline{\rm MS}}) \left[1 + \sum_n r_n \alpha_s^n(m_{\overline{\rm MS}})\right]. \tag{119}$$

The pole mass has many nice features: it is infrared finite, gauge independent and independent on the renormalization scheme. However, it turns out [41] that the perturbative series in Eq. (119) is divergent and the sum of the series is ambiguous by an amount of order $\Lambda_{\rm QCD}$. It follows that the quark pole mass is perturbatively defined only to an accuracy

$$\delta m_{\rm pole} \sim \Lambda_{\rm QCD}\,, \tag{120}$$

the same as the 'physical' quark mass in Fig. 35. Roughly speaking, this means that it is not possible to cheat: since the 'physical' quark mass does not exist, there is no unique way to define it in the theory.

At first sight, this result looks very intriguing: how does the perturbation theory know about quark confinement? Let us postpone this question to the next section. For now, notice that this uncertainty has important practical implications as it means that the pole mass has to be eliminated as a parameter in calculations of physical observables, if these observables are less sensitive to the infrared region than the pole mass itself.[5] Important examples of such observables are inclusive heavy quark decays and top quark production near threshold in e^+e^- annihilation that we have discussed in Sect. 5.1. The 'PS' (potential-subtracted) top quark mass used in Fig. 24 corresponds to a particular version of the mass definition using a 'physical' cutoff $\mu = 20$ GeV in the spirit of Fig. 35. The value $\mu = 20$ GeV is chosen because this is the characteristic size of the would-be bound $t\bar{t}$ states. It is much less that $m_t \sim 175$ GeV so that the mass dependence on μ is linear and no large logarithms are generated in the calculation of the cross section, and, on the other hand, much bigger than $\Lambda_{\rm QCD}$ so that nonperturbative uncertainities are negligible.

6.3 High orders of perturbation theory

Let us return once more to the calculation of QCD corrections to the polarization operator in Fig. 32 and consider it from a somewhat different perspective. The three diagrams in Fig. 32 all have a single gluon line. Let k_μ be the gluon momentum and let us assume that the calculation of the diagrams is organized in such a way that the integration over the gluon momentum is the last one. Before this integration is

[5] Note also that a quantity that cannot be defined *in principle*, is not a good candidate for the particle data tables.

done, the (intermediate) expression for each diagram (or the sum of all the three) can look as follows:

$$\int_0^\infty \frac{dk^2}{k^2}\, f(k,Q)\, \alpha_s(k)\,. \tag{121}$$

where $f(k,Q)$ is a certain function. The factor $1/k^2$ can be thought of as coming from the gluon propagator, it is written explicitly for convenience and in order to emphasize that the whole expression has to carry no dimension. The expression in (121) parallels that in (97) (apart from that it is written in momentum space) and the only new point is that we indicate the argument of the QCD coupling explicitly: The scale of the coupling played no rôle in the discussion in Sect. 6.1, but it will be important for what follows.

As we discussed in Sect. 3.2, using the QCD coupling at the scale of the gluon virtuality (more generally, transverse momentum) is nothing but a compact way to take into account a certain class of Feynman diagrams in higher orders. This can be made explicit by the expansion

$$\alpha_s(k) = \frac{\alpha_s(Q)}{1+\beta_0\alpha_s(Q)\ln k^2/Q^2} = \sum_{n=0}^{\infty} \alpha_s^{n+1}(Q)(-\beta_0\,\ln k^2/Q^2)^n, \tag{122}$$

where $\beta_0 = b/(4\pi)$, $b = 11-2/3n_f$. The n-th term in the sum corresponds to the n-th order perturbative correction to the gluon propagator coming from the 'bubble' diagrams like those shown in Fig. 36[6]. In

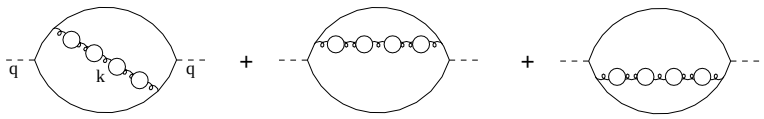


Fig. 36: The set of 'bubble' diagrams consists of all diagrams with any number of loops inserted into a single gluon line

fact, our aim in this section is exactly to investigate generic features of Feynman diagrams in high orders. The 'bubble' chains in Fig. 36 provide a convenient example because of their simple structure; their interpretation as part of the running coupling only serves as a motivation.

For fixed n, that is for the contribution of diagrams with n bubbles we get

$$\alpha_s^{n+1}(Q)(-\beta_0)^n \int_0^\infty \frac{dk^2}{k^2}\, f(k,Q)\ln^n(k^2/Q^2)\,. \tag{123}$$

We expect that large contributions to the integral come from either $k \ll Q$ or $k \gg Q$ because of the logarithm. Let us concentrate on the small momentum region. To this end, we can cut the integration from above at $k^2_{\max} = Q^2$ and expand the function $f(k,Q)$ in a Taylor series

$$f(k,Q) = \sum_m f_m \left(\frac{k^2}{Q^2}\right)^m\,. \tag{124}$$

For fixed m, the remaining integral is easily calculated

$$f_m\alpha_s^{n+1}(-\beta_0)^n \int_0^{Q^2} \frac{dk^2}{k^2}\left(\frac{k^2}{Q^2}\right)^m \ln^n(k^2/Q^2) \;= f_m\alpha_s^{n+1}(-\beta_0)^n \int_0^\infty dt\, e^{-mt}(-t)^n = \frac{f_m}{m}\,\alpha_s^{n+1}\left(\frac{\beta_0}{m}\right)^n n! \tag{125}$$

where $\alpha_s = \alpha_s(Q)$ and I have made the substitution $t = \ln Q^2/k^2$. Let us discuss this result in some detail.

[6]This interpretation is not exact in a nonabelain gauge theory like QCD because the coupling renormalization involves parts of the vertex corrections. One can show, however, that this is not a 'problem', see [42] for details.

We have found that a generic perturbative contribution to the polarization operator has the structure

$$q^2\frac{d}{dq^2}\Pi(q^2) = \Pi_0\left[1+\alpha_s\sum_{n=0}^{\infty} r_n\alpha_s^n\right], \tag{126}$$

and the coefficients r_n depend on the order of perturbation theory as

$$r_n \sim \text{const}\cdot a^n n!\,, \tag{127}$$

with $a=\beta_0/m$ and integer m. The most important thing to notice is $n!$ which is a very steep function.

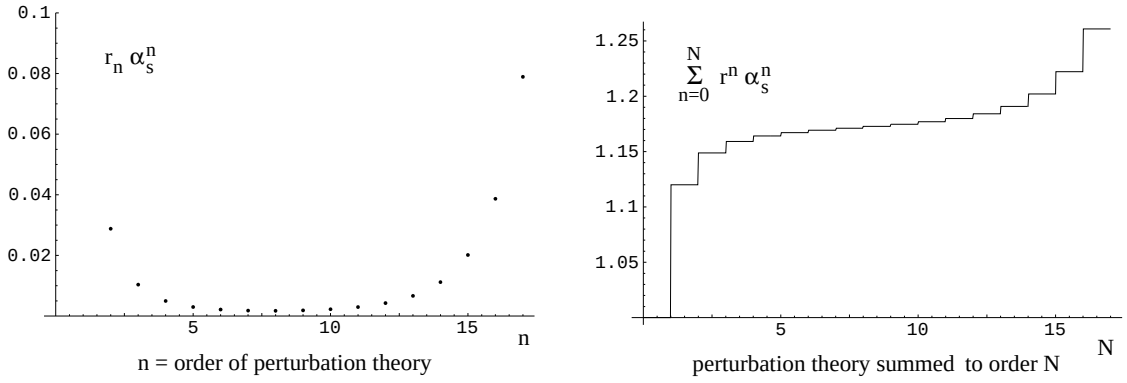


Fig. 37: Divergence of a perturbation theory. A toy-model example using (127) and the numerical value $a\cdot\alpha_s = 0.12$

For large values of n, sooner or later it will happen that the growth of $n!$ will overcome the smallness of the coupling and the perturbative series starts to diverge, as illustrated in Fig. 37. The divergence means that the sum of the series is not well defined; the best what we can do is to stop calculating further corrections as soon as $r_n\alpha_s^n$ reaches the minimum value and ascribe an error (uncertainty) to the sum of the series, of the order of the minimum term[7]. Assuming (127), the minimum term in the series occurs at $n\sim 1/\alpha_s$ and appears to be of the order (this is an elementary calculation that I advice the reader to do)

$$[r_n\alpha_s^n]_{\min} \sim e^{-\beta_0\alpha_s(Q)/m} = \left(\frac{\Lambda^2_{\rm QCD}}{Q^2}\right)^m. \tag{128}$$

We conclude that the sum of all orders in QCD perturbation theory is only defined to a power accuracy in the hard scale. This means, among other things, that it is not possible to distingush between perturbative and nonperturbative effects in a rigorous way: Nonperturbative corrections are needed in order to make the perturbative series well-defined.

Divergencies of the perturbative series that are related to (coupling) renormalization are called renormalons. Let us try to find a physical explanation for their appearance.

The Taylor expansion in (124) was motivated under the assumption that k^2/Q^2 is a good expansion parameter. Let us check whether this is indeed the case. To this end, calculate the average value of $\ln k^2/Q^2$ in the integral (125):

$$\langle\ln k^2/Q^2\rangle = \frac{\int_0^{Q^2}(dk^2/k^2)\,(k^2/Q^2)^m\ln^n(k^2/Q^2)\cdot\ln(k^2/Q^2)}{\int_0^{Q^2}(dk^2/k^2)\,(k^2/Q^2)^m\ln^n(k^2/Q^2)} = -\frac{n+1}{m}. \tag{129}$$

In other words, the average $k^2\sim Q^2\exp(-n/m)$ decreases exponentially with the order of perturbation theory. A perturbative calculation is justified for the polarization operator because the internal gluon

[7]This involves an assumption that the QCD perturbative series is an *asymptotic series* in mathematical sense.

momenta have to be of the order of the hard scale Q simply because there is no other dimensionful parameter. This argument applies to perturbative diagrams of any order. However, the *coefficient of proportionality* between k and Q can and does depend on n; what we have found is that this coefficient decreases rapidly so that even if Q is huge, in high orders one unavoidably hits the nonperturbative region.

Another way to see this: Imagine the perturbation theory is defined with an explicit cutoff at low gluon momenta as in (105): $k^2 > \mu^2$. Introducing this cutoff in (125) results in the integral $\int_0^{t_0} dt\, t^n\, e^{-mt}$ with $t_0 = \ln Q^2/\mu^2$ and for large n this integral is of order $\sim (\mu^2/Q^2)^m \ln^n(\mu^2/Q^2)/n$, without any $n!$ enhancement. It folows that (infrared) renormalons are in one-to-one correspondence with the contribution of low momenta in Feynman diagrams. In particular, if the perturbation theory is defined with an explicit infrared cutoff, there are no renormalons whatsoever, the series becomes convergent (for a sufficiently small coupling). Note also that the phenomenon of the renormalon divergence has nothing to do with the behavior of the strong coupling at large scales, but rather derives from the logarithmic scale behavior in the perturbative regime $\Lambda_{\rm QCD} \ll k \ll Q$.

As I have mentioned, there exists the second possibility of having large perturbative corrections from the large $k \gg Q$ in (123). Such contributions give rise to the so-called ultraviolet renormalons that I will not discuss in these lectures. They are not related to nonperturbative corrections but can be important in a different context, see [42] for a review and further references.

The connection between infrared renormalons and contributions of small momenta in Feynman diagrams can be used in both directions. As discussed in Sect. 6.1, small momentum contributions to the polarization operator can be isolated in terms of vacuum matrix elements of local operator. It follows that the leading infrared renormalon (the one corresponding to the largest uncertainty (128) alias the smallest m) must be in one-to-one correspondence to the contribution of the local operator of the lowest dimension, the gluon condensate. We have found that the gauge invariance forbids the existence of the gluon condensate of dimension two, so that the leading nonperturbative correction has to be suppressed by four powers of Q. Translating to the renormalon language, this result means that the gauge invariance forbids having a nonzero value for the coefficient $f_{m=1}$ in (124) so that the $m = 1$ contribution is missing and the ambiguity in the summation of the series (128) starts with $m = 2$, i.e. it is of order $\sim 1/Q^4$. The same result can of course be obtained by the direct calculation of the function $f(k, Q)$ that enters Eq. (121) but the whole point is that this calculation is not needed. Moreover, the absence of $m = 1$ term in the Taylor expansion of the gluon virtuality distribution function $f(k, Q)$ is a general property of *any* gauge-invariant set of Feynman diagrams of arbitrary order.

The argument can be reversed and used in the opposite direction: Any divergence of perturbation theory observed in the calculation of a certain class of Feynman diagrams (typically having chain structure) in general necessitates the existence of a power-suppressed nonperturbative correction unless there is some conspiration. In fact any divergence of the perturbative series generated through the bubble chain insertions is in one-to-one correspondence to the calculation of the simple one-gluon exchange diagram with an explicit infrared cutoff. From the technical point of view, the language of renormalons allows to introduce infrared regularization in the theory without the need to calculate integrals with an explicit infrared cutoff, in very much the same spirit as the dimensional regularization is used to handle logarithmic ultraviolet (and infrared) divergences. For example, the function $c(x)$ in (112) is calculated most easily using this trick.

Any phenomenological analysis of higher-twist corrections is done by a comparison of the measured value of a certain observable to the fixed-order (or resummed or etc.) perturbative prediction, and fitting the difference to a power-suppressed contribution, as indicated in Eq. (110). As follows from the above discussion, the higher-twist correction defined in this way (and this definition is the only possible one) presents an effective parametrisation of *both* nonperturbative effects and higher-order perturbative corrections. Because of this, the size of power corrections that are necessary to describe the data may depend significantly on the order of the perturbative series involved in the analysis — large power cor-

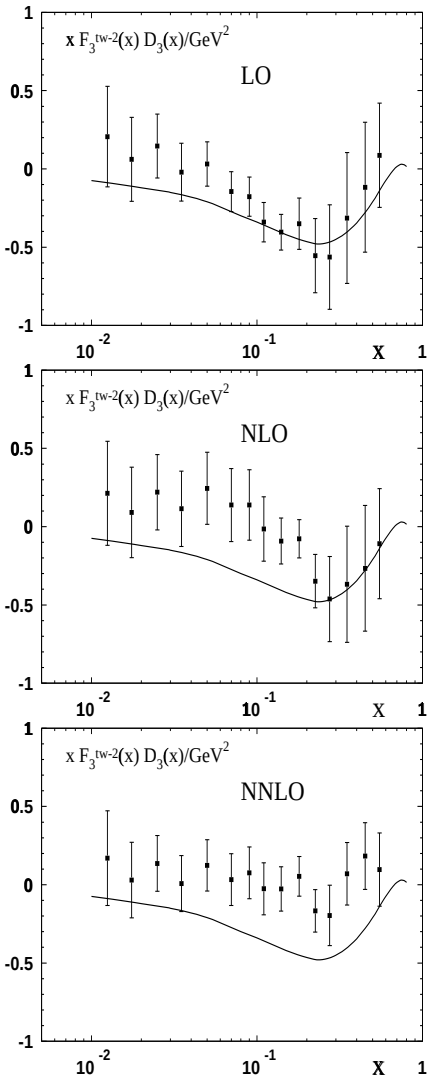


Fig. 38: Twist-4 correction to xF_3 as extracted from the (revised) CCFR data. The three plots show the effect of including leading order (LO), next-to-leading order (NLO) and next-to-next-to-leading order (NNLO) QCD corrections in the twist-2 term. The data points [43] are overlaid with the shape obtained from the "renormalon model" for the $1/Q^2$ power correction (the normalization is arbitatry).

rections can in many cases be interpreted as missing (uncalculated) contributions of higher orders of perturbation theory. The general trend for the necessary power corrections decreasing with the increasing accuracy of the perturbative description has indeed been observed in deep inelastic scattering, see Fig. 38.

6.4 Instantons

One may ask whether the QCD perturbation theory 'knows' through the renormalon divergences about all nonperturbative effects? The answer to this question is 'no'. Moreover, the basic assumption that *all* nonperturbative effects in QCD come from large distances is probably incorrect. We know that because we know an example of a nonperturbative phenomenon in QCD that is in fact short-distance.

Nonperturbative corrections can be thought of as being due to the complicated structure of the nonperturbative QCD vacuum, filled with virtual quanta. A fast quark propagates through the vacuum and acquires, e.g. the effective mass in the similar fashion as electrons acquire an effective mass in metals. The usual logic says that nonperturbative vacuum fluctuations are due to slowly varying color

fields and in this case they can be taken into account in some kind of a mean-field approximation: The operator product expansion is nothing but the mean-field approximation, adapted for a field theory.

Existence of instantons show that this logic has a caveat. An instanton is a nonperturbative vacuum fluctuation (nonperturbative means that it cannot be built from a finite number of quanta) that is 'almost' gauge artefact: The instanton field looks as a gauge transformation so one would be prepared to declare it gauge-equivalent to a zero field if there would be not a problem: this transformation is *singular*. We are not allowed to make singular gauge transformations, so we cannot get rid of an instanton: the best we could do is to move the singularity to another point or to the spacial infinity, but it does not help. Instantons have an integer quantum number which is called topological charge and many intriguing features. For example, the Dirac equation for a massless left-handed quark in the field of an instanton has a solution with zero energy, while for the antiinstanton there is a similar solution for a right-handed quark. This means that if an instanton is added (or removed) from the vacuum, the total amount of available left-handed minus right-handed quark states changes by one: the vacuum is changing its chirality. A detailed discussion of instantons goes beyond the scope of these lectures, see [44] for some very good reviews.

Instantons have a size and the probability to find an instanton of small size $\rho = 1/Q$ in the QCD vacuum can be calculated:

$$d(\rho = 1/Q) \sim \exp\left[-\frac{2\pi}{\alpha_s(Q)}\right] = \left(\frac{\Lambda^2_{\rm QCD}}{Q^2}\right)^{b/2}. \qquad (130)$$

Since $b = 11 - 2/3n_f$ is a large number, this probability turns out to be very small, of order $1/Q^8$ – $1/Q^9$, and this is what saves us: instanton-induced contributions to a generic hard process are negligible compared to the 'usual' nonperturbative effects related to slow-varying vacuum fluctuations[8].

It turns out, however, that nonperturbative production of final states with a very large hadron multiplicity that I have mentioned in Sect. 2.3 can be mediated by an instanton field [45] and the corresponding experimental program is being pursued at HERA [46].

ACKNOWLEDGEMENTS

I would like to thank the organizers for the invitation to give these lectures, and for providing a very nice atmosphere during the school. Special thanks are due to A. Lenz and E. Stein for the reading and the comments on the manuscript.

References

recommended reading:

[1] J. C. Collins, D. E. Soper and G. Sterman, *Factorization Of Hard Processes In QCD*, In: Mueller, A.H. (ed.), Perturbative Quantum Chromodynamics, (World Scientific Publ., 1989).

[2] G. Sterman, *Partons, factorization and resummation*, hep-ph/9606312.

[3] B. R. Webber, *Hadronization*, hep-ph/9411384.

[4] R. K. Ellis, W. J. Stirling and B. R. Webber, *Cambridge, UK: Univ. Pr. (1996) 435 p. (Cambridge monographs on particle physics, nuclear physics and cosmology: 8)*.

[8] In the present context we are not interested in possible effects of large instantons which are just part of an overall nonperturbative background.

[5] M. Beneke, in *Proc. of the 19th Intl. Symp. on Photon and Lepton Interactions at High Energy LP99* ed. J.A. Jaros and M.E. Peskin, eConf**C990809** (2000) 632 [hep-ph/0001134].

[6] A. H. Mueller, *Small-x physics, high parton densities and parton saturation in QCD*, hep-ph/9911289.

other references:

[7] R. Akers *et al.* [OPAL Collaboration], Z. Phys. **C68** (1995) 531.

[8] P. L. Anthony *et al.* [E155 Collaboration], Phys. Lett. **B493** (2000) 19.

[9] J. Yu [CCFR / NuTeV Collaboration], hep-ex/9806030.

[10] S. Bethke, J. Phys. G**G26** (2000) R27.

[11] F. V. Tkachev, Phys. Lett. **B100** (1981) 65; K. G. Chetyrkin and F. V. Tkachev, Nucl. Phys. **B192** (1981) 159.

[12] K. G. Chetyrkin and M. Steinhauser, Nucl. Phys. **B573** (2000) 617.

[13] K. Melnikov and T. V. Ritbergen, Phys. Lett. **B482** (2000) 99.

[14] V. A. Smirnov and O. L. Veretin, Nucl. Phys. **B566** (2000) 469.

[15] S. A. Larin, P. Nogueira, T. van Ritbergen and J. A. Vermaseren, Nucl. Phys. **B492** (1997) 338.

[16] A. Berera and D. E. Soper, Phys. Rev. D **53** (1996) 6162;
M. Grazzini, L. Trentadue and G. Veneziano, Nucl. Phys. **B519** (1998) 394;
J. C. Collins, Phys. Rev. D **57** (1998) 3051.

[17] V. N. Gribov, Sov. Phys. JETP**30** (1970) 709.

[18] X. Ji, Phys. Rev. D **55** (1997) 7114;
A. V. Radyushkin, Phys. Lett. **B380** (1996) 417.
J. C. Collins and A. Freund, Phys. Rev. D **59** (1999) 074009.

[19] J. C. Collins, L. Frankfurt and M. Strikman, Phys. Rev. D **56** (1997) 2982.

[20] M. Beneke, G. Buchalla, M. Neubert and C. T. Sachrajda, Phys. Rev. Lett. **83** (1999) 1914; Nucl. Phys. **B591** (2000) 313.

[21] V. S. Fadin and V. A. Khoze, JETP Lett. **46** (1987) 525; Sov. J. Nucl. Phys. **48** (1988) 309.

[22] M. Beneke, hep-ph/9911490.

[23] M. Beneke, A. Signer and V.A. Smirnov, *Phys. Lett.* B **454** 137 (1999).

[24] G. P. Korchemsky and S. Tafat, JHEP**0010** (2000) 010.

[25] V. V. Sudakov, Sov. Phys. JETP**3** (1956) 65.

[26] M. Derrick *et al.* [ZEUS Collaboration], Z. Phys. **C72** (1996) 399.

[27] A. De Rujula, S. L. Glashow, H. D. Politzer, S. B. Treiman, F. Wilczek and A. Zee, Phys. Rev. D **10** (1974) 1649.

[28] E. A. Kuraev, L. N. Lipatov and V. S. Fadin, Sov. Phys. JETP**45** (1977) 199.

[29] I. I. Balitsky and L. N. Lipatov, Sov. J. Nucl. Phys. **28** (1978) 822.

[30] V. S. Fadin and L. N. Lipatov, Nucl. Phys. **B406** (1993) 259; Phys. Lett. **B429** (1998) 127.

[31] L. N. Lipatov, Phys. Lett. **B309** (1993) 394; L. D. Faddeev and G. P. Korchemsky, Phys. Lett. **B342** (1995) 311.

[32] L. N. Lipatov, Phys. Rept. **320** (1999) 249; I. Balitsky, hep-ph/0101042.

[33] N. Nikolaev and B. G. Zakharov, Z. Phys. **C53** (1992) 331;
A. H. Mueller, Nucl. Phys. **B415** (1994) 373;
A. H. Mueller and B. Patel, Nucl. Phys. **B425** (1994) 471.

[34] L. McLerran and R. Venugopalan, Phys. Rev. D **49** (1994) 2233; Phys. Rev. D **50** (1994) 2225;

[35] A. H. Mueller, Phys. Lett. **B104** (1981) 161;
B. I. Ermolaev and V. S. Fadin, JETP Lett. **33** (1981) 269.

[36] M. A. Shifman, A. I. Vainshtein and V. I. Zakharov, Nucl. Phys. **B147** (1979) 385.

[37] V. M. Braun and A. V. Kolesnichenko, Nucl. Phys. **B283** (1987) 723.

[38] M. Beneke and V. M. Braun, Nucl. Phys. **B454** (1995) 253;
Y. L. Dokshitzer, G. Marchesini and B. R. Webber, Nucl. Phys. **B469** (1996) 93.

[39] M. Virchaux and A. Milsztajn, Phys. Lett. **B274** (1992) 221.

[40] P.A. Movilla Fernández *et al.* [The JADE Collaboration], *Eur. Phys. J.* C **1**, 461 (1998).

[41] I. I. Bigi, M. A. Shifman, N. G. Uraltsev and A. I. Vainshtein, Phys. Rev. D **50** (1994) 2234;
M. Beneke and V. M. Braun, Nucl. Phys. **B426** (1994) 301.

[42] M. Beneke, Phys. Rept. **317** (1999) 1;
M. Beneke and V. M. Braun, hep-ph/0010208.

[43] A.L. Kataev, A.V. Kotikov, G. Parente and A.V. Sidorov, in *Proceedings of the 32nd Rencontres de Moriond: QCD and High-Energy Hadronic Interactions*, Les Arcs, France, 22-29 Mar 1997 [hep-ph/9706534].

[44] S. Coleman, *The uses of instantons*, in Aspects of Symmetry: selected Erice lectures of Sidney Coleman. Cambridge Univ. Press, 1985;
A. I. Vainshtein, V. I. Zakharov, V. A. Novikov and M. A. Shifman, *ABC of instantons*, in: Instantons in Gauge Theories. Edited by M. Shifman. World Scientific, 1994;
T. Schafer and E. V. Shuryak, Rev. Mod. Phys. **70** (1998) 323.

[45] A. Ringwald, Nucl. Phys. **B330** (1990) 1;
I. I. Balitsky and V. M. Braun, Phys. Lett. **B314** (1993) 237;
A. Ringwald and F. Schrempp, Phys. Lett. **B438** (1998) 217.

[46] E. A. Wolf, hep-ex/0101038.

FLAVOUR PHYSICS AND CP VIOLATION

Gustavo C. Branco
CERN, Geneva, Switzerland and Instituto Superior Tecnico, Lisboa, Portugal

Abstract
These lectures present selected topics in Flavour Physics, including fermion masses and mixing and CP violation. We discuss CP violation in the Standard Model and Beyond and emphasize the important role of CP asymmetries in neutral B mesons decays in providing stringent tests of the Standard Model, with the potential for discovering New Physics.

1 INTRODUCTION

The Standard Model (SM) is at present in agreement with all experimental data, with the exception of the recently discovered neutrino oscillations, pointing towards non-vanishing neutrino masses [1]. However, independently of the question of neutrino masses and mixing, it is clear that the SM cannot be regarded as the final theory of the fundamental interactions. A good part of the motivation to go beyond the SM, has to do with the so called flavour problem and CP violation. These lectures describe some aspects of Flavour Physics and CP Violation, covering a selected number of topics. A thorough coverage of the subject can be found in the recent books [2], [3], [4]. as well as in working group reports [5], [6] and in recent reviews [7].

In order to construct a gauge theory unifying the electromagnetic, weak and strong interactions, one has to choose the gauge symmetry group, the representations of the fermion and scalar fields and the pattern of symmetry breaking. In the Standard Model the choices are:

• Gauge group: $U(1) \times SU(2)_L \times SU(3)_C$

• Representations:

Quarks:

$$Q_L^0 \equiv \begin{bmatrix} u^0 \\ d^0 \end{bmatrix}_{jL} \qquad (1/6, 2, 3)$$

Leptons:

$$L^0 \equiv \begin{bmatrix} \nu^0 \\ l^0 \end{bmatrix}_{jL} \qquad (-1/2, 2, 1)$$

Scalars:

$$\Phi \equiv \begin{bmatrix} \phi^+ \\ \phi^0 \end{bmatrix} \qquad (1/2, 2, 1)$$

where L, R denote left-handed (lh) and right-handed (rh) fermions, with chirality (-1), (+1) respectively. The numbers in parenthesis indicate the transformation properties under $U(1), SU(2)_L$ and $SU3)_C$, respectively. Note that no right-handed neutrinos are introduced in the Standard Model. The index j is a family or generation index. The replication of fermion families, with all families transforming in the same way under the gauge group, is one of the striking features of the fundamental interactions. The quark fields are written with the upper index (0) to emphasize that u^0, d^0 are eigenstates of the gauge interactions, i.e. all gauge currents are flavour diagonal when expressed in terms of these fields.

• Pattern of Symmetry Breaking

The most general Higgs potential consistent with gauge invariance and renormalizability is given by:

$$V(\Phi) = \frac{1}{2}\mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \tag{1}$$

where $\lambda > 0$ so that the potential is bounded from below. If $\mu^2 > 0$ the minimum of the potential is at $< 0|\Phi^\dagger\Phi|0 >= 0$. In this case the minimum respects the gauge symmetry. For $\mu^2 < 0$ the minimum is at $< 0|\Phi^\dagger\Phi|0 >= v^2/2$ and without loss of generality one may write:

$$< 0|\Phi|0 >= \begin{bmatrix} 0 \\ v/\sqrt{2} \end{bmatrix} \tag{2}$$

In this case the gauge symmetry is spontaneously broken into $U(1)_{em} \times SU(3)_C$. Gauge invariance does not allow for the introduction of direct mass terms for the gauge fields B^μ, ${W_k}^\mu (k = 1, 2, 3), {G_\alpha}^\mu (\alpha = 1, ...8)$ associated to $U(1), SU(2)_L$ and, $SU(3)_C$, respectively. After spontaneous symmetry breaking, the fields :

$$W^{\mu^\pm} = \frac{{W_1}^\mu \mp i{W_2}^\mu}{\sqrt{2}}; \;\; Z_\mu = -\sin\theta_W B^\mu + \cos\theta_W {W_3}^\mu \tag{3}$$

acquire masses ${M_W}^2 = 1/2g^2v^2$, ${M_Z}^2 = 1/2(g^2 + {g'}^2)v^2$ while the photon:

$$A_\mu = \cos\theta_W B^\mu + \sin\theta_W {W_3}^\mu \tag{4}$$

together with the gluon fields ${G_\alpha}^\mu$ remain massless.

2 STANDARD MODEL INTERACTIONS

2.1 Gauge Interactions

In these lectures we will concentrate our attention on the electroweak sector of the SM. $SU(2)$ gauge invariance requires the introduction of a covariant derivative given by:

$$D^\mu \equiv \partial^\mu - igW_j{}^\mu T_j - ig\prime B^\mu Y \tag{5}$$

where $T_j \equiv 1/2\tau_j$ and Y denote the $SU(2)$ and $U(1)$ generators, respectively. From the kinetic energy part of the Lagrangian, one obtains the electromagnetic interactions, as well as the charged current and neutral current interactions, which are given by:

$$\begin{gathered} \mathcal{L}_{em} = -eJ^\mu_{em}A_\mu \\ J^\mu_{em} = \textstyle\sum_j \frac{2}{3}\left(\overline{u^\circ_{Lj}}\gamma^\mu u^\circ_{Lj} + \overline{u^\circ_{Rj}}\gamma^\mu u^\circ_{Rj}\right) - \frac{1}{3}\left(\overline{d^\circ_{Lj}}\gamma^\mu d^\circ_{Lj} + \overline{d^\circ_{Rj}}\gamma^\mu d^\circ_{Rj}\right) - \left(\overline{l^\circ_{Lj}}\gamma^\mu l^\circ_{Lj} + \overline{l^\circ_{Rj}}\gamma^\mu l^\circ_{Rj}\right) \end{gathered} \tag{6}$$

$$\mathcal{L}_Z = \frac{g}{cos\theta_W}\left[J^\mu_3 - 2\sin^2\theta_W J^\mu_{em}\right] Z_\mu \tag{7}$$

$$J^\mu_3 = \sum_j \frac{1}{2}\left(\overline{u^\circ_{Lj}}\gamma^\mu u^\circ_{Lj} - \overline{d^\circ_{Lj}}\gamma^\mu d^\circ_{Lj}\right)$$

$$\mathcal{L}_W = \frac{g}{\sqrt{2}}\sum_j \left(\overline{u^\circ_{Lj}}\gamma^\mu d^\circ_{Lj} + \overline{\nu^\circ_{Lj}}\gamma^\mu l^\circ_{Lj}\right) W^+_\mu + h.c. \tag{8}$$

2.2 The Yukawa Interactions

The Yukawa interactions can be written as :

$$\mathcal{L}_Y = \sum_{i,j}\left\{ Y^d_{ij}(\overline{u}^0_i\overline{d}^0_i)_L \begin{bmatrix} \phi^+ \\ \phi^0 \end{bmatrix} d^0_{jR} + Y^u_{ij}(\overline{u}^0_i\overline{d}^0_i)_L \begin{bmatrix} \phi^{0*} \\ -\phi^- \end{bmatrix} u^0_{jR} + Y^\ell_{ij}(\overline{\nu}^0_i\overline{\ell}^0_i)_L \begin{bmatrix} \phi^+ \\ \phi^0 \end{bmatrix} \ell^0_{jR} \right\} + h.c. \tag{9}$$

The three 3×3 matrices Y^d, Y^u, Y^ℓ are arbitrary complex matrices in flavour space. One of the most important features of the SM is the fact that gauge invariance prevents the introduction of fermion mass terms in the Lagrangian. This results from the fact that under $SU(2)$, lh fermions transform as doublets while their rh components transform as singlets. Fermion masses are generated from Yukawa interactions, once spontaneous symmetry breaking takes place. Making the expansion of Φ around the vacuum expectation value (vev):

$$\Phi = \begin{bmatrix} \phi^+ \\ v + H + iG \end{bmatrix} \tag{10}$$

from Eqs. (9), (10) one obtains the mass terms for the up quarks, the down quarks, and the charged leptons, as well as the interactions of the physical Higgs particle H with quarks and leptons. The fields ϕ^+ and G are absorbed as longitudinal components of W^+ and Z_μ , through the Higgs mechanism. The fermion mass terms can be written as

$$\mathcal{L}_{\text{mass}} = \overline{d}^0_L M_d d^0_R + \overline{u}^0_L M_u u^0_R + \overline{\ell}^0_L M_\ell \ell^0_R + h.c. \tag{11}$$

with $M_d = Y^d \frac{v}{\sqrt{2}}; M_u = Y^u \frac{v}{\sqrt{2}}; M_\ell = Y^\ell \frac{v}{\sqrt{2}}$, where we have suppressed the flavour indices. Due to the arbitrariness of Yukawa couplings, the matrices $M_d, M_u,$ are arbitrary complex matrices which are diagonalized by the bi-unitary transformations:

$$\begin{array}{lll} u^0_L = U^u_L u_L & ; & d^0_L = U^d_L d_L \\ u^0_R = U^u_R u_R & ; & d^0_R = U^d_R d_R \end{array} \tag{12}$$

where u, d denote the mass eigenstates:

$$\begin{array}{lll} U^{u\dagger}_L M_u U^u_R = D_u & ; & U^{u\dagger}_L H_u U^u_L = D^2_u \\ U^{d\dagger}_L M_d U^d_R = D_d & ; & U^{d\dagger}_L H_d U^u_L = D^2_d \end{array} \tag{13}$$

with $D_u = \text{diag}(m_u, m_c, m_t)$, $D_d = \text{diag}(m_d, m_s, m_b)$, and $H_u = M_u M^\dagger_u$, $H_d = M_d M^\dagger_d$.

2.3 Weak-Basis Transformations

In the limit where one switches off the Yukawa interactions, the Lagrangian of the SM acquires a very large global symmetry, which reflects the replication of fermion families. It is useful to define "Weak-Basis transformations" given by:

$$\begin{array}{lll} u^0_L \to u^{0\prime}_L = V^q_L u^0_L & ; & u^0_R \to u^{0\prime}_R = V^u_R u^0_R \\ d^0_L \to d^{0\prime}_L = V^q_L d^0_L & ; & d^0_R \to d^{0\prime}_R = V^d_R d^0_R \\ l^0_L \to l^{0\prime}_L = V^l_L l^0_L & ; & l^0_R \to l^{0\prime}_R = V^l_R l^0_R \\ \nu^0_L \to \nu^{0\prime}_L = V^l_L \nu^0_L & ; & \end{array} \tag{14}$$

Since V^q_L, V^u_R, V^d_R, V^l_L, V^l_R are arbitrary unitary matrices, the transformations of Eq. (14) form a $U(3)^5$ group. The crucial point is that all the gauge currents remain diagonal under a Weak-Basis (WB) transformation, while the Yukawa couplings transform in the following way:

$$\begin{array}{c} Y^d \to Y'^d = V^{q\dagger}_L Y^d V^d_R \\ Y^u \to Y'^u = V^{q\dagger}_L Y^d V^u_R \\ Y^l \to Y'^l = V^{l\dagger}_L Y^d V^l_R \end{array} \tag{15}$$

It is clear that the set of Yukawa matrices $Y'^{d,u,l}$ contains the same physics as the set $Y^{d,u,l}$. Since the fermion mass matrices are proportional to the Yukawa couplings, it is obvious that under a WB transformation $M_{d,u,l}$ transform in the same way as the corresponding Yukawa matrices. The Yukawa matrices are complex and arbitrary, so they contain 18x3=54 parameters. Of course, the number of physical parameters contained in the Yukawa matrices is much smaller, due to the redundancy resulting from the freedom to make WB transformations.

2.4 Quark Mixing in Charged Currents: the V_{CKM} Matrix

In a WB all gauge currents are diagonal in flavour space. Let us now express the quark charged currents in terms of mass eigenstates. From Eqs. (11) and (12) one obtains:

$$\mathcal{L}_{c.c.} = g\ [\overline{u}\ \overline{c}\ \overline{t}]_L V_{CKM} \begin{bmatrix} d \\ s \\ b \end{bmatrix}_L + h.c. \tag{16}$$

where $V_{CKM} = U_L^{u^\dagger} U_L^d$. Using standard notation, V_{CKM} can be written as:

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \tag{17}$$

From Eq. (12) it is clear that in the SM, the appearance of a non-trivial V_{CKM} is entirely natural, just reflecting the fact that M_u, M_d are independent matrices and thus diagonalized by different unitary matrices. In the mass eigenstate basis the quark mass terms are:

$$\mathcal{L}_{\text{mass}} = \sum_i m_i \overline{d}_i d_i + \sum_k m_k \overline{u}_k u_k \tag{18}$$

where $i = d, s, b$ and $k = u, c, t$. One has the freedom to rephase the quark fields by a phase transformation:

$$u_i = exp(i\varphi_i^u) u_i'; \quad d_i = exp(i\varphi_i^d) d_i' \tag{19}$$

Under these phase transformations the mass terms remain invariant while V_{CKM} transforms as:

$$V_{jk} = exp\left[-i(\varphi_j^u - \varphi_k^d)\right] V_{jk}' \tag{20}$$

Obviously, the individual phases of V_{CKM} have no physical meaning. Only functions of V_{CKM} which are rephasing invariant have physical meaning and can be measured. The simplest examples are the moduli and the invariant quartets:

$$Q_{ijk\ell} = V_{ij} V_{k\ell} V_{i\ell}^* V_{kj}^* \tag{21}$$

It can be readily seen that the invariants of higher order can be written as a product of quartets divided by some moduli. For example:

$$V_{ij} V_{k\ell} V_{mn} V_{i\ell}^* V_{kn}^* V_{mj}^* = Q_{ijk\ell} Q_{kjmn} / |V_{kj}|^2 \tag{22}$$

2.4.1 Physical Parameters in V_{CKM}

For n generations, the V_{CKM} matrix is an n-dimensional unitary matrix therefore it has n^2 independent parameters. However, due to the freedom of making rephasing transformations and taking into account that an overall phase transformation of all quarks leaves V_{CKM} invariant, one can eliminate $(2n-1)$ phases from V_{CKM}. Taking into account that $1/2n(n-1)$ angles are needed to define an orthogonal matrix, one concludes that the number of physical phases contained in V is given by

$$N_\phi^{SM} = n^2 - \frac{1}{2}n(n-1) - (2n-1) = \frac{1}{2}(n-1)(n-2) \tag{23}$$

The total number of physical parameters entering in V is:

$$N_p^{SM} = \frac{1}{2}n(n-1) + N_\phi^{SM} = (n-1)^2 \tag{24}$$

Note that in the derivation of Eq. (23) we have made the implicit assumption that no two quarks of equal charge have degenerate masses. The point is that if two quarks of the same charge have degenerate masses, the the most general set of transformations which leaves the mass terms invariant is no longer given by Eq. (19) since it should include the possibility of making a U(2) transformation acting on the two degenerate quarks. It can be readily seen that in the three generations SM all phases can be removed from V_{CKM} in the limit where two quarks of the same charge become degenerate in mass. We will come back to this question in Sec. 2.5, where we study WB invariants and CP Violation. It should be emphasized that the counting of physical phases given by Eq. (23) is still valid in the chiral limit, where $m_u = m_d = 0$.

From Eq. (23), it follows that in the SM for $n \geq 3$ there are non-trivial phase(s) in V_{CKM} . In particular, for n=3 , there is one non-trivial phase, the so called Kobayashi-Maskawa (KM) phase. We will show next that the appearance of a non-trivial physical phase in V_{CKM} does imply CP violation.

2.4.2 CP Violation in Charged Weak-Current Interactions

In order to check the CP properties of a Lagrangian, it is convenient to separate it in two parts, one containing the kinetic energy terms and the interactions which conserve CP like the electromagnetic interactions, and the other part containing the remaining terms of the Lagrangian. Then one constructs the most general CP transformation which leaves invariant the first part of the Lagrangian and checks whether CP invariance of the remaining Lagrangian leads to non-trivial constraints. This procedure may be done either in a WB basis or in a mass-eigenstate-basis (MEB). In this section, we will work in a MEB and in Sec. 2.5 we will discuss CP violation in a WB.

The most general CP trasformation which leaves invariant all terms of the Lagrangian, except $\mathcal{L}_W$, is given by:

$$\begin{array}{c}
(CP)u_\alpha(t,\vec{r})(CP)^\dagger = e^{i\xi_\alpha}\gamma^0 C\overline{u}_\alpha^T(t,-\vec{r}) \\
(CP)\overline{u}_\alpha(t,\vec{r})(CP)^\dagger = -e^{-i\xi_\alpha}u_\alpha^T(t,-\vec{r})C^{-1}\gamma^0 \\
(CP)d_k(t,\vec{r})(CP)^\dagger = e^{i\xi_k}\gamma^0 C\overline{d}_k^T(t,-\vec{r}) \\
(CP)\overline{d}_k(t,\vec{r})(CP)^\dagger = -e^{-i\xi_k}d_k^T(t,-\vec{r})C^{-1}\gamma^0 \\
(CP)W^{+\mu}(t,\vec{r})(CP)^\dagger = -e^{i\xi_W}W_\mu^-(t,-\vec{r})
\end{array} \tag{25}$$

Using Eq. (25) one can verify that CP invariance of $\mathcal{L}_W$ constrains V_{CKM} to satisfy:

$$V_{\alpha k}^* = e^{i(\xi_W+\xi_k-\xi_\alpha)}V_{\alpha k} \tag{26}$$

which in turn implies:

$$\mathrm{Im}Q_{\alpha i\beta j} = \mathrm{Im}(V_{\alpha i}V_{\beta j}V_{\alpha j}^*V_{\beta i}^*) = 0 \tag{27}$$

For two generations, the constraints of unitarity imply the automatic vanishing of $\mathrm{Im}Q_{\alpha i\beta j}$, while for three or more generations, this quantity does not vanish for a generic V_{CKM} , thus implying the possibility of having CP violation in the three generations SM.

2.4.3 Experimental Value of Quark Masses and Mixings

Due to the arbitrariness of Yukawa couplings, in the SM quark masses and mixings are free parameters, to be determined by experiment. There has been a continuing effort in measuring V_{CKM} with increasing accuracy. As we have seen in 2.4.1 for $n = 3$, V_{CKM} is characterized by four independent parameters. There are many possible choices for these parameters from which one can construct the full V_{CKM} matrix, using the constraints of unitarity. In section $2.4.5.$ we will describe some of the possible parametrizations of V_{CKM}. From an experimental point of view, it is important to determine all elements of V_{CKM} without using the assumption of unitarity and through as many different physical processes as possible. This leads to an overdetermination of V_{CKM} which provides a crucial test of the correctness

of our various theoretical assumptions, including that of unitarity of V_{CKM} . Most of the elements of the V_{CKM} matrix can be obtained by direct measurement of decay rates corresponding to SM tree level processes. The extraction of these V_{CKM} elements is expected to be valid even in the presence of New Physics, since eventual new contributions to these processes are negligible compared to tree level SM contributions. Next we will briefly review our present experimental knowledge [8] on the various V_{CKM} matrix elements. Nuclear beta decays and the analysis of Ke3 decays allow the determination of $|V_{ud}|$ and $|V_{us}|$, respectively, yielding:

$$|V_{ud}| = 0.9735 \pm 0.0008 \qquad |V_{us}| = 0.2196 \pm 0.0023 \tag{28}$$

Neutrino and anti-neutrino production of charm off valence d-quarks and charmed-tagged W decays determine $|V_{cd}|$ and $|V_{cs}|$, giving:

$$|V_{cd}| = 0.224 \pm 0.016 \qquad |V_{cs}| = 0.97 \pm 0.009(stat.) \pm 0.07(syst.) \tag{29}$$

$|V_{cb}|$ - Lep data on both exclusive and inclusive decays gives:

$$|V_{cb}| = 0.0402 \pm 0.0019 \tag{30}$$

$\frac{|V_{ub}|}{|V_{cb}|}$ - The value of $\frac{|V_{ub}|}{|V_{cb}|}$ can be extracted from an analysis of the endpoint spectrum of semileptonic B-decays or from exclusive decays such as $B \to \pi l \nu$ or $B \to \rho l \nu$. Taking into account the various theoretical uncertainties and the difference between the results from inclusive decays, the Particle Data Group [8] quotes the following value, with a rather conservative error:

$$\frac{|V_{ub}|}{|V_{cb}|} = 0.09 \pm 0.025 \tag{31}$$

$|V_{tb}|$ - The study of the decays $t \to b l^+ \nu$ leads to:

$$\frac{|V_{tb}|^2}{|V_{td}|^2 + |V_{ts}|^2 + |V_{tb}|^2} = 0.99 \pm 0.29 \tag{32}$$

Finally, we consider the values $|V_{td}|$ and $|V_{ts}|$ which, within the framework of the SM, can be extracted from the experimental values of $B_d - \bar{B}_d$ and $B_s - \bar{B}_s$ mixings. In the SM, the dominant contribution to these mixings arises from W-mediated box diagrams which lead to:

$$\Delta M_{B_q} = \frac{G_f^2 m_W^2}{6\pi^2} \eta_B m_{B_q} f_{B_d}^2 B_{B_q} S_0(x_t) |V_{tq} V_{tb}|^2 \tag{33}$$

where $q = d, s$, η_B is a QCD correction factor, $S_0(x_t \equiv m_t^2/m_W^2)$ is an Inami-Lim function [9] and $f_{B_d}^2 B_{B_q}$ parametrizes the hadronic matrix element which has been estimated by lattice QCD calculations to be for $q = d$, $f_{B_d}^2 B_{B_q} = (210 \pm 40 Mev)^2$. Using the measured value of $\Delta M_{B_d} = 0.473 \pm 0.016 ps^{-1}$, one obtains:

$$|V_{td} V_{tb}| = 0.0083 \pm 0.0016 \tag{34}$$

In the ratio of ΔM_{B_s} and ΔM_{B_d} many of the common factors cancel, so one has:

$$\frac{\Delta M_{B_s}}{\Delta M_{B_d}} = \frac{m_{B_s}}{m_{B_d}} \frac{B_{B_s}}{B_{B_d}} \frac{f_{B_s}^2}{f_{B_d}^2} \frac{|V_{ts}|^2}{|V_{td}|^2} \tag{35}$$

The uncertainty in the ratio of the above mass differences is smaller than the uncertainty in the numerator and denominator, taken separately. From the experimental lower bound $\Delta M_{B_s} > 14.3 ps^{-1}$ one obtains:

$$|V_{td}|/|V_{ts}| < 0.24 \tag{36}$$

Taking into account that for $n = 3$, unitarity of V_{CKM}, implies $|V_{ts}| \approx |V_{cb}|$, one obtains

$$|V_{td}| \leq 0.01 \tag{37}$$

which is more restrictive than the upper bound obtained from Eq. (34). It should be noted that within the three generations SM, constraints of unitarity imply bounds on the V_{CKM} matrix elements which, in some cases, impose more sensitive restrictions than direct measurements. For example, from normalization of the second row, one obtains:

$$|V_{cs}| = \left[1 - |V_{cd}|^2 - |V_{cb}|^2\right]^{1/2} \tag{38}$$

which implies

$$0.973 < |V_{cs}| < 0.978 \tag{39}$$

Similarly, from normalization of the third column one has:

$$|V_{tb}| = \left[1 - |V_{cb}|^2 - |V_{ub}|^2\right]^{1/2} \tag{40}$$

which leads to:

$$0.9990 < |V_{tb}| < 0.9993 \tag{41}$$

On the other hand, orthogonality of the first and third columns lead to:

$$|V_{cd}||V_{cb}| - |V_{ud}||V_{ub}| \leq |V_{td}| \leq |V_{ud}||V_{ub}| + |V_{cd}||V_{cb}| \tag{42}$$

which implies

$$0.004 < |V_{td}| < 0.015 \tag{43}$$

Although this restriction on $|V_{td}|$ is not as good as the one obtained in Eq. (34) it has the interest of being independent of $B_d - \bar{B}_d$ mixing.

2.4.4 Unitarity Triangles

We have seen that in the SM, the V_{CKM} matrix is unitary. For $n = 3$, orthogonality of rows and columns lead to the following six orthogonality realations:

$$\begin{array}{c} V_{ud}V_{us}^* + V_{cd}V_{cs}^* + V_{td}V_{ts}^* = 0 \\ V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0 \\ V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* = 0 \\ V_{ud}V_{cd}^* + V_{us}V_{cs}^* + V_{ub}V_{cb}^* = 0 \\ V_{ud}V_{td}^* + V_{us}V_{ts}^* + V_{ub}V_{tb}^* = 0 \\ V_{cd}V_{td}^* + V_{cs}V_{ts}^* + V_{cb}V_{tb}^* = 0 \end{array} \tag{44}$$

The above orthogonality relations may be represented in the complex plane by six triangles. All unitarity triangles have the same area. This can be seen by first noting that the imaginary parts of all invariant quartets have the same absolute value. For example, if one considers the first of Eqs. (44), multiplies by $V_{cd}^* V_{cs}$ and takes imaginary parts, one obtains:

$$\text{Im}(V_{ud}V_{cs}V_{us}^*V_{cd}^*) = \text{Im}(V_{cd}V_{ts}V_{cs}^*V_{td}^*) \tag{45}$$

Analogous arguments can be applied to other quartets, showing that $|\text{Im}Q_{\alpha i \beta j}|$ are all equal. The quantities $|\text{Im}Q_{\alpha i \beta j}|$ have a simple geometrical interpretation. Consider the orthogonality relation involving the first and third column of V_{CKM}:

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0 \tag{46}$$

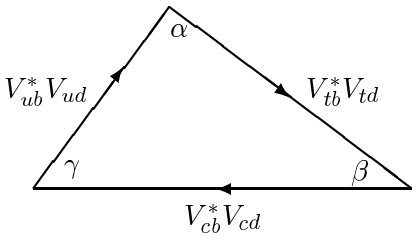


Fig. 1: Unitarity triangle

which can be represented in the complex plane by the triangle of Fig.1. The area of the triangle is given by:

$$\text{Area} = \frac{1}{2}|V_{cd}||V_{cb}|h \tag{47}$$

with

$$h = |V_{ud}||V_{ub}||\sin\arg(-V_{ud}V_{cb}V_{ub}^*V_{cd}^*)| \tag{48}$$

one thus obtains:

$$\text{Area} = \frac{1}{2}|\text{Im}(V_{ud}V_{cb}V_{ub}^*V_{cd}^*)| \tag{49}$$

Analogous considerations apply to the other orthogonality relations and the corresponding triangles. The equality of the various $|\text{Im}Q_{\alpha i\beta j}|$ then implies that all unitarity triangles have the same area. Most of the unitarity triangles have one of the sides much smaller than the other two. For example, in the triangle arising from the orthogonality of the first two columns of V, the sides have lengths $|V_{ud}||V_{us}| \approx 0.215$, $|V_{cd}||V_{cs}| \approx 0.214$ and $|V_{td}||V_{ts}| \approx 3.6 \times 10^{-4}$, where we have taken the central experimental values [8] of $|V_{ij}|$. Only two of the six unitarity triangles have all sides of comparable length: they correspond to orthogonality of the first and third column and orthogonality of the first and third line. The unitarity triangle represented in Fig. 1 is specially important, since it plays a crucial rôle in the current tests of the SM and its KM mechanism of CP violation. For this reason this triangle is usually refered to as "the unitarity triangle"(UT). Under the a rephasing of quark fields of Eq. (19) the unitarity triangle rotates and thus the orientation of the UT in the complex plane has no physical meaning. However, the angles of the UT are rephasing invariant and do have physical meaning. This is readily seen by noting that α, β, γ are defined by:

$$\begin{aligned}
\alpha &\equiv \arg\left(-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right) = \arg\left(-Q_{ubtd}\right) \\
\beta &\equiv \arg\left(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right) = \arg\left(-Q_{tbcd}\right) \\
\gamma &\equiv \arg\left(-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right) = \arg\left(-Q_{cbud}\right)
\end{aligned} \tag{50}$$

By definition, the angles α, β, γ always satisfy the relation:

$$\alpha + \beta + \gamma = \arg\left(-1\right) = \pi \pmod{2\pi} \tag{51}$$

In the literature it is sometimes stated that current measurements of CP asymmetries in B-meson factories will test the "validity" of Eq. (51). This is a rather misleading statement, obviously Eq. (51) is always valid, since it follows from the definition of α, β and γ as given by Eq. (50). We will return to this point in Sec.4.

2.4.5 *Parametrization of* V_{CKM}

The V_{CKM} matrix can be parametrized in different ways.

• *The Standard Parametrization*

The Particle Data Group advocates the following parametrization:

$$V = \begin{bmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{13}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta_{13}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta_{13}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{13}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & c_{23}c_{13} \end{bmatrix} \tag{52}$$

where c_{ij}, s_{ij} are shorthands for $\cos\theta_{ij}$ and $\sin\theta_{ij}$ respectively. If one allows for the phase δ_{13} to be free, all the three rotation angles θ_{ij} may be restricted to lie in the first quadrant, without loss of generality. This parametrization has the advantage that all the s_{ij} are simply related to directly measured quantities:

$s_{13} = |V_{ub}|$; $\quad s_{12} = \frac{|V_{us}|}{\sqrt{1-|V_{ub}|^2}} \approx |V_{us}|$; $\quad s_{23} = \frac{|V_{cb}|}{\sqrt{1-|V_{ub}|^2}}$

• *Rephasing Invariant Parametrizations*

It is possible to parametrize V_{CKM}, using only rephasing invariant input parameters, i.e. quantities which remain invariant when one makes a phase redefinition of the quark fields. The following rephasing invariant parametrizations have been proposed:

i) *Bjorken-Dunietz (BD) Parametrization*

In this parametrization [10], one chooses the following phase convention: V_{ud}, V_{us}, V_{cb} and V_{tb} are chosen to be real and positive. As input parameters BD choose:

$|V_{us}|$, $|V_{ub}|$, $|V_{cb}|$ and $\arg Q_{uscb}$

The remaining elements of V_{CKM} can be readily obtained from these input parameters, using unitarity of V_{CKM}.

ii) *Branco-Lavoura Parametrization*

This parametrization [11] has the special feature of using, as input parameters only moduli of V_{CKM} . A convenient set of input moduli is:

$|V_{us}|$, $|V_{ub}|$, $|V_{cb}|$, $|V_{td}|$

From these four moduli, one can trivially obtain all other moduli by using the fact that unitarity constrains columns and rows to be normalized. The real part of any invariant quartet can be obtained through the relation:

$$2ReQ_{\alpha i\beta j} = 1 - U_{\alpha i} - U_{\beta j} - U_{\alpha j} - U_{\beta i} + U_{\alpha i}U_{\beta j} + U_{\alpha j}U_{\beta i} \tag{53}$$

where $U_{\alpha i} \equiv |V_{\alpha i}|^2$. The strength of CP violation, measured by the imaginary part of any invariant quartet (recall that in the three generations SM, the imaginary part of all invariant quartets have the same modulus) is given by:

$$|ImQ_{\alpha i\beta j}| = \left[U_{\alpha i}U_{\beta j}U_{\alpha j}U_{\beta i} - (2ReQ_{\alpha i\beta j})^2\right]^{1/2} \tag{54}$$

It should be emphasized that in the context of the SM and as a result of unitarity of V_{CKM}, the strength of CP violation can be obtained from the knowledge of any four independent moduli [11]. This is a very special feature of the three generations SM.

iii) *Aleksan-Kayser-London (AKL) Parametrization*

Let us recall that in the SM with n generations, one may eliminate $(2n-1)$ phases from the initial n^2 phases of V_{CKM}, through rephasing of quark fields. As a result, the number of rephasing invariant phases is:

$$n_{\text{phases}} = n^2 - (2n-1) = (n-1)^2 \tag{55}$$

Of course, at this stage, we are not imposing unitarity. Note that the number of rephasing invariant phases equals the number of parameters which are necessary to parametrize V_{CKM}. AKL have suggested [12] that the following four independent rephasing invariant phases be used as input parameters:

$w_{tbcd},\ w_{cbud},\ w_{tbcs},\ w_{csud}$

with the definition $w_{\alpha i\beta j} = \arg Q_{\alpha i\beta j}$. Using unitarity, one can derive the whole of V_{CKM} matrix, including the value of its moduli.

• *Wolfenstein Parametrization*

Using the experimental fact that there is a strong hierarchy in off-diagonal elements of V_{CKM}, Wolfenstein [13] has suggested the following approximate parametrization:

$$V \simeq \begin{bmatrix} 1-\frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho-i\eta) \\ -\lambda & 1-\frac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{bmatrix} \tag{56}$$

The parameter $\lambda \approx 0.22$ is small and thus can be used as an expansion parameter. The great advantage of the Wolfenstein parametrization is that one may estimate the order of magnitude of any function of the matrix elements of V_{CKM}, by taking the leading term in its expansion in λ. It is clear from Eq. (56) that in this parametrization unitarity of V_{CKM} is satisfied up to order λ^3. Branco and Lavoura (BL) [14] have suggested an exact version of the Wolfenstein parametrization by adopting the phase convention of Bjorken and Dunietz and defining λ, A, μ, ϕ by the following relations:

$$\lambda = V_{us};\ \ A = \frac{|V_{cb}|}{|V_{us}|^2};\ \ \mu = \frac{|V_{ub}|}{|V_{cb}||V_{us}|};\ \ \phi = w_{uscb} \tag{57}$$

It should be emphasized that Eqs. (57) are exact by definition and therefore the expressions for $|V_{us}|$, $|V_{cb}|$ and $|V_{ub}|$ are not corrected by terms of higher order in λ. The expansion of V_{CKM} in powers of λ can thus be derived to any desired order and one obtains

$$\begin{aligned} &|V_{ud}| = 1-\tfrac{1}{2}\lambda^2-\tfrac{1}{8}\lambda^4+O(\lambda^6), \\ &|V_{cd}| = -\lambda+A^2(\tfrac{1}{2}-\rho-i\eta)\lambda^5+O(\lambda^7), \\ &|V_{cs}| = 1-\tfrac{1}{2}\lambda^2-\tfrac{1}{8}(1+4A^2)\lambda^4+O(\lambda^6), \\ &|V_{td}| = A(1-\rho-i\eta)\lambda^3+\tfrac{1}{2}A(\rho+i\eta)\lambda^5+O(\lambda^7), \\ &|V_{ts}| = -A\lambda^2+A(\tfrac{1}{2}-\rho-i\eta)\lambda^4+O(\lambda^6), \\ &|V_{tb}| = 1-\tfrac{1}{2}A^2\lambda^4+O(\lambda^6). \end{aligned} \tag{58}$$

where we have put $\rho \equiv \mu\cos\phi$, $\eta \equiv \mu\sin\phi$. Obviously Eqs. (58) coincide with Eq. (56) up to order λ^3.

Another way of evaluating in an exact way higher order terms in λ, is the one proposed by Buras, Lautenbacher and Ostermaier (BLO) [15] who suggested that the parameters λ, A, ρ, η be defined through the exact relations:

$$s_{12} = \lambda, \quad s_{23} = A\lambda^2, \quad s_{13}e^{-i\delta} = A\lambda^3(\rho-i\eta) \tag{59}$$

From which it follows that

$$c_{12} = \sqrt{1-\lambda^2}; \quad c_{23} = \sqrt{1-A^2\lambda^4} \quad c_{13} = \sqrt{1-A^2\lambda^6(\rho^2+\eta^2}} \tag{60}$$

Substituting these expressions in the standard parametrization of Eq. (52) one obtains an exact parametrization in terms of λ, A, ρ, η. One can then expand each element in powers of λ. In practice the differences between the BLO and BL parametrizations first arise only at order λ^6. It is very useful to introduce the parameters $\bar{\rho}$, $\bar{\eta}$ [15], defined by:

$$\bar{\rho} = \rho(1-\frac{\lambda^2}{2}); \quad \bar{\eta} = \eta(1-\frac{\lambda^2}{2}) \tag{61}$$

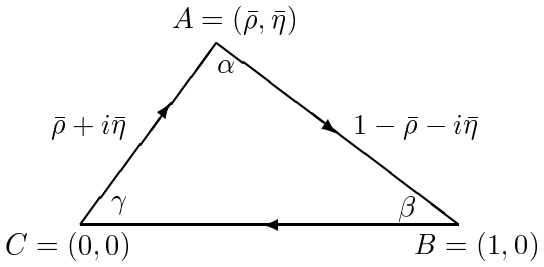


Fig. 2: Normalised unitarity triangle

The advantage of the introduction of these parameters is that one can obtain a normalised unitarity triangle (see Fig. 2) which is valid beyond leading approximation. We will see in Sec.4 that constraints on the SM from various experimental data can be easily translated into constraints on the location of the vertex of the normalised triangle.

2.5 Weak-Basis Invariants and CP Violation

In the previous section, we have studied the CP properties of the SM in a basis where the quark mass matrices have been diagonalized, which led to a non-trivial V_{CKM} matrix in the charged currents. We have seen that CP invariance requires that all rephasing invariant combination of V_{CKM} matrix elements be real. In this section, we present an alternative way of studying the CP properties of any gauge theory. We will consider the Lagrangian in its initial form, i.e. in a weak-basis where, as we have seen, all gauge currents are diagonal, real. We then consider the most general CP transformation which leaves invariant the part of the Lagrangian containing the gauge interactions and then check whether the CP transformation thus defined implies any restrictions on the remaining of the Lagrangian.

The most general CP transformation which leaves $\mathcal{L}_{\text{gauge}}$ invariant is [16]:

$$\begin{array}{lll} u_L^0 \to W_L C u_L^{0*} & ; & u_R^0 \to W_R^u C u_R^{0*} \\ d_L^0 \to W_L C d_L^{0*} & ; & d_R^0 \to W_R^d C d_R^{0*} \end{array} \tag{62}$$

where $C = i\gamma^2\gamma^0$ and W_L, W_R^u, W_R^d are n-dimensional unitary matrices acting in flavour space.

Note that the existence of ℓh charged current interactions constrains u_L, d_L to transform in the same way under CP, while the absence of rh charged currents in the SM allows u_R, d_R to transform differently under CP. It can be readily verified from Eqs. (11),(62) that in order for $\mathcal{L}_{\text{mass}}$ to be invariant under CP, the matrices M_u, M_d should satisfy the relations [16]:

$$\begin{array}{ll} W_L^\dagger M_u W_R^u & = M_u^* \\ W_L^\dagger M_d W_R^d & = M_d^* \end{array} \tag{63}$$

from which it follows:

$$\begin{array}{lll} W_L^\dagger H_u W_L & = H_u^* & = H_u^T \\ W_L^\dagger H_d W_L & = H_d^* & = H_d^T \end{array} \tag{64}$$

The existence of unitary matrices W_L satisfying Eqs. (64) is a necessary and sufficient condition for $\mathcal{L}^{SM}$ to be CP invariant. However, written in this form, these conditions are of little practical use. One would like to have non-trivial CP constraints on H_u, H_d, written in terms of WB invariants. In order to achieve this, we note that from Eq. (64) one has:

$$\begin{aligned} W_L^\dagger H_u H_d W_L &= H_u^T H_d^T \\ W_L^\dagger H_d H_u W_L &= H_d^T H_u^T \end{aligned} \tag{65}$$

Subtracting these two equations, one obtains:

$$W_L^\dagger \left[H_u, H_d\right] W_L = -\left[H_u, H_d\right]^T \tag{66}$$

If one evaluates the traces of both sides of Eq. (66), one finds that they vanish identically and no constraint is obtained. In order to obtain a non-trivial constraint, we multiply Eq. (66) by itself an odd number of times, to obtain:

$$W_L^\dagger \left[H_u, H_d\right]^r W_L = -\left\{\left[H_u, H_d\right]^r\right\}^T \quad (r\ odd) \tag{67}$$

Taking traces, one finally obtains [16]:

$$tr\left[H_u, H_d\right]^r = 0 \quad (r\ odd) \tag{68}$$

Eq. (68) gives CP constraints on H_u, H_d, valid for an arbitrary number of generations. For $n = 1$ and $n = 2$, the constraints of Eq. (68) are automatically satisfied for arbitrary hermitian matrices H_u, H_d. Of course, this is to be expected from the counting of physical phases in the V_{CKM} matrix, given by Eq. (23). For $n \geq 3$, Eq. (68) provides non-trivial CP constraints on H_u, H_d. For $n = 3$, it can be readily shown [16] that the condition

$$tr\left[H_u, H_d\right]^3 = 0 \tag{69}$$

is a necessary and sufficient condition for CP invariance in the SM. It is useful to express the invariant of Eq. (69) in terms of physically measurable quantities. One obtains:

$$tr\left[H_u, H_d\right]^3 = 6i\Delta_{21}\Delta_{31}\Delta_{32}\mathrm{Im}Q \tag{70}$$

where:

$$\Delta_{21} = (m_s^2 - m_d^2)(m_c^2 - m_u^2); \Delta_{31} = (m_b^2 - m_d^2)(m_t^2 - m_u^2); \Delta_{32} = (m_b^2 - m_s^2)(m_t^2 - m_c^2) \tag{71}$$

From Eq. (70), (71), it follows that in the SM, CP violation vanishes in the limit where any two quarks of the same charge become degenerate. But it does not necessarily vanish in the limit where one quark is massless (e.g., $m_u = 0$) or even in the chiral limit ($m_u = m_d = 0$).

At this stage, it is worth examining the significance of our analysis. We have started by considering the most general CP transformation which leaves the SM gauge interations invariant. We have then derived the restrictions which the Hermitian mass matrices H_u, H_d should satisfy in order for $\mathcal{L}_{\text{mass}}$ to be invariant. The fact that for $n \geq 3$ one obtains non-trivial CP restrictions on H_u, H_d, implies, of course, that in the SM with three generations, CP can be violated. In the SM, CP violation arises as a clash between the CP properties of the gauge interactions and the mass terms [17].

We have chosen to present our analysis in terms of $\mathcal{L}_{\text{gauge}}$ and $\mathcal{L}_{\text{mass}}$. Alternatively, one can consider $\mathcal{L}_{\text{gauge}}$ and $\mathcal{L}_Y$. One can then study the CP properties of $\mathcal{L}^{SM}$ even before spontaneous gauge symmetry breaking and therefore before fermion mass terms are generated. In this case, the non-trivial CP restrictions would be written in terms of the Yukawa coupling matrices Y^d, Y^u, with $Y^d(Y^d)^\dagger, Y^u(Y^u)^\dagger$ substituting H_u, H_d in Eq. (68). Therefore, for three generations, the necessary and sufficient condition for the Lagrangian of the SM to conserve CP is:

$$tr\left[Y^d(Y^d)^\dagger, Y^u(Y^u)^\dagger\right]^3 = 0 \tag{72}$$

Conversely, a non-vanishing value of the invariant of Eq. (72) is an indication that $\mathcal{L}^{SM}$ violates CP.

We have presented a general method to analyse the CP properties of $\mathcal{L}^{SM}$. It is clear that the method can be readily extended to theories beyond the SM [3]. The only difference is that in the presence of other interactions, beyond those already present in the SM, the allowed CP transformations of Eq. (62) are modified. For example, in the left-right symmetric model (LRS), due to the presence of rh charged-current interactions, the fields u_R, d_R must transform in the same way under CP, i.e., $W_R^u = W_R^d$ in Eq. (62). These new CP transformation properties lead to new CP constraints [18] and, as a result, CP can be violated in the LRS model even for one or two generations.

3 SPONTANEOUS CP VIOLATION

The origin of CP violation is still an open question in Particle Physics. There are many different models of CP violation, which can be classified in two broad classes:

• *Models with explicit CP violation*

In this class of models CP is explicitly broken in the Lagrangian through the introduction of complex couplings. The simplest example of this class of models is, of course, the three generations SM. As we have seen in the previous section, CP violation arises from complex Yukawa couplings in the Lagrangian. In the SM, Yukawa couplings are the only ones which can be complex. All other couplings, like for example the parameters of the one Higgs doublet potential, are constrained to be real by the requirement of Hermiticity of the Hamiltonian. In most of the extensions of the SM, like for example the minimal supersymmetric standard model (MSSM) or general multi-Higgs extensions of the SM, there are other couplings which can be complex thus providing other sources of CP violation.

• *Models with spontaneous CP violation*

In this class of models, CP and T are good symmetries of the Lagrangian, but the vacuum does not respect these symmetries. In this section we will study spontaneous CP violation (SCPV) in general and describe some of the minimal extensions of the SM where it can occur.

In order to achieve spontaneous CP violation, the following two conditions should be satisfied:

(i) There is a transformation which may be physically interpreted as a CP transformation, under which the Lagrangian is invariant.

(ii) There is no transformation that may be physically interpreted as CP under which both the vacuum and the Lagrangian are invariant.

The CPT theorem tells us that in the context of relativistic quantum field theory, if a Lagrangian is CP invariant, it is also T invariant. Also if a vacuum breaks CP, it will necessarily break T too. The idea of spontaneous CP and T breaking was suggested by T. D. Lee [19] a long time ago, at a time when only two (and incomplete) generations were known. The scenario of spontaneous CP and T violation is specially attractive from a theoretical point of view, since if time reversal and space inversion (identified as CP, not P) are good symmetries of the Lagrangian, then the full Poincaré group of space time transformations, including the discrete ones, is a good symmetry of the Lagrangian. The idea of spontaneus CP and T breaking has the further appeal of putting the breakdown of these symmetries on the same footing as the breaking of the gauge symmetry which is also spontaneous in order to preserve renormalizability.

Next we derive simple criteria which can be used to determine whether a set of vacua violate CP or not. For defineteness, let us consider a simple extension of the SM, with an arbitrary number n_d of Higgs doublets, which under CP transform as :

$$CP\Phi_\alpha CP^\dagger = U^{CP}_{\alpha\beta}\Phi^*_\beta \tag{73}$$

where U is an $n_d \times n_d$ unitary matrix, mixing the scalar doublets. If no extra symmetries beyond

SU(2)xU(1) are present in the Lagrangian, U^{CP} reduces to a diagonal unitary matrix. In general, the neutral components of the Higgs doublets have the following vacuum expectation values:

$$< 0|\phi_\alpha^0|0 >= v_\alpha \exp{(i\theta_\alpha)} \tag{74}$$

Let us assume that the vacuum is CP invariant, i.e. :

$$CP|0 >= |0 > \tag{75}$$

From Eqs. (73), (74) and (75) one has then:

$$< 0|\phi_\alpha^0|0 >=< 0|CP^\dagger CP\phi_\alpha^0 CP^\dagger CP|0 >=< 0|CP\phi_\alpha^0 CP^\dagger|0 >= U_{\alpha\beta}^{CP} v_\beta \exp{(-i\theta_\beta)} \tag{76}$$

leading to:

$$v_\alpha \exp{(i\theta_\alpha)} = U_{\alpha\beta}^{CP} v_\beta \exp(-i\theta_\beta) \tag{77}$$

In order for a vacuum to be CP invariant, the set of vevs should satisfy Eq. (77). If it is not possible to find a unitary matrix U^{CP} which on the one hand corresponds to a CP transformation allowed by the Lagrangian and on the other hand satisfies Eq. (77), then the vacuum leads to spontaneous CP violation. It can be readily seen that in the SM with only one Higgs doublet it is not possible to achieve spontaneous CP violation. The vev of Eq. (74) should be constant over space-time in order not to break Poincaré invariance, and one can always make an SU(2) transformation which brings the vev into the form:

$$< 0|\Phi|0 >= \begin{bmatrix} 0 \\ v\exp(i\theta) \end{bmatrix} \tag{78}$$

Obviously, in this case, the most general CP transformation of Φ is:

$$CP\Phi CP^\dagger = \exp(i\alpha)\Phi^* \tag{79}$$

By choosing $\alpha = 2\theta$ Eq. (77) is satisfied and therefore the vacuum corresponding to Eq. (78) is CP invariant. At this stage, one may ask what is the minimal extension(s) of the SM which can lead to spontaneous CP violation. It will be shown that the required minimal Higgs structure leading to CP violation depends on whether one imposes natural flavour conservation (NFC) in the Higgs sector or not and whether non-standard quarks are introduced or not. Next we will describe each one of these scenarios.

3.1 The Lee Model: only standard quarks without NFC in the Higgs sector

Let us consider the SM with an arbitrary number of standard fermion families . Since we want to investigate the possibility of generating spontaneous CP violation, we will impose CP invariance at the Lagrangian level and thus assume, without loss of generality, that all couplings are real. The minimal Higgs structure which can lead to spontaneous CP violation without NFC in the Higgs sector and only standard quarks consists of two Higgs doublets. The most general Higgs potential consistent with renormalizability and gauge invariance is:

$$V(\Phi_1, \Phi_2) = V_0 + \Big[\lambda_1(\Phi_1^\dagger\Phi_2)(\Phi_1^\dagger\Phi_2) + \lambda_2(\Phi_1^\dagger\Phi_2)(\Phi_1^\dagger\Phi_1) + \lambda_3(\Phi_2^\dagger\Phi_2)(\Phi_1^\dagger\Phi_2) + h.c.\Big] \tag{80}$$

where V_0 stands for the terms of the scalar potential like $\Phi_i^\dagger\Phi_i, (\Phi_i^\dagger\Phi_i)(\Phi_j^\dagger\Phi_j), \Phi_i^\dagger\Phi_j\Phi_j^\dagger\Phi_i$ which do not depend on the relative phase of Φ_1 and Φ_2. Since we want to obtain spontaneous CP violation, the Lagrangian will be assumed to be CP, T invariant. In this case, one may choose, without loss of

generality, all the coupling constants real. It has been shown [19] that there is a region of the parameter space of the Higgs potential where the absolute minimum is at:

$$\theta = \arccos\left[\frac{-\lambda_2 v_1^2 - \lambda_3 v_2^2}{4\lambda_1 v_1 v_2}\right] \tag{81}$$

where $\theta = (\theta_2 - \theta_1)$, with θ_j, v_j defined by:

$$< 0|\phi_j^0|0 >= \frac{v_j}{\sqrt{2}} exp(i\theta_j) \tag{82}$$

In general, the vacuum of Eq. (82) leads to spontaneous T, CP violation. This can be seen using the general criteria which we have previously derived. Due to the fact that in the Higgs potential of Eq. (80), the parameters λ_i are arbitrary, the scalar fields transform under CP as:

$$CP\Phi_i CP^\dagger = \exp(i\alpha_i)\Phi_i^* \tag{83}$$

On the other hand, the quartic couplings with coefficients λ_2, λ_3 further lead to the constraint $\alpha_1 = \alpha_2 = \alpha$. It can then be readily verified that for the vacuum of Eq. (82), one cannot find an α such that Eq. (77) is obeyed both by Φ_1 and Φ_2. This completes the proof that the vacuum of Eq. (82) in general violates T and CP. In order to analyse the sources of CP violation in this model, let us write the quark Yukawa interactions:

$$\mathcal{L}_Y = \sum_{k=1}^{2} \left\{ Y_k^d \overline{Q}_L^0 \Phi_k d_R^0 + Y_k^u \overline{Q}_L^0 \widetilde{\Phi}_k u_R^0 \right\} + h.c. \tag{84}$$

where Q_L^0 stand for the quark doublets. Upon spontaneous symmetry breaking, the quark mass matrices are given by:

$$\begin{aligned} M_d &= \tfrac{1}{\sqrt{2}} \left[v_1 e^{i\theta_1} Y_1^d + v_2 e^{i\theta_2} Y_2^d \right] \\ M_u &= \tfrac{1}{\sqrt{2}} \left[v_1 e^{-i\theta_1} Y_1^u + v_2 e^{-i\theta_2} Y_2^u \right] \end{aligned} \tag{85}$$

One obtains for the Hermitian quark mass matrices:

$$\begin{aligned} H_d \equiv M_d M_d^\dagger &= \frac{1}{2}\Big[v_1^2 Y_1^d Y_1^{dT} + v_2^2 Y_2^d Y_2^{dT} + v_1 v_2 (Y_1^d Y_2^{dT} + Y_2^d Y_1^{dT}) \cos\theta - \\ &\quad - i v_1 v_2 (Y_1^d Y_2^{dT} - Y_2^d Y_1^{dT}) \sin\theta \Big] \end{aligned} \tag{86}$$

$$\begin{aligned} H_u \equiv M_u M_u^\dagger &= \frac{1}{2}\Big[v_1^2 Y_1^u Y_1^{uT} + v_2^2 Y_2^u Y_2^{uT} + v_1 v_2 (Y_1^u Y_2^{uT} + Y_2^u Y_1^{uT}) \cos\theta + \\ &\quad + i v_1 v_2 (Y_1^u Y_2^{uT} - Y_2^u Y_1^{uT}) \sin\theta \Big] \end{aligned} \tag{87}$$

Although there is only one physical phase, namely $\theta = (\theta_2 - \theta_1)$, it is clear from Eqs. (86) and (87) that due to the arbitrariness of Y_k^d, Y_k^u, the matrices H_u, H_d are arbitrary Hermitian matrices. As a result, there will be in general a non-trivial CP violating phase in the V_{CKM} matrix. For any specific choice of Y_k^d, Y_k^u this can be explicitly verified by computing the invariant $tr[H_u, H_d]^3$, which we have previously described. In this model, the origin of CP violation is a non-trivial vacuum phase θ and not complex Yukawa couplings, as is the case in the SM. However, in both models the KM mechanism exists in the charged weak currents. From a phenomenological point of view, the distinctive feature of the two Higgs

doublet model without NFC, is the existence of other contributions to CP violation arising from tree-level FCNC mediated by neutral physical scalars. Of course, these FCNC interactions result from the fact that since NFC has not been imposed, quarks of a given charge receive contributions to their masses from Yukawa couplings to two different Higgs multiplets. In order to derive the quark interactions with the physical neutral Higgs, it is useful to make the following expansion:

$$\Phi_j = e^{i\theta_j} \left[\begin{array}{c} \phi_j^+ \\ \frac{1}{\sqrt{2}}(v_j + \rho_j + i\eta_j) \end{array} \right] \tag{88}$$

The pseudo-Goldstone bosons G^+, G^0 are obtained through the transformations:

$$\left[\begin{array}{c} G^+ \\ H^+ \end{array} \right] = O \left[\begin{array}{c} \phi_1^+ \\ \phi_2^+ \end{array} \right] \quad ; \quad \left[\begin{array}{c} G^0 \\ I \end{array} \right] = O \left[\begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right] \quad ; \quad \left[\begin{array}{c} H^0 \\ R \end{array} \right] = O \left[\begin{array}{c} \rho_1 \\ \rho_2 \end{array} \right] \tag{89}$$

where

$$O = \frac{1}{v} \left(\begin{array}{cc} v_1 & v_2 \\ v_2 & -v_1 \end{array} \right) \tag{90}$$

with $v = (v_1^2 + v_2^2)^{1/2} = (\sqrt{2} G_F)^{-1/2} \simeq 246$ Gev. There are three physical neutral fields which are orthogonal combinations of $H^0,\ R$ and I. The Yukawa interactions of $H^0,\ R$ and I with the quark mass eigenstates can be written as [20]:

$$\begin{aligned} \mathcal{L}_Y^N &= -\frac{1}{v}(\overline{u} D_u u + \overline{d} D_d d) H^0 - \\ &- \frac{1}{v}\left[\overline{d}(N_d \gamma_R + N_d^\dagger \gamma_L) d + \overline{u}(N_u \gamma_R + N_u^\dagger \gamma_L) u\right] R \\ &- \frac{i}{v}\left[\overline{d}(N_d \gamma_R - N_d^\dagger \gamma_L) d - \overline{u}(N_u \gamma_R - N_u^\dagger \gamma_L) u\right] I \end{aligned} \tag{91}$$

where we have defined $\gamma_{R,L} \equiv \frac{1}{2}(1 \pm \gamma_5)$ and D_u, D_d are the diagonal quark mass matrices and N_d, N_u are given by:

$$\begin{aligned} N_d &= U_L^{d\dagger}\left[\frac{v_2}{\sqrt{2}} Y_1^d - \frac{v_1}{\sqrt{2}} e^{i\theta} Y_2^d\right] U_R^d \\ N_u &= U_L^{u\dagger}\left[\frac{v_2}{\sqrt{2}} Y_1^u - \frac{v_1}{\sqrt{2}} e^{i\theta} Y_2^u\right] U_R^u \end{aligned} \tag{92}$$

It is clear from Eqs. (91) and (92) that the couplings of H^0 conserve flavour while those of R and I do violate flavour. Therefore in this model new contributions to CP violation in the kaon sector arise from $\Delta S = 2$ transitions generated at tree-level through diagrams like those of Fig. 3, where R and I are exchanged. Obviously, in this model there are analogous tree-level contributions to $\Delta C = 2$ and $\Delta B = 2$ transitions which may have an impact on $D^0 - \overline{D}^0$ and $B^0 - \overline{B}^0$ mixing, respectively.

One of the disadvantages of multi-Higgs models without NFC is the following: due to the smallness of the $K^0 - \overline{K}^0$ and $B^0 - \overline{B}^0$ mass differences as well as the CP violating parameter ε of the neutral kaon system, if one does not assume any special suppression of FCNC couplings, neutral scalars must be very heavy, with masses of order of at least a few Tev. Over the past years, it has been suggested by various authors [21] that there may be flavour-dependent suppression factors in the neutral couplings which could allow for lighter Higgs. More recently [22] a class of two Higgs doublet models has been

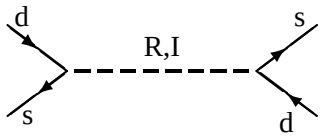


Fig. 3: Higgs contribution to the $\Delta S = 2$ amplitude

proposed in which the flavour changing couplings of the neutral scalars are related in an exact way to elements of the V_{CKM} matrix. In some of these models, the mass of the Higgs particles could be of order $100 - 200$ GeV, in spite of the existence of FCNC couplings.

3.2 Models with NFC and only Standard Quarks

The simplest way to recover NFC in the Higgs sector is to impose a discrete Z_2 symmetry [23] under which

$$\begin{aligned} \Phi_2 &\rightarrow -\Phi_2 \\ d_{jR} &\rightarrow -d_{jR} \end{aligned} \tag{93}$$

while all other fields remain unchanged. In this case, NFC in the Higgs sector is guaranteed by the fact that the symmetry of Eqs. (93) implies $Y_1^d = Y_2^u = 0$ and as a result the down quarks only receive mass from $< \phi_2^0 >$ while the mass of up quarks arises only from $< \phi_1^0 >$. The symmetry of Eq. (93) has to be imposed on the full Lagrangian, which implies that the terms in λ_2, λ_3 are forbidden. As a result, the only minimum of the Higgs potential, with a non-vanishing phase corresponds to:

$$< \phi_1^0 >= v_1 \qquad ; \qquad < \phi_2^0 >= v_2 \exp(i\pi/2) \tag{94}$$

It can be readily shown [24] that in spite of the appearance of a relative phase of $\pi/2$, the vacuum of Eq. (94) is T and CP invariant. This is due to the fact that the Z_2 symmetry of the Lagrangian allows the scalar fields to transform under CP as:

$$CP\Phi_1 CP^\dagger = \Phi_1^* \quad ; \quad CP\Phi_2 CP^\dagger = -\Phi_2^* \tag{95}$$

thus implying

$$U^{CP} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{96}$$

The matrix U^{CP} does satisfy Eq. (77), which implies that the vacuum of Eq. (94) is T, CP invariant. In conclusion, the same symmetry which guarantees NFC in the Yukawa interactions, also prevents the vacuum from breaking T or CP. It has been pointed out [25] that if one allows for terms like $\Phi_1^\dagger \Phi_2$ which softly break the Z_2 symmetry, one can obtain spontaneous CP violation. On the other hand, the simplest way of achieving spontaneous CP violation, while maintaing NFC through an exact symmetry of the Lagrangian, consists of introducing a third Higgs doublet Φ_3 which does not couple to quarks. This is achieved in a natural way by introducing an extra symmetry Z_2' under which Φ_3 is odd, while all other fields are even. In this case, the Higgs potential is of the form:

$$V = V_0 + c_{12}(\Phi_1^\dagger \Phi_2)(\Phi_1^\dagger \Phi_2) + c_{13}(\Phi_1^\dagger \Phi_3)(\Phi_1^\dagger \Phi_3) + c_{23}(\Phi_2^\dagger \Phi_3)(\Phi_2^\dagger \Phi_3) + h.c. \tag{97}$$

It is clear from Eq. (97) that there are three terms which have phase dependence and it has been shown [24] that there is a region of parameters where the minimum of the Higgs potential does lead to spontaneous T, CP violation. In this model CP violation arises exclusively from charged Higgs [26], since

the charged current interactions conserve CP, i. e., the V_{CKM} matrix can be made real by an appropriate choice of the phases of quark fields [27].

3.3 Models with Isosinglet Quarks

We have seen that if only standard quarks are introduced, the minimal Higgs structure that can generate spontaneous CP violation, consists of two or three Higgs doublets, depending on whether NFC is imposed in the Higgs sector or not. In models with isosinglet quarks, i.e. quarks whose ℓh and rh components are both singlets under weak isospin, it is possible to generate spontaneous CP violation with a rather simple Higgs structure, namely the standard Higgs doublet and a complex $SU(2) \times U(1)$ scalar singlet. The minimal quark field content of a model of this type is

$$\left[\begin{array}{c} u_j \\ d_j \end{array} \right]_L \quad , \quad u_{jR}, \quad d_{jR}, \quad D_R, \quad D_L \tag{98}$$

where j is a family index and D is a charge $-1/3$ isosinglet quark. Models with isosinglet quarks have been suggested within the framework of grand-unified theories such as E_6 and they have also been proposed in models which provide a possible solution [28] to the strong CP problem [29]. Here we will consider a minimal model [30] of this class, with the quark field content of Eq. (98) and a Higgs system with the standard doublet Φ and a complex isosinglet scalar S. We will introduce a Z_2 symmetry under which all new fields D_L, D_R, S are odd, while all other fields are even. This symmetry is not necessary to achieve spontaneous CP breaking but it is essential to obtain a simple solution to the strong CP problem. The most general, renormalizable $SU(2) \times U(1) \times Z_2$ invariant potential is given by:

$$V = V_\Phi + V_S + V_{\Phi,S} \tag{99}$$

where V_Φ is the SM Higgs potential and

$$\begin{aligned} V_S &= S^*S(a_1 + b_1 S^*S) + (S^2 + S^{*^2})(a_2 + b_2 S^*S) + b_3(S^4 + S^{*^4}) \\ V_{\Phi,S} &= \left[c_1(S^2 + S^{*^2}) + c_2 S^*S\right] \Phi^\dagger \Phi \end{aligned} \tag{100}$$

Although the Higgs sector contains only a doublet and a singlet, the Higgs potential has various terms which exhibit a non-trivial phase dependence. As a result, there is a region of parameter space where the minimum of the potential is at:

$$< \phi^0 >= \frac{v}{\sqrt{2}} \qquad ; \qquad < S >= \frac{1}{\sqrt{2}} V e^{i\alpha} \tag{101}$$

This minimum violates both T and CP. In order to see how the phase α will introduce CP violation relevant to the quark sector, we have to analyse the Yukawa couplings which are constrained by the $SU(2) \times U(1) \times Z_2$ symmetry to have the form:

$$\mathcal{L}_Y = \mathcal{L}_Y^{SM} + M\overline{D}_L D_R + (f_i S + f_i' S^*)\overline{D}_L d_{Ri} + h.c. \tag{102}$$

where $\mathcal{L}_Y^{SM}$ stands for the Yukawa couplings of the SM. We have included a mass term for the isosinglet quark, since it is gauge and also Z_2 invariant. Upon spontaneous symmetry breaking, the 4×4 mass matrix of the $Q = -1/3$ quarks is given by:

$$M_d = \left[\begin{array}{cc} m_d & 0 \\ M_D & M \end{array} \right] \tag{103}$$

where m_d is a 3×3 mass matrix involving only standard quarks and

$$(M_D)_i = \frac{V}{\sqrt{2}} \left[f_i \text{exp}(i\alpha) + f_i' \text{exp}(-i\alpha) \right] \tag{104}$$

The diagonalization of M_d is through the bi-unitary transformation:

$$U_L^\dagger M_d U_R = \begin{bmatrix} \overline{m} & 0 \\ 0 & \overline{M} \end{bmatrix} \tag{105}$$

where $\overline{m} = \text{diag}(m_d, m_s, m_b)$ and $\overline{M}$ is the mass of the heavy $Q = -1/3$ quark. It is useful to write the unitary matrix U_L in block form:

$$U_L = \begin{bmatrix} K & R \\ S & T \end{bmatrix} \tag{106}$$

Multiplying Eq. (105) by its Hermitian conjugate and using Eq. (106), one obtains:

$$(m_d m_d^\dagger) K + m_d M_D^\dagger S = K \overline{m}^2 \tag{107}$$

$$(m_d m_d^\dagger) R + m_d M_D^\dagger T = R \overline{M}^2 \tag{108}$$

$$M_D m_d^\dagger K + (M_D M_D^\dagger + M M^\dagger) S = S \overline{m}^2 \tag{109}$$

$$M_D m_d^\dagger R + (M_D M_D^\dagger + M M^\dagger) T = T \overline{M}^2 \tag{110}$$

So far, we have not done any approximations. Recalling that V is the vacuum expectation value of an $SU(2) \times U(1)$ singlet, it is natural to assume that $V \gg v$. One then obtains from Eq. (109):

$$S \simeq -\frac{M_D m_d^\dagger K}{(M_D M_D^\dagger + M M^\dagger)} \tag{111}$$

Finally, using Eq. (107) and Eq. (111), one obtains:

$$K^{-1} m_0^2 K \simeq \overline{m}^2 \tag{112}$$

where

$$m_0^2 = m_d m_d^\dagger - \frac{m_d M_D^\dagger M_D m_d^\dagger}{\overline{M}^2} \tag{113}$$

It is clear that the phase α originated from $< S >$, enters in m_0^2 unsuppressed by the ratio v/V and, as a result, there will be a non-trivial KM phase in the 3×3 mixing matrix K which connects the standard quarks. It is important to emphasize the rôle played by the isosinglet quark D. It is through the mixings of D with standard quarks that CP violation generated at a high energy scale by $< S >$ can appear unsuppressed at low energies in the standard quark mixings. Another interesting feature of this model is the appearance of FCNC as well as deviations from unitarity in the 3×3 V_{CKM} matrix. These two phenomena are closely related and they are both suppressed by the ratio v/V. This can be seen by noting that unitarity of the 4×4 matrix U_L defined by Eq. (106) implies:

$$K^\dagger K = 1 - S^\dagger S \tag{114}$$

It follows from Eq. (114) that deviations from unitarity of the U_L block connecting standard quarks (K) are of order $S^\dagger S \simeq v^2/V^2$. On the other hand, the neutral currents in this model are given by

$$\mathcal{L}_Z = \frac{g}{2\cos\theta_w}\left[\overline{u}_{Li}\gamma^\mu u_{Li} - Z^d_{\alpha\beta}\overline{d}_{L\alpha}\gamma^\mu d_{L\beta} - \sin^2\theta_w J^\mu_{em}\right] Z_\mu + h.c. \tag{115}$$

with,

$$Z^d_{\alpha\beta} = \delta_{\alpha\beta} - (U^*_L)_{4\alpha}(U_L)_{4\beta} \tag{116}$$

where α, β denote the four $Q = -1/3$ quarks, with $d_4 \equiv D$ standing for the new heavy quark. Of special importance are the couplings involving the standard quarks which are given by:

$$Z^d_{ij} = \delta_{ij} - S^*_i S_j \tag{117}$$

Using Eq. (114) and Eq. (117) one obtains:

$$\begin{array}{rcl} V^*_{ud}V_{us} + V^*_{cd}V_{cs} + V^*_{td}V_{ts} & = & Z^d_{ds} \\ V^*_{ud}V_{ub} + V^*_{cd}V_{cb} + V^*_{td}V_{tb} & = & Z^d_{db} \\ V^*_{us}V_{ub} + V^*_{cs}V_{cb} + V^*_{ts}V_{tb} & = & Z^d_{sb} \end{array} \tag{118}$$

Eqs. (118) show the important connection between deviations of 3×3 unitarity and FCNC. Although this relation was derived within the context of a minimal extension of the SM with an isosinglet quark, it is clear that it is valid in a larger class of models. The Z - mediated FCNC can give significant contributions to $B_d - \overline{B}_d$ and $B_s - \overline{B}_s$ mixings which in turn can produce important departures from the SM predictions for CP asymmetries in B^0 decays.

3.4 Summary

It is worthwhile summarizing the main points of this section:

• In the framework of unified gauge theories, CP violation can be introduced explicitly at the Lagrangian level or through spontaneous CP breaking.

• The minimal model with explicit CP violation is the Standard Model, where CP violation is introduced through complex Yukawa couplings, which for three or more generations lead to non-real rephasing invariant functions of V_{CKM} which in turn imply CP violation.

• Within the framework of $SU(2) \times U(1) \times SU(3)$ gauge theories with no extra symmetries and with only standard quarks, the minimal Higgs structure which leads to spontaneous CP violation consists of two Higgs doublets. This is the Lee model, where CP originates in a single relative phase between the vevs of the two scalar doublets. This relative phase enters in the quark mass matrices in such a way that it generates a complex V_{CKM} in charged gauge interactions, leading to CP violation. The Lee model has another source of CP violation which arises from tree level flavour-changing-neutral-currents mediated by physical neutral scalars.

• If one imposes exact NFC in the Higgs sector, the minimal structure which leads to spontaneous CP violation consists of three Higgs doublets. In this case V_{CKM} is real and CP violation arises exclusively from physical Higgs exchange.

• If one introduces an extra isosinglet quark, it is possible to have spontaneous CP violation with a simple Higgs structure, consisting of the standard higgs doublet plus one complex singlet scalar. CP violation originates in the phase of the vev of the complex singlet. Through mixing mass terms this phase appears unsuppressed in the 3×3 sector of V_{CKM} connecting standard quarks. A key feature of this class of models is the presence of deviations of 3×3 unitarity in V_{CKM} which in turn lead to Z-mediated flavour-changing-neutral currents.

4 TESTING THE SM AND SEARCHING FOR NEW PHYSICS

4.1 Introductory Remarks

CP violation is closely related to the least understood and least experimentally tested sectors of the SM, namely the Yukawa and Higgs sectors. The measurement at B-factories (BaBar, Belle, LHCB) of CP asymmetries in B decays, together the study of rare decays, will subject the SM to a very stringent test, with the potential for discovering New Physics. At this stage, it is worth recalling some of the main motivations for considering Physics Beyond the SM. In fact some of these motivations are closely related to Flavour Physics and CP Violation.

(i) In the SM, neutrinos are strictly massless. Present experimental evidence for neutrino oscillatons provides strong motivation to consider non-vanishing neutrino masses. It is possible to have neutrino masses and oscillations, in the framework of a minimal extension of the SM, where one simply adds right-handed neutrinos ν_R. Since ν_R are $SU(2) \times U(1)$ singlets, their mass terms can be arbitrarily large, not being protected by the low energy gauge symmetry. One is then naturally led to three light neutrinos, their mass being suppressed by the seesaw mechanism. With the introduction of non-vanishing neutrino masses, the general flavour problem becomes even more accute, since apart from the six quark masses, four V_{CKM} parameters and three charged-lepton masses, one has three light neutrino masses and the six parameters characterizing the leptonic mixing matrix (three angles, one Dirac CP violating phase and two Majorana CP violation phases). Therefore, altogether, one has twenty two free parameters characterizing fermion masses and mixing! It is clear that a theory with such a large number of free parameters cannot be the final theory. The need to solve the Flavour Problem, i.e. to understand the origin of family replication and the pattern of fermion masses and mixings, is one of the major motivations to consider Physics Beyond the SM.

(ii) CP violation is one of the crucial ingredients necessary to generate the observed matter-antimatter asymmetry in the Universe (Baryogenesis). By now, it has been established that the size of CP violation in the SM is not sufficient to generate the observed baryon asymmetry, thus hinting at New Physics, in particular new sources of CP violation. Whether the new sources of CP violation will manifest themselves at low energies, through deviations from the predictions of the SM and its KM mechanism of CP violation, is at present an entirely open question.

(iii) Finally, another important motivation to consider Physics Beyond the SM is the need to solve the gauge hierarchy problem. In the SM, the Higgs mass term is not protected by any symmetry from becoming arbitrarily large, of the order of the GUT or Planck scale. Low energy supersymmetry (SUSY) is an elegant way of solving the gauge hierarchy problem. The so called minimal supersymmetric standard model (MSSM) has in principle many new sources of CP violation which may lead to deviations of the predictions of the SM for CP asymmetries in B decays. The simplest way of obtaining these deviations within the MSSM, is by having sizable new SUSY contributions to B_d-$\bar{B}_d$ and B_s-$\bar{B}_s$ mixing.

4.2 Overdetermination of the Unitarity Triangle and New Physics

It is useful to consider a "normalised" unitarity triangle which is obtained by dividing the sides of the unitarity triangle of Fig.1 by $|V_{cd}V_{cb}^*|$. Taking into account that to an excellent approximation one has:

$$\begin{aligned} \tfrac{1}{A\lambda^3} V_{ud} V_{ub}^* &= \bar{\rho} + i\bar{\eta} \\ \tfrac{1}{A\lambda^3} V_{td} V_{tb}^* &= 1 - (\bar{\rho} + i\bar{\eta}) \end{aligned} \tag{119}$$

one obtains the unitarity triangle show in Fig.2 in the $\bar{\rho}$, $\bar{\eta}$ plane. One can then obtain, within the SM, the various constraints on the allowed location of the vertex of the triangle. In an ideal world with no experimental and theoretical uncertainties, one would need only two experimental inputs in order to determine the values of $\bar{\rho}$ and $\bar{\eta}$ and therefore the position of the vertex of the unitarity triangle. Unfortunately, one cannot avoid experimental errors and also theoretical uncertainties in the extraction

of constraints on $\bar{\rho}$ $\bar{\eta}$, from input experimental data. The great challenge for the SM is finding a region in the $\bar{\rho}$, $\bar{\eta}$ plane which is allowed by all experimental data. This is a very important test of the SM which will become increasingly stricter when more and better data becomes available from the on-going experiments on CP asymmetries in B^0 decays, as well as on rare kaon decays. At present, the most important experiments which constrain the location of the vertex of the UT are:

- The value of $|\frac{V_{ub}}{V_{cb}}|$
- The experimental value of $B_d - \bar{B}_d$ and $B_s - \bar{B}_s$ mixings
- The experimental value of the CP violation parameter ϵ

The experimental data on the above quantities is in agreement with the SM and restricts the vertex of the UT to be in a small region in the $\bar{\rho}$, $\bar{\eta}$ plane. The extraction of $\bar{\rho}$, $\bar{\eta}$ from experiment, within the SM enables one to make predictions for $\sin(2\alpha)$, $\sin(2\beta)$, and $\sin(2\gamma)$, since from simple trigonometry one obtains:

$$\begin{aligned} \sin 2\alpha &= \tfrac{2\bar{\eta}(\bar{\eta}^2+\bar{\rho}^2-\bar{\rho})}{(\bar{\rho}^2+\bar{\eta}^2)[(1-\bar{\rho})^2+\bar{\eta}^2]} \\ \sin 2\beta &= \tfrac{2\bar{\eta}(1-\bar{\rho})}{(1-\bar{\rho})^2+\bar{\eta}^2} \\ \sin 2\gamma &= \tfrac{2\bar{\rho}\bar{\eta}}{\bar{\rho}^2+\bar{\eta}^2} \end{aligned} \tag{120}$$

The derivation of the actual allowed values for α , β, γ is quite involved, since it depends on how one treats the theoretical uncertainties [31], [32], [33], [34], [35]. In a recent review [32], Buras suggests the following conservative bounds:

$$\begin{gathered} 78.8^\circ \leq \alpha \leq 120^\circ; \quad 15.1^\circ \leq \beta \leq 28.6^\circ; \quad 37.9^\circ \leq \gamma \leq 76.5^\circ \\ \sin 2\beta = 0.67 \pm 0.17 \end{gathered} \tag{121}$$

This is to be compared with the recent experimental results from BaBar [36]:

$$\sin(2\beta) = 0.34 \pm 0.20 \text{ (stat)} \pm 0.05 \text{ (syst)} \tag{122}$$

and Belle [37]:

$$\sin(2\beta) = 0.58^{+0.32}_{-0.34} \text{ (stat)}^{+0.09}_{-0.10} \text{ (syst)} \tag{123}$$

It is clear from the above equations that there is no disagreement between the SM predictions and the present experimental value of $\sin 2\beta$. However, it is conceivable that with more accurate data, a discrepancy may arise between experiment and the SM predictions. For example, the experimental value of $\sin 2\beta$ may turn out to be smaller than the lowest value allowed by the SM , which would signal the presence of New Physics. There are various ways of accounting for such deviations, which have been extensively discussed in the literature [38], [39], [40], [41], [42], [43]. Some time ago, it has been pointed out [44], [45] that one of the simplest ways of having deviations from the SM predictions for CP asymmetries is by assuming that there are new contributions to $B_d - \bar{B}_d$ mixing arising from Physics Beyond the SM. A simple example of New Physics would be Z-mediated flavour-changing-neutral-currents which could either dominate [44] or more generically give a significant contribution [45], [46] to $B_d - \bar{B}_d$ mixing. Let us consider the decays of neutral B mesons into final states f which are common to B^0 and $\overline{B}^0$. Due to $B^0 - \overline{B}^0$ mixing, CP violation arises through the interference of the two amplitudes $B^0 \to f$ and $B^0 \to \overline{B}^0 \to f$. In order to allow for Physics Beyond the SM we will write the off-diagonal element of $B_q - \overline{B}_q$ as [45]:

$$M_{12} = M_{12}^{(0)} \Delta_{qb} \qquad (q = d, s) \tag{124}$$

where $M_{12}^{(0)}$ denotes the box diagram contribution. Therefore $\Delta_{qb} \neq 1$ indicates new contributions to $B_q - \overline{B}_q$ mixing. Assuming that f is a CP eigenstate and that all amplitudes contributing to the decay have the same KM phase, the time dependent asymmetry is given by:

Table 1: The predicted values for the angles ϕ_{id}. The values shown are for CP even final states. Thus $\phi_{id} = -2\beta = -\phi_{\psi K_S}$.

Initial state	Quark subprocess	Final state	ϕ	Standard model	Beyond standard model
B_d	$\overline{b} \to \overline{c}c\overline{s}$	ψK_S	ϕ_{1d}	-2β	$-2\beta + \arg\Delta_{bd}$
	$\overline{b} \to \overline{c}c\overline{d}$	D^+D^-	ϕ_{2d}	-2β	$-2\beta + \arg\Delta_{bd}$
	$\overline{b} \to \overline{u}u\overline{d}$	$\pi^+\pi^-$	ϕ_{3d}	2α	$2\alpha + \arg\Delta_{bd}$

$$a \equiv \frac{\Gamma(B^0 \to f) - \Gamma(\overline{B}^0 \to f)}{\Gamma(B^0 \to f) + \Gamma(\overline{B}^0 \to f)} = -\xi \sin(\Delta M t) \sin\theta \tag{125}$$

where ξ is the CP eigenvalue of f and:

$$\begin{aligned} \phi \quad &= \phi^{(0)} + \arg\Delta_{bq} \qquad \Delta_{bq} = \Delta^*_{qb} \\ \phi^{(0)} \quad &= \arg\left[\left(\tfrac{q}{p}\right)^{(0)} \tfrac{\overline{A}(f)}{A(f)}\right] \quad ; \quad \left(\tfrac{q}{p}\right)^{(0)} = \left[\tfrac{M_{12}^{(0)*}}{M_{12}^{(0)}}\right]^{1/2} \end{aligned} \tag{126}$$

The index (0) denotes the contributions arising within the three generation SM, while $A(f)$, $\overline{A}(f)$ stand for the decay amplitudes from the initial state $|B^0>, |\overline{B}^0>$ to a CP eigenstate $|f>$. From Eq. (126) it follows that there are two possible sources which may lead to departures from the SM predictions for the CP asymmetries:

(i) The presence of the phase of Δ_{bq} which determines the deviation from the box diagram contribution $\phi^{(0)}$.

(ii) One may have a situation where the expression for $\phi^{(0)}$ is the one given by the SM, but the actual numerical value of $\phi^{(0)}$ differs from the SM prediction. This is due to the fact that models beyond the SM allow in general for a different range of the V_{CKM} matrix elements. In Table 1 we give explicitly the value of ϕ for various final states, in the SM and in models beyond the SM with $\arg\Delta bq \neq 0$.

So far our analysis has been to a great extent model independent. Let us now consider the special case where the new contribution to the $\Delta B = 2$ effective Lagrangian arises from Z exchange. We have seen that these Z mediated FCNC naturally arise in models with isosinglet quarks. One obtains in this case [45]:

$$\begin{aligned} \Delta_{bd} \quad &= 1 + r_d e^{i\theta_{bd}} \\ r_d \quad &= \tfrac{1}{\nu|\overline{E}(x_t)|} \left|\tfrac{Z_{bd}}{V_{td}V_{tb}^*}\right|^2 \\ \theta_{bd} \quad &= \arg\left[\tfrac{Z_{bd}}{V_{td}V_{tb}^*}\right] \end{aligned} \tag{127}$$

where $\nu = \alpha/4\pi\sin^2\theta_w$, $x_t \equiv (m_t/m_W)^2$ and $\overline{E}(x_t)$ is an Inami-Lim function for the top quark box diagram. It has been shown [45] that even for a relatively small contribution from the Z exchange diagram (i.e. $r_d < 1$) can produce drastic deviations from the SM prediction. For example, for $r_d \approx 0.2$, it is possible for a range of θ_{bd} to obtain the asymmetry $a(\psi K_S)$ with a sign opposite to the SM prediction [45]. The asymmetry $a(\psi K_S)$ is specially important since it is the only one whose sign is predicted with certainty by the SM. It is clear that a similar analysis can be done for other new physics contributions to $B_d - \overline{B}_d$ mixing like for example those arising from flavour violating neutral Higgs couplings. It

should be emphasized that the drastic deviations from the SM predictions for the CP asymmetries we have described, do not result from having very different values for the angles α, β, γ but are due to the fact that the various CP asymmetries no longer measure the angles α, β, γ [44], [45].

4.3 Concluding Remarks

Flavour Physics is an important subject which will certainly play a crucial rôle in the development of Particle Physics. Understanding the origin of the replication of fermion families and the pattern of fermion masses and mixings remains one of the fundamental open questions in Particle Physics. The on-going and the planned experiments on CP violation and rare decays will test the flavour sector of the SM and have the potential of uncovering New Physics. This New Physics could appear, for example, as :

- New contributions to $B_d - \bar{B}_d$ and/or to $B_s - \bar{B}_s$ mixing
- New contributions to ϵ
- New contributions to rare decays

The appearance of any of the above New Physics can lead to the impossibility of finding a location of the vertex of the unitary triangle where all data can be fitted, within the framework of the Standard Model. This would signal the presence of New Physics whose source could be, for example SUSY or multi-Higgs models with Spontaneous CP violation, or Z-mediated flavour-changing neutral currents, or, of course, something else. Hopefully, this New Physics will guide towards a solution of the Flavour Puzzle.

Acknowledgements

I am grateful to the organizers of the 2000 European School of Physics for having invited me to give these lectures, for the excellent organization and for making the School a very enjoyable experience. I would like to thank M. N. Rebelo for a careful reading of the manuscript.

References

[1] For a review see A. De Rujula, these Proceedings.

[2] Heavy Flavours II, eds. A. J. Buras and M. Lindner, World Scientific, (1998).

[3] G. C. Branco, L. Lavoura and J. Silva, (1999), CP Violation, Oxford Science Publications, Clarendon Press, Oxford.

[4] I. I. Bigi and A. I. Sanda, (2000), CP Violation, Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology, Cambridge University Press, Cambridge.

[5] The BaBar Physics Book, eds. P. Harrison and H. Quinn, (1998), SLAC report 504.

[6] B Decays at the LHC, eds. P. Ball, R. Fleischer, G. F. Tartarelli, P. Vikas and G. Wilkinson, hep-ph/0003238.

[7] A. J. Buras, Lectures given at the Erice International School of Subnuclear Physics (2000): "Theory and Experiment Heading for New Physics", hep-ph/0101336;
Y. Nir, hep-ph/0008226;
D. Wyler, hep-ph/0101259;
M. Gronau, hep-ph/0011392;
L. Wolfenstein hep-ph/0011400..

[8] Particle Data Group, *Eur. Phys. J.* **C 15** (2000) 1.

[9] T. Inami and C. S. Lim *Prog. Theor. Phys.* **65** (1981) 297, Erratum-ibid. **65** (1981) 1772.

[10] J. D. Bjorken and I. Dunietz, *Phys. Rev.* **D36** (1987) 2109.

[11] G. C. Branco and L. Lavoura, *Phys. Lett.* **B208** (1988) 123.

[12] R. Aleksan, B. Kayser and L. London, *Phys. Rev. Lett.* **73** (1994) 18.

[13] L. Wolfenstein, *Phys. Rev. Lett.* **51** (1983) 1945.

[14] G. C. Branco and L. Lavoura, *Phys. Rev.* **D38** (1988) 2295.

[15] A. J. Buras, M. E. Lautenbacher and G. Ostermaier, *Phys. Rev.* **D50** (1994) 3433.

[16] J. Bernabéu, G. C. Branco and M. Gronau, *Phys. Lett.* **169B** (1986) 243.

[17] W. Grimus and M. N. Rebelo, *Phys. Rept.* **281** (1997) 239.

[18] G. C. Branco and M. N. Rebelo, *Phys. Lett.* **B173** (1986) 313.

[19] T. D. Lee, *Phys. Rep.* **9 C** (1974) 143; *Phys. Rev.* **D8** (1973) 1226.

[20] L. Lavoura, *Int. J. Mod. Phys.* **A9** (1994) 1873.

[21] A. Antaramian, L. J. Hall and A. Rasin, Phy. Rev. Lett. 69 (1992) 1871;
L. J. Hall and S. Weinberg, *Phys. Rev.* **D48** (1993) 979; A. S. Joshipura, S. D. Ridani, *Phys. Lett.* **B260** (1991) 149.

[22] G. C. Branco, W. Grimus and L. Lavoura, *Phys. Lett.* **B380** (1996) 119.

[23] S. Weinberg, *Phys. Rev. Lett.* **37** (1976) 657.

[24] G. C. Branco, *Phys. Rev.* **D22** (1980) 2901.

[25] G. C. Branco and M. N. Rebelo, *Phys. Lett.* **160B** (1985) 117.

[26] G. C. Branco, A. J. Buras and J. M. Gérard, *Nucl. Phys.* **B259** (1985) 306.

[27] G. C. Branco, *Phys. Rev. Lett.* **44** (1980) 504.

[28] A. Nelson, *Phys. Lett.* **B136** (1984) 387;
S. M. Barr, *Phys. Rev. Lett.* **53** (1984) 329.

[29] For reviews, see R. D. Peccei, DESY Report 88-109 (1988), in CP Violation (World Scientific, Singapore);
J. Kim, *Phys. Rep.* **150** (1987) 1;
H. Y. Cheng, *Phys. Rep.* **158** (1988) 1.

[30] L. Bento, G. C. Branco and P. A. Parada, *Phys. Lett.* **B267** (1991) 95.

[31] M. Ciuchini *et al.*, hep-ph/0012308.

[32] A. J. Buras in [7].

[33] A. Ali and D. London, *Eur. Phys. J.* **C9** (1999) 687.

[34] Y. Grossman, Y. Nir, S. Plaszcynski and M.-H. Schune, *Nucl. Phys.* **B511** (1998) 69.

[35] J. L. Rosner, hep-ph/0011355.

[36] B. Aubert *et al.*, hep-ex/0102030.

[37] A. Abashian *et al.*, hep-ex/0102018.

[38] J. P. Silva and L. Wolfenstein, hep-ph/0008004.

[39] G. Eyal, Y. Nir and G. Perez, *JHEP* **0008** (2000) 028.

[40] A. L. Kagan and M. Neubert, **Phys. Lett. B429** (2000) 115.

[41] A. J. Buras and R. Buras, hep-ph/0008273.

[42] Z. Xing, hep-ph/0008018.

[43] G. Barenboim, F. J. Botella and O. Vives, hep-ph/0012197.

[44] Y. Nir and J. D. Silverman, *Phys. Rev.* **D42** (1990) 1477.

[45] G. C. Branco, T. Morozumi, P. A. Parada and M. N. Rebelo, *Phys. Rev.* **D48** (1993) 1167.

[46] G. Barenboim, F. J. Botella, G. C. Branco and O. Vives, *Phys. Lett.* **B422** (1998) 277.

BEYOND THE STANDARD MODEL
(in search of supersymmetry)

D. I. Kazakov
BLTP, JINR, Dubna and ITEP, Moscow
e-mail: kazakovd@thsun1.jinr.ru

Abstract
The present lectures contain an introduction to low energy supersymmetry, a new symmetry that relates bosons and fermions, in particle physics. The Standard Model of fundamental interactions is briefly reviewed, and the motivation to introduce supersymmetry is discussed. The main notions of supersymmetry are introduced. The supersymmetric extension of the Standard Model - the Minimal Supersymmetric Standard Model - is considered in more detail. Phenomenological features of the MSSM as well as possible experimental signatures of SUSY are described. An intriguing situation with the supersymmetric Higgs boson is discussed.

1 INTRODUCTION. THE STANDARD MODEL AND BEYOND

The Standard Model (SM) of fundamental interactions describes *strong, weak* and *electromagnetic* interactions of elementary particles [1]. It is based on a *gauge principle,* according to which all the forces of Nature are mediated by an exchange of the gauge fields of the corresponding local symmetry group. The symmetry group of the SM is

$$SU_{colour}(3) \otimes SU_{left}(2) \otimes U_{hypercharge}(1), \tag{1.1}$$

whereas the field content is the following:

Gauge sector : Spin = 1

The gauge bosons are spin 1 vector particles belonging to the adjoint representation of the group (1.1). Their quantum numbers with respect to $SU(3) \otimes SU(2) \otimes U(1)$ are

$$\begin{array}{lllll} \textit{gluons} & G^a_\mu : & (\underline{8},\underline{1},0) & SU_c(3) & g_s, \\ \textit{intermediate weak bosons} & W^i_\mu : & (\underline{1},\underline{3},0) & SU_L(2) & g, \\ \textit{abelian boson} & B_\mu : & (\underline{1},\underline{1},0) & U_Y(1) & g', \end{array} \tag{1.2}$$

where the coupling constants are usually denoted by g_s, g and g', respectively.

Fermion sector : Spin = 1/2

The matter fields are fermions belonging to the fundamental representation of the gauge group. These are believed to be quarks and leptons of at least of three generations. The SM is left-right asym-

metric. Left-handed and right-handed fermions have different quantum numbers

$$
\begin{array}{lrcccccc}
\textit{quarks} & & & & & & & \\
Q^i_{\alpha L} = & \begin{pmatrix} U^i_\alpha \\ D^i_\alpha \end{pmatrix}_L & = & \begin{pmatrix} u^i \\ d^i \end{pmatrix}_L, & \begin{pmatrix} c^i \\ s^i \end{pmatrix}_L, & \begin{pmatrix} t^i \\ b^i \end{pmatrix}_L, & \ldots & (\underline{3}, \underline{2}, 1/3) \\
 & U^i_{\alpha R} & = & u_{iR}, & c_{iR}, & t_{iR}, & \ldots & (\underline{3}^*, \underline{1}, 4/3) \\
 & D^i_{\alpha R} & = & d_{iR}, & s_{iR}, & b_{iR}, & \ldots & (\underline{3}^*, \underline{1}, -2/3) \\
\textit{leptons} & L_{\alpha L} & = & \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, & \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, & \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L, & \ldots & (\underline{1}, \underline{2}, -1) \\
 & E_{\alpha R} & = & e_R, & \mu_R, & \tau_R, & \ldots & (\underline{1}, \underline{1}, -2)
\end{array}
\tag{1.3}
$$

$i = 1, 2, 3$ - colour, $\alpha = 1, 2, 3, \ldots$ - generation.

Higgs sector : Spin = 0

In the minimal version of the SM there is one doublet of Higgs scalar fields

$$H = \begin{pmatrix} H^0 \\ H^- \end{pmatrix} \quad (\underline{1}, \underline{2}, -1), \tag{1.4}$$

which is introduced in order to give masses to quarks, leptons and intermediate weak bosons via spontaneous breaking of electroweak symmetry.

In the framework of Quantum Field Theory the SM is described by the following Lagrangian:

$$\mathcal{L} = \mathcal{L}_{gauge} + \mathcal{L}_{Yukawa} + \mathcal{L}_{Higgs}, \tag{1.5}$$

$$
\begin{aligned}
\mathcal{L}_{gauge} &= -\frac{1}{4} G^a_{\mu\nu} G^a_{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W^i_{\mu\nu} - \frac{1}{4} B_{\mu\nu} B_{\mu\nu} \\
&+ i\overline{L}_\alpha \gamma^\mu D_\mu L_\alpha + i\overline{Q}_\alpha \gamma^\mu D_\mu Q_\alpha + i\overline{E}_\alpha \gamma^\mu D_\mu E_\alpha \\
&+ i\overline{U}_\alpha \gamma^\mu D_\mu U_\alpha + i\overline{D}_\alpha \gamma^\mu D_\mu D_\alpha + (D_\mu H)^\dagger (D_\mu H),
\end{aligned}
\tag{1.6}
$$

where

$$
\begin{aligned}
G^a_{\mu\nu} &= \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^{abc} G^b_\mu G^c_\nu, \\
W^i_{\mu\nu} &= \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g \epsilon^{ijk} W^j_\mu W^k_\nu, \\
B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu, \\
D_\mu L_\alpha &= (\partial_\mu - i\frac{g}{2} \tau^i W^i_\mu + i\frac{g'}{2} B_\mu) L_\alpha, \\
D_\mu E_\alpha &= (\partial_\mu + i g' B_\mu) E_\alpha, \\
D_\mu Q_\alpha &= (\partial_\mu - i\frac{g}{2} \tau^i W^i_\mu - i\frac{g'}{6} B_\mu - i\frac{g_s}{2} \lambda^a G^a_\mu) Q_\alpha, \\
D_\mu U_\alpha &= (\partial_\mu - i\frac{2}{3} g' B_\mu - i\frac{g_s}{2} \lambda^a G^a_\mu) U_\alpha, \\
D_\mu D_\alpha &= (\partial_\mu + i\frac{1}{3} g' B_\mu - i\frac{g_s}{2} \lambda^a G^a_\mu) D_\alpha.
\end{aligned}
$$

$$\mathcal{L}_{Yukawa} = y^L_{\alpha\beta} \overline{L}_\alpha E_\beta H + y^D_{\alpha\beta} \overline{Q}_\alpha D_\beta H + y^U_{\alpha\beta} \overline{Q}_\alpha U_\beta \tilde{H} + h.c., \tag{1.7}$$

where $\tilde{H} = i\tau_2 H^\dagger$.

$$\mathcal{L}_{Higgs} = -V = m^2 H^\dagger H - \frac{\lambda}{2} (H^\dagger H)^2. \tag{1.8}$$

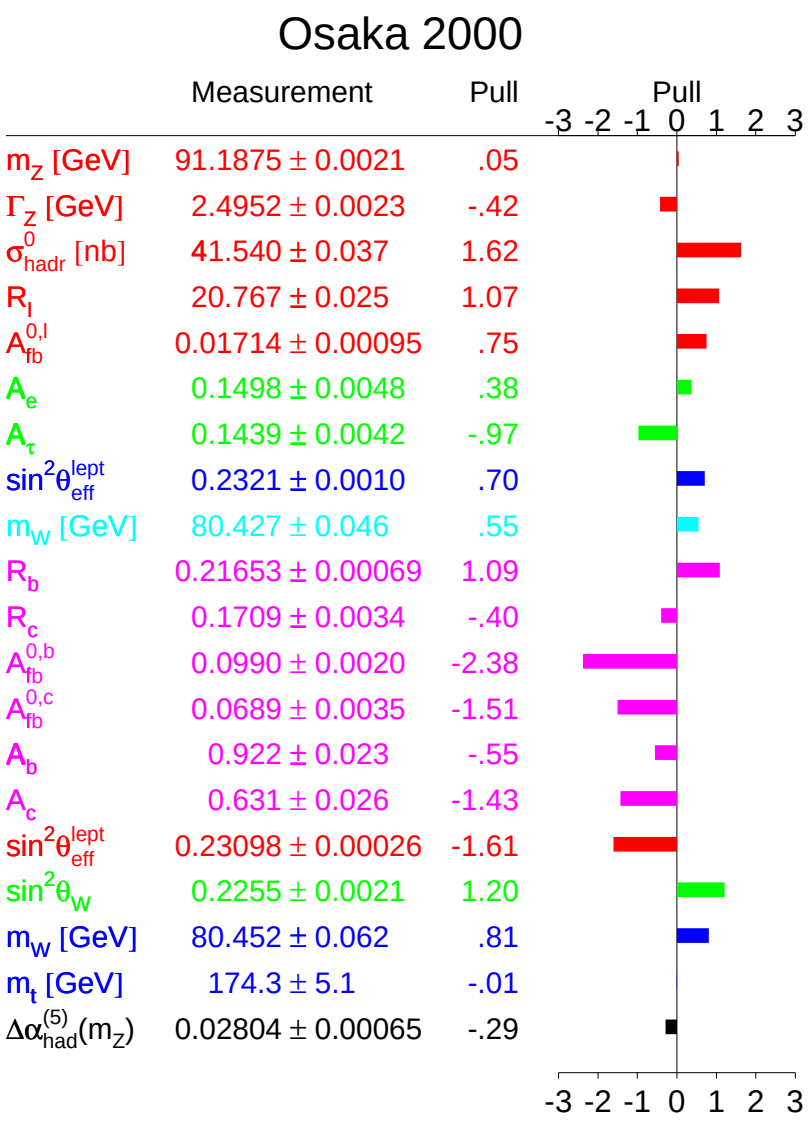


Fig. 1: Global Fit of the Standard Model

Here $\{y\}$ are the Yukawa and λ is the Higgs coupling constants, both dimensionless, and m is the only dimensional mass parameter[1].

The Lagrangian of the SM contains the following set of free parameters:

- 3 gauge couplings g_s, g, g';
- 3 Yukawa matrices $y^L_{\alpha\beta}, y^D_{\alpha\beta}, y^U_{\alpha\beta}$;
- Higgs coupling constant λ;
- Higgs mass parameter m^2;
- number of matter fields (generations).

All the particles obtain their masses due to spontaneous breaking of $SU_{left}(2)$ symmetry group

[1]We use the usual for particle physics units $c = \hbar = 1$

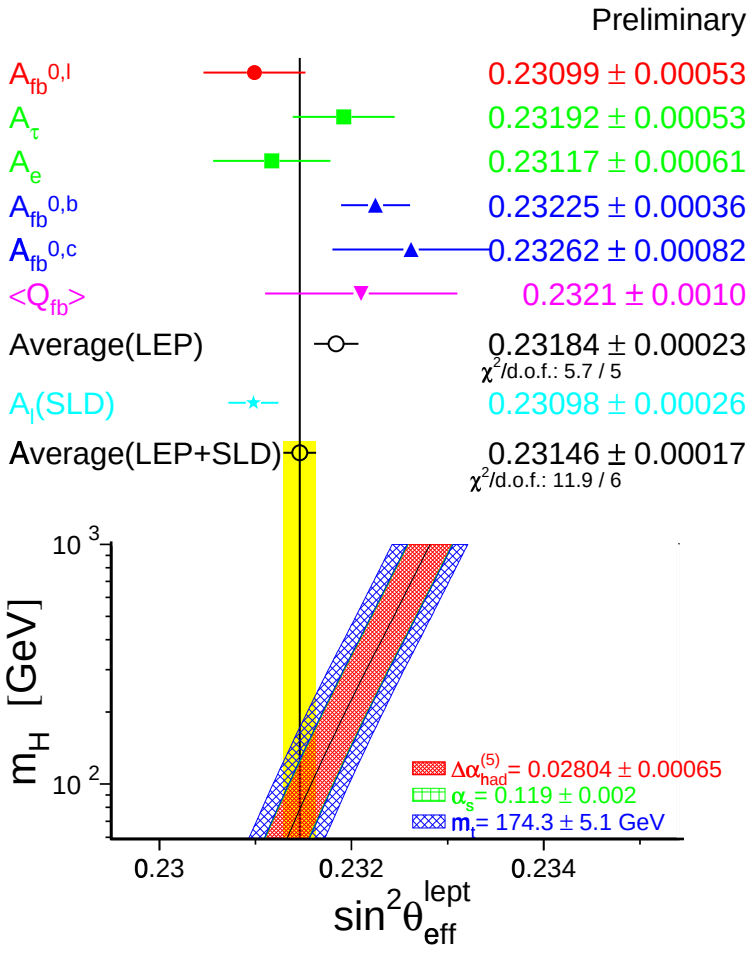


Fig. 2: Weak mixing angle and the Higgs boson mass

via a non-zero vacuum expectation value (v.e.v.) of the Higgs field

$$< H >= \begin{pmatrix} v \\ 0 \end{pmatrix}, \quad v = m/\sqrt{\lambda}. \tag{1.9}$$

As a result, the gauge group of the SM is spontaneously broken down to

$$SU_c(3) \otimes SU_L(2) \otimes U_Y(1) \Rightarrow SU_c(3) \otimes U_{EM}(1).$$

The physical weak intermediate bosons are linear combinations of the gauge ones

$$W_\mu^\pm = \frac{W_\mu^1 \mp iW_\mu^2}{\sqrt{2}}, \quad Z_\mu = -\sin\theta_W B_\mu + \cos\theta_W W_\mu^3 \tag{1.10}$$

with masses

$$m_W = \frac{1}{\sqrt{2}} gv, \quad m_Z = m_W/\cos\theta_W, \quad \tan\theta_W = g'/g, \tag{1.11}$$

while the photon field

$$\gamma_\mu = \cos\theta_W B_\mu + \sin\theta_W W_\mu^3 \tag{1.12}$$

remains massless.

The matter fields acquire masses proportional to the corresponding Yukawa couplings:

$$M^u_{\alpha\beta} = y^u_{\alpha\beta} v,\ M^d_{\alpha\beta} = y^d_{\alpha\beta} v,\ M^l_{\alpha\beta} = y^l_{\alpha\beta} v,\ m_H = \sqrt{2} m. \tag{1.13}$$

Explicit mass terms in the Lagrangian are forbidden because they are not $SU_{left}(2)$ symmetric and would destroy the renormalizability of the Standard Model.

The SM has been constructed as a result of numerous efforts both theoretical and experimental. At present, the SM is extraordinary successful, the achieved accuracy of its predictions corresponds to experimental data within 5 % [1, 2]. The combined results of the Global SM fit are shown in Fig.1 [2]. All the particles, except for Higgs boson, have been discovered experimentally. And the mass of the Higgs boson is severely constrained from precision electroweak data (see Fig.2 [2]).

However, the SM has its natural drawbacks and unsolved problems. Among them are

- inconsistency of the SM as a QFT (Landau pole),
- large number of free parameters,
- formal unification of strong and electroweak interactions,
- still unclear mechanism of EW symmetry breaking: The Higgs boson has not yet been observed and it is not clear whether it is fundamental or composite,
- the problem of CP-violation is not well understood including CP-violation in a strong interaction,
- flavour mixing and the number of generations are arbitrary,
- the origin of the mass spectrum is unclear.

The answer to these problems lies beyond the SM. There are two possible ways of going beyond the SM

$\Rightarrow$ To consider the *same* fundamental fields with *new* interactions. This way leads us to supersymmetry, Grand Unification, String Theory, etc. It seems to be favoured by modern experimental data.

$\Rightarrow$ To consider *new* fundamental fields with *new* interactions. This way leads us to compositeness, fermion-antifermion condensates, Technicolour, extended Technicolour, preons, etc. It is not favoured by data at the moment.

There are also possible exotic ways out of the SM: gravity at TeV energies, large extra dimensions, brane world, etc. We do not consider them here. In what follows we go along the lines of the first possibility and describe supersymmetry as a nearest option for the new physics on TeV scale.

2 WHAT IS SUPERSYMMETRY? MOTIVATION IN PARTICLE PHYSICS

Supersymmetry or fermion-boson symmetry has not yet been observed in Nature. This is a purely theoretical invention [3]. Its validity in particle physics follows from the common belief in unification. Over 30 years thousands of papers have been written on supersymmetry. For reviews see, e.g. Refs.[4]-[8].

2.1 Unification with gravity

The *general idea* is a unification of all forces of Nature. It defines the *strategy* : increasing unification towards smaller distances up to $l_{Pl} \sim 10^{-33}$ cm including quantum gravity. However, the graviton has spin 2, while the other gauge bosons (photon, gluons, W and Z weak bosons) have spin 1. Therefore, they correspond to different representations of the Poincaré algebra. Attempts to unify all four forces within the same algebra face a problem. Due to no-go theorems [9], unification of spin 2 and spin 1 gauge fields within a unique algebra is forbidden. The only exception from this theorem is supersymmetry algebra. The *uniqueness* of SUSY is due to a strict mathematical statement that algebra of SUSY is the *only* graded (i.e. containing anticommutators as well as commutators) Lie algebra possible within relativistic field theory [9].

If Q is a generator of SUSY algebra, then

$$Q|boson>=|fermion> \quad \text{and} \quad Q|fermion>=|boson> .$$

Hence, starting with the graviton state of spin 2 and acting by SUSY generators we get the following chain of states:

$$spin\ 2 \ \rightarrow \ spin\ 3/2 \ \rightarrow \ spin\ 1 \ \rightarrow \ spin\ 1/2 \ \rightarrow \ spin\ 0.$$

Thus, a partial unification of matter (fermions) with forces (bosons) naturally arises from an attempt to unify gravity with other interactions.

SUSY algebra appears as a generalization of Poincaré algebra (see next section) and links together various representations with different spins. The key relation is given by the anticommutator

$$\{Q_\alpha, \bar{Q}_{\dot{\alpha}}\} = 2\sigma^\mu_{\alpha,\dot{\alpha}} P_\mu.$$

Taking infinitesimal transformations $\delta_\epsilon = \epsilon^\alpha Q_\alpha,\ \bar{\delta}_{\bar{\epsilon}} = \bar{Q}_{\dot{\alpha}} \bar{\epsilon}^{\dot{\alpha}}$, one gets

$$\{\delta_\epsilon, \bar{\delta}_{\bar{\epsilon}}\} = 2(\epsilon\sigma^\mu\bar{\epsilon})P_\mu, \tag{2.1}$$

where ϵ is a transformation parameter. Choosing ϵ to be local, i.e. a function of a space-time point $\epsilon = \epsilon(x)$, one finds from eq.(2.1) that an anticommutator of two SUSY transformations is a local coordinate translation. And a theory which is invariant under the general coordinate transformation is General Relativity. Thus, making SUSY local, one obtains General Relativity, or a theory of gravity, or supergravity [10].

Theoretical attractiveness of SUSY field theories is explained by remarkable properties of SUSY models. This is first of all cancellation of ultraviolet divergencies in rigid SUSY theories which is the origin of

- possible solution of the hierarchy problem in GUTs;
- vanishing of the cosmological constant;
- integrability, allowing for an exact non-perturbative solution.

It is believed that along these lines one can also obtain the unification of all forces of Nature including quantum (super)gravity.

What is essential, the standard concepts of QFT allow SUSY without any further assumptions, it is straightforward to construct the supersymmetric generalization of the SM. Moreover, it can be checked experimentally! In recent years, supersymmetry became a subject of intensive experimental tests. Its predictions can be verified at modern and future colliders.

2.2 Unification of gauge couplings

Since the main motivation for SUSY is related with the unification theory, let us briefly recall the main ideas of the Grand Unification [11].

The philosophy of Grand Unification is based on a *hypothesis*: Gauge symmetry increases with energy. Having in mind unification of all forces of Nature on a common basis and neglecting gravity for the time being due to its weakness, the idea of GUTs is the following:

All known interactions are different branches of a unique interaction associated with a simple gauge group. The unification (or splitting) occurs at high energy

Low energy			$\Rightarrow$	High energy
$SU_c(3)\otimes$	$SU_L(2)\otimes$	$U_Y(1)$	$\Rightarrow$	G_{GUT} $(or\ G^n + \text{discrete symmetry})$
gluons	W, Z	photon	$\Rightarrow$	gauge bosons
quarks	leptons		$\Rightarrow$	fermions
g_3	g_2	g_1	$\Rightarrow$	g_{GUT}

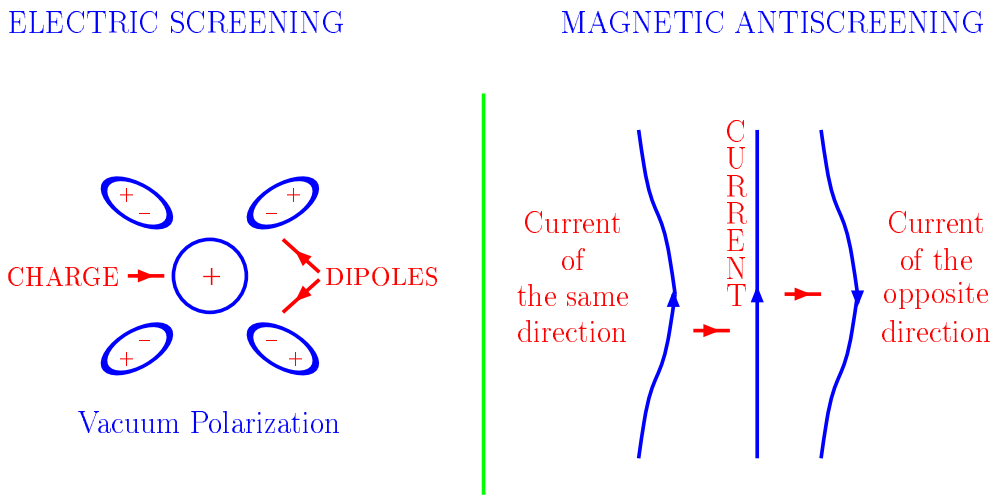


Fig. 3: Electric screening and magnetic antiscreening

At first sight this is impossible due to a big difference in the values of the couplings of strong, weak and electromagnetic interactions. However, this is not so. The crucial point here is the running coupling constants. It is a generic property of quantum field theory which has an analogy in classical physics.

Indeed, consider electric and magnetic phenomena. Let us take some dielectric medium and put a sample electric charge in it. What happens is that the medium is polarized. It contains electric dipoles which are arranged in such a way as to screen the charge (see Fig.3). It is a consequence of the Coulomb law: attraction of the opposite charges and repulsion of the same ones. This is the origin of electric screening.

The opposite situation occurs in a magnetic medium. According to the Biot-Savart law, electric currents of the same direction are attracted to each other, while those of the opposite one are repulsed (see Fig.3). This leads to antiscreening of electric currents in a magnetic medium.

In QFT, the role of the medium is played by the vacuum. Vacuum is polarized due to the presence of virtual pairs of particles in it. The matter fields and transverse quanta of vector fields in this case behave like dipoles in a dielectric medium and cause screening, while the longitudinal quanta of vector fields behave like currents and cause antiscreening. These two effects compete each other (see eq.(2.6) below).

Thus, the couplings become the functions of a distance or an energy scale

$$\alpha_i = \alpha_i(\frac{Q^2}{\Lambda^2}) = \alpha_i(\text{distance}), \qquad \alpha_i \equiv g_i^2/4\pi.$$

This dependence is described by the renormalization group equations and is confirmed experimentally (see Fig.4).

In the SM the strong and weak couplings associated with non-Abelian gauge groups decrease with energy, while the electromagnetic one associated with the Abelian group on the contrary increases. Thus, it becomes possible that at some energy scale they become equal. According to the GUT idea, this equality is not occasional but is a manifestation of a unique origin of these three interactions. As a result of spontaneous symmetry breaking, the unifying group is broken and the unique interaction is splitted into three branches which we call strong, weak and electromagnetic interactions. This happens at a very high energy of an order of $10^{15\div16}$ GeV. Of course, this energy is out of the range of accelerators; however, some crucial predictions follow from the very fact of unification.

After the precise measurement of the $SU(3) \times SU(2) \times U(1)$ coupling constants, it has become

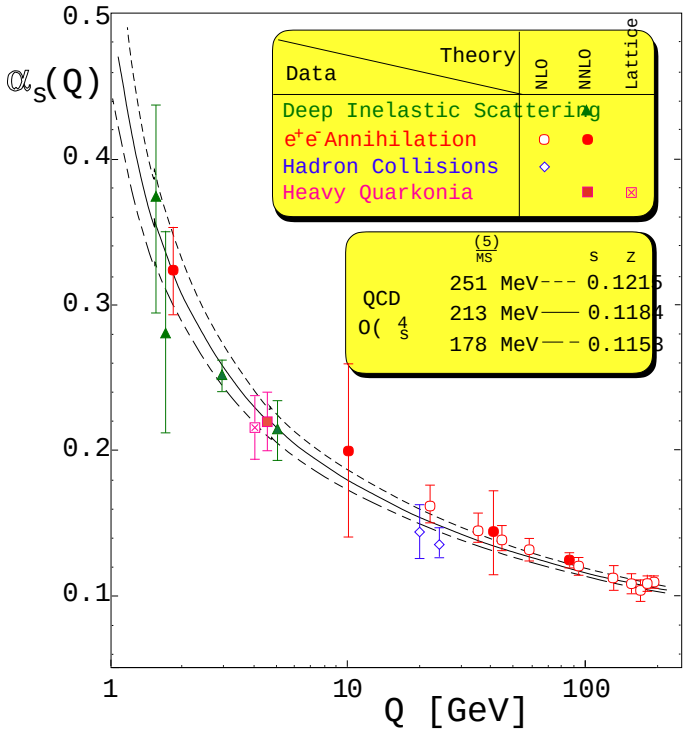


Fig. 4: Summary of running of the strong coupling α_s [12]

possible to check the unification numerically.

The three coupling constants to be compared are

$$\begin{aligned} \alpha_1 &= (5/3)g'^2/(4\pi) = 5\alpha/(3\cos^2\theta_W), \\ \alpha_2 &= g^2/(4\pi) = \alpha/\sin^2\theta_W, \\ \alpha_3 &= g_s^2/(4\pi) \end{aligned} \tag{2.2}$$

where g', g and g_s are the usual $U(1)$, $SU(2)$ and $SU(3)$ coupling constants and α is the fine structure constant. The factor of 5/3 in the definition of α_1 has been included for proper normalization of the generators.

The couplings, when defined as renormalized values including loop corrections require the specification of a renormalization prescription for which the modified minimal subtraction ($\overline{MS}$) scheme [13] is used.

In this scheme, the world averaged values of the couplings at the Z^0 energy are obtained from a fit to the LEP and Tevatron data [14],[2],[12]:

$$\begin{aligned} \alpha^{-1}(M_Z) &= 128.978 \pm 0.027 \\ \sin^2\theta_{\overline{MS}} &= 0.23146 \pm 0.00017 \\ \alpha_s &= 0.1184 \pm 0.0031, \end{aligned} \tag{2.3}$$

that gives

$$\alpha_1(M_Z) = 0.017, \quad \alpha_2(M_Z) = 0.034, \quad \alpha_3(M_Z) = 0.118 \pm 0.003. \tag{2.4}$$

Assuming that the SM is valid up to the unification scale, one can then use the known RG equations for the three couplings. They are the following:

$$\frac{d\tilde{\alpha}_i}{dt} = b_i\tilde{\alpha}_i^2, \quad \tilde{\alpha}_i = \frac{\alpha_i}{4\pi}, \quad t = log(\frac{Q^2}{\mu^2}), \tag{2.5}$$

where for the SM the coefficients b_i are

$$b_i = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -22/3 \\ -11 \end{pmatrix} + N_{Fam} \begin{pmatrix} 4/3 \\ 4/3 \\ 4/3 \end{pmatrix} + N_{Higgs} \begin{pmatrix} 1/10 \\ 1/6 \\ 0 \end{pmatrix}. \tag{2.6}$$

Here N_{Fam} is the number of generations of matter multiplets and N_{Higgs} is the number of Higgs doublets. We use $N_{Fam} = 3$ and $N_{Higgs} = 1$ for the minimal SM, which gives $b_i = (41/10, -19/6, -7)$.

Notice a positive contribution (screening) from the matter multiplets and negative one (antiscreening) from the gauge fields. For the Abelian group $U(1)$ this contribution is absent due to the absence of a self-interaction of Abelian gauge fields.

The solution to eq.(2.5) is very simple

$$\frac{1}{\tilde{\alpha}_i(Q^2)} = \frac{1}{\tilde{\alpha}_i(\mu^2)} - b_i log(\frac{Q^2}{\mu^2}). \tag{2.7}$$

The result is demonstrated in Fig.5 showing the evolution of the inverse of the couplings as a function of the logarithm of energy. In this presentation, the evolution becomes a straight line in first order. The second order corrections are small and do not cause any visible deviation from a straight line. Fig.5 clearly demonstrates that within the SM the coupling constant unification at a single point is impossible. It is excluded by more than 8 standard deviations. This result means that the unification can only be obtained if new physics enters between the electroweak and the Planck scales!

Since we do not know what kind of new physics it may be, there is a lot of arbitrariness. In this situation, some guiding idea is needed. It is very tempting to try to check whether unification is possible within a supersymmetric generalization of the SM. In the SUSY case, the slopes of the RG evolution curves are modified. The coefficients b_i in eq.(2.5) now are

$$b_i = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -6 \\ -9 \end{pmatrix} + N_{Fam} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + N_{Higgs} \begin{pmatrix} 3/10 \\ 1/2 \\ 0 \end{pmatrix}, \tag{2.8}$$

where we use $N_{Fam} = 3$ and $N_{Higgs} = 2$ in the minimal SUSY model which gives $b_i = (33/5, 1, -3)$.

It turns out that within the SUSY model a perfect unification can be obtained if the SUSY masses are of an order of 1 TeV. This is shown in Fig.6; the SUSY particles are assumed to effectively contribute to the running of the coupling constants only for energies above the typical SUSY mass scale, which causes the change in the slope of the lines near 1 TeV. From the fit requiring unification one finds for the break point M_{SUSY} and the unification point M_{GUT} [15]

$$\begin{aligned} M_{SUSY} &= 10^{3.4\pm0.9\pm0.4}\ GeV, \\ M_{GUT} &= 10^{15.8\pm0.3\pm0.1}\ GeV, \\ \alpha_{GUT}^{-1} &= 26.3 \pm 1.9 \pm 1.0, \end{aligned} \tag{2.9}$$

where $\alpha_{GUT} = g_5^2/4\pi$. The first error originates from the uncertainty in the coupling constant, while the second one is due to the uncertainty in the mass splittings between the SUSY particles. The χ^2 distributions of M_{SUSY} and M_{GUT} are shown in Fig.6 [15], where

$$\chi^2 = \sum_{i=1}^{3} \frac{(\alpha_i^{-1} - \alpha_{GUT}^{-1})^2}{\sigma_i^2}. \tag{2.10}$$

For SUSY models, the dimensional reduction $\overline{DR}$ scheme is a more appropriate renormalization scheme [16]. In this scheme, all thresholds are treated by simple step approximations, and unification

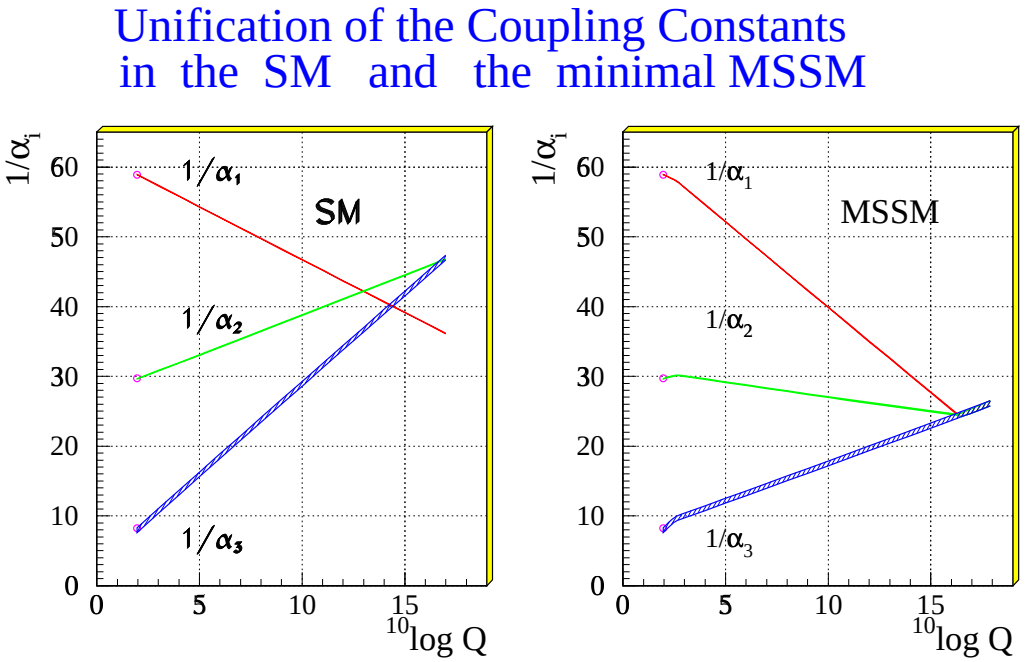


Fig. 5: Evolution of the inverse of the three coupling constants in the Standard Model (left) and in the supersymmetric extension of the SM (MSSM) (right). Only in the latter case unification is obtained. The SUSY particles are assumed to contribute only above the effective SUSY scale M_{SUSY} of about 1 TeV, which causes a change in the slope in the evolution of couplings. The thickness of the lines represents the error in the coupling constants [15].

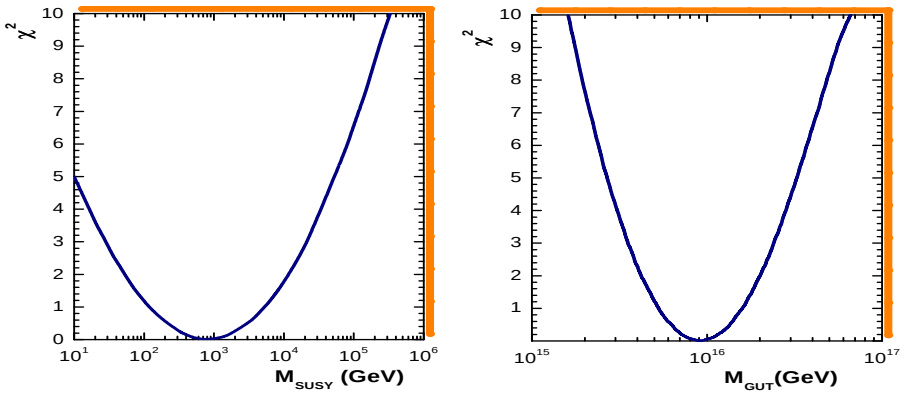


Fig. 6: The χ^2 distributions of M_{SUSY} and M_{GUT}

occurs if all three α's meet exactly at one point. This crossing point corresponds to the mass of the heavy gauge bosons. The $\overline{MS}$ and $\overline{DR}$ couplings differ by a small offset

$$\frac{1}{\alpha_i^{\overline{DR}}} = \frac{1}{\alpha_i^{\overline{MS}}} - \frac{C_i}{12\pi}, \tag{2.11}$$

where C_i are the quadratic Casimir operators of the group ($C_i = N$ for SU(N) and 0 for U(1) so α_1 remains the same).

This observation was considered as the first "evidence" for supersymmetry, especially since M_{SUSY} was found in the range preferred by the fine-tuning arguments.

It should be noted that the unification of the three curves at a single point is not that trivial as it

may seem from the existence of three free parameters (M_{SUSY}, M_{GUT} and α_{GUT}). Out of more than a thousand models tried, only a handful yielded unification. The reason is simple: Introducing new particles one influences all three curves simultaneously, thus giving rise to strong correlations between the slopes of the three lines. For example, adding new generations and/or new Higgs doublets never yields unification! Nevertheless, unification does not prove supersymmetry. The real proof would be the observation of the sparticles.

2.3 Solution of the hierarchy problem

The appearance of two different scales $V \gg v$ in a GUT theory, namely, M_W and M_{GUT}, leads to a very serious problem which is called the *hierarchy problem*. There are two aspects of this problem.

The first one is the very existence of the hierarchy. To get the desired spontaneous symmetry breaking pattern, one needs

$$\begin{array}{ccccc} m_H & \sim & v & \sim & 10^2 \text{ GeV} \\ m_\Sigma & \sim & V & \sim & 10^{16} \text{ GeV} \end{array} \qquad \frac{m_H}{m_\Sigma} \sim 10^{-14} \ll 1, \tag{2.12}$$

where H and Σ are the Higgs fields responsible for the spontaneous breaking of the $SU(2)$ and the GUT groups, respectively.

The question arises of how to get so small number in a natural way. One needs some kind of fine tuning in a theory, and we don't know if there anything behind it.

The second aspect of the hierarchy problem is connected with the preservation of a given hierarchy. Even if we choose the hierarchy like eq.(2.12) the radiative corrections will destroy it! To see this, consider the radiative correction to the light Higgs mass. It is given by the Feynman diagram shown in Fig.7 and is proportional to the mass squared of the heavy particle. This correction obviously spoils the

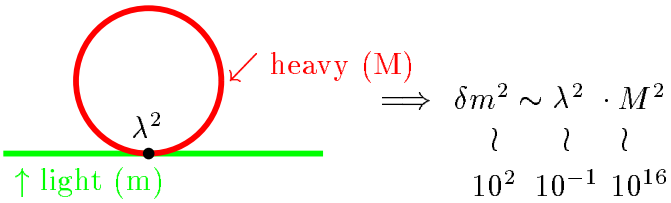


Fig. 7: Radiative correction to the light Higgs boson mass

hierarchy if it is not cancelled. This very accurate cancellation with a precision $\sim 10^{-14}$ needs a fine tuning of the coupling constants.

The only known way of achieving this kind of cancellation of quadratic terms (also known as the cancellation of the quadratic divergencies) is supersymmetry. Moreover, SUSY automatically cancels quadratic corrections in all orders of PT. This is due to the contributions of superpartners of ordinary particles. The contribution from boson loops cancels those from the fermion ones because of an additional factor (-1) coming from Fermi statistics, as shown in Fig.8. One can see here two types of contribution. The first line is the contribution of the heavy Higgs boson and its superpartner. The strength of interaction is given by the Yukawa coupling λ. The second line represents the gauge interaction proportional to the gauge coupling constant g with the contribution from the heavy gauge boson and heavy gaugino.

In both the cases the cancellation of quadratic terms takes place. This cancellation is true in the case of unbroken supersymmetry due to the following sum rule relating the masses of superpartners

$$\sum_{bosons} m^2 = \sum_{fermions} m^2 \tag{2.13}$$

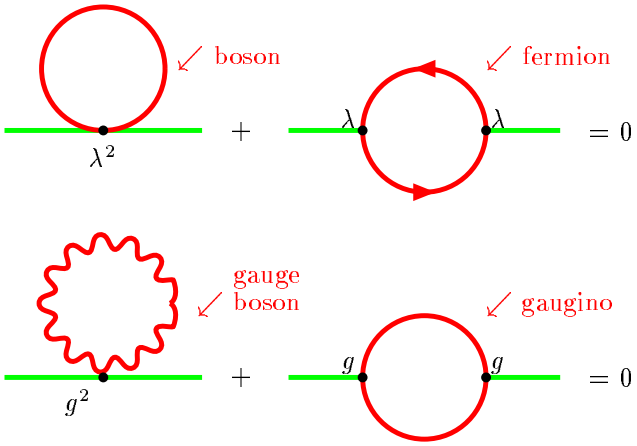


Fig. 8: Cancellation of quadratic terms (divergencies)

and is violated when SUSY is broken. Then, the cancellation is true up to the SUSY breaking scale, M_{SUSY}, since

$$\sum_{bosons} m^2 - \sum_{fermions} m^2 = M^2_{SUSY}, \tag{2.14}$$

which should not be very large (≤ 1 TeV) to make the fine-tuning natural. Indeed, let us take the Higgs boson mass. Requiring for consistency of perturbation theory that the radiative corrections to the Higgs boson mass do not exceed the mass itself gives

$$\delta M_h^2 \sim g^2 M^2_{SUSY} \sim M_h^2. \tag{2.15}$$

So, if $M_h \sim 10^2$ GeV and $g \sim 10^{-1}$, one needs $M_{SUSY} \sim 10^3$ GeV in order that the relation (2.15) is valid. Thus, we again get the same rough estimate of $M_{SUSY} \sim 1$ TeV as from the gauge coupling unification above. Two requirements match together.

That is why it is usually said that supersymmetry solves the hierarchy problem. Moreover, sometimes it is said that: "There is no GUT without SUSY". However, this is only the second aspect of the problem, the preservation of the hierarchy. The origin of the hierarchy is the other part of the problem. We show below how SUSY can explain this part as well.

2.4 Beyond GUTs: superstring

Another motivation for supersymmetry follows from even more radical changes of basic ideas related to the ultimate goal of construction of consistent unified theory of everything. At the moment the only viable conception is the superstring theory [17], which pretends to be a self-consistent quantum field theory in a non-perturbative sense allowing exact non-perturbative solutions in the quantum case. In the superstring theory, strings are considered as fundamental objects, closed or open, and are nonlocal in nature. Ordinary particles are considered as string excitation modes. String interactions, which are local, generate proper interactions of usual particles, including gravitational ones.

To be consistent, the string theory should be conformal invariant in D-dimensional target space and have a stable vacuum [18]. The first requirement is valid in classical theory but may be violated by quantum anomalies. Cancellation of quantum anomalies takes place when space-time dimension of a target space equals a critical one. For a bosonic string the critical dimension is $D = 26$, and for a fermionic one it is $D = 10$.

The second requirement is that the massless string excitations (the particles of the SM) are stable. This assumes the absence of tachyons, the states with imaginary mass, which can be guaranteed only in supersymmetric string theories!

Thus, the superstring theory proves to be the only known consistent quantum theory. This serves as justification of research in spite of absence of even a shred of experimental evidence. However, many ingredients of this theory are still unclear.

3 BASICS OF SUPERSYMMETRY

Supersymmetry transformations differ from ordinary global transformations as far as they convert bosons into fermions and vice versa. Indeed, if we symbolically write SUSY transformation as

$$\delta B = \varepsilon \cdot f,$$

where B and f are boson and fermion fields, respectively, and ε is an infinitesimal transformation parameter, then from the usual (anti)commutation relations for (fermions) bosons

$$\{f, f\} = 0, \quad [B, B] = 0$$

we immediately find

$$\{\varepsilon, \varepsilon\} = 0.$$

This means that all the generators of SUSY must be *fermionic*, i.e. they must change the spin by a half-odd amount and change the statistics.

3.1 Algebra of SUSY

Combined with the usual Poincaré and internal symmetry algebra the Super-Poincaré Lie algebra contains additional SUSY generators Q^i_α and $\bar{Q}^i_{\dot\alpha}$ [3]

$$\begin{array}{l}
[P_\mu, P_\nu] = 0, \\
[P_\mu, M_{\rho\sigma}] = i(g_{\mu\rho}P_\sigma - g_{\mu\sigma}P_\rho), \\
[M_{\mu\nu}, M_{\rho\sigma}] = i(g_{\nu\rho}M_{\mu\sigma} - g_{\nu\sigma}M_{\mu\rho} - g_{\mu\rho}M_{\nu\sigma} + g_{\mu\sigma}M_{\nu\rho}), \\
[B_r, B_s] = iC^t_{rs}B_t, \\
[B_r, P_\mu] = [B_r, M_{\mu\sigma}] = 0, \\
[Q^i_\alpha, P_\mu] = [\bar{Q}^i_{\dot\alpha}, P_\mu] = 0, \\
[Q^i_\alpha, M_{\mu\nu}] = \frac{1}{2}(\sigma_{\mu\nu})^\beta_\alpha Q^i_\beta, \quad [\bar{Q}^i_{\dot\alpha}, M_{\mu\nu}] = -\frac{1}{2}\bar{Q}^i_{\dot\beta}(\bar\sigma_{\mu\nu})^{\dot\beta}_{\dot\alpha}, \\
[Q^i_\alpha, B_r] = (b_r)^i_j Q^j_\alpha, \quad [\bar{Q}^i_{\dot\alpha}, B_r] = -\bar{Q}^j_{\dot\alpha}(b_r)^i_j, \\
\{Q^i_\alpha, \bar{Q}^j_{\dot\beta}\} = 2\delta^{ij}(\sigma^\mu)_{\alpha\dot\beta}P_\mu, \\
\{Q^i_\alpha, Q^j_\beta\} = 2\epsilon_{\alpha\beta}Z^{ij}, \quad Z_{ij} = a^r_{ij}b_r, \quad Z^{ij} = Z^+_{ij}, \\
\{\bar{Q}^i_{\dot\alpha}, \bar{Q}^j_{\dot\beta}\} = -2\epsilon_{\dot\alpha\dot\beta}Z^{ij}, \quad [Z_{ij}, anything] = 0, \\
\alpha, \dot\alpha = 1, 2 \quad i, j = 1, 2, \ldots, N.
\end{array} \tag{3.1}$$

Here P_μ and $M_{\mu\nu}$ are four-momentum and angular momentum operators, respectively, B_r are the internal symmetry generators, Q^i and $\bar{Q}^i$ are the spinorial SUSY generators and Z_{ij} are the so-called central charges; $\alpha, \dot\alpha, \beta, \dot\beta$ are the spinorial indices. In the simplest case one has one spinor generator Q_α (and the conjugated one $\bar{Q}_{\dot\alpha}$) that corresponds to an ordinary or N=1 supersymmetry. When $N > 1$ one has an extended supersymmetry.

A natural question arises: how many SUSY generators are possible, i.e. what is the value of N? To answer this question, consider massless states [5]. Let us start with the ground state labeled by energy and helicity, i.e. projection of a spin on the direction of momenta, and let it be annihilated by Q_i

$$\text{Vacuum} = |E, \lambda>, \qquad Q_i|E, \lambda> = 0.$$

Then one and more particle states can be constructed with the help of a creation operators as

State	Expression	# of States
vacuum	$\|E,\lambda>$	1
$1-$particle state	$\bar{Q}_i\|E,\lambda>=\|E,\lambda+1/2>_i$	$\left(\begin{array}{c} N \\ 1 \end{array}\right)=N$
$2-$particle state	$\bar{Q}_i\bar{Q}_j\|E,\lambda>=\|E,\lambda+1>_{ij}$	$\left(\begin{array}{c} N \\ 2 \end{array}\right)=\frac{N(N-1)}{2}$
...	...	...
$N-$particle state	$\bar{Q}_1\bar{Q}_2...\bar{Q}_N\|E,\lambda>=\|E,\lambda+N/2>$	$\left(\begin{array}{c} N \\ N \end{array}\right)=1$
Total # of States	$\sum_{k=0}^{N}\left(\begin{array}{c} N \\ k \end{array}\right)=2^N=2^{N-1}$ bosons $+\,2^{N-1}$ fermions,	

where the energy E is not changed, since according to (3.1) the operators $\bar{Q}_i$ commute with the Hamiltonian.

Thus, one has a sequence of bosonic and fermionic states and the total number of bosons equals that of fermions. This is a generic property of any supersymmetric theory. However, in CPT invariant theories the number of states is doubled, since CPT transformation changes the sign of helicity. Hence, in CPT invariant theories, one has to add the states with opposite helicity to the above mentioned ones.

Consider some examples. Let us take $N=1$ and $\lambda=0$. Then one has the following set of states:

$$N=1 \quad \lambda=0 \qquad \begin{array}{lccc} \text{helicity} & 0 & 1/2 \\ \# \text{ of states} & 1 & 1 \end{array} \quad \stackrel{CPT}{\Longrightarrow} \quad \begin{array}{lccc} \text{helicity} & 0 & -1/2 \\ \# \text{ of states} & 1 & 1 \end{array}$$

Hence, a complete $N=1$ multiplet is

$$N=1 \qquad \begin{array}{lccc} \text{helicity} & -1/2 & 0 & 1/2 \\ \# \text{ of states} & 1 & 2 & 1 \end{array}$$

which contains one complex scalar and one spinor with two helicity states.

This is an example of the so-called self-conjugated multiplet. There are also self-conjugated multiplets with $N>1$ corresponding to extended supersymmetry. Two particular examples are the $N=4$ super Yang-Mills multiplet and the $N=8$ supergravity multiplet

$$\begin{array}{llccccccccc} N=4 & \text{SUSY YM} & \text{helicity} & & & -1 & -1/2 & 0 & 1/2 & 1 & \\ & \lambda=-1 & \# \text{ of states} & & & 1 & 4 & 6 & 4 & 1 & \end{array}$$

$$\begin{array}{llccccccccc} N=8 & \text{SUGRA} & \text{helicity} & -2 & -3/2 & -1 & -1/2 & 0 & 1/2 & 1 & 3/2 & 2 \\ & \lambda=-2 & \# \text{ of states} & 1 & 8 & 28 & 56 & 70 & 56 & 28 & 8 & 1 \end{array}$$

One can see that the multiplets of extended supersymmetry are very rich and contain a vast number of particles.

The constraint on the number of SUSY generators comes from a requirement of consistency of the corresponding QFT. The number of supersymmetries and the maximal spin of the particle in the multiplet are related by

$$N \leq 4S,$$

where S is the maximal spin. Since the theories with spin greater than 1 are non-renormalizable and the theories with spin greater than 5/2 have no consistent coupling to gravity, this imposes a constraint on the number of SUSY generators

$$\begin{array}{ll} N \leq 4 & \text{for renormalizable theories (YM),} \\ N \leq 8 & \text{for (super)gravity.} \end{array}$$

In what follows, we shall consider simple supersymmetry, or $N = 1$ supersymmetry, contrary to extended supersymmetries with $N > 1$. In this case, one has two types of supermultiplets: the so-called chiral multiplet with $\lambda = 0$, which contains two physical states (ϕ, ψ) with spin 0 and 1/2, respectively, and the vector multiplet with $\lambda = 1/2$, which also contains two physical states (λ, A_μ) with spin 1/2 and 1, respectively.

3.2 Superspace and superfields

An elegant formulation of supersymmetry transformations and invariants can be achieved in the framework of superspace [7]. Superspace differs from the ordinary Euclidean (Minkowski) space by adding of two new coordinates, θ_α and $\bar{\theta}_{\dot{\alpha}}$, which are Grassmannian, i.e. anticommuting, variables

$$\{\theta_\alpha, \theta_\beta\} = 0, \;\; \{\bar{\theta}_{\dot{\alpha}}, \bar{\theta}_{\dot{\beta}}\} = 0, \;\; \theta_\alpha^2 = 0, \;\; \bar{\theta}_{\dot{\alpha}}^2 = 0, \;\; \alpha, \beta, \dot{\alpha}, \dot{\beta} = 1, 2.$$

Thus, we go from space to superspace

$$\begin{array}{ccc} Space & \Rightarrow & Superspace \\ x_\mu & & x_\mu, \theta_\alpha, \bar{\theta}_{\dot{\alpha}} \end{array}$$

A SUSY group element can be constructed in superspace in the same way as an ordinary translation in the usual space

$$G(x, \theta, \bar{\theta}) = e^{i(-x^\mu P_\mu + \theta Q + \bar{\theta}\bar{Q})}. \tag{3.2}$$

It leads to a supertranslation in superspace

$$\begin{array}{rcl} x_\mu & \rightarrow & x_\mu + i\theta\sigma_\mu\bar{\varepsilon} - i\varepsilon\sigma_\mu\bar{\theta}, \\ \theta & \rightarrow & \theta + \varepsilon, \\ \bar{\theta} & \rightarrow & \bar{\theta} + \bar{\varepsilon}, \end{array} \tag{3.3}$$

where ε and $\bar{\varepsilon}$ are Grassmannian transformation parameters. From eq.(3.3) one can easily obtain the representation for the supercharges (3.1) acting on the superspace

$$Q_\alpha = \frac{\partial}{\partial\theta_\alpha} - i\sigma^\mu_{\alpha\dot{\alpha}}\bar{\theta}^{\dot{\alpha}}\partial_\mu, \quad \bar{Q}_{\dot{\alpha}} = -\frac{\partial}{\partial\bar{\theta}_{\dot{\alpha}}} + i\theta_\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu. \tag{3.4}$$

Taking the Grassmannian transformation parameters to be local, or space-time dependent, one gets a local translation. As has already been mentioned, this leads to a theory of (super) gravity.

To define the fields on a superspace, consider representations of the Super-Poincaré group (3.1) [5]. The simplest one is a scalar superfield $F(x, \theta, \bar{\theta})$ which is SUSY invariant. Its Taylor expansion in θ and $\bar{\theta}$ has only several terms due to the nilpotent character of Grassmannian parameters. However, this superfield is a reducible representation of SUSY. To get an irreducible one, we define a *chiral* superfield which obeys the equation

$$\bar{D}F = 0, \quad \text{where } \bar{D} = -\frac{\partial}{\partial\bar{\theta}} - i\theta\sigma^\mu\partial_\mu \tag{3.5}$$

is a superspace covariant derivative.

For the chiral superfield Grassmannian Taylor expansion looks like ($y = x + i\theta\sigma\bar{\theta}$)

$$\begin{aligned}\Phi(y,\theta) &= A(y) + \sqrt{2}\theta\psi(y) + \theta\theta F(y) \\ &= A(x) + i\theta\sigma^\mu\bar{\theta}\partial_\mu A(x) + \frac{1}{4}\theta\theta\bar{\theta}\bar{\theta}\Box A(x) \\ &+ \sqrt{2}\theta\psi(x) - \frac{i}{\sqrt{2}}\theta\theta\partial_\mu\psi(x)\sigma^\mu\bar{\theta} + \theta\theta F(x). \end{aligned} \tag{3.6}$$

The coefficients are ordinary functions of x being the usual fields. They are called the *components* of a superfield. In eq.(3.6) one has 2 bosonic (complex scalar field A) and 2 fermionic (Weyl spinor field ψ) degrees of freedom. The component fields A and ψ are called the *superpartners*. The field F is an *auxiliary* field, it has the "wrong" dimension and has no physical meaning. It is needed to close the algebra (3.1). One can get rid of the auxiliary fields with the help of equations of motion.

Thus, a superfield contains an equal number of bosonic and fermionic degrees of freedom. Under SUSY transformation they convert into one another

$$\begin{aligned}\delta_\varepsilon A &= \sqrt{2}\varepsilon\psi, \\ \delta_\varepsilon\psi &= i\sqrt{2}\sigma^\mu\bar{\varepsilon}\partial_\mu A + \sqrt{2}\varepsilon F, \\ \delta_\varepsilon F &= i\sqrt{2}\bar{\varepsilon}\sigma^\mu\partial_\mu\psi. \end{aligned} \tag{3.7}$$

Notice that the variation of the F-component is a total derivative, i.e. it vanishes when integrated over the space-time.

One can also construct an antichiral superfield Φ^+ obeying the equation

$$D\Phi^+ = 0, \quad \text{with } D = \frac{\partial}{\partial\theta} + i\sigma^\mu\bar{\theta}\partial_\mu.$$

The product of chiral (antichiral) superfields Φ^2, Φ^3, etc is also a chiral (antichiral) superfield, while the product of chiral and antichiral ones $\Phi^+\Phi$ is a general superfield.

For any arbitrary function of chiral superfields one has

$$\begin{aligned}\mathcal{W}(\Phi_i) &= \mathcal{W}(A_i + \sqrt{2}\theta\psi_i + \theta\theta F) \\ &= \mathcal{W}(A_i) + \frac{\partial\mathcal{W}}{\partial A_i}\sqrt{2}\theta\psi_i + \theta\theta\left(\frac{\partial\mathcal{W}}{\partial A_i}F_i - \frac{1}{2}\frac{\partial^2\mathcal{W}}{\partial A_i\partial A_j}\psi_i\psi_j\right). \end{aligned} \tag{3.8}$$

The $\mathcal{W}$ is usually referred to as a superpotential which replaces the usual potential for the scalar fields.

To construct the gauge invariant interactions, one needs a real vector superfield $V = V^+$. It is not chiral but rather a general superfield with the following Grassmannian expansion:

$$\begin{aligned}V(x,\theta,\bar{\theta}) &= C(x) + i\theta\chi(x) - i\bar{\theta}\bar{\chi}(x) \\ &+ \frac{i}{2}\theta\theta[M(x) + iN(x)] - \frac{i}{2}\bar{\theta}\bar{\theta}[M(x) - iN(x)] \\ &- \theta\sigma^\mu\bar{\theta}v_\mu(x) + i\theta\theta\bar{\theta}[\lambda(x) + \frac{i}{2}\bar{\sigma}^\mu\partial_\mu\chi(x)] \\ &- i\bar{\theta}\bar{\theta}\theta[\lambda + \frac{i}{2}\sigma^\mu\partial_\mu\bar{\chi}(x)] + \frac{1}{2}\theta\theta\bar{\theta}\bar{\theta}[D(x) + \frac{1}{2}\Box C(x)]. \end{aligned} \tag{3.9}$$

The physical degrees of freedom corresponding to a real vector superfield V are the vector gauge field v_μ and the Majorana spinor field λ. All other components are unphysical and can be eliminated. Indeed, under the Abelian (super)gauge transformation the superfield V is transformed as

$$V \quad \to \quad V + \Phi + \Phi^+,$$

where Φ and Φ^+ are some chiral superfields. In components it looks like

$$\begin{aligned} C &\rightarrow C + A + A^*, \\ \chi &\rightarrow \chi - i\sqrt{2}\psi, \\ M + iN &\rightarrow M + iN - 2iF, \\ v_\mu &\rightarrow v_\mu - i\partial_\mu(A - A^*), \\ \lambda &\rightarrow \lambda, \\ D &\rightarrow D, \end{aligned} \tag{3.10}$$

and corresponds to ordinary gauge transformations for physical components. According to eq.(3.10), one can choose a gauge (the Wess-Zumino gauge [19]) where $C = \chi = M = N = 0$, leaving one with only physical degrees of freedom except for the auxiliary field D. In this gauge

$$\begin{aligned} V &= -\theta\sigma^\mu\bar{\theta}v_\mu(x) + i\theta\theta\bar{\theta}\bar{\lambda}(x) - i\bar{\theta}\bar{\theta}\theta\lambda(x) + \frac{1}{2}\theta\theta\bar{\theta}\bar{\theta}D(x), \\ V^2 &= -\frac{1}{2}\theta\theta\bar{\theta}\bar{\theta}v_\mu(x)v^\mu(x), \\ V^3 &= 0, \quad etc. \end{aligned} \tag{3.11}$$

One can define also a field strength tensor (as analog of $F_{\mu\nu}$ in gauge theories)

$$\begin{aligned} W_\alpha &= -\frac{1}{4}\bar{D}^2 e^V D_\alpha e^{-V}, \\ \bar{W}_{\dot{\alpha}} &= -\frac{1}{4}D^2 e^V \bar{D}_\alpha e^{-V}, \end{aligned} \tag{3.12}$$

which is a polynomial in the Wess-Zumino gauge. (Here Ds are the supercovariant derivatives.)

The strength tensor is a chiral superfield

$$\bar{D}_{\dot{\beta}} W_\alpha = 0, \quad D_\beta \bar{W}_{\dot{\alpha}} = 0.$$

In the Wess-Zumino gauge it is a polynomial over component fields:

$$W_\alpha = T^a \left(-i\lambda^a_\alpha + \theta_\alpha D^a - \frac{i}{2}(\sigma^\mu\bar{\sigma}^\nu\theta)_\alpha F^a_{\mu\nu} + \theta^2\sigma^\mu D_\mu\bar{\lambda}^a \right), \tag{3.13}$$

where

$$F^a_{\mu\nu} = \partial_\mu v^a_\nu - \partial_\nu v^a_\mu + f^{abc}v^b_\mu v^c_\nu, \quad D_\mu\bar{\lambda}^a = \partial\bar{\lambda}^a + f^{abc}v^b_\mu\bar{\lambda}^c.$$

In Abelian case eqs.(3.12) are simplified and take form

$$W_\alpha = -\frac{1}{4}\bar{D}^2 D_\alpha V, \quad \bar{W}_{\dot{\alpha}} = -\frac{1}{4}D^2\bar{D}_\alpha V.$$

3.3 Construction of SUSY Lagrangians

Let us start with the Lagrangian which has no local gauge invariance. In the superfield notation SUSY invariant Lagrangians are the polynomials of superfields. Having in mind that for component fields one should have ordinary terms and the above mentioned property of SUSY invariance of the highest dimension components of a superfield, the general SUSY invariant Lagrangian has the form

$$\mathcal{L} = \Phi^+_i\Phi_i|_{\theta\theta\bar{\theta}\bar{\theta}} + [(\lambda_i\Phi_i + \frac{1}{2}m_{ij}\Phi_i\Phi_j + \frac{1}{3}g_{ijk}\Phi_i\Phi_j\Phi_k)|_{\theta\theta} + h.c.]. \tag{3.14}$$

Hereafter the vertical line means the corresponding term of a Taylor expansion.

The first term is a kinetic term. It contains both the chiral and antichiral superfields Φ_i and Φ_i^+, respectively, and is a function of Grassmannian parameters θ and $\bar{\theta}$. Being expanded over θ and $\bar{\theta}$ it leads to the usual kinetic terms for the corresponding component fields.

The terms in the bracket form the superpotential. It is composed of the chiral fields only (plus the hermitian conjugated counterpart composed of antichiral superfields) and is a chiral superfield. Since the products of a chiral superfield and antichiral one produce a general superfield, they are not allowed in a superpotential. The last coefficient of its expansion over the parameter θ is supersymmetrically invariant and gives the usual potential after getting rid of the auxiliary fields, as it will be clear later.

The Lagrangian (3.14) can be written in a much more elegant way in superspace. The same way as an ordinary action is an integral over space-time of Lagrangian density, in supersymmetric case the action is an integral over the superspace. The space-time Lagrangian density then is [5, 6, 7]

$$\mathcal{L} = \int d^2\theta d^2\bar{\theta}\ \Phi_i^+\Phi_i + \int d^2\theta\ [\lambda_i\Phi_i + \frac{1}{2}m_{ij}\Phi_i\Phi_j + \frac{1}{3}y_{ijk}\Phi_i\Phi_j\Phi_k] + h.c. \tag{3.15}$$

where the first part is a kinetic term and the second one is a superpotential $\mathcal{W}$. Here instead of taking the proper components we use integration over the superspace according to the rules of Grassmannian integration [20]

$$\int d\theta_\alpha = 0, \quad \int \theta_\alpha\, d\theta_\beta = \delta_{\alpha\beta}.$$

Performing explicit integration over the Grassmannian parameters, we get from eq.(3.15)

$$\begin{aligned} \mathcal{L} &= i\partial_\mu\bar{\psi}_i\bar{\sigma}^\mu\psi_i + A_i^*\Box A_i + F_i^*F_i \\ &+ [\lambda_iF_i + m_{ij}(A_iF_j - \frac{1}{2}\psi_i\psi_j) + y_{ijk}(A_iA_jF_k - \psi_i\psi_jA_k) + h.c.]. \end{aligned} \tag{3.16}$$

The last two terms are the interaction ones. To obtain a familiar form of the Lagrangian, we have to solve the constraints

$$\frac{\partial\mathcal{L}}{\partial F_k^*} = F_k + \lambda_k^* + m_{ik}^*A_i^* + y_{ijk}^*A_i^*A_j^* = 0, \tag{3.17}$$

$$\frac{\partial\mathcal{L}}{\partial F_k} = F_k^* + \lambda_k + m_{ik}A_i + y_{ijk}A_iA_j = 0. \tag{3.18}$$

Expressing the auxiliary fields F and F^* from these equations, one finally gets

$$\begin{aligned} \mathcal{L} &= i\partial_\mu\bar{\psi}_i\bar{\sigma}^\mu\psi_i + A_i^*\Box A_i - \frac{1}{2}m_{ij}\psi_i\psi_j - \frac{1}{2}m_{ij}^*\bar{\psi}_i\bar{\psi}_j \\ &\quad -y_{ijk}\psi_i\psi_jA_k - y_{ijk}^*\bar{\psi}_i\bar{\psi}_jA_k^* - V(A_i, A_j), \end{aligned} \tag{3.19}$$

where the scalar potential $V = F_k^*F_k$. We will return to the discussion of the form of the scalar potential in SUSY theories later.

Consider now the gauge invariant SUSY Lagrangians. They should contain gauge invariant interaction of the matter fields with the gauge ones and the kinetic term and the self-interaction of the gauge fields.

Let us start with the gauge field kinetic terms. In the Wess-Zumino gauge one has

$$W^\alpha W_\alpha|_{\theta\theta} = -2i\lambda\sigma^\mu D_\mu\bar{\lambda} - \frac{1}{2}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}D^2 + i\frac{1}{4}F^{\mu\nu}F^{\rho\sigma}\epsilon_{\mu\nu\rho\sigma}, \tag{3.20}$$

where $D_\mu = \partial_\mu + ig[v_\mu,]$ is the usual covariant derivative and the last, the so-called topological θ term,[2] is the total derivative.

[2]Terminology comes from the θ term of QCD [21] and has nothing to do with the Grassmannian parameter θ.

The gauge invariant Lagrangian now has a familiar form

$$\begin{aligned}\mathcal{L} &= \frac{1}{4}\int d^2\theta\, W^\alpha W_\alpha + \frac{1}{4}\int d^2\bar\theta\, \bar W^{\dot\alpha}\bar W_{\dot\alpha} \\ &= \frac{1}{2}D^2 - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - i\lambda\sigma^\mu D_\mu\bar\lambda. \end{aligned} \tag{3.21}$$

To obtain a gauge-invariant interaction with matter chiral superfields, consider their gauge transformation (Abelian)

$$\Phi \;\to\; e^{-ig\Lambda}\Phi, \quad \Phi^+ \;\to\; \Phi^+ e^{ig\Lambda^+}, \quad V \;\to\; V + i(\Lambda - \Lambda^+),$$

where Λ is a gauge parameter (chiral superfield).

It is clear now how to construct both the SUSY and gauge invariant kinetic term (compare with the covariant derivative in a usual gauge theory)

$$\Phi_i^+\Phi_i|_{\theta\theta\bar\theta\bar\theta} \Rightarrow \Phi_i^+ e^{gV}\Phi_i|_{\theta\theta\bar\theta\bar\theta} \tag{3.22}$$

A complete SUSY and gauge invariant Lagrangian then looks like

$$\begin{aligned}\mathcal{L}_{inv} &= \frac{1}{4}\int d^2\theta\, W^\alpha W_\alpha + \frac{1}{4}\int d^2\bar\theta\, \bar W^{\dot\alpha}\bar W_{\dot\alpha} + \int d^2\theta d^2\bar\theta\, \Phi_i^+ e^{gV}\Phi_i \\ &+ \int d^2\theta\, (\frac{1}{2}m_{ij}\Phi_i\Phi_j + \frac{1}{3}y_{ijk}\Phi_i\Phi_j\Phi_k) + h.c. \end{aligned} \tag{3.23}$$

In particular, the SUSY generalization of QED looks as follows:

$$\begin{aligned}\mathcal{L}_{SUSY\,QED} &= \frac{1}{4}\int d^2\theta\, W^\alpha W_\alpha + \frac{1}{4}\int d^2\bar\theta\, \bar W^{\dot\alpha}\bar W_{\dot\alpha} \\ &+ \int d^4\theta\, (\Phi_+^+ e^{gV}\Phi_+ + \Phi_-^+ e^{-gV}\Phi_-) \\ &+ \int d^2\theta\, m\, \Phi_+\Phi_- + \int d^2\bar\theta\, m\, \Phi_+^+\Phi_-^+, \end{aligned} \tag{3.24}$$

where two superfields Φ_+ and Φ_- have been introduced in order to have both left- and right-handed fermions.

The non-Abelian generalization is straightforward

$$\begin{aligned}\mathcal{L}_{SUSY\,YM} &= \frac{1}{4}\int d^2\theta\, Tr(W^\alpha W_\alpha) + \frac{1}{4}\int d^2\bar\theta\, Tr(\bar W^\alpha \bar W_\alpha) \\ &+ \int d^2\theta d^2\bar\theta\, \bar\Phi_{ia}(e^{gV})^a_b\Phi_i^b + \int d^2\theta\, \mathcal{W}(\Phi_i) + \int d^2\bar\theta\, \bar{\mathcal{W}}(\bar\Phi_i), \end{aligned} \tag{3.25}$$

where $\mathcal{W}$ is a superpotential, which should be invariant under the group of symmetry of a particular model.

In terms of component fields the above Lagrangian takes the form

$$\begin{aligned}\mathcal{L}_{SUSY\,YM} &= -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - i\lambda^a\sigma^\mu D_\mu\bar\lambda^a + \frac{1}{2}D^aD^a \\ &+ (\partial_\mu A_i - igv^a_\mu T^a A_i)^\dagger(\partial_\mu A_i - igv^{a\mu}T^aA_i) - i\bar\psi_i\bar\sigma^\mu(\partial_\mu\psi_i - igv^{a\mu}T^a\psi_i) \\ &- D^aA_i^\dagger T^aA_i - i\sqrt{2}A_i^\dagger T^a\lambda^a\psi_i + i\sqrt{2}\bar\psi_i T^aA_i\bar\lambda^a + F_i^\dagger F_i \\ &+ \frac{\partial\mathcal{W}}{\partial A_i}F_i + \frac{\partial\bar{\mathcal{W}}}{\partial A_i^\dagger}F_i^\dagger - \frac{1}{2}\frac{\partial^2\mathcal{W}}{\partial A_\partial A_j}\psi_i\psi_j - \frac{1}{2}\frac{\partial^2\bar{\mathcal{W}}}{\partial A_i^\dagger\partial A_j^\dagger}\bar\psi_i\bar\psi_j. \end{aligned} \tag{3.26}$$

Integrating out the auxiliary fields D^a and F_i, one reproduces the usual Lagrangian.

3.4 The scalar potential

Contrary to the SM, where the scalar potential is arbitrary and is defined only by the requirement of the gauge invariance, in supersymmetric theories it is completely defined by the superpotential. It consists of the contributions from the D-terms and F-terms. The kinetic energy of the gauge fields (recall eq.(3.21) yields the $1/2D^aD^a$ term, and the matter-gauge interaction (recall eq.(3.23) yields the $gD^aT^a_{ij}A^*_iA_j$ one. Together they give

$$\mathcal{L}_D = \frac{1}{2}D^aD^a + gD^aT^a_{ij}A^*_iA_j. \tag{3.27}$$

The equation of motion reads

$$D^a = -gT^a_{ij}A^*_iA_j. \tag{3.28}$$

Substituting it back into eq.(3.27) yields the D-term part of the potential

$$\mathcal{L}_D = -\frac{1}{2}D^aD^a \quad \Longrightarrow V_D = \frac{1}{2}D^aD^a, \tag{3.29}$$

where D is given by eq.(3.28).

The F-term contribution can be derived from the matter field self-interaction eq.(3.16). For a general type superpotential W one has

$$\mathcal{L}_F = F^*_iF_i + (\frac{\partial W}{\partial A_i}F_i + h.c.). \tag{3.30}$$

Using the equations of motion for the auxiliary field F_i

$$F^*_i = -\frac{\partial W}{\partial A_i} \tag{3.31}$$

yields

$$\mathcal{L}_F = -F^*_iF_i \quad \Longrightarrow V_F = F^*_iF_i, \tag{3.32}$$

where F is given by eq.(3.31). The full potential is the sum of the two contributions

$$V = V_D + V_F. \tag{3.33}$$

Thus, the form of the Lagrangian is practically fixed by symmetry requirements. The only freedom is the field content, the value of the gauge coupling g, Yukawa couplings y_{ijk} and the masses. Because of the renormalizability constraint $V \leq A^4$ the superpotential should be limited by $\mathcal{W} \leq \Phi^3$ as in eq.(3.15). All members of a supermultiplet have the same masses, i.e. bosons and fermions are degenerate in masses. This property of SUSY theories contradicts the phenomenology and requires supersymmetry breaking.

3.5 Spontaneous breaking of SUSY

Since supersymmetric algebra leads to mass degeneracy in a supermultiplet, it should be broken to explain the absence of superpartners at modern energies. There are several ways of supersymmetry breaking. It can be broken either explicitly or spontaneously. Performing SUSY breaking one has to be careful not to spoil the cancellation of quadratic divergencies which allows one to solve the hierarchy problem. This is achieved by spontaneous breaking of SUSY.

Apart from non-supersymmetric theories in SUSY models the energy is always nonnegative definite. Indeed, according to quantum mechanics

$$E = <0|\ H\ |0>$$

and due to SUSY algebra eq.(3.1)

$$\{Q_\alpha, \bar{Q}_{\dot{\beta}}\} = 2(\sigma^\mu)_{\alpha\dot{\beta}} P_\mu,$$

taking into account that $tr(\sigma^\mu P_\mu) = 2P_0$, one gets

$$E = \frac{1}{4} \sum_{\alpha=1,2} < 0|\{Q_\alpha, \bar{Q}_\alpha\}|0 >= \frac{1}{4} \sum_\alpha |Q_\alpha|0 > |^2 \geq 0.$$

Hence

$$E =< 0|\ H\ |0 >\neq 0 \quad if\ and\ only\ if \quad Q_\alpha|0 >\neq 0.$$

Therefore, supersymmetry is spontaneously broken, i.e. vacuum is not invariant $(Q_\alpha|0 >\neq 0)$, *if and only if* the minimum of the potential is positive $(i.e.\ E > 0)$.

The situation is illustrated in Fig.9. The SUSY ground state has $E = 0$, while a non-SUSY one has $E > 0$. On the right-hand side a non-SUSY potential is shown. It does not appear even in spontaneously broken SUSY theories. However, just this type of the potential is used for spontaneous breaking of the gauge invariance via the Higgs mechanism. This property has crucial consequences for the spontaneous breaking of the gauge invariance. Indeed, as will be seen later, in the MSSM spontaneous breaking of $SU(2)$ invariance takes place only after SUSY is broken.

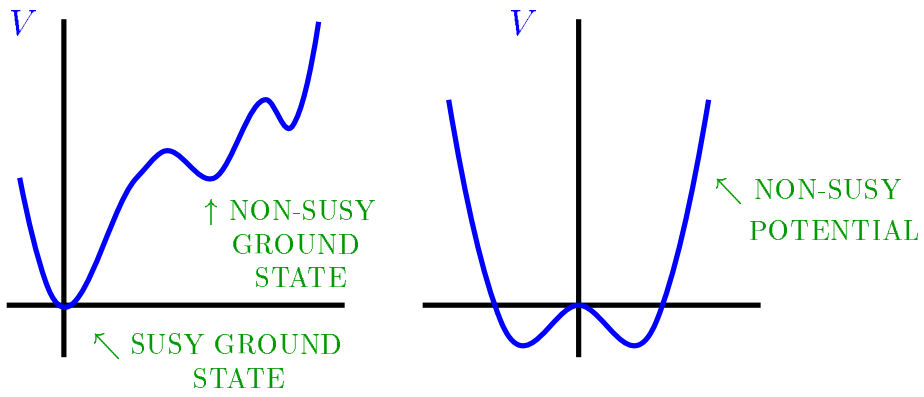


Fig. 9: Scalar potential in supersymmetric and non-supersymmetric theories

Spontaneous breaking of supersymmetry is achieved in the same way as the electroweak symmetry breaking. One introduces the field whose vacuum expectation value is nonzero and breaks the symmetry. However, due to a special character of SUSY, this should be a superfield whose auxiliary F and D components acquire nonzero v.e.v.'s. Thus, among possible spontaneous SUSY breaking mechanisms one distinguishes the F and D ones.

i) Fayet-Iliopoulos (D-term) mechanism [22].
In this case the, the linear D-term is added to the Lagrangian

$$\Delta\mathcal{L} = \xi V|_{\theta\theta\bar{\theta}\bar{\theta}} = \xi \int d^4\theta\ V. \tag{3.34}$$

It is gauge and SUSY invariant by itself; however, it may lead to spontaneous breaking of both of them depending on the value of ξ. We show in Fig.10a the sample spectrum for two chiral matter multiplets. The drawback of this mechanism is the necessity of $U(1)$ gauge invariance. It can be used in SUSY generalizations of the SM but not in GUTs.

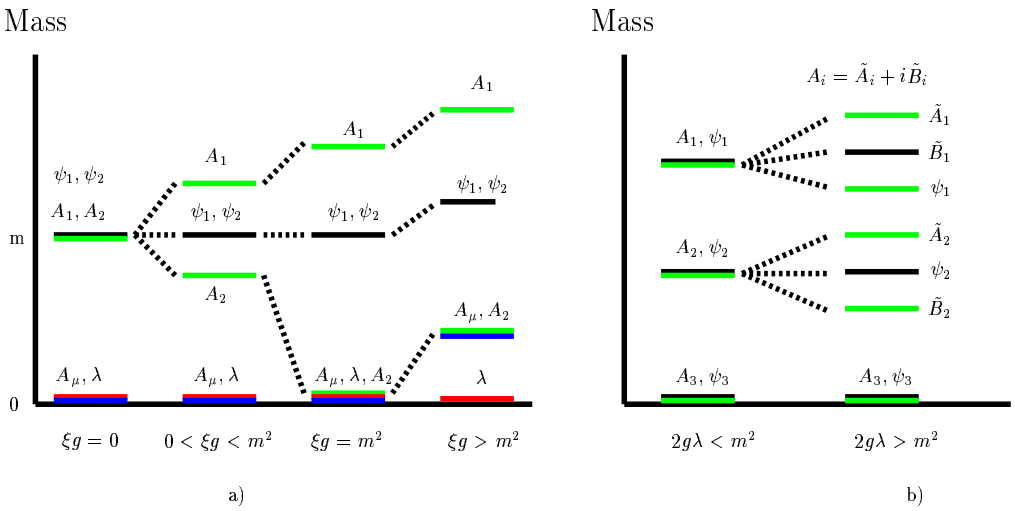


Fig. 10: Spectrum of spontaneously broken SUSY theories

The mass spectrum also causes some troubles since the following sum rule is always valid

$$\sum_{boson\ states} m_i^2 = \sum_{fermion\ states} m_i^2, \tag{3.35}$$

which is bad for phenomenology.

ii) O'Raifeartaigh (F-term) mechanism [23].
In this case, several chiral fields are needed and the superpotential should be chosen in a way that trivial zero v.e.v.s for the auxiliary F-fields be absent. For instance, choosing the superpotential to be

$$\mathcal{W}(\Phi) = \lambda\Phi_3 + m\Phi_1\Phi_2 + g\Phi_3\Phi_1^2,$$

one gets the equations for the auxiliary fields

$$\begin{aligned} F_1^* &= mA_2 + 2gA_1A_3, \\ F_2^* &= mA_1, \\ F_3^* &= \lambda + gA_1^2, \end{aligned}$$

which have no solutions with $< F_i >= 0$ and SUSY is spontaneously broken. The sample spectrum is shown in Fig.10b.

The drawbacks of this mechanism is a lot of arbitrariness in the choice of potential. The sum rule (3.35) is also valid here.

Unfortunately, none of these mechanisms explicitly works in SUSY generalizations of the SM. None of the fields of the SM can develop nonzero v.e.v.s for their F or D components without breaking $SU(3)$ or $U(1)$ gauge invariance since they are not singlets with respect to these groups. This requires the presence of extra sources of spontaneous SUSY breaking, which we consider below. They are based, however, on the same F and D mechanisms.

4 SUSY GENERALIZATION OF THE STANDARD MODEL. THE MSSM

As has been already mentioned, in SUSY theories the number of bosonic degrees of freedom equals that of fermionic. At the same time, in the SM one has 28 bosonic and 90 fermionic degrees of freedom (with massless neutrino, otherwise 96). So the SM is to a great extent non-supersymmetric. Trying to add some new particles to supersymmetrize the SM, one should take into account the following observations:

1. There are no fermions with quantum numbers of the gauge bosons;
2. Higgs fields have nonzero v.e.v.s; hence they cannot be superpartners of quarks and leptons since this would induce spontaneous violation of baryon and lepton numbers;
3. One needs at least two complex chiral Higgs multiplets to give masses to Up and Down quarks.

The latter is due to the form of a superpotential and chirality of matter superfields. Indeed, the superpotential should be invariant under the $SU(3) \times SU(2) \times U(1)$ gauge group. If one looks at the Yukawa interaction in the Standard Model, eq.(1.7), one finds that it is indeed $U(1)$ invariant since the sum of hypercharges in each vertex equals zero. In the last term this is achieved by taking the conjugated Higgs doublet $\tilde{H} = i\tau_2 H^\dagger$ instead of H. However, in SUSY H is a chiral superfield and hence a superpotential, which is constructed out of chiral fields, can contain only H but not $\tilde{H}$ which is an antichiral superfield.

Another reason for the second Higgs doublet is related to chiral anomalies. It is known that chiral anomalies spoil the gauge invariance and, hence, the renormalizability of the theory. They are canceled in the SM between quarks and leptons in each generation.

Indeed, chiral (or triangle anomaly) is proportional to the trace of three hypercharges. In the SM one has

$$\begin{array}{rcccccccccc} TrY^3 & = & 3 & \left(\frac{1}{27}\right. & +\frac{1}{27} & -\frac{64}{27} & \left.+\frac{8}{27}\right) & -1 & -1 & +8 & = 0. \\ & & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \\ & & colour & u_L & d_L & u_R & d_R & \nu_L & e_L & e_R & \end{array}$$

However, if one introduces a chiral Higgs superfield, it contains higgsinos, which are chiral fermions, and contain anomalies. To cancel them one has to add the second Higgs doublet with the opposite hypercharge. Therefore, the Higgs sector in SUSY models is inevitably enlarged, it contains an even number of doublets.

Conclusion: In SUSY models supersymmetry associates *known* bosons with *new* fermions and *known* fermions with *new* bosons.

4.1 The field content

Consider the particle content of the Minimal Supersymmetric Standard Model [24]. According to the previous discussion, in the minimal version we double the number of particles (introducing a superpartner to each particle) and add another Higgs doublet (with its superpartner). The particle content of the MSSM then appears as [25]

Particle Content of the MSSM

Superfield	Bosons		Fermions		$SU_c(3)$	$SU_L(2)$	$U_Y(1)$
Gauge							
$\mathbf{G^a}$	gluon	g^a	gluino	$\tilde{g}^a$	8	0	0
$\mathbf{V^k}$	Weak	W^k $(W^\pm, Z)$	wino, zino	$\tilde{w}^k$ $(\tilde{w}^\pm, \tilde{z})$	1	3	0
$\mathbf{V'}$	Hypercharge	B (γ)	bino	$\tilde{b}(\tilde{\gamma})$	1	1	0
Matter							
$\mathbf{L_i}$	sleptons	$\tilde{L}_i = (\tilde{\nu}, \tilde{e})_L$	leptons	$L_i = (\nu, e)_L$	1	2	-1
$\mathbf{E_i}$		$\tilde{E}_i = \tilde{e}_R$		$E_i = e_R$	1	1	2
$\mathbf{Q_i}$	squarks	$\tilde{Q}_i = (\tilde{u}, \tilde{d})_L$	quarks	$Q_i = (u, d)_L$	3	2	$1/3$
$\mathbf{U_i}$		$\tilde{U}_i = \tilde{u}_R$		$U_i = u^c_R$	3^*	1	$-4/3$
$\mathbf{D_i}$		$\tilde{D}_i = \tilde{d}_R$		$D_i = d^c_R$	3^*	1	$2/3$
Higgs							
$\mathbf{H_1}$	Higgses	H_1	higgsinos	$\tilde{H}_1$	1	2	-1
$\mathbf{H_2}$		H_2		$\tilde{H}_2$	1	2	1

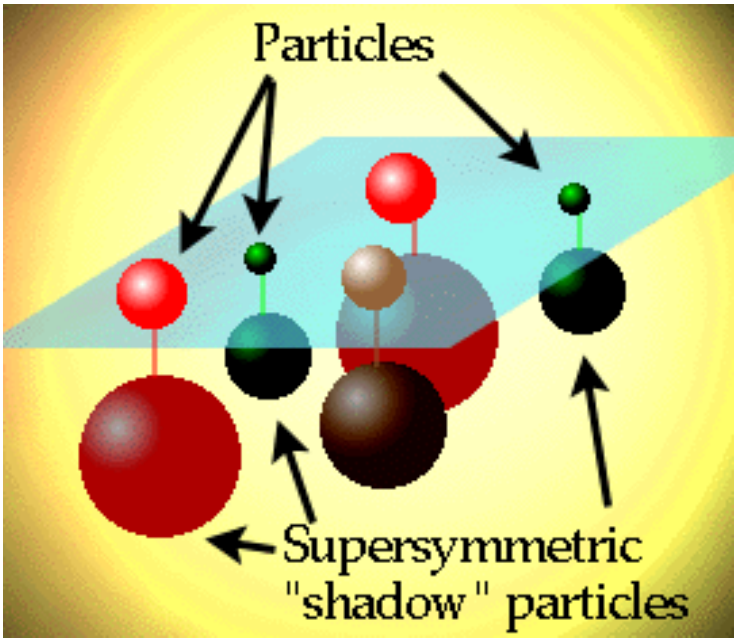


Fig. 11: The shadow world of SUSY particles [26]

where $a = 1, 2, ..., 8$ and $k = 1, 2, 3$ are the $SU(3)$ and $SU(2)$ indices, respectively, and $i = 1, 2, 3$ is the generation index. Hereafter, tilde denotes a superpartner of an ordinary particle.

Thus, the characteristic feature of any supersymmetric generalization of the SM is the presence of superpartners (see Fig.11). If supersymmetry is exact, superpartners of ordinary particles should have the same masses and have to be observed. The absence of them at modern energies is believed to be explained by the fact that their masses are very heavy, that means that supersymmetry should be broken. Hence, if the energy of accelerators is high enough, the superpartners will be created.

The presence of an extra Higgs doublet in SUSY model is a novel feature of the theory. In the MSSM one has two doublets with the quantum numbers (1,2,-1) and (1,2,1), respectively:

$$H_1 = \begin{pmatrix} H_1^0 \\ H_1^- \end{pmatrix} = \begin{pmatrix} v_1 + \frac{S_1 + iP_1}{\sqrt{2}} \\ H_1^- \end{pmatrix}, \quad H_2 = \begin{pmatrix} H_2^+ \\ H_2^0 \end{pmatrix} = \begin{pmatrix} H_2^+ \\ v_2 + \frac{S_2 + iP_2}{\sqrt{2}} \end{pmatrix}, \tag{4.1}$$

where v_i are the vacuum expectation values of the neutral components.

Hence, one has 8=4+4=5+3 degrees of freedom. As in the case of the SM, 3 degrees of freedom can be gauged away, and one is left with 5 physical states compared to 1 state in the SM.

Thus, in the MSSM, as actually in any of two Higgs doublet models, one has five physical Higgs bosons: two CP-even neutral, one CP-odd neutral and two charged. We consider the mass eigenstates below.

4.2 Lagrangian of the MSSM

The Lagrangian of the MSSM consists of two parts; the first part is SUSY generalization of the Standard Model, while the second one represents the SUSY breaking as mentioned above.

$$\mathcal{L} = \mathcal{L}_{SUSY} + \mathcal{L}_{Breaking}, \tag{4.2}$$

where

$$\mathcal{L}_{SUSY} = \mathcal{L}_{Gauge} + \mathcal{L}_{Yukawa} \tag{4.3}$$

and

$$\mathcal{L}_{Gauge} = \sum_{SU(3),SU(2),U(1)} \frac{1}{4}\left(\int d^2\theta\, TrW^\alpha W_\alpha + \int d^2\bar\theta\, Tr\bar W^{\dot\alpha}\bar W_{\dot\alpha}\right) + \sum_{Matter}\int d^2\theta d^2\bar\theta\, \Phi_i^\dagger e^{g_3\hat V_3 + g_2\hat V_2 + g_1\hat V_1}\Phi_i, \quad (4.4)$$

$$\mathcal{L}_{Yukawa} = \int d^2\theta\, (\mathcal{W}_R + \mathcal{W}_{NR}) + h.c. \quad (4.5)$$

The index R in a superpotential refers to the so-called R-parity [27] which adjusts a "+" charge to all the ordinary particles and a "−" charge to their superpartners. The first part of $\mathcal{W}$ is R-symmetric

$$W_R = \epsilon_{ij}(y^U_{ab}Q^j_a U^c_b H^i_2 + y^D_{ab}Q^j_a D^c_b H^i_1 + y^L_{ab}L^j_a E^c_b H^i_1 + \mu H^i_1 H^j_2), \quad (4.6)$$

where $i,j = 1,2,3$ are the $SU(2)$ and $a,b = 1,2,3$ are the generation indices; colour indices are suppressed. This part of the Lagrangian almost exactly repeats that of the SM except that the fields are now the superfields rather than the ordinary fields of the SM. The only difference is the last term which describes the Higgs mixing. It is absent in the SM since there is only one Higgs field there.

The second part is R-nonsymmetric

$$W_{NR} = \epsilon_{ij}(\lambda^L_{abd}L^i_a L^j_b E^c_d + \lambda^{L\prime}_{abd}L^i_a Q^j_b D^c_d + \mu'_a L^i_a H^j_2) + \lambda^B_{abd}U^c_a D^c_b D^c_d. \quad (4.7)$$

These terms are absent in the SM. The reason is very simple: one can not replace the superfields in eq.(4.7) by the ordinary fields like in eq.(4.6) because of the Lorentz invariance. These terms have a different property, they violate either lepton (the first line in eq.(4.7)) or baryon number (the second line). Since both effects are not observed in Nature, these terms must be suppressed or be excluded. One can avoid such terms if one introduces special symmetry called the R-symmetry [28]. This is the global $U(1)_R$ invariance

$$U(1)_R: \quad \theta \to e^{i\alpha}\theta, \quad \Phi \to e^{in\alpha}\Phi, \quad (4.8)$$

i.e., the superfield has the quantum number $R = n$. To preserve $U(1)_R$ invariance the superpotential W must have $R = 2$. Thus, to get $W_{NR} = 0$ one must choose $R = 1$ for all the Higgs superfields and $R = 1/2$ for quark and lepton ones. However, this property happens to be too restrictive. Indeed, the gaugino mass term, which is Lorentz and gauge invariant and is introduced while supersymmetry breaking, happens to be R-invariant only for $\alpha = \pm\pi$. This reduces the R-symmetry to the discrete group Z_2, called the R-parity [27]. The R-parity quantum number is given by

$$R = (-1)^{3(B-L)+2S} \quad (4.9)$$

for particles with spin S. Thus, all the ordinary particles have the R-parity quantum number equal to $R = +1$, while all the superpartners have R-parity quantum number equal to $R = -1$. The R-parity obviously forbids the W_{NR} terms. It is usually assumed that they are absent in the MSSM, i.e. R-parity is preserved. However, there is no physical principle behind it. It may well be that these terms are present, though experimental limits on the couplings are very severe [29]

$$\lambda^L_{abc},\ \lambda^{L\prime}_{abc} < 10^{-4}, \qquad \lambda^B_{abc} < 10^{-9}.$$

4.3 Properties of interactions

If one assumes that the R-parity is preserved, then the interactions of superpartners are essentially the same as in the SM, but two of three particles involved into an interaction at any vertex are replaced by superpartners. The reason for it, as we discussed earlier, is the R-parity. According to eq.(4.9), all the ordinary particles are R-even, while all the superpartners are R-odd.

Conservation of the R-parity has two consequences

- the superpartners are created in pairs;
- the lightest superparticle (LSP) is stable.

Usually it is photino $\tilde{\gamma}$, the superpartner of a photon with some admixture of neutral higgsino.

Typical vertices are shown in Figs.12-14. The tilde above a letter denotes the corresponding superpartner. Note that the coupling is the same in all the vertices involving superpartners.

In the case of R-parity violation one has additional vertices with new types of interaction. As has been already mentioned, they violate either the lepton or baryon number. The typical ones are

$$\mathcal{L}_{LLE} = \lambda' \{\tilde{\nu}_L e_L e_R^c - \tilde{e}_L \nu_L e_R^c + \tilde{e}_R^* \nu_L e_R + \ldots\}, \tag{4.10}$$

$$\mathcal{L}_{LQD} = \lambda \left\{\tilde{\nu}_L d_L \bar{d}_R - \tilde{e}_L u_L \bar{d}_R + \tilde{d}_L \nu_L \bar{d}_R - \tilde{u}_L e_L \bar{d}_R + \ldots\right\}. \tag{4.11}$$

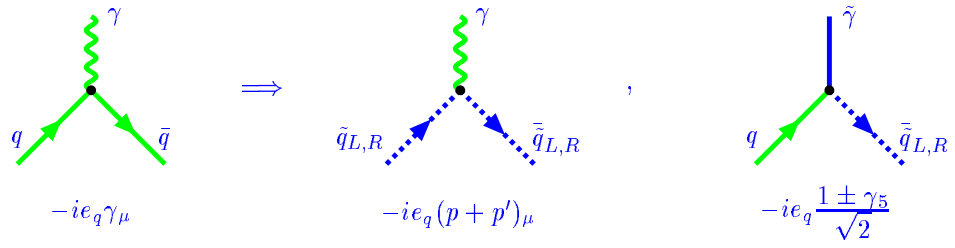


Fig. 12: Gauge-matter interaction

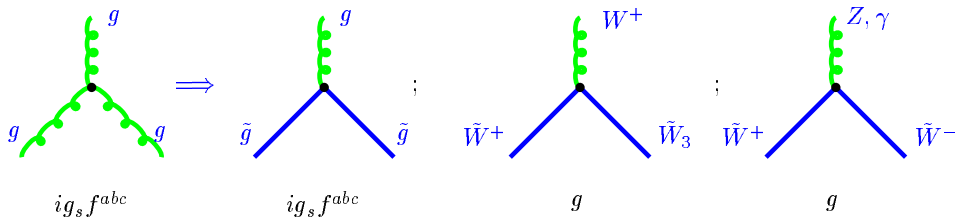


Fig. 13: Gauge self-interaction

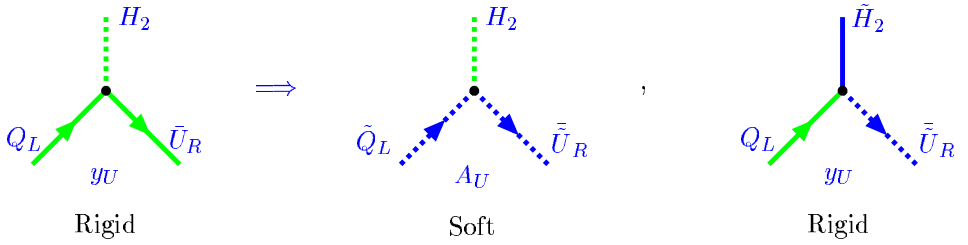


Fig. 14: Yukawa-type interaction

There are also UDD terms which violate the baryon number. These terms together lead to a fast proton decay via the process shown in Fig.15. To avoid it, one usually leaves either L or B violating interactions.

The limits on R-parity violating couplings come from non-observation of various processes, like proton decay, $\nu_\mu e$ scattering, etc and also from the charged current universality: $\Gamma(\pi \to e\nu)/\Gamma(\pi \to \mu\nu), \Gamma(\tau \to e\nu\bar{\nu})/\Gamma(\tau \to \mu\nu\bar{\nu})$, etc.

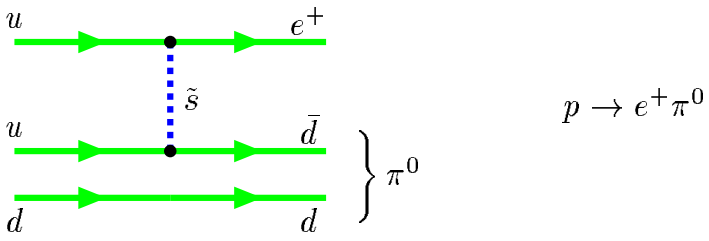


Fig. 15: Proton decay in R-parity violating models

4.4 Creation and decay of superpartners

The above-mentioned rule together with the Feynman rules for the SM enables us to draw diagrams describing creation of superpartners. One of the most promising processes is the e^+e^- annihilation (see Fig.16). The usual kinematic restriction is given by the centre of mass energy

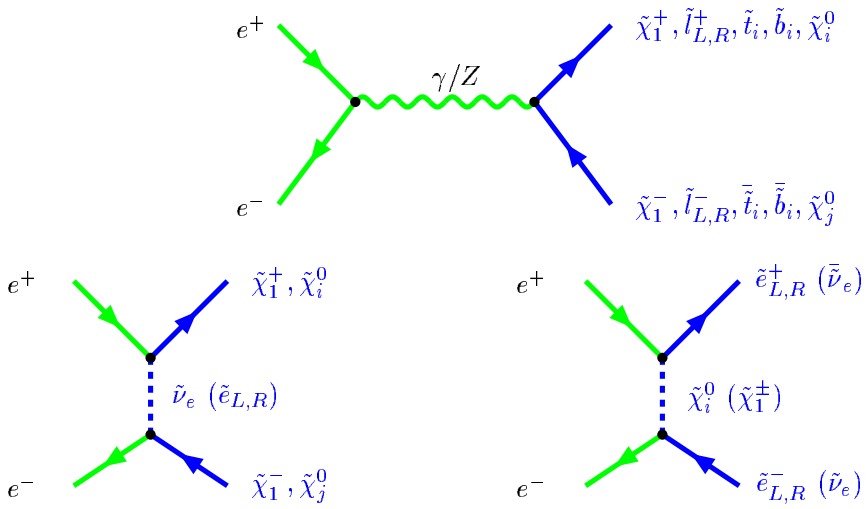


Fig. 16: Creation of superpartners

$$m^{max}_{sparticle} \leq \frac{\sqrt{s}}{2}.$$

Similar processes take place at hadron colliders with electrons and positrons being replaced by quarks and gluons.

Creation of superpartners can be accompanied by creation of ordinary particles as well. We consider various experimental signatures for e^+e^- and hadron colliders below. They crucially depend on SUSY breaking pattern and on the mass spectrum of superpartners.

The decay properties of superpartners also depend on their masses. For the quark and lepton superpartners the main processes are shown in Fig.17.

When the R-parity is conserved, new particles will eventually end up giving neutralinos (the lightest superparticle) whose interactions are comparable to those of neutrinos and they leave undetected. Therefore, their signature would be missing energy and transverse momentum.

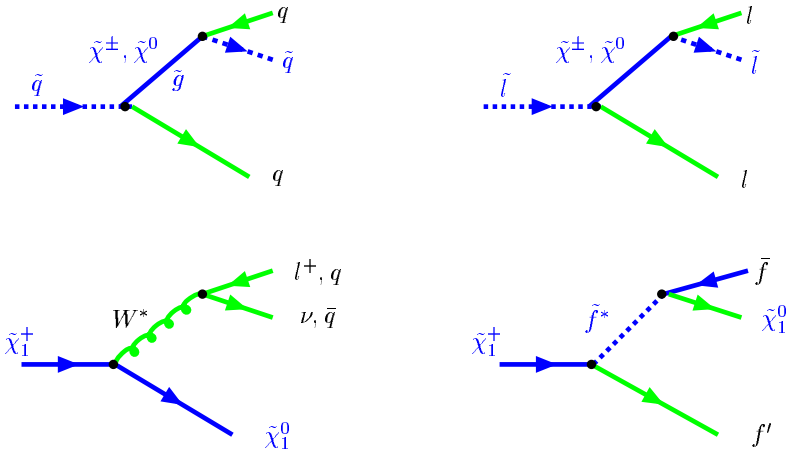


Fig. 17: Decay of superpartners

Examples. Consider some explicit examples of superpartner decays.

squarks :	$\tilde{q}_{L,R}$	$\rightarrow$	$q + \tilde{\chi}_i^0$	(quark + photino)
	$\tilde{q}_L$	$\rightarrow$	$q' + \tilde{\chi}_i^{\pm}$	(quark + chargino)
	$\tilde{q}$	$\rightarrow$	$q + \tilde{g}$	(quark + gluino) $for\ m_{\tilde{q}} > m_{\tilde{g}}$
	$\tilde{t}_1$	$\rightarrow$	$c + \tilde{\chi}_1^0$	(main decay) signal: 2 acollinear jets $+ \not{E}_T$
	$\tilde{t}_1$	$\rightarrow$	$b + \tilde{\chi}_1^+$	signal: 2 b jets + 2 leptons $+ \not{E}_T$
			$\hookrightarrow \tilde{\chi}_1^0 f \bar{f}'$ $(f\bar{f}' = l\bar{\nu}, q\bar{q})$	(4 jets) $+ \not{E}_T$

sleptons :	$\tilde{l}$	$\rightarrow$	$l + \tilde{\chi}_i^0$	(lepton + photino)
	$\tilde{l}_L$	$\rightarrow$	$\nu_l + \tilde{\chi}_i^{\pm}$	(neutrino + chargino)
gluino :	$\tilde{g}$	$\rightarrow$	$q + \bar{q} + \tilde{\gamma}$	(quark + antiquark + photino)
	$\tilde{g}$	$\rightarrow$	$g + \tilde{\gamma}$	(gluon + photino)
chargino :	$\tilde{\chi}_i^{\pm}$	$\rightarrow$	$e + \nu_e + \tilde{\chi}_i^0$	(electron + neutrino + photino)
	$\tilde{\chi}_i^{\pm}$	$\rightarrow$	$q + \bar{q}' + \tilde{\chi}_i^0$	(quark + antiquark + photino)
neutralino :	$\tilde{\chi}_2^0$	$\rightarrow$	$\tilde{\chi}_1^0 + X$	

In the last case there are many possible channels both visible and invisible.

		Visible Channels	Final States
$\tilde{\chi}_2^0$	$\rightarrow$	$\tilde{\chi}_1^0 l^+ l^-$ $(l = e, \mu, \tau)$	
	$\rightarrow$	$\tilde{\chi}_1^{\pm} l^{\mp} \nu_l$	$l^+ l^- + \not{E}_T$
		$\hookrightarrow \tilde{\chi}_1^0 l^{\pm} \nu_l$	
	$\rightarrow$	$\tilde{\chi}_1^0 q\bar{q}$	2 jets $+ \not{E}_T$
	$\rightarrow$	$\tilde{\chi}_\gamma^0$	$\gamma + \not{E}_T$

$$
\begin{array}{lll}
\rightarrow & \tilde{\chi}_1^{\pm} q\bar{q}' & \\
 & \hookrightarrow \tilde{\chi}_1^0 l^{\pm} q\bar{q}' & 2 \text{ jets} + \not{E}_T \\
\rightarrow & \tilde{\chi}_1^{\pm} l^{\mp} \nu_l & \\
 & \hookrightarrow \tilde{\chi}_1^0 q\bar{q}' & l^{\pm} + 2 \text{ jets} + \not{E}_T \\
\rightarrow & \tilde{\chi}_1^{\pm} q\bar{q}' & \\
 & \hookrightarrow \tilde{\chi}_1^0 l^{\pm} \nu_l & l^{\pm} + 2 \text{ jets} + \not{E}_T
\end{array}
$$

Invisible Channel	Final State
$\rightarrow \tilde{\chi}_1^0 \nu_l \bar{\nu}_l$	$\not{E}_T$

Thus, if supersymmetry exists in Nature and if it is broken somewhere below 1 TeV, then it will be possible to detect it in the nearest future.

5 BREAKING OF SUSY IN THE MSSM

Since none of the fields of the MSSM can develop non-zero v.e.v. to break SUSY without spoiling the gauge invariance, it is supposed that spontaneous supersymmetry breaking takes place via some other fields. The most common scenario for producing low-energy supersymmetry breaking is called the *hidden sector* one [30]. According to this scenario, there exist two sectors: the usual matter belongs to the "visible" one, while the second, "hidden" sector, contains fields which lead to breaking of supersymmetry. These two sectors interact with each other by exchange of some fields called *messengers*, which mediate SUSY breaking from the hidden to the visible sector (see Fig.18). There might be various types of messenger fields: gravity, gauge, etc. Below we consider four possible scenarios.

The hidden sector is the weakest part of the MSSM. It contains a lot of ambiguities and leads to uncertainties of the MSSM predictions considered below.

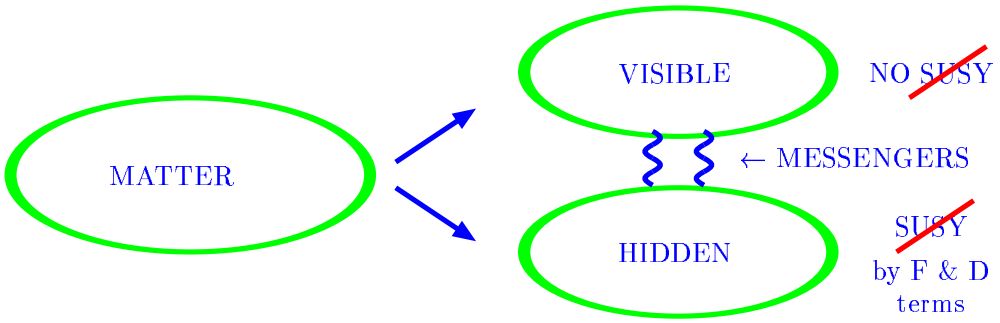


Fig. 18: Hidden Sector Scenario

5.1 The hidden sector: four scenarios

So far there are known four main mechanisms to mediate SUSY breaking from a hidden to a visible sector:

- Gravity mediation (SUGRA);
- Gauge mediation;
- Anomaly mediation;
- Gaugino mediation.

Consider them in more detail.

SUGRA

This mechanism is based on effective nonrenormalizable interactions arising as a low-energy limit of supergravity theories [31]. In this case, two sectors interact with each other via gravity. There are two types of scalar fields that develop nonzero v.e.v.s, namely moduli fields T, which appear as a result of compactification from higher dimensions, and the dilaton field S, part of SUGRA supermultiplet. These fields obtain nonzero v.e.v.s for their F components: $< F_T > \neq 0, < F_S > \neq 0$, which leads to spontaneous SUSY breaking. Since in SUGRA theory supersymmetry is local, spontaneous breaking leads to Goldstone particle which is a Goldstone fermion in this case. With the help of a super-Higgs effect this particle may be absorbed into an additional component of a spin 3/2 particle, called *gravitino,* which becomes massive.

SUSY breaking is then mediated to a visible sector via gravitational interaction leading to the following SUSY breaking scale:

$$M_{\not{SUSY}} \sim \frac{< F_T >}{M_{PL}} + \frac{< F_S >}{M_{PL}} \sim m_{3/2},$$

where $m_{3/2}$ is the gravitino mass.

The effective low-energy theory, which emerges, contains explicit soft supersymmetry breaking terms

$$\mathcal{L}_{soft} = -\sum_i m_i^2 |A_i|^2 - \sum_i M_i(\lambda_i\lambda_i + \bar{\lambda}_i\bar{\lambda}_i) - \mathcal{B}\,\mathcal{W}^{(2)}(A) - \mathcal{A}\,\mathcal{W}^{(3)}(A), \tag{5.1}$$

where $\mathcal{W}^{(2)}$ and $\mathcal{W}^{(3)}$ are the quadratic and cubic terms of a superpotential, respectively. The mass parameters are

$$\begin{aligned} m_i^2 &\sim \left(\frac{< F_S >}{M_{PL}}\right)^2 \sim m_{3/2}^2, \quad M_i \sim \frac{< F_S >}{M_{PL}} \sim m_{3/2}, \\ \mathcal{B} &\sim \left(\frac{< F_T >}{M_{PL}}\right)^2 \sim m_{3/2}^2, \quad \mathcal{A} \sim \frac{< F_{T,S} >}{M_{PL}} \sim m_{3/2}. \end{aligned}$$

To have SUSY masses of an order of 1 TeV, one needs $\sqrt{< F_{T,S} >} \sim 10^{11}$ GeV.

In spite of attractiveness of these mechanism in general, since we know that gravity exists anyway, it is not truly substantiated due to the lack of a consistent theory of quantum (super)gravity. Among the problems of a supergravity mechanism also are the large freedom of parameters and the absence of automatic suppression of flavour violation.

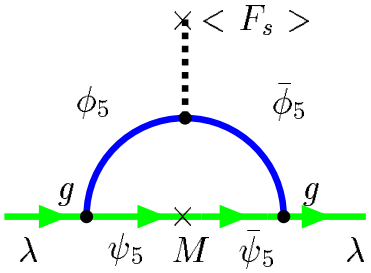
Gauge Mediation

In this version of a hidden sector scenario, the SUSY breaking effects are mediated to the observable world not via gravity but via gauge interactions [32]. The messengers are the gauge bosons and matter fields of the SM and of some GUT theory. The hidden sector is necessary since the dynamical SUSY breaking requires the fields with quantum numbers not compatible with the SM. The advantage of this scenario is that one can construct a renormalizable model with dynamic SUSY breaking, where in principle all the parameters can be calculated.

Consider some simplest possibility where in a hidden sector one has a singlet scalar superfield S with nonzero v.e.v. $< F_S > \neq 0$. The messenger sector consists of some superfield Φ, for instance, $\bar{5}$ of SU(5), that couples to S and to the SM fields with a superpotential

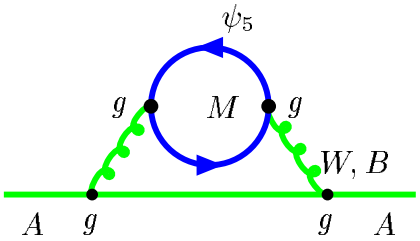
$$\mathcal{W} \sim S\Phi^\dagger\Phi, \quad < S >= M \neq 0. \tag{5.2}$$

Integrating out the messenger fields gives mass to gauginos at the one loop level (see Fig.19) and to the scalar fields (squarks and sleptons) at the two loop one (see Fig.20). So, in gauge mediated scenario all



$$\Longrightarrow \quad M_{\lambda_i} \sim c_i \, N \, \frac{\alpha_i}{4\pi} \, \frac{< F_s >}{M}$$

Fig. 19: Gaugino mass generation



$$\Longrightarrow \quad \tilde{m}_A^2 \sim \left(\frac{< F_s >}{M}\right)^2 \, N \sum_{i=1,2,3} c_i^A \left(\frac{\alpha_i}{4\pi}\right)^2$$

Fig. 20: Squark mass generation

the soft masses are correlated to the gauge couplings and in this sense this scenario is more restrictive than the SUGRA one. There is no problem with flavour violating processes as well, since the soft terms automatically repeat the rigid sector.

It is remarkable that in this scenario the LSP happens to be the gravitino. The mass of the gravitino is given by

$$m_{\tilde{G}} \sim \frac{< F_S >}{M} \cdot \frac{M}{M_{PL}} \sim 10^{-14} \frac{M}{[GeV]}, \tag{5.3}$$

that leads to a very light gravitino field.

The problem of the gauge mediated SUSY breaking scenario emerges in the Higgs sector since the Higgs mass mixing parameters, which break an unwanted Peccei-Quin symmetry, cannot be generated by gauge interactions only. In order to parameterize some new unknown interactions, two new inputs have to be introduced (μ and B in SUGRA conventions).

Anomaly Mediation

An anomaly mediation mechanism assumes no SUSY breaking at the tree level. SUSY breaking is generated due to conformal anomaly. This mechanism refers to a hidden sector of a multidimensional theory with the couplings being dynamic fields which may acquire v.e.v.s. for their F components [33]. The external field or scale dependence of the couplings emerges as a result of conformal anomaly and that is why is proportional to the corresponding β functions. In the leading order one has

$$\begin{aligned} M_i(\Lambda) &\sim b_i \alpha_i(\Lambda) \, \frac{< F_{T,S} >}{M_{PL}} \sim b_i \, \alpha_i \, m_{3/2}, \\ m^2(\Lambda) &\sim b_i^2 \, \alpha_i^2(\Lambda) \, m_{3/2}^2, \end{aligned} \tag{5.4}$$

where b_i are the one-loop RG coefficients (see eq.(2.8)).

This reminds supergravity mediation mechanism but with fixed coefficients. It leads to two main differences:
i) the inverted relation between the gaugino masses at high energy scale

$$M_1 : M_2 : M_3 = b_1 : b_2 : b_3,$$

ii) negative slepton mass squared (tachyons!) at the tree level.
This problem has to be cured.

Gaugino Mediation

At last we would like to mention the gaugino mediation mechanism of SUSY breaking [34]. This is a less developed scenario so far. It is based on a paradigm of a brane world. According to this paradigm, there exists a multidimensional world where our four dimensional space-time represents a brane of 4 dimensions. The fields of the SM live on the brane, while gravity and some other fields can propagate in the bulk. There also exists another brane where supersymmetry is broken. SUSY breaking is mediated to our brane via the fields propagating in the bulk. It is assumed that the gaugino field plays an essential role in this mechanism (see Fig.21)

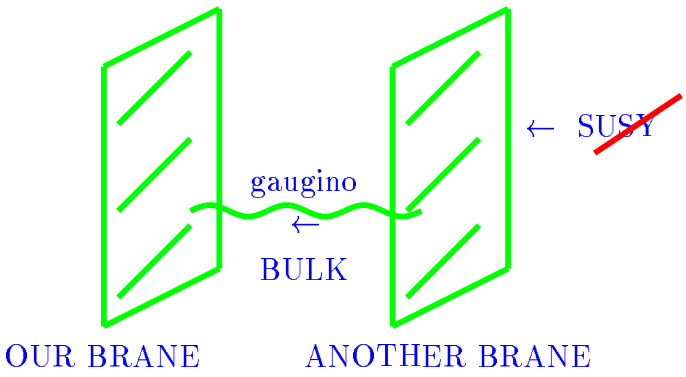


Fig. 21: Gaugino mediated SUSY breaking

All four mechanisms of soft SUSY breaking are different in details but are common in results. They generate gauge invariant soft SUSY breaking operators of dimension ≤ 4 of the form

$$\mathcal{L}_{soft} = -\sum_i m_i^2 |A_i|^2 - \sum_i M_i(\lambda_i\lambda_i + \bar{\lambda}_i\bar{\lambda}_i) - \sum_{ij} B_{ij} A_i A_j - \sum_{ijk} A_{ijk} A_i A_j A_k + h.c., \quad (5.5)$$

where the bilinear and trilinear couplings B_{ij} and A_{ijk} are such that not to break the gauge invariance. These are the only possible soft terms that do not break renormalizability of a theory and preserve SUSY Ward identities for the rigid terms [35].

Predictions for the sparticle spectrum depend on the mechanism of SUSY breaking. For comparison of four above-mentioned mechanisms we show in Fig.22 the sample spectra as the ratio to the gaugino mass M_2 [36].

In what follows, to calculate the mass spectrum of superpartners, we need an explicit form of SUSY breaking terms. Applying eq.(5.5) to the MSSM and avoiding the R-parity violation gives

$$\begin{aligned} -\mathcal{L}_{Breaking} &= \sum_i m_{0i}^2 |\varphi_i|^2 + \left(\frac{1}{2} \sum_\alpha M_\alpha \tilde{\lambda}_\alpha \tilde{\lambda}_\alpha + B H_1 H_2 \right. \quad (5.6) \\ &+ \left. A^U_{ab} \tilde{Q}_a \tilde{U}^c_b H_2 + A^D_{ab} \tilde{Q}_a \tilde{D}^c_b H_1 + A^L_{ab} \tilde{L}_a \tilde{E}^c_b H_1 + h.c. \right), \end{aligned}$$

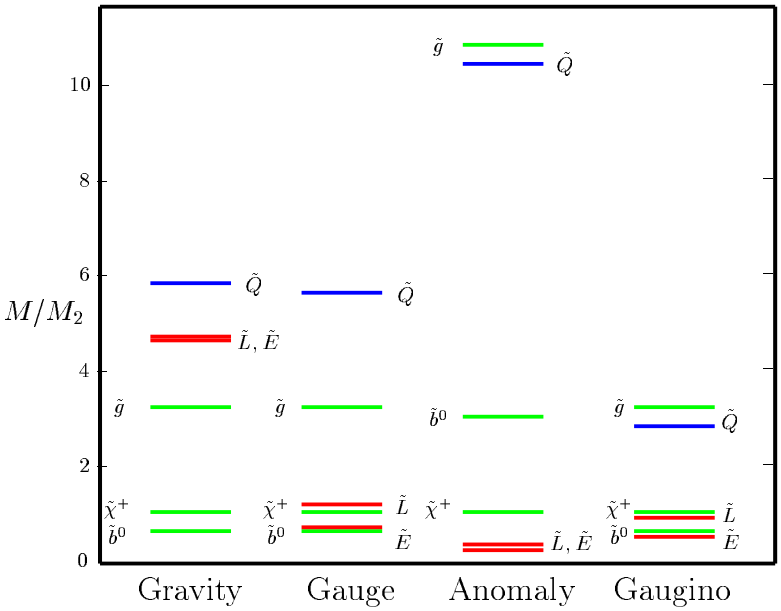


Fig. 22: Superparticle spectra for various mediation mechanisms

where we have suppressed the $SU(2)$ indices. Here φ_i are all scalar fields, $\tilde{\lambda}_\alpha$ are the gaugino fields, $\tilde{Q}, \tilde{U}, \tilde{D}$ and $\tilde{L}, \tilde{E}$ are the squark and slepton fields, respectively, and $H_{1,2}$ are the SU(2) doublet Higgs fields.

Eq.(5.6) contains a vast number of free parameters which spoils the prediction power of the model. To reduce their number, we adopt the so-called universality hypothesis, i.e., we assume the universality or equality of various soft parameters at a high energy scale, namely, we put all the spin 0 particle masses to be equal to the universal value m_0, all the spin 1/2 particle (gaugino) masses to be equal to $m_{1/2}$ and all the cubic and quadratic terms, proportional to A and B, to repeat the structure of the Yukawa superpotential (4.6). This is an additional requirement motivated by the supergravity mechanism of SUSY breaking. Universality is not a necessary requirement and one may consider nonuniversal soft terms as well. However, it will not change the qualitative picture presented below; so for simplicity, in what follows we consider the universal boundary conditions. In this case, eq.(5.6) takes the form

$$\begin{aligned} -\mathcal{L}_{Breaking} &= m_0^2 \sum_i |\varphi_i|^2 + \left(\frac{1}{2} m_{1/2} \sum_\alpha \tilde{\lambda}_\alpha \tilde{\lambda}_\alpha \right. \\ &+ \left. A[y^U_{ab} \tilde{Q}_a \tilde{U}^c_b H_2 + y^D_{ab} \tilde{Q}_a \tilde{D}^c_b H_1 + y^L_{ab} \tilde{L}_a \tilde{E}^c_b H_1] + B[\mu H_1 H_2] + h.c. \right), \end{aligned} \tag{5.7}$$

It should be noted that supergravity induced universality of the soft terms is more likely to be valid at the Planck scale rather than at the GUT one. This is because a natural scale for gravity is M_{Planck} while M_{GUT} is the scale for gauge interactions. However, due to a small difference between these two scales, it is usually ignored in the first approximation resulting in minor uncertainties in the low-energy predictions [37].

The soft terms explicitly break supersymmetry. As will be shown later, they lead to the mass spectrum of superpartners different from that of ordinary particles. Remind that the masses of quarks and leptons remain zero until $SU(2)$ invariance is spontaneously broken.

5.2 The soft terms and the mass formulas

There are two main sources of the mass terms in the Lagrangian: the D terms and soft ones. With given values of $m_0, m_{1/2}, \mu, Y_t, Y_b, Y_\tau, A$, and B one can construct the mass matrices for all the particles. Knowing them at the GUT scale, one can solve the corresponding RG equations, thus linking the values at the GUT and electroweak scales. Substituting these parameters into the mass matrices, one can predict the mass spectrum of superpartners [24, 38, 39].

5.2.1 Gaugino-higgsino mass terms

The mass matrix for gauginos, the superpartners of the gauge bosons, and for higgsinos, the superpartners of the Higgs bosons, is nondiagonal, thus leading to their mixing. The mass terms look like

$$\mathcal{L}_{Gaugino-Higgsino} = -\frac{1}{2} M_3 \bar{\lambda}_a \lambda_a - \frac{1}{2} \bar{\chi} M^{(0)} \chi - (\bar{\psi} M^{(c)} \psi + h.c.), \tag{5.8}$$

where $\lambda_a, a = 1, 2, \ldots, 8$, are the Majorana gluino fields and

$$\chi = \begin{pmatrix} \tilde{B}^0 \\ \tilde{W}^3 \\ \tilde{H}_1^0 \\ \tilde{H}_2^0 \end{pmatrix}, \quad \psi = \begin{pmatrix} \tilde{W}^+ \\ \tilde{H}^+ \end{pmatrix} \tag{5.9}$$

are, respectively, the Majorana neutralino and Dirac chargino fields.

The neutralino mass matrix is

$$M^{(0)} = \begin{pmatrix} M_1 & 0 & -M_Z \cos\beta \sin_W & M_Z \sin\beta \sin_W \\ 0 & M_2 & M_Z \cos\beta \cos_W & -M_Z \sin\beta \cos_W \\ -M_Z \cos\beta \sin_W & M_Z \cos\beta \cos_W & 0 & -\mu \\ M_Z \sin\beta \sin_W & -M_Z \sin\beta \cos_W & -\mu & 0 \end{pmatrix}, \tag{5.10}$$

where $\tan\beta = v_2/v_1$ is the ratio of two Higgs v.e.v.s and $\sin_W = \sin\theta_W$ is the usual sinus of the weak mixing angle. The physical neutralino masses $M_{\tilde{\chi}_i^0}$ are obtained as eigenvalues of this matrix after diagonalization.

For charginos one has

$$M^{(c)} = \begin{pmatrix} M_2 & \sqrt{2} M_W \sin\beta \\ \sqrt{2} M_W \cos\beta & \mu \end{pmatrix}. \tag{5.11}$$

This matrix has two chargino eigenstates $\tilde{\chi}_{1,2}^{\pm}$ with mass eigenvalues

$$M_{1,2}^2 = \frac{1}{2}\left[M_2^2 + \mu^2 + 2M_W^2 \mp \sqrt{(M_2^2 - \mu^2)^2 + 4M_W^4 \cos^2 2\beta + 4M_W^2 (M_2^2 + \mu^2 + 2M_2 \mu \sin 2\beta)} \right]. \tag{5.12}$$

5.2.2 Squark and slepton masses

Non-negligible Yukawa couplings cause a mixing between the electroweak eigenstates and the mass eigenstates of the third generation particles. The mixing matrices for $\tilde{m}_t^2, \tilde{m}_b^2$ and $\tilde{m}_\tau^2$ are

$$\begin{pmatrix} \tilde{m}_{tL}^2 & m_t(A_t - \mu \cot\beta) \\ m_t(A_t - \mu \cot\beta) & \tilde{m}_{tR}^2 \end{pmatrix}, \tag{5.13}$$

$$\begin{pmatrix} \tilde{m}_{bL}^2 & m_b(A_b - \mu \tan\beta) \\ m_b(A_b - \mu \tan\beta) & \tilde{m}_{bR}^2 \end{pmatrix}, \tag{5.14}$$

$$\begin{pmatrix} \tilde{m}^2_{\tau L} & m_\tau(A_\tau - \mu\tan\beta) \\ m_\tau(A_\tau - \mu\tan\beta) & \tilde{m}^2_{\tau R} \end{pmatrix} \tag{5.15}$$

with

$$\begin{aligned}
\tilde{m}^2_{tL} &= \tilde{m}^2_Q + m^2_t + \frac{1}{6}(4M^2_W - M^2_Z)\cos 2\beta, \\
\tilde{m}^2_{tR} &= \tilde{m}^2_U + m^2_t - \frac{2}{3}(M^2_W - M^2_Z)\cos 2\beta, \\
\tilde{m}^2_{bL} &= \tilde{m}^2_Q + m^2_b - \frac{1}{6}(2M^2_W + M^2_Z)\cos 2\beta, \\
\tilde{m}^2_{bR} &= \tilde{m}^2_D + m^2_b + \frac{1}{3}(M^2_W - M^2_Z)\cos 2\beta, \\
\tilde{m}^2_{\tau L} &= \tilde{m}^2_L + m^2_\tau - \frac{1}{2}(2M^2_W - M^2_Z)\cos 2\beta, \\
\tilde{m}^2_{\tau R} &= \tilde{m}^2_E + m^2_\tau + (M^2_W - M^2_Z)\cos 2\beta
\end{aligned}$$

and the mass eigenstates are the eigenvalues of these mass matrices. For the light generations the mixing is negligible.

The first terms here ($\tilde{m}^2$) are the soft ones, which are calculated using the RG equations starting from their values at the GUT (Planck) scale. The second ones are the usual masses of quarks and leptons and the last ones are the D terms of the potential.

5.3 The Higgs potential

As has already been mentioned, the Higgs potential in the MSSM is totally defined by superpotential (and the soft terms). Due to the structure of $\mathcal{W}$ the Higgs self-interaction is given by the D-terms while the F-terms contribute only to the mass matrix. The tree level potential is

$$\begin{aligned}
V_{tree}(H_1, H_2) &= m^2_1|H_1|^2 + m^2_2|H_2|^2 - m^2_3(H_1H_2 + h.c.) \\
&+ \frac{g^2 + g^{'2}}{8}(|H_1|^2 - |H_2|^2)^2 + \frac{g^2}{2}|H^+_1 H_2|^2,
\end{aligned} \tag{5.16}$$

where $m^2_1 = m^2_{H_1} + \mu^2, m^2_2 = m^2_{H_2} + \mu^2$. At the GUT scale $m^2_1 = m^2_2 = m^2_0 + \mu^2_0,\ m^2_3 = -B\mu_0$. Notice that the Higgs self-interaction coupling in eq.(5.16) is fixed and defined by the gauge interactions as opposed to the SM.

The potential (5.16), in accordance with supersymmetry, is positive definite and stable. It has no nontrivial minimum different from zero. Indeed, let us write the minimization condition for the potential (5.16)

$$\frac{1}{2}\frac{\delta V}{\delta H_1} = m^2_1 v_1 - m^2_3 v_2 + \frac{g^2 + g'^2}{4}(v^2_1 - v^2_2)v_1 = 0, \tag{5.17}$$

$$\frac{1}{2}\frac{\delta V}{\delta H_2} = m^2_2 v_2 - m^2_3 v_1 + \frac{g^2 + g'^2}{4}(v^2_1 - v^2_2)v_2 = 0, \tag{5.18}$$

where we have introduced the notation

$$< H_1 > \equiv v_1 = v\cos\beta,\ < H_2 > \equiv v_2 = v\sin\beta, \quad v^2 = v^2_1 + v^2_2, \quad \tan\beta \equiv \frac{v_2}{v_1}.$$

Solution of eqs.(5.17),(5.18) can be expressed in terms of v^2 and $\sin 2\beta$

$$v^2 = \frac{4(m^2_1 - m^2_2\tan^2\beta)}{(g^2 + g'^2)(\tan^2\beta - 1)}, \quad \sin 2\beta = \frac{2m^2_3}{m^2_1 + m^2_2}. \tag{5.19}$$

One can easily see from eq.(5.19) that if $m_1^2 = m_2^2 = m_0^2 + \mu_0^2$, v^2 happens to be negative, i.e. the minimum does not exist. In fact, real positive solutions to eqs.(5.17),(5.18) exist only if the following conditions are satisfied [25]:

$$m_1^2 + m_2^2 > 2m_3^2, \quad m_1^2 m_2^2 < m_3^4, \tag{5.20}$$

which is not the case at the GUT scale. This means that spontaneous breaking of the $SU(2)$ gauge invariance, which is needed in the SM to give masses for all the particles, does not take place in the MSSM.

This strong statement is valid, however, only at the GUT scale. Indeed, going down with energy, the parameters of the potential (5.16) are renormalized. They become the "running" parameters with the energy scale dependence given by the RG equations. The running of the parameters leads to a remarkable phenomenon known as *radiative spontaneous symmetry breaking* to be discussed below.

Provided conditions (5.20) are satisfied, the mass matrices at the tree level are
CP-odd components P_1 and P_2 :

$$\mathcal{M}^{odd} = \left.\frac{\partial^2 V}{\partial P_i \partial P_j}\right|_{H_i = v_i} = \begin{pmatrix} \tan\beta & 1 \\ 1 & \cot\beta \end{pmatrix} m_3^2, \tag{5.21}$$

CP-even neutral components S_1 and S_2:

$$\mathcal{M}^{even} = \left.\frac{\partial^2 V}{\partial S_i \partial S_j}\right|_{H_i = v_i} = \begin{pmatrix} \tan\beta & -1 \\ -1 & \cot\beta \end{pmatrix} m_3^2 + \begin{pmatrix} \cot\beta & -1 \\ -1 & \tan\beta \end{pmatrix} M_Z \cos\beta \sin\beta, \tag{5.22}$$

Charged components H^- and H^+:

$$\mathcal{M}^{charged} = \left.\frac{\partial^2 V}{\partial H_i^+ \partial H_j^-}\right|_{H_i = v_i} = \begin{pmatrix} \tan\beta & 1 \\ 1 & \cot\beta \end{pmatrix} (m_3^2 + M_W \cos\beta \sin\beta). \tag{5.23}$$

Diagonalizing the mass matrices, one gets the mass eigenstates [25]:

$$\begin{cases} G^0 &= -\cos\beta P_1 + \sin\beta P_2, & Goldstone\ boson\ \to Z_0, \\ A &= \sin\beta P_1 + \cos\beta P_2, & Neutral\ CP = -1\ Higgs, \end{cases}$$

$$\begin{cases} G^+ &= -\cos\beta (H_1^-)^* + \sin\beta H_2^+, & Goldstone\ boson\ \to W^+, \\ H^+ &= \sin\beta (H_1^-)^* + \cos\beta H_2^+, & Charged\ Higgs, \end{cases}$$

$$\begin{cases} h &= -\sin\alpha S_1 + \cos\alpha S_2, & SM\ Higgs\ boson\ CP = 1, \\ H &= \cos\alpha S_1 + \sin\alpha S_2, & Extra\ heavy\ Higgs\ boson, \end{cases}$$

where the mixing angle α is given by

$$\tan 2\alpha = -\tan 2\beta \left(\frac{m_A^2 + M_Z^2}{m_A^2 - M_Z^2}\right).$$

The physical Higgs bosons acquire the following masses [24]:

$$\begin{aligned} \text{CP-odd neutral Higgs } A: &\qquad m_A^2 = m_1^2 + m_2^2, \\ \text{Charge Higgses } H^\pm: &\qquad m_{H^\pm}^2 = m_A^2 + M_W^2, \end{aligned} \tag{5.24}$$

CP-even neutral Higgses H, h:

$$m_{H,h}^2 = \frac{1}{2}\left[m_A^2 + M_Z^2 \pm \sqrt{(m_A^2 + M_Z^2)^2 - 4 m_A^2 M_Z^2 \cos^2 2\beta}\right], \tag{5.25}$$

where, as usual,

$$M_W^2 = \frac{g^2}{2}v^2, \;\; M_Z^2 = \frac{g^2+g'^2}{2}v^2.$$

This leads to the once celebrated SUSY mass relations

$$\begin{aligned} m_{H^\pm} &\geq M_W, \\ m_h \leq m_A &\leq M_H, \\ m_h \leq M_Z|\cos 2\beta| &\leq M_Z, \\ m_h^2 + m_H^2 &= m_A^2 + M_Z^2. \end{aligned} \tag{5.26}$$

Thus, the lightest neutral Higgs boson happens to be lighter than the Z boson, which clearly distinguishes it from the SM one. Though we do not know the mass of the Higgs boson in the SM, there are several indirect constraints leading to the lower boundary of $m_h^{SM} \geq 135$ GeV [40]. After including the radiative corrections, the mass of the lightest Higgs boson in the MSSM, m_h, however increases. We consider it in more detail below.

5.4 Renormalization group analysis

To calculate the low energy values of the soft terms, we use the corresponding RG equations. The one-loop RG equations for the rigid MSSM couplings are [41]

$$\begin{aligned} \frac{d\tilde{\alpha}_i}{dt} &= b_i\tilde{\alpha}_i^2, \quad t \equiv \log Q^2/M_{GUT}^2 \\ \frac{dY_U}{dt} &= -Y_L\left(\frac{16}{3}\tilde{\alpha}_3 + 3\tilde{\alpha}_2 + \frac{13}{15}\tilde{\alpha}_1 - 6Y_U - Y_D\right), \\ \frac{dY_D}{dt} &= -Y_D\left(\frac{16}{3}\tilde{\alpha}_3 + 3\tilde{\alpha}_2 + \frac{7}{15}\tilde{\alpha}_1 - Y_U - 6Y_D - Y_L\right), \\ \frac{dY_L}{dt} &= -Y_L\left(3\tilde{\alpha}_2 + \frac{9}{5}\tilde{\alpha}_1 - 3Y_D - 4Y_L\right), \end{aligned} \tag{5.27}$$

where we use the notation $\tilde{\alpha} = \alpha/4\pi = g^2/16\pi^2,\ Y = y^2/16\pi^2$.

For the soft terms one finds

$$\begin{aligned} \frac{dM_i}{dt} &= b_i\tilde{\alpha}_i M_i. \\ \frac{dA_U}{dt} &= \frac{16}{3}\tilde{\alpha}_3 M_3 + 3\tilde{\alpha}_2 M_2 + \frac{13}{15}\tilde{\alpha}_1 M_1 + 6Y_U A_U + Y_D A_D, \\ \frac{dA_D}{dt} &= \frac{16}{3}\tilde{\alpha}_3 M_3 + 3\tilde{\alpha}_2 M_2 + \frac{7}{15}\tilde{\alpha}_1 M_1 + 6Y_D A_D + Y_U A_U + Y_L A_L, \\ \frac{dA_L}{dt} &= 3\tilde{\alpha}_2 M_2 + \frac{9}{5}\tilde{\alpha}_1 M_1 + 3Y_D A_D + 4Y_L A_L, \\ \frac{dB}{dt} &= 3\tilde{\alpha}_2 M_2 + \frac{3}{5}\tilde{\alpha}_1 M_1 + 3Y_U A_U + 3Y_D A_D + Y_L A_L. \\ \frac{d\tilde{m}_Q^2}{dt} &= -\Big[(\frac{16}{3}\tilde{\alpha}_3 M_3^2 + 3\tilde{\alpha}_2 M_2^2 + \frac{1}{15}\tilde{\alpha}_1 M_1^2) - Y_U(\tilde{m}_Q^2 + \tilde{m}_U^2 + m_{H_2}^2 + A_U^2) \\ &\quad -Y_D(\tilde{m}_Q^2 + \tilde{m}_D^2 + m_{H_1}^2 + A_D^2)\Big], \\ \frac{d\tilde{m}_U^2}{dt} &= -\Big[(\frac{16}{3}\tilde{\alpha}_3 M_3^2 + \frac{16}{15}\tilde{\alpha}_1 M_1^2) - 2Y_U(\tilde{m}_Q^2 + \tilde{m}_U^2 + m_{H_2}^2 + A_U^2)\Big], \\ \frac{d\tilde{m}_D^2}{dt} &= -\Big[(\frac{16}{3}\tilde{\alpha}_3 M_3^2 + \frac{4}{15}\tilde{\alpha}_1 M_1^2) - 2Y_D(\tilde{m}_Q^2 + \tilde{m}_D^2 + m_{H_1}^2 + A_D^2)\Big], \end{aligned}$$

$$
\begin{aligned}
\frac{d\tilde{m}_L^2}{dt} &= -\left[3(\tilde{\alpha}_2 M_2^2 + \frac{1}{5}\tilde{\alpha}_1 M_1^2) - Y_L(\tilde{m}_L^2 + \tilde{m}_E^2 + m_{H_1}^2 + A_L^2)\right], \\
\frac{d\tilde{m}_E^2}{dt} &= -\left[(\frac{12}{5}\tilde{\alpha}_1 M_1^2) - 2Y_L(\tilde{m}_L^2 + \tilde{m}_E^2 + m_{H_1}^2 + A_L^2)\right], \\
\frac{d\mu^2}{dt} &= -\mu^2\left[3(\tilde{\alpha}_2 + \frac{1}{5}\tilde{\alpha}_1) - (3Y_U + 3Y_D + Y_L)\right], \\
\frac{dm_{H_1}^2}{dt} &= -\Big[3(\mathrm{a}_2 M_2^2 + \frac{1}{5}\mathrm{a}_1 M_1^2) - 3Y_D(\tilde{m}_Q^2 + \tilde{m}_D^2 + m_{H_1}^2 + A_D^2) \\
&\quad -Y_L(\tilde{m}_L^2 + \tilde{m}_E^2 + m_{H_1}^2 + A_L^2)\Big], \\
\frac{dm_{H_2}^2}{dt} &= -\left[3(\mathrm{a}_2 M_2^2 + \frac{1}{5}\mathrm{a}_1 M_1^2) - 3Y_U(\tilde{m}_Q^2 + \tilde{m}_U^2 + m_{H_2}^2 + A_U^2)\right].
\end{aligned}
\tag{5.28}
$$

Having all the RG equations, one can now find the RG flow for the soft terms. To see what happens at lower scales, one has to run the RG equations for the mass parameters in the opposite direction from the GUT to the EW scale. Let us take some initial values of the soft masses at the GUT scale in the interval between $10^2 \div 10^3$ GeV consistent with the SUSY scale suggested by unification of the gauge couplings (2.9). This leads to the following RG flow of the soft terms shown in Fig.23. [38, 39]

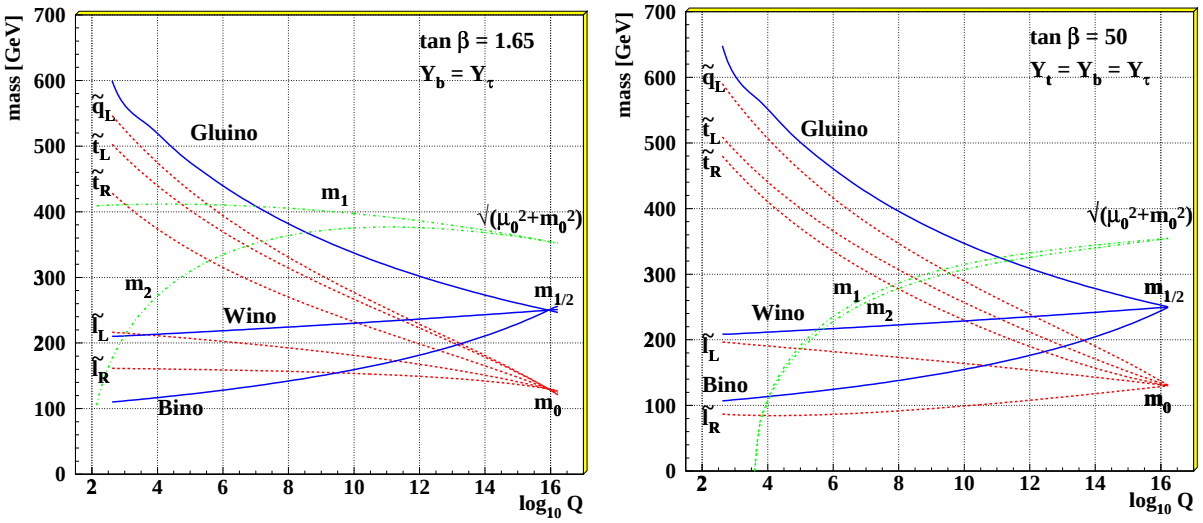


Fig. 23: An example of evolution of sparticle masses and soft supersymmetry breaking parameters $m_1^2 = m_{H_1}^2 + \mu^2$ and $m_2^2 = m_{H_2}^2 + \mu^2$ for low (left) and high (right) values of $\tan\beta$

One should mention the following general features common to any choice of initial conditions:

i) The gaugino masses follow the running of the gauge couplings and split at low energies. The gluino mass is running faster than the others and is usually the heaviest due to the strong interaction.

ii) The squark and slepton masses also split at low energies, the stops (and sbottoms) being the lightest due to relatively big Yukawa couplings of the third generation.

iii) The Higgs masses (or at least one of them) are running down very quickly and may even become negative.

To calculate the masses one has also to take into account the mixing between various states (see eqs.(5.10,5.11, 5.13-5.15).

Numerical solutions allow one to understand the significance of different initial conditions for the evolution down to low energies. As an example we present below the results of a numerical solution to the

RG equations for the soft terms in the case of low values of $\tan\beta$. In this case, one can ignore the bottom and tau Yukawa couplings and keep only the top one. Taking $M_{GUT} = 2.0 \cdot 10^{16}$ GeV, $\alpha(M_{GUT}) \approx 1/24.3,\ Y_t(M_{GUT}) \approx \tilde{\alpha}(M_{GUT}),\ tan\beta = 1.65$, one gets the following numerical results [39]:

$$
\begin{aligned}
M_3(M_Z) &= 2.7\, m_{1/2}, \\
M_2(M_Z) &= 0.8\, m_{1/2}, \\
M_1(M_Z) &= 0.4\, m_{1/2}, \\
\mu(M_Z) &= 0.63\, \mu_0, \\
A_t(M_Z) &= 0.009\, A_t(0) - 1.7\, m_{1/2}, \\
\tilde{m}^2_{E_L}(M_Z) &= m_0^2 + 0.52\, m_{1/2}^2 - 0.27\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{\nu_L}(M_Z) &= m_0^2 + 0.52\, m_{1/2}^2 + 0.5\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{E_R}(M_Z) &= m_0^2 + 0.15\, m_{1/2}^2 - 0.23\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{U_L}(M_Z) &= m_0^2 + 6.6\, m_{1/2}^2 + 0.35\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{D_L}(M_Z) &= m_0^2 + 6.6\, m_{1/2}^2 - 0.42\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{U_R}(M_Z) &= m_0^2 + 6.2\, m_{1/2}^2 + 0.15\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{D_R}(M_Z) &= m_0^2 + 6.1\, m_{1/2}^2 - 0.07\, \cos(2\beta) M_Z^2, \\
\tilde{m}^2_{b_R}(M_Z) &= \tilde{m}^2_{D_R}, \\
\tilde{m}^2_{b_L}(M_Z) &= \tilde{m}^2_{D_L} - 0.48\, m_0^2 - 1.21\, m_{1/2}^2, \\
\tilde{m}^2_{t_R}(M_Z) &= \tilde{m}^2_{U_R} - 0.96\, m_0^2 - 2.42\, m_{1/2}^2, \\
\tilde{m}^2_{t_L}(M_Z) &= \tilde{m}^2_{U_L} - 0.48\, m_0^2 - 1.21\, m_{1/2}^2, \\
m_1^2(M_Z) &= m_0^2 + 0.40\, \mu_0^2 + 0.52\, m_{1/2}^2, \\
m_2^2(M_Z) &= -0.44\, m_0^2 + 0.40\, \mu_0^2 - 3.11\, m_{1/2}^2 - 0.09\, A_0 m_{1/2} - 0/02\, A_0^2.
\end{aligned}
$$

Typical dependence of the mass spectra on the initial conditions (m_0) is also shown in Fig.24 [42]. For a given value of $m_{1/2}$ the masses of the lightest particles are practically independent of m_0, while the heavier ones increase with it monotonically as it follows also from the numerical solutions given above. One can see that the lightest neutralinos and charginos as well as the stop squark may be rather light.

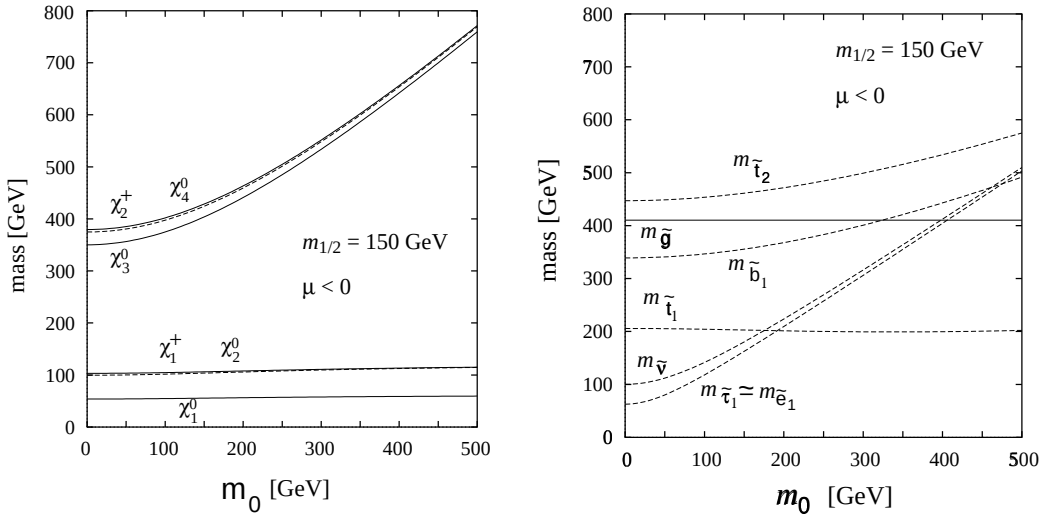


Fig. 24: The masses of sparticles as functions of the initial value m_0

5.5 Radiative electroweak symmetry breaking

The running of the Higgs masses leads to the phenomenon known as *radiative electroweak symmetry breaking*. By this we mean the following: at the GUT energy scale both the Higgs mass parameters m_1^2 and m_2^2 are positive, and the Higgs potential has no nontrivial minima. However, when running down to the EW scale due to the radiative corrections they may change the sign so that the potential develops a nontrivial minimum. At this minimum the electroweak symmetry happens to be spontaneously broken. Thus, contrary to the SM, where one has to choose the negative sign of the Higgs mass squared "by hand", in the MSSM the effect of spontaneous symmetry breaking is triggered by the radiative corrections.

Indeed, one can see in Fig.23 that m_2^2 (or both m_1^2 and m_2^2) decreases when going down from the GUT scale to the M_Z scale and can even become negative. This is the effect of the large top (and bottom) Yukawa couplings in the RG equations. As a result, at some value of Q^2 the conditions (5.20) are satisfied, so that the nontrivial minimum appears. This triggers spontaneous breaking of the $SU(2)$ gauge invariance. The vacuum expectations of the Higgs fields acquire nonzero values and provide masses to quarks, leptons and $SU(2)$ gauge bosons, and additional masses to their superpartners.

In this way one also obtains the explanation of why the two scales are so much different. Due to the logarithmic running of the parameters, one needs a long "running time" to get m_2^2 (or both m_1^2 and m_2^2) to be negative when starting from a positive value of the order of $M_{SUSY} \sim 10^2 \div 10^3$ GeV at the GUT scale.

6 CONSTRAINED MSSM

6.1 Parameter space of the MSSM

The Minimal Supersymmetric Standard Model has the following free parameters:

- Three gauge couplings α_i.
- The matrices of the Yukawa couplings y^i_{ab}, where $i = L, U, D$.
- The Higgs field mixing parameter μ.
- The soft supersymmetry breaking parameters.

Compared to the SM there is an additional Higgs mixing parameter, but the Higgs self-coupling, which is arbitrary in the SM, is fixed by supersymmetry. The main uncertainty comes from the unknown soft terms.

With universality hypothesis one is left with the following set of 5 free parameters defining the mass scales

$$\mu,\ m_0,\ m_{1/2},\ A \text{ and } B.$$

Parameter B is usually traded for $\tan\beta$, the ratio of the v.e.v.s of the two Higgs fields.

In particular models, like in SUGRA or gauge and anomaly mediation, some of soft parameters may be related to each other. However, since the mechanism of SUSY breaking is unknown, in what follows we consider them as free phenomenological parameters to be fitted by experiment. The experimental constraints are sufficient to determine these parameters, albeit with large uncertainties. The statistical analysis yields the probability for every point in the SUSY parameter space, which allows one to calculate the cross sections for the expected new physics of the MSSM at the existing or future accelerators (LEP II, Tevatron, LHC).

While choosing parameters and making predictions, one has two possible ways to proceed:

i) take the low-energy parameters as input, impose the constraints, define the allowed parameter space and calculate the spectrum and cross-sections as functions of these parameters. They might be the superparticle masses $\tilde{m}_{t1}, \tilde{m}_{t2}, m_A$, $\tan\beta$, mixings X_{stop}, μ, etc.

ii) take the high-energy parameters as input, run the RG equations, find the low-energy values, then impose the constrains and define the allowed parameter space for initial values. Now the calculations

can be carried out in terms of the initial parameters. They might be, for example, the above mentioned 5 soft parameters.

Both the ways are used in a phenomenological analysis. We show below how it works in practice.

6.2 The choice of constraints

Among the constraints that we are going to impose on the MSSM model are those which follow from the comparison of the SM with experimental data, from the experimental limits on the masses of as yet unobserved particles, etc, and also those that follow from the ideas of unification and from SUSY GUT models. Some of them look very obvious while the others depend on a choice. Perhaps, the most remarkable fact is that all of them can be fulfilled simultaneously. The only model where one can do it is proved to be the MSSM.

In our analysis we impose the following constraints on the parameter space of the MSSM:

• Gauge coupling constant unification;

This is one of the most restrictive constraints, which we have discussed in Sect 2. It fixes the scale of SUSY breaking of an order of 1 TeV.

• M_Z from electroweak symmetry breaking;

Radiative corrections trigger spontaneous symmetry breaking in the electroweak sector. In this case, the Higgs potential does not have its minimum for all fields equal to zero, but the minimum is obtained for nonzero vacuum expectation values of the fields. Solving M_Z from eq.(5.19) yields

$$M_Z^2 = 2\frac{m_1^2 - m_2^2 \tan^2\beta}{\tan^2\beta - 1}. \tag{6.1}$$

To get the right value of M_Z requires proper adjustment of parameters. This condition determines the value of μ for given values of m_0 and $m_{1/2}$.

• Yukawa coupling constant unification;

The masses of top, bottom and τ can be obtained from the low energy values of the running Yukawa couplings

$$m_t = y_t\, v \sin\beta, \;\; m_b = y_b\, v \cos\beta, \;\; m_\tau = y_\tau\, v \cos\beta. \tag{6.2}$$

Eq.(6.2) is written for the so-called running masses. They can be translated to the pole masses with account taken of the radiative corrections. For the pole masses of the third generation the following values are taken [43], [1]

$$\begin{aligned} M_t &= 174.3 \pm 5.1 \text{ GeV}/c^2, \\ M_b &= 4.94 \pm 0.15 \text{ GeV}/c^2, \\ M_\tau &= 1.7771 \pm 0.0005 \text{ GeV}/c^2. \end{aligned} \tag{6.3}$$

The requirement of bottom-tau Yukawa coupling unification strongly restricts the possible solutions in m_t versus $\tan\beta$ plane [44]-[49] as it can be seen from Fig.25.

• Branching ratio $BR(b \to s\gamma)$;

The branching ratio $BR(b \to s\gamma)$ has been measured by the CLEO [50] collaboration and later by ALEPH [51] and yields the world average of $BR(b \to s\gamma) = (3.14 \pm 0.48) \cdot 10^{-4}$. The Standard Model contribution to this process comes from the $W - t$ loop and gives a prediction which is very close to the experimental value leaving little space for SUSY. In the MSSM, this flavour changing neutral current (FCNC) receives additional contributions from the $H^\pm - t$, $\tilde{\chi}^\pm - \tilde{t}$ and $\tilde{g} - \tilde{q}$ loops. The $\tilde{\chi}^0 - \tilde{t}$ loops are much smaller [52, 53]. In the leading order, SUSY contribution may be rather big, exceeding the experimental value by several standard deviations. However, the NLO corrections are essential.

This requirement imposes severe restrictions on the parameter space, especially for the case of large $\tan\beta$.

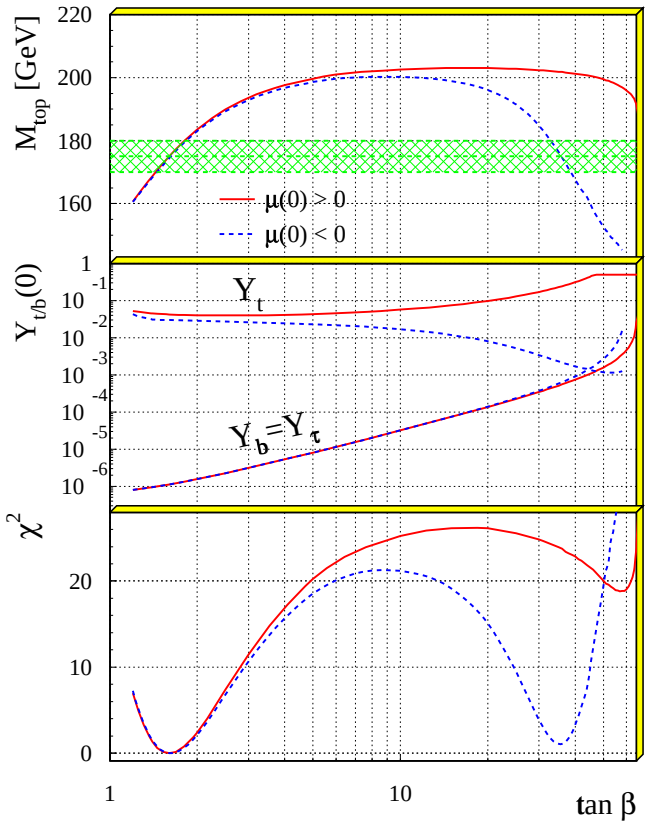


Fig. 25: The upper part shows the top quark mass as a function of $\tan\beta$ for $m_0 = 600$ GeV, $m_{1/2} = 400$ GeV. The middle part shows the corresponding values of the Yukawa couplings at the GUT scale and the lower part of the χ^2 values.

• Experimental lower limits on SUSY masses;

SUSY particles have not been found so far and from the searches at LEP one knows the lower limit on the charged lepton and chargino masses of about half of the centre of mass energy [54]. The lower limit on the neutralino masses is smaller. The lower limit on the Higgs mass is roughly given by the c.m.e. minus the Z-boson mass. These limits restrict the minimal values for the SUSY mass parameters. There exist also limits on squark and gluino masses from the hadron colliders [55], but these limits depend on the assumed decay modes. Furthermore, if one takes the limits given above into account, the constraints from the limits on all other particles are usually fulfilled, so they do not provide additional reductions of the parameter space in the case of the minimal SUSY model.

• Dark Matter constraint;

Abundant evidence of the existence of nonrelativistic, neutral, nonbaryonic dark matter exists in our Universe [56, 57]. The lightest supersymmetric particle (LSP) is supposedly stable and would be an ideal candidate for dark matter.

The present lifetime of the universe is at least 10^{10} years, which implies an upper limit on the expansion rate and correspondingly on the total relic abundance. Assuming $h_0 > 0.4$ one finds that the contribution of each relic particle species χ has to obey [57]

$$\Omega_\chi h_0^2 < 1,$$

where $\Omega_\chi h^2$ is the ratio of the relic particle density of particle χ and the critical density, which overcloses the Universe. This bound can only be met, if most of the LSP's annihilated into fermion-antifermion pairs, which in turn would annihilate into photons again.

Since the neutralinos are mixtures of gauginos and higgsinos, the annihilation can occur both,

via s-channel exchange of the Z^0 and Higgs bosons and t-channel exchange of a scalar particle, like a selectron [58]. This constrains the parameter space, as discussed by many groups [59]-[62].

• Proton life-time constraint;

There are two sources of proton decay in SUSY GUTs. The first one is the same as in non-SUSY theories and is related to the s-channel exchange of heavy gauge bosons. To avoid contradiction with experiment, the unification scale has to be above 10^{15} GeV which is usually satisfied in any SUSY GUT.

The second source is more specific to SUSY models. The proton decay in this case takes place due to the loop diagrams with the exchange of heavy higgsino triplets. The preferable decay mode in this case is $p \to \bar{\nu}K$ or $p \to \mu^+ K$ instead of $p \to e^+\pi$ in non-SUSY GUTs. The decay rate in this case depends on a particular GUT model and it is not so easy to satisfy the experimental requirements.

Having in mind the above mentioned constraints one can try to fix the arbitrariness in the parameters. In a kind of a statistical analysis, in which all the constraints are implemented in a χ^2 definition, one can find the most probable region of the parameter space by minimizing the χ^2 function. For the purpose of this analysis the following χ^2 definition is used [39]:

$$\begin{aligned}
\chi^2 \;=\; & \sum_{i=1}^{3} \frac{(\alpha_i^{-1}(M_Z) - \alpha^{-1}_{MSSM_i}(M_Z))^2}{\sigma_i^2} \\
& + \frac{(M_Z - 91.18)^2}{\sigma_Z^2} + \frac{(M_t - 174)^2}{\sigma_t^2} \\
& + \frac{(M_b - 4.94)^2}{\sigma_b^2} + \frac{(M_\tau - 1.7771)^2}{\sigma_\tau^2} \\
& + \frac{(Br(b \to s\gamma) - 3.14 \times 10^{-4})^2}{\sigma(b \to s\gamma)^2} \\
& + \frac{(\Omega h^2 - 1)^2}{\sigma_\Omega^2} \qquad (for\ \Omega h^2 > 1) \\
& + \frac{(\tilde{M} - \tilde{M}_{exp})^2}{\sigma_{\tilde{M}}^2} \qquad (for\ \tilde{M} < \tilde{M}_{exp}) \\
& + \frac{(\tilde{m}_{LSP} - \tilde{m}_\chi)^2}{\sigma_{LSP}^2} \qquad (for\ \tilde{m}_{LSP}\ charged).
\end{aligned} \tag{6.4}$$

The first six terms are used to enforce gauge coupling unification, electroweak symmetry breaking and $b - \tau$ Yukawa coupling unification, respectively. The following two terms impose the constraints from $b \to s\gamma$ and the relic density, while the last terms require the SUSY masses to be above the experimental lower limits and the lightest supersymmetric particle (LSP) to be a neutralino since a charged stable LSP would have been observed. The input and fitted output variables have been summarized in Table 1.

		Fit parameters	
exp. input data	$\Rightarrow$	low $\tan\beta$	high $\tan\beta$
$\alpha_1, \alpha_2, \alpha_3$	minimize χ^2	$M_{GUT},\ \alpha_{GUT}$	$M_{GUT},\ \alpha_{GUT}$
m_t		$Y_t^0,\ Y_b^0 = Y_\tau^0$	$Y_t^0 = Y_b^0 = Y_\tau^0$
m_b		$m_0, m_{1/2}$	$m_0, m_{1/2}$
m_τ		$\tan\beta$	$\tan\beta$
M_Z		μ	μ
$b \to s\gamma$		(A_0)	A_0
$\tau_{universe}$			

Table 1: Summary of fit input and output variables.

The five-dimensional parameter space of the MSSM is big enough to be represented in a two- or three-dimensional picture. To make our analysis more clear, we consider various projections of the parameter space.

We first choose the value of the Higgs mixing parameter μ from the requirement of radiative EW symmetry breaking, then we take the values of $\tan\beta$ from the requirement of Yukawa coupling unification (see Fig.25). One finds two possible solutions: low $\tan\beta$ solution corresponding to $\tan\beta \approx 1.7$ and high $\tan\beta$ solution corresponding to $\tan\beta \approx 30 \div 60$. In what follows, we refer to these two solutions as low and high $\tan\beta$ scenarios, respectively.

What is left are the values of the soft parameters $A,\ m_0$ and $m_{1/2}$. However, the role of the trilinear coupling A is not essential since at low energies it runs to the infra-red fixed point and is almost independent of initial conditions. Therefore, imposing the above-mentioned constraints, the parameter space of the MSSM is reduced to a two dimensional one. In what follows, we consider the plane $m_0, m_{1/2}$ and find the allowed region in this plane. Each point at this plane corresponds to a fixed set of parameters and allows one to calculate the spectrum, the cross-sections and other quantities of interest.

We present the allowed regions of the parameter space for low and high $\tan\beta$ scenarios in Fig.26. This plot demonstrates the role of various constraints in the χ^2 function. The contours enclose domains by the particular constraints used in the analysis [63]. In case when the requirement of the $b \to s\gamma$ decay

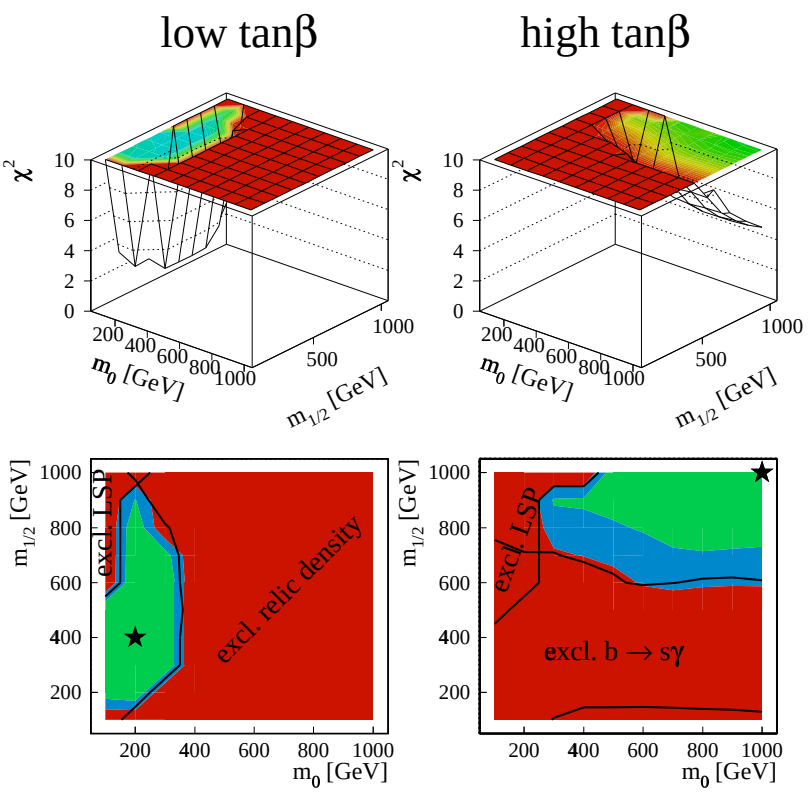


Fig. 26: The χ^2-distribution for low and high $\tan\beta$ solutions. The different shades in the projections indicate steps of $\Delta\chi^2 = 4$, so basically only the light shaded region is allowed. The stars indicate the optimum solution. Contours enclose domains by the particular constraints used in the analysis.

rate is not taken into account (due to uncertainties of the high order contributions), the allowed region of parameter space becomes much wider, as it is illustrated in Fig.27. Now much lower values of m_0 and $m_{1/2}$ are allowed which lead to lower values of sparticle masses.

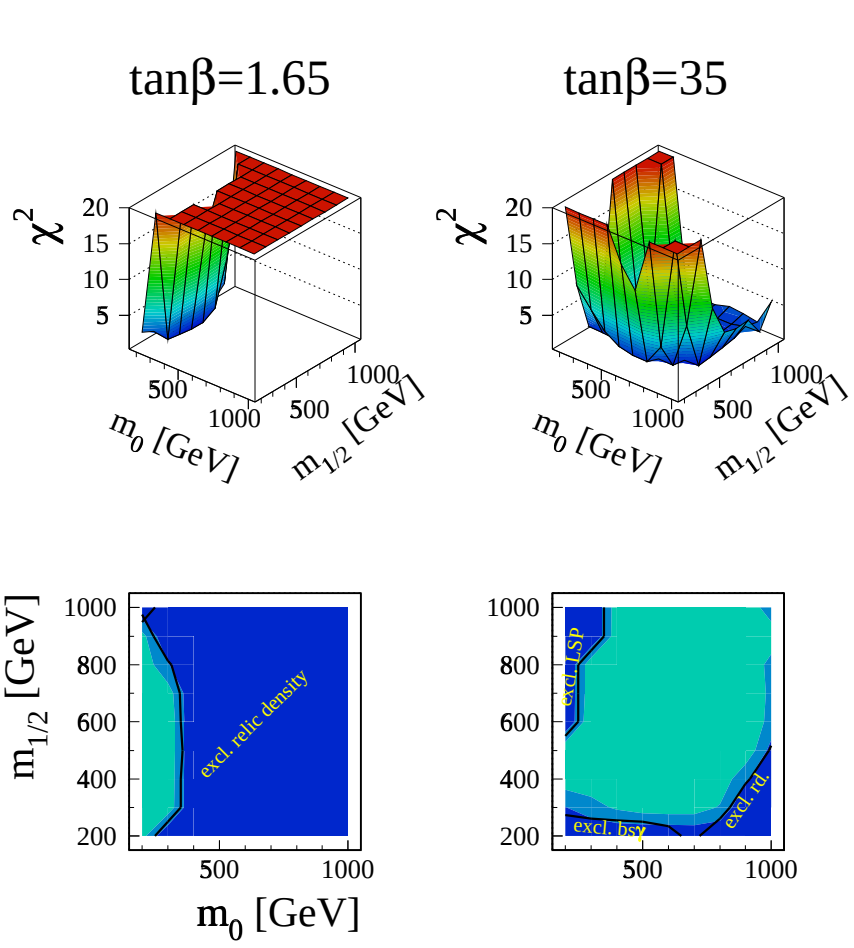


Fig. 27: The same as Fig.26 but with the $b \to s\gamma$ constraint released with account taken of the higher order corrections [64].

6.3 The mass spectrum of superpartners

When the parameter set is fixed, one can calculate the mass spectrum of superpartners. Below we show the set of parameters and the predicted mass spectrum corresponding to the best fit values indicated by stars in Fig.26 [39].

Fitted SUSY parameters		
Symbol	**low** $\tan\beta$	**high** $\tan\beta$
$\tan\beta$	1.71	35.0
m_0	200	600
$m_{1/2}$	500	400
$\mu(0)$	1084	-558
$A(0)$	0	0
$1/\alpha_{GUT}$	24.8	24.8
M_{GUT}	$1.6\ 10^{16}$	$1.6\ 10^{16}$

Table 2: Values of the fitted SUSY parameters for low and high $\tan\beta$ (in GeV, when applicable).

To demonstrate the dependence of masses of the lightest particles on the choice of parameters, we show below in Figs.28,29 their values in the whole $m_0, m_{1/2}$ plane for the case of low and high $tan\beta$ solutions, respectively [63]. One can see that the masses of gauginos (charginos and neutralinos) and Higgses basically depend on $m_{1/2}$, while those of squarks and sleptons on m_0.

SUSY masses in [GeV]		
Symbol	low $\tan\beta$	high $\tan\beta$
$\tilde{\chi}_1^0(\tilde{B})$, $\tilde{\chi}_2^0(\tilde{W}^3)$	214, 413	170, 322
$\tilde{\chi}_3^0(\tilde{H}_1)$,$\tilde{\chi}_4^0(\tilde{H}_2)$	1028, 1016	481, 498
$\tilde{\chi}_1^\pm(\tilde{W}^\pm)$, $\tilde{\chi}_2^\pm(\tilde{H}^\pm)$	413, 1026	322, 499
$\tilde{g}$	1155	950
$\tilde{e}_L$, $\tilde{e}_R$	303, 270	663, 621
$\tilde{\nu}_L$	290	658
$\tilde{q}_L$, $\tilde{q}_R$	1028, 936	1040, 1010
$\tilde{\tau}_1$, $\tilde{\tau}_2$	279, 403	537, 634
$\tilde{b}_1$, $\tilde{b}_2$	953, 1010	835, 915
$\tilde{t}_1$, $\tilde{t}_2$	727, 1017	735, 906
h, H	95, 1344	119, 565
A, $H^\pm$	1340, 1344	565, 571

Table 3: Values of the SUSY mass spectra for the low and high $\tan\beta$ solutions given in Table 2.

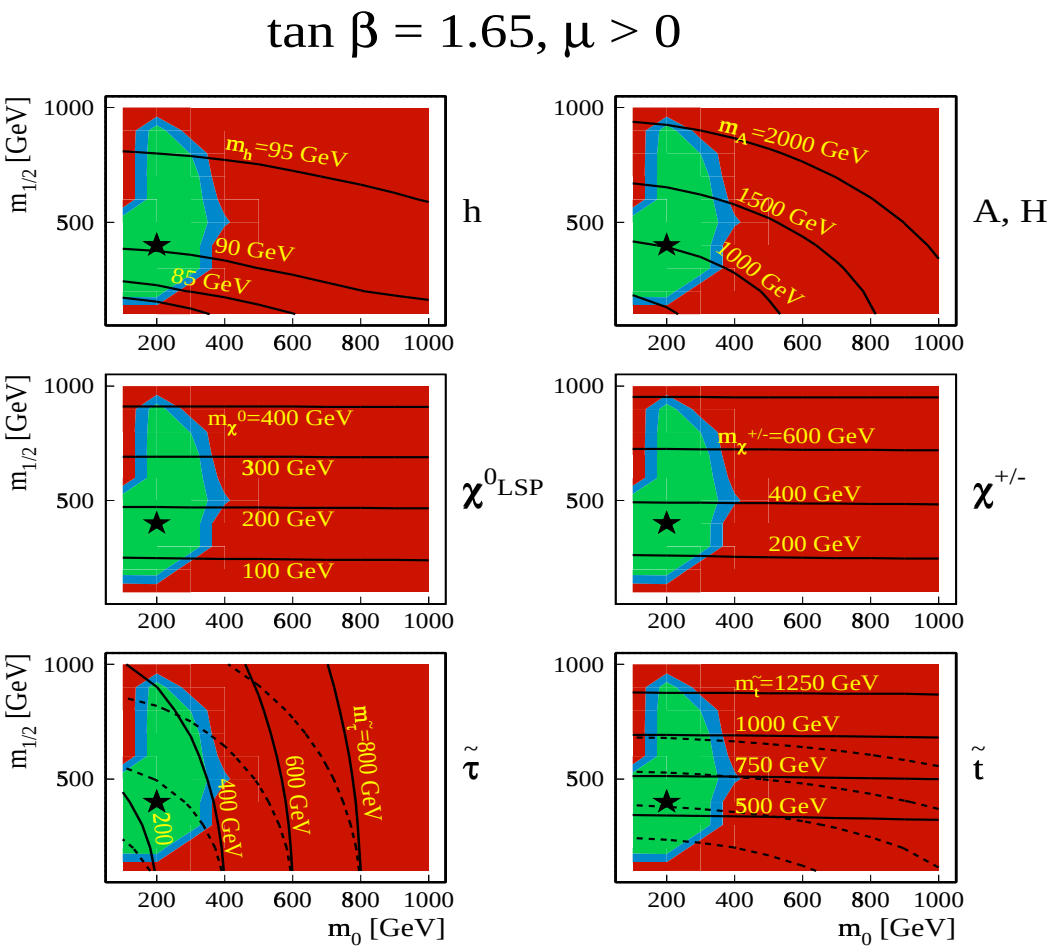


Fig. 28: The masses of the lightest particles in the CMSSM for the low $\tan\beta$ scenario. The contours show the fixed mass values of the corresponding particles.

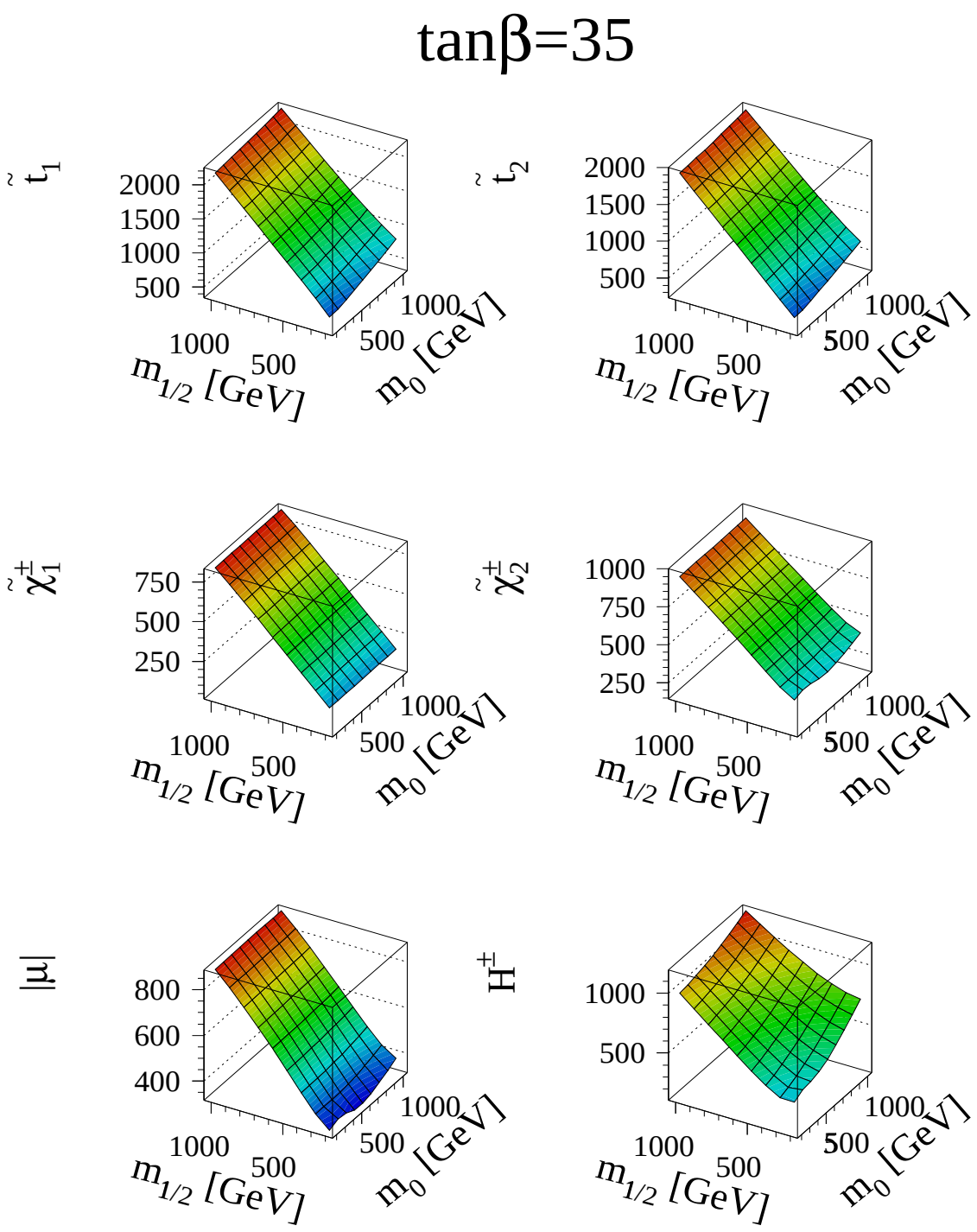


Fig. 29: The masses of the particles and the Higgs mixing parameter μ for $\tan\beta = 35, \mu < 0$.

6.4 Experimental signatures at e^+e^- colliders

Experiments are finally beginning to push into a significant region of supersymmetry parameter space. We know the sparticles and their couplings, but we do not know their masses and mixings. Given the mass spectrum one can calculate the cross-sections and consider the possibilities of observing new particles at modern accelerators. Otherwise, one can get restrictions on unknown parameters.

We start with e^+e^- colliders and, first of all, with LEP II. In the leading order creation of superpartners is given by the diagrams shown in Fig.16 above. For a given center of mass energy the cross-sections depend on the mass of created particles and vanish at the kinematic boundary. For a sample example of c.m. energy of LEP II equal to 183 GeV, they are shown at Fig.30.

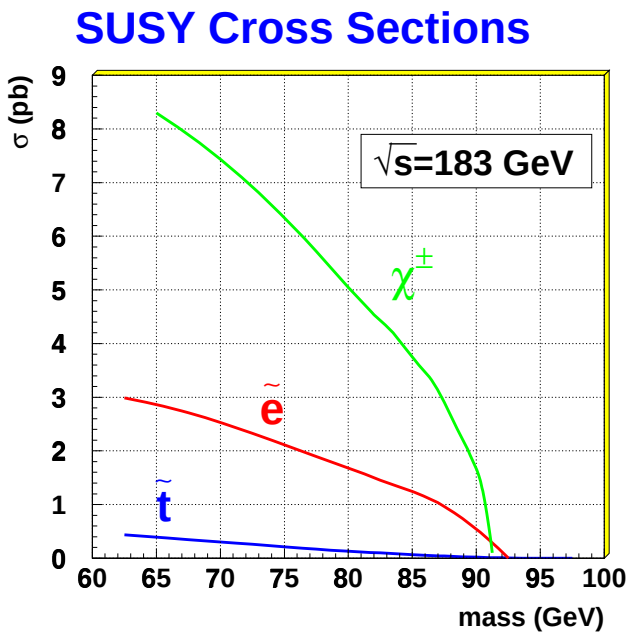


Fig. 30: The cross-section of sparticle production at LEP II as functions of sparticle masses

Experimental signatures are defined by the decay modes which vary with the mass spectrum. The main ones are summarized below.

Production	Key Decay Modes	Signatures
• $\tilde{l}_{L,R}\tilde{l}_{L,R}$	$\tilde{l}_R^{\pm} \to l^{\pm}\tilde{\chi}_i^0 \searrow$ cascade	acomplanar pair of
	$\tilde{l}_L^{\pm} \to l^{\pm}\tilde{\chi}_i^0 \nearrow$ decays	charged leptons + $\not{E}_T$
• $\tilde{\nu}\tilde{\nu}$	$\tilde{\nu} \to l^{\pm}\tilde{\chi}_1^0$	$\not{E}_T$
• $\tilde{\chi}_1^{\pm}\tilde{\chi}_1^{\pm}$	$\tilde{\chi}_1^{\pm} \to \tilde{\chi}_1^0 l^{\pm}\nu,\ \tilde{\chi}_1^0 q\bar{q}'$	isolated lepton + 2 jets + $\not{E}_T$
	$\tilde{\chi}_1^{\pm} \to \tilde{\chi}_2^0 f\bar{f}'$	pair of acomplanar
	$\tilde{\chi}_1^{\pm} \to l\tilde{\nu}_l \to l\nu_l\tilde{\chi}_1^0$	leptons + $\not{E}_T$
	$\tilde{\chi}_1^{\pm} \to \nu_l\tilde{l} \to \nu_l l\tilde{\chi}_1^0$	4 jets + $\not{E}_T$

- $\tilde{\chi}_i^0\tilde{\chi}_j^0$ $\quad \tilde{\chi}_i^0 \to \tilde{\chi}_1^0 X, \tilde{\chi}_j^0 \to \tilde{\chi}_1^0 X'$ $\quad X = \nu_l\bar{\nu}_l$ invisible
  $= \gamma, 2l$, 2 jets
  $2l + \not{E}_T, l + 2j + \not{E}_T$
- $\tilde{t}_i\tilde{t}_j$ $\quad \tilde{t}_1 \to c\tilde{\chi}_1^0$ $\quad$ 2 jets $+ \not{E}_T$
  $\tilde{t}_1 \to b\tilde{\chi}_1^\pm \to bf\bar{f}'\tilde{\chi}_1^0$ $\quad$ 2 b jets + 2 leptons $+ \not{E}_T$
  2 b jets + 2 jets + lepton $+ \not{E}_T$
- $\tilde{b}_i\tilde{b}_j$ $\quad \tilde{b}_i \to b\tilde{\chi}_1^0$ $\quad$ 2 b jets $+ \not{E}_T$
  $\tilde{b}_i \to b\tilde{\chi}_2^0 \to bf\bar{f}'\tilde{\chi}_1^0$ $\quad$ 2 b jets + 2 leptons $+ \not{E}_T$
  2 b jets + 2 jets $+ \not{E}_T$

A characteristic feature of all possible signatures is the missing energy and transverse momenta, which is a trade mark of a new physics.

Numerous attempts to find superpartners at LEP II gave no positive result thus imposing the lower bounds on their masses [54]. They are shown on the parameter plane in Figs.31,32.

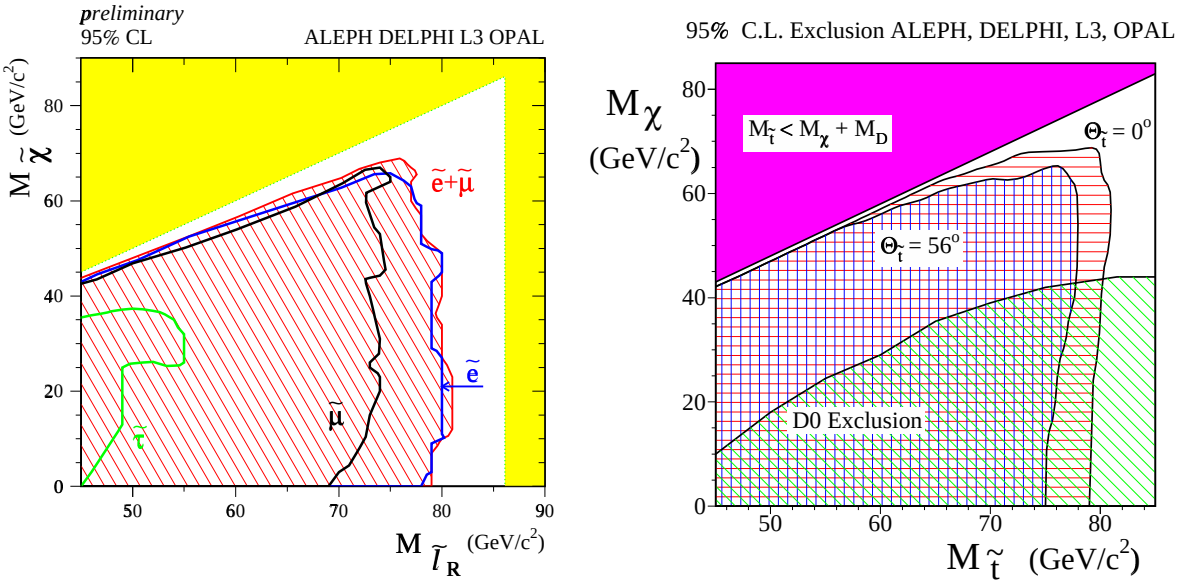


Fig. 31: The excluded region in chargino-slepton and chargino-stop mass plane

In the case of stop masses, the result depends on the stop mixing angle $\Theta_{\tilde{t}}$ calculated from the stop mixing matrix. It defines the mass eigenstates basis $\tilde{t}_1$ and $\tilde{t}_2$

$$\begin{pmatrix} \tilde{t}_1 \\ \tilde{t}_2 \end{pmatrix} = \begin{pmatrix} \cos\Theta_{\tilde{t}} & \sin\Theta_{\tilde{t}} \\ -\sin\Theta_{\tilde{t}} & \cos\Theta_{\tilde{t}} \end{pmatrix} \begin{pmatrix} \tilde{t}_L \\ \tilde{t}_R \end{pmatrix}.$$

Nonobservation of charginos at the maximal LEP II energy defines the lower limit on chargino masses as shown in Fig.32 [54].

Typical LEP II limits on the masses of superpartners are

$$\begin{array}{lll} m_{\chi_1^0} > 40\ GeV & m_{\tilde{e}_{L,R}} > 105\ GeV & m_{\tilde{t}} > 90\ GeV \\ m_{\chi_1^\pm} > 100\ GeV & m_{\tilde{\mu}_{L,R}} > 100\ GeV & m_{\tilde{b}} > 80\ GeV \\ & m_{\tilde{\tau}_{L,R}} > 80\ GeV & \end{array} \qquad (6.5)$$

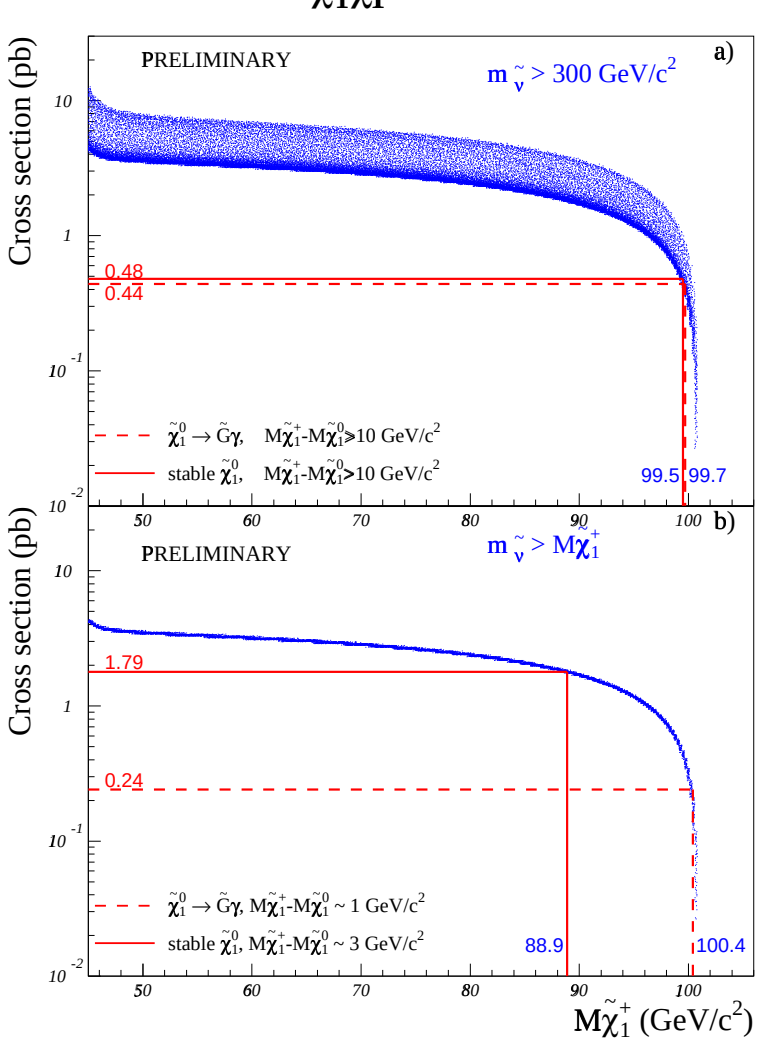


Fig. 32: Cross section of chargino production at LEP and experimental limits on chargino mass

6.5 Experimental signatures at hadron colliders

Experimental signatures at hadron colliders are similar to those at e^+e^- machines; however, here one has much wider possibilities. Besides the usual annihilation channel identical to e^+e^- one with the obvious replacement of electrons by quarks (see Fig.33), one has numerous processes of gluon fusion, quark-antiquark and quark-gluon scattering (see Fig.34).

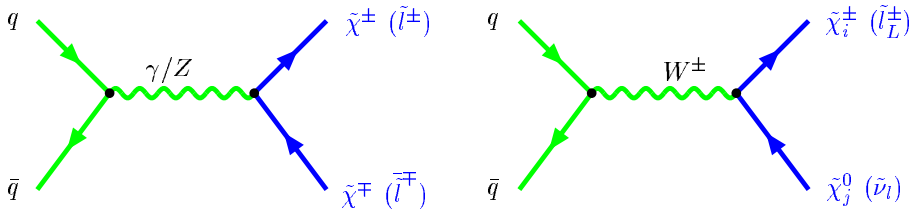


Fig. 33: Annihilation channel

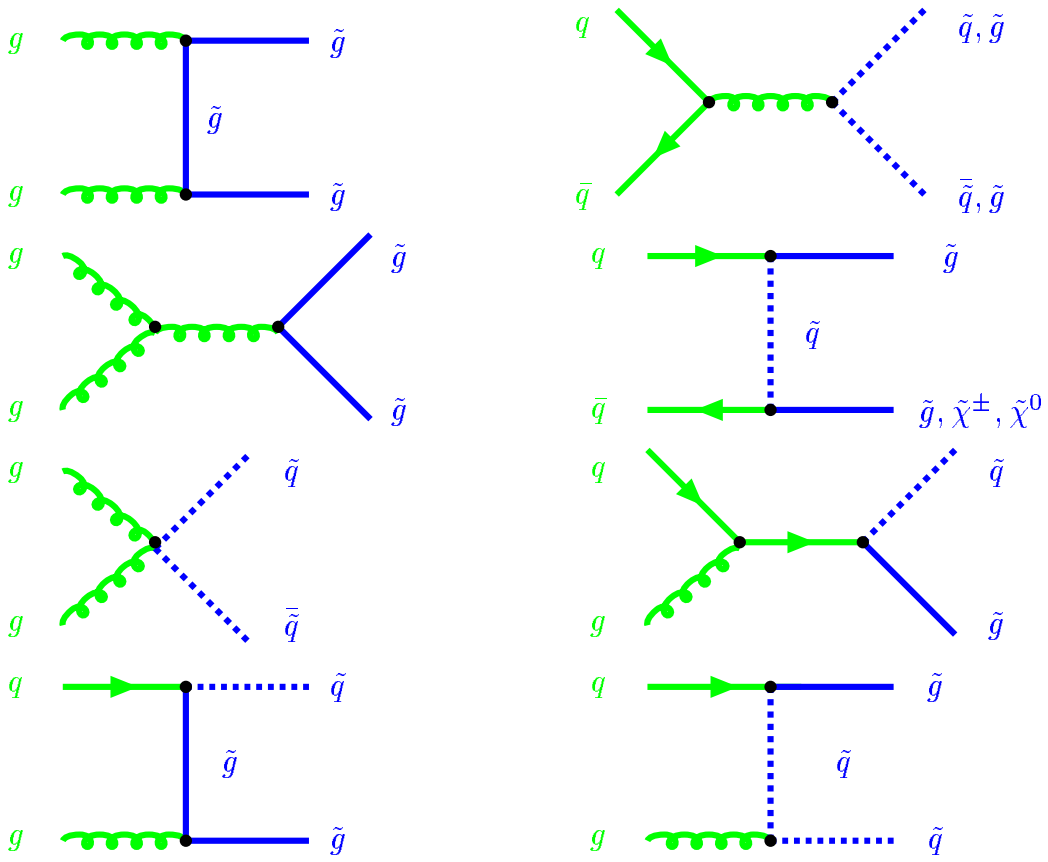


Fig. 34: Gluon fusion, $q\bar{q}$ scattering, quark-gluon scattering

The final states depend on gluino decay modes. If squarks are heavier, i.e. $m_{\tilde{q}} > m_{\tilde{g}}$, then the main gluino decay modes are

$$\tilde{g} \to t + \bar{t} + \tilde{\chi}_i^0, \quad \tilde{g} \to t + \bar{b} + \tilde{\chi}_i^-, \quad \tilde{g} \to t + b + \tilde{\chi}_i^+,$$

otherwise gluino can decay into quarks and squarks with further decay of the latter.

Experimental SUSY signatures at the Tevatron (and LHC) are

Production	Key Decay Modes		Signatures
• $\tilde{g}\tilde{g}, \tilde{q}\tilde{q}, \tilde{g}\tilde{q}$	$\tilde{g} \to q\bar{q}\tilde{\chi}_1^0$ $q\bar{q}'\tilde{\chi}_1^{\pm}$ $g\tilde{\chi}_1^0$	$m_{\tilde{q}} > m_{\tilde{g}}$	$\not{E}_T$ + multijets (+leptons)
	$\tilde{q} \to q\tilde{\chi}_i^0$ $\tilde{q} \to q'\tilde{\chi}_i^{\pm}$	$m_{\tilde{g}} > m_{\tilde{q}}$	
• $\tilde{\chi}_1^{\pm}\tilde{\chi}_2^0$	$\tilde{\chi}_1^{\pm} \to \tilde{\chi}_1^0 l^{\pm}\nu$, $\tilde{\chi}_2^0 \to \tilde{\chi}_1^0 ll$		Trilepton + $\not{E}_T$
	$\tilde{\chi}_1^{\pm} \to \tilde{\chi}_1^0 q\bar{q}'$, $\tilde{\chi}_2^0 \to \tilde{\chi}_1^0 ll$,		Dilepton + jet + $\not{E}_T$

•	$\tilde{\chi}_1^+\tilde{\chi}_1^-$	$\tilde{\chi}_1^+ \to l\tilde{\chi}_1^0 l^\pm\nu$	Dilepton + $\not{E}_T$
•	$\tilde{\chi}_i^0\tilde{\chi}_i^0$	$\tilde{\chi}_i^0 \to \tilde{\chi}_1^0 X, \tilde{\chi}_i^0 \to \tilde{\chi}_1^0 X'$	$\not{E}_T$ + Dilepton + (jets) + (leptons)
•	$\tilde{t}_1\tilde{t}_1$	$\tilde{t}_1 \to c\tilde{\chi}_1^0$	2 acollinear jets + $\not{E}_T$
		$\tilde{t}_1 \to b\tilde{\chi}_1^\pm, \tilde{\chi}_1^\pm \to \tilde{\chi}_1^0 l^\pm\nu, \tilde{\chi}_1^\pm \to \tilde{\chi}_1^0 q\bar{q}'$	single lepton + $\not{E}_T + b's$
		$\tilde{t}_1 \to b\tilde{\chi}_1^\pm, \tilde{\chi}_1^\pm \to \tilde{\chi}_1^0 l^\pm\nu, \tilde{\chi}_1^\pm \to \tilde{\chi}_1^0 l^\pm\nu$	Dilepton + $\not{E}_T + b's$
•	$\tilde{l}\tilde{l}, \tilde{l}\tilde{\nu}, \tilde{n}u\tilde{\nu}$	$\tilde{l}^\pm \to l \pm \tilde{\chi}_i^0, \tilde{l}^\pm \to \nu_l\tilde{\chi}_i^\pm$	Dilepton + $\not{E}_T$
		$\tilde{\nu} \to \nu\tilde{\chi}_1^0$	Single lepton + $\not{E}_T + (jets)$
			$\not{E}_T$

Note again the characteristic missing energy and transverse momenta events.

Contrary to e^+e^- colliders, at hadron machines the background is extremely rich and essential.

6.6 The lightest superparticle

One of the crucial questions is the properties of the lightest superparticle. Different SUSY breaking scenarios lead to different experimental signatures and different LSP.

• Gravity mediation

In this case, the LSP is the lightest neutralino $\tilde{\chi}_1^0$, which is almost 90% photino for a low $\tan\beta$ solution and contains more higgsino admixture for high $\tan\beta$. The usual signature for LSP is missing energy; $\tilde{\chi}_1^0$ is stable and is the best candidate for the cold dark matter in the Universe. Typical processes, where the LSP is created, end up with jets + $\not{E}_T$, or leptons + $\not{E}_T$, or both jest + leptons + $\not{E}_T$.

• Gauge mediation

In this case the LSP is the gravitino $\tilde{G}$ which also leads to missing energy. The actual question here is what the NLSP, the next lightest particle, is. There are two possibilities:

i) $\tilde{\chi}_1^0$ is the NLSP. Then the decay modes are

$$\tilde{\chi}_1^0 \to \gamma\tilde{G},\ h\tilde{G},\ Z\tilde{G}.$$

As a result, one has two hard photons + $\not{E}_T$, or jets + $\not{E}_T$.

ii) $\tilde{l}_R$ is the NLSP. Then the decay mode is $\tilde{l}_R \to \tau\tilde{G}$ and the signature is a charged lepton and the missing energy.

• Anomaly mediation

In this case, one also has two possibilities:

i) $\tilde{\chi}_1^0$ is the LSP and wino-like. It is almost degenerate with the NLSP.

ii) $\tilde{\nu}_L$ is the LSP. Then it appears in the decay of chargino $\tilde{\chi}^+ \to \tilde{\nu}l$ and the signature is the charged lepton and the missing energy.

• R-parity violation

In this case, the LSP is no longer stable and decays into the SM particles. It may be charged (or even colored) and may lead to rare decays like neutrinoless double β-decay, etc.

Experimental limits on the LSP mass follow from non-observation of the corresponding events. Modern low limit is around 40 GeV (see Fig.35).

7 THE HIGGS BOSON IN THE SM AND THE MSSM

One of the hottest topics in the SM now is the search for the Higgs boson. It is also a window to a new physics. Below we consider the situation with the Higgs boson search and the properties of the Higgs boson in the MSSM.

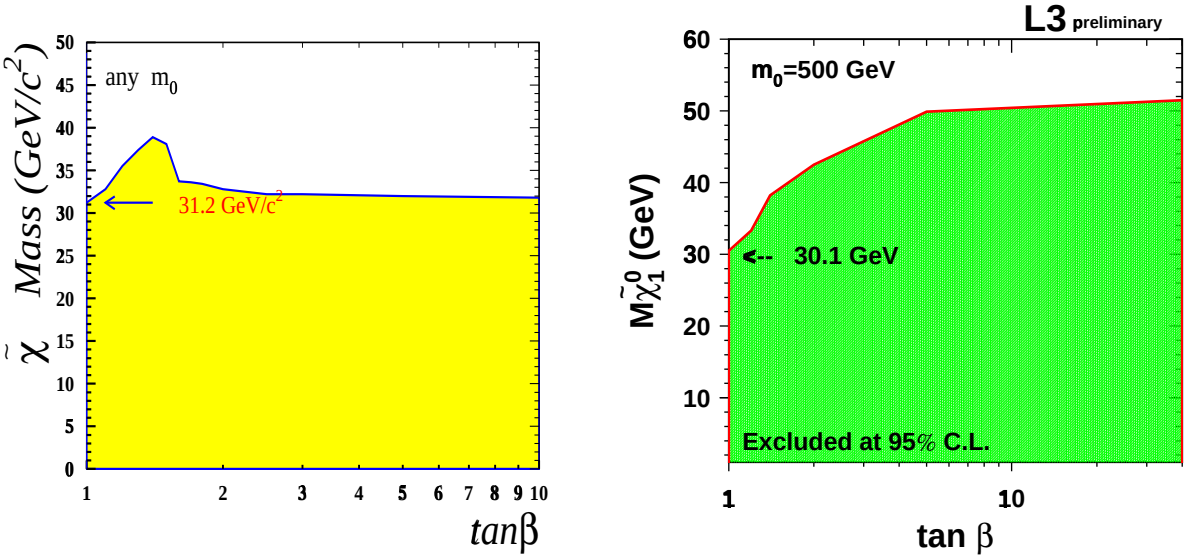


Fig. 35: The LSP mass limits within the MSSM [54]

7.1 Allowed mass range in the SM

The last unobserved particle from the Standard Model is the Higgs boson [65]. Its discovery would allow one to complete the SM paradigm and confirm the mechanism of spontaneous symmetry breaking. On the contrary, the absence of the Higgs boson would awake doubts about the whole picture and would require new concepts.

Experimental limits on the Higgs boson mass come from a direct search at LEP II and Tevatron and from indirect fits of electroweak precision data, first of all from the radiative corrections to the W and top quark masses. A combined fit of modern experimental data gives [66]

$$m_h = 90^{+55}_{-47}\ \mathrm{GeV}, \tag{7.1}$$

which at the 95% confidence level leads to the upper bound of 200 GeV (see Fig.36). At the same time, recent direct searches at LEP II for the c.m. energy of 209 GeV give the lower limit of 113.4 GeV[66]. From a theoretical point of view a low Higgs mass could be a hint for physics beyond the SM, in particular, for the supersymmetric extension of the SM.

Within the Standard Model the value of the Higgs mass m_h is not predicted. However, one can get the bounds on the Higgs mass [40, 67]. They follow from the behaviour of the quartic coupling which is related to the Higgs mass by eqs.(1.9,1.13) $m_h^2 = 2\lambda v$ and obeys the following renormalization group equation describing the change of λ with a scale:

$$\frac{d\lambda}{dt} = \frac{1}{16\pi^2}\left(6\lambda^2 + 6\lambda y_t^2 - 6y_t^4 + \text{gauge terms}\right) \tag{7.2}$$

with $t = \ln(Q^2/\mu^2)$. Here y_t is the top-quark Yukawa coupling.

Since the quartic coupling grows with rising energy infinitely and reaches the Landau pole, the upper bound on m_h follows from the requirement that the theory be valid up to the scale M_{Planck} or up to a given cut-off scale Λ below M_{Planck} [40]. The scale Λ could be identified with the scale at which the Landau pole develops. The upper bound on m_h depends mildly on the top-quark mass through the impact of the top-quark Yukawa coupling on the running of the quartic coupling λ in eq.(7.2).

On the other hand, the requirement of vacuum stability in the SM (positivity of λ) imposes a lower bound on the Higgs boson mass, which crucially depends on both the top-quark mass and the cut-off Λ [40, 67]. Again, the dependence of this lower bound on m_t is due to the effect of the top-quark

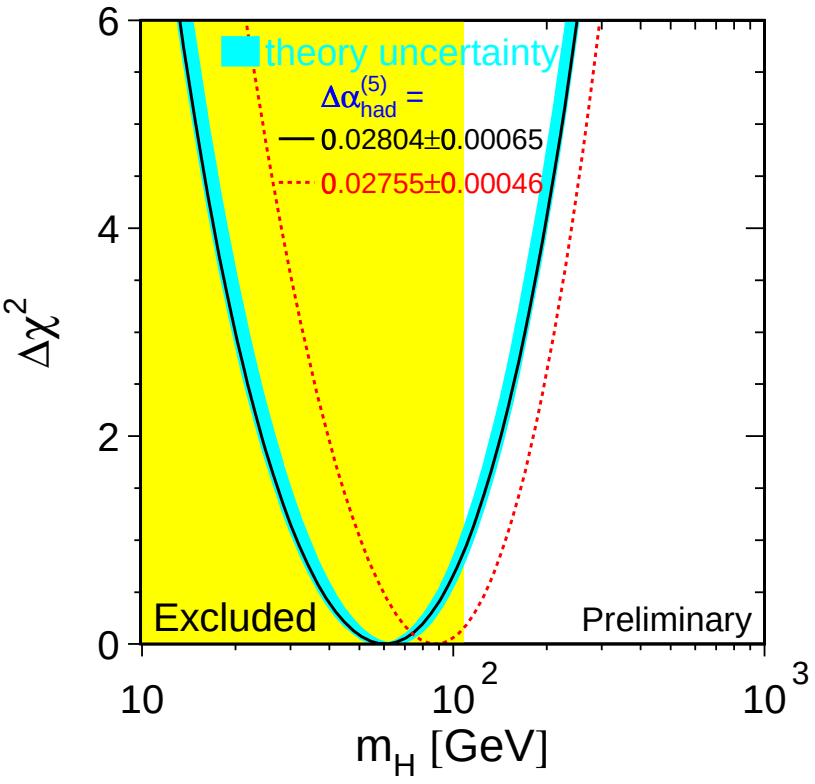


Fig. 36: The χ^2 distribution as a function of the Higgs mass from the SM fit to the electroweak precision observables and the top mass. The shaded area is excluded by the direct searches.

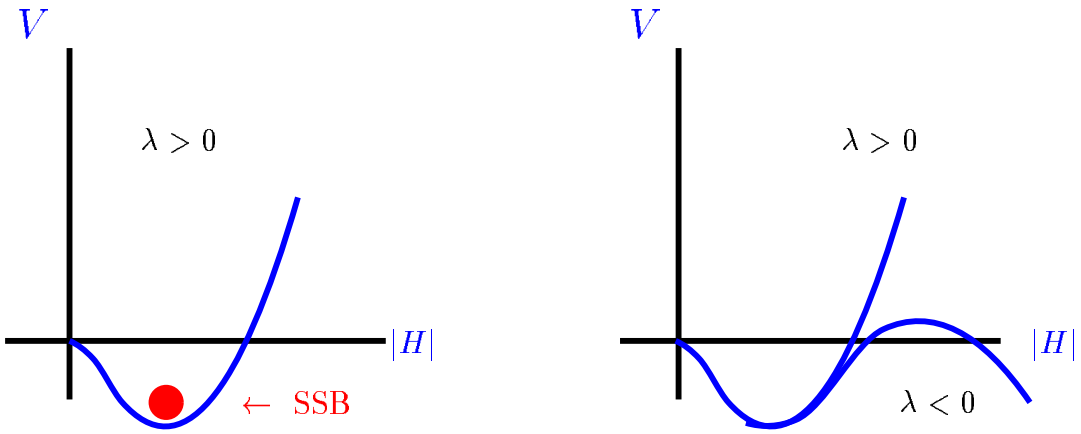


Fig. 37: The shape of the Higgs potential

Yukawa coupling on the quartic coupling in eq.(7.2), which drives λ to negative values at large scales, thus destabilizing the standard electroweak vacuum (see Figs.37).

From the point of view of LEP and Tevatron physics, the upper bound on the SM Higgs boson mass does not pose any relevant restriction. The lower bound on m_h, instead, is particularly important in view of the search for a Higgs boson at LEP II and Tevatron. For $m_t \sim 174$ GeV and $\alpha_s(M_Z) = 0.118$ the running of the Higgs quartic coupling is shown in Fig.38. The results at $\Lambda = 10^{19}$ GeV or at $\Lambda = 1$

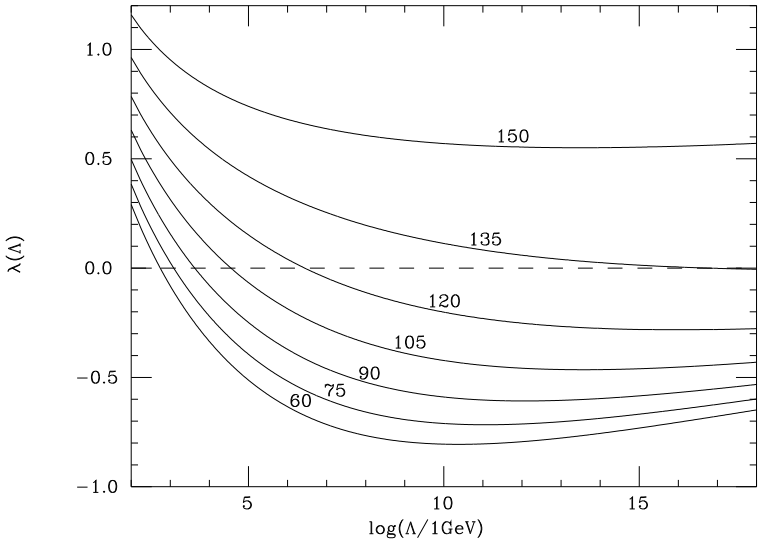


Fig. 38: The running of the Higgs quartic coupling. Numbers shown above the lines indicate the value of the Higgs mass in GeV.

TeV can be given by the approximate formulae [67]

$$m_h \;>\; 135 + 2.1[m_t - 174] - 4.5\left[\frac{\alpha_s(M_Z) - 0.118}{0.006}\right], \quad \Lambda = 10^{19}\ \mathrm{GeV}, \tag{7.3}$$

$$m_h \;>\; 72 + 0.9[m_t - 174] - 1.0\left[\frac{\alpha_s(M_Z) - 0.118}{0.006}\right], \quad \Lambda = 1\ \mathrm{TeV}, \tag{7.4}$$

where the masses are in units of GeV.

Fig.39 [68] shows the perturbativity and stability bounds on the Higgs boson mass of the SM for different values of the cut-off Λ at which new physics is expected. We see from Fig.39 and eqs.(7.3,7.4)

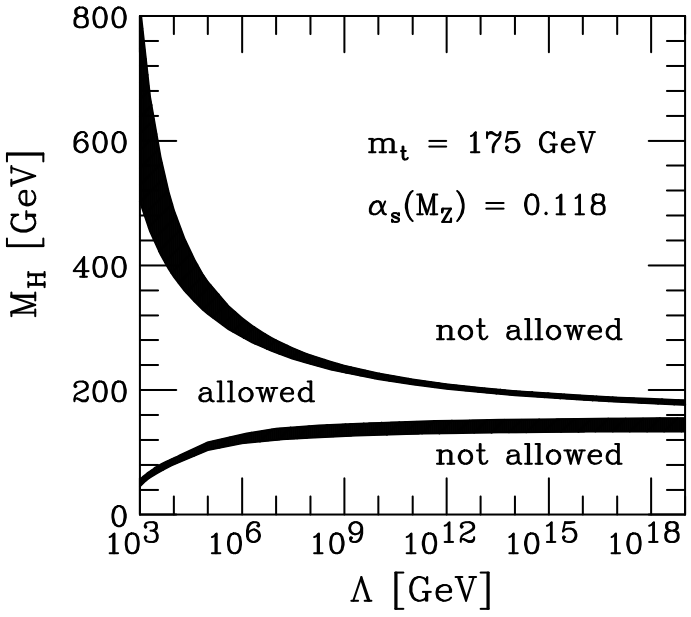


Fig. 39: Strong interaction and stability bounds on the SM Higgs boson mass. Λ denotes the energy scale up to which the SM is valid.

that indeed for $m_t \sim 174$ GeV the discovery of a Higgs particle at LEP II would imply that the Standard Model breaks down at a scale Λ well below M_{GUT} or M_{Planck}, smaller for lighter Higgs. Actually, if the SM is valid up to $\Lambda \sim M_{GUT}$ or M_{Planck}, for $m_t \sim 174$ GeV only a small range of values is allowed: $134 < m_h <\sim 200$ GeV. For m_t = 174 GeV and $m_h < 100$ GeV [i.e. in the LEP II range] new physics should appear below the scale $\Lambda \sim$ a few to 100 TeV. The dependence on the top-quark mass however is noticeable. A lower value, $m_t \simeq 170$ GeV, would relax the previous requirement to $\Lambda \sim 10^3$ TeV, while a heavier value $m_t \simeq 180$ GeV would demand new physics at an energy scale as low as 10 TeV.

7.2 SM Higgs production at LEP

The dominant mechanism for the Higgs boson production at LEP is the Higgsstrahlung. The Higgs boson is produced together with the Z^0 boson. A small contribution to the cross section comes also from the WW- and ZZ- fusion processes (see Fig.40). The cross section depends on the Higgs boson mass

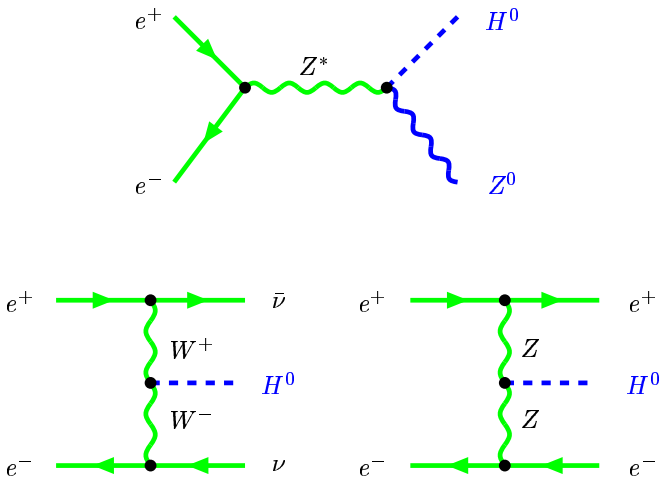


Fig. 40: SM Higgs production at LEP: Higgsstrahlung (above) and WW- and ZZ- fusion (below)

and decreases with increase of the latter. On the other hand, it grows with the centre of mass energy, as shown in Fig.41 [69]. Kinematic limit on the Higgs production is given by the c.m. energy minus the Z-boson mass.

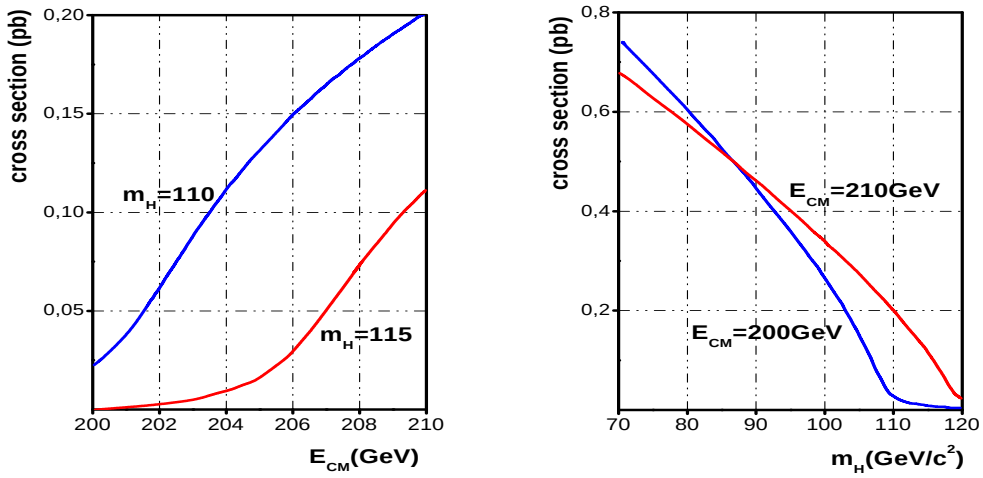


Fig. 41: The cross section of the Higgs production at LEP II

However, one of the main problems is to distinguish the final products of the Higgs boson decay from the background, mainly the ZZ pair production. The branching ratios for the Higgs boson decay are shown in Fig.42. The Z boson has the same decay modes with different branchings. In the final

HZ final states (Higgsstrahlung production)

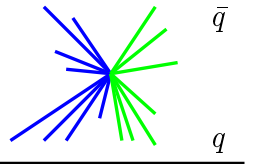
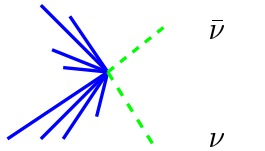
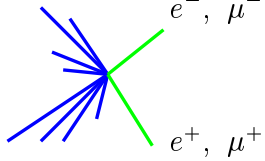
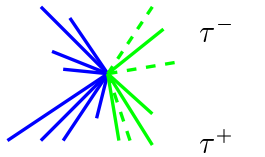
Channel	BR	Topology
$Hq\bar{q}$	64.6%	
$H\nu\bar{\nu}$	20.0%	
$He^+e^-,\ H\mu^+\mu^-$	6.7%	
$H\tau^+\tau^-,\ \tau^+\tau^- q\bar{q},$	8.7%	

Fig. 42: The final states of the Higgs boson decay with the branching ratios

states, one has either four hadronic jets, or two jets and two leptons, or four leptons. The most probable is the four jet configuration, which is the most difficult from the point of view of unwanted background. A two-jet and two-lepton final state is more clean though less probable.

Attempts to find the Higgs boson have not met success so far. All the data are consistent with the background. An interesting four-jet event is shown in Fig.43 and is most likely a ZZ candidate [70]. A reconstructed invariant mass of two jets does not show noticeable deviation from background expectation. For the 68.1 background events expected, there are 70 events observed. The reconstructed Higgs mass for four-jet events is shown in Fig.44. At this kind of plots the real Higgs boson should give a peak above the background, as is shown for a would be Higgs mass of 110 GeV in Fig.44 [70].

Combined results from four LEP collaborations (ALEPH, DELPHI, L3 and OPAL) in the energy interval $\sqrt{s} = 200 - 210$ GeV allow one to get a lower limit on the Higgs mass. As it follows from Fig.45, at the 95% confidence level it is [66]

$$m_h > 113.3 \text{ GeV/c}^2 \quad @\ 95\%\ C.L. \tag{7.5}$$

Recent hot news from the LEP II accelerator show slight excess of events in hadronic channels. For the hard cuts keeping only "really good" events one can achieve the signal/background ratio of 2 with a few signal events indicating the 114 GeV Higgs boson (see Fig.46). Deviation from the background

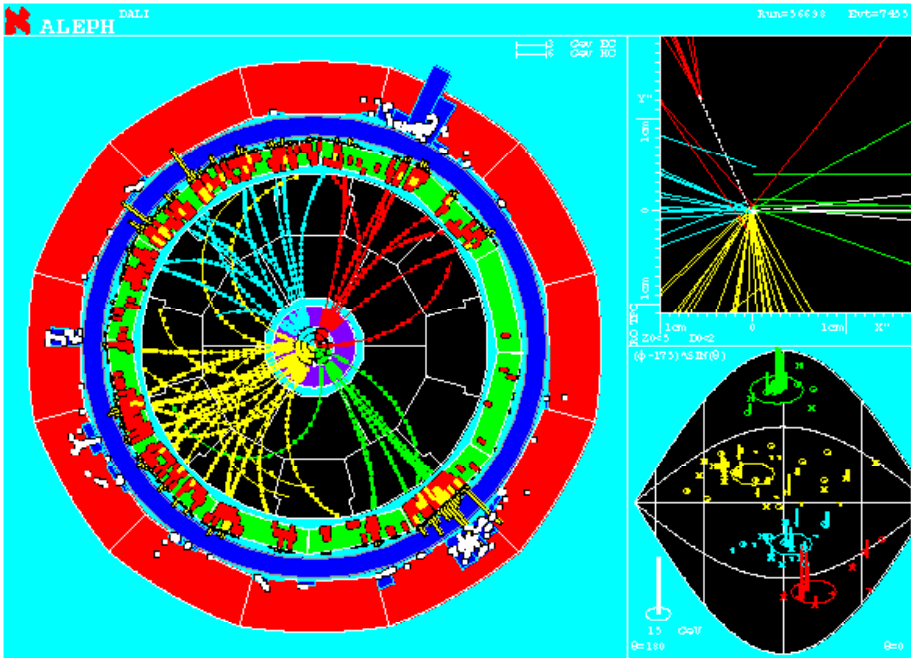


Fig. 43: Typical four jet event

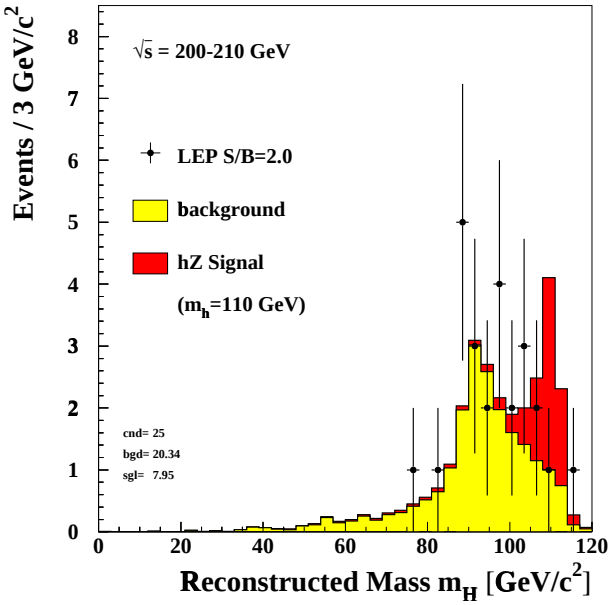


Fig. 44: Reconstructed Higgs mass for four jet events. The peak shown in red corresponds to a would be Higgs boson with mass of 110 GeV

achieves 2.9 standard deviations and is better seen in the confidence level plots [70]. There are also some events in leptonic channel [71]. However, statistics is not enough to make definite conclusions.

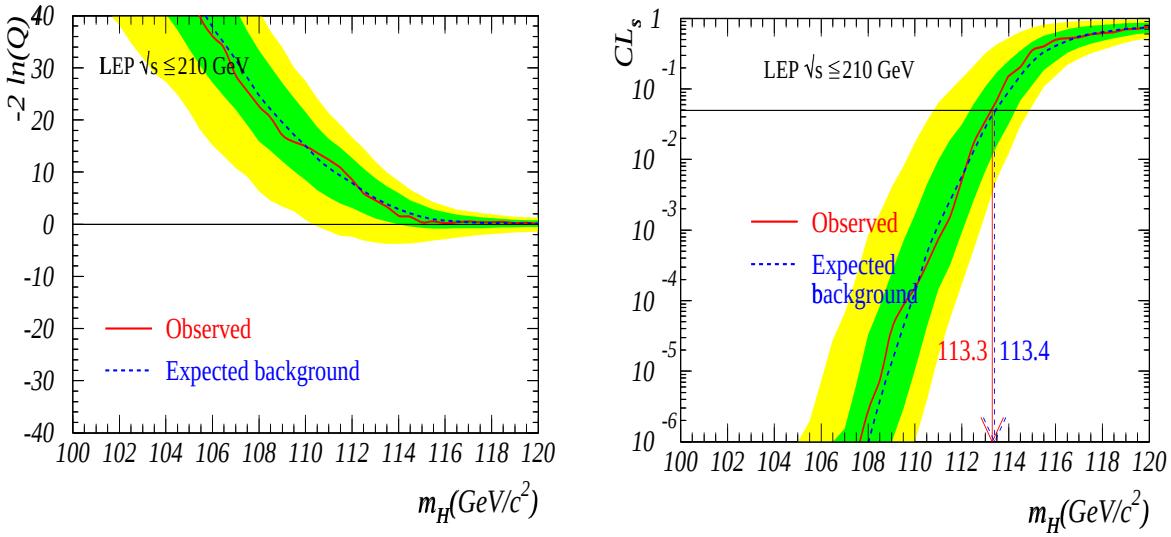
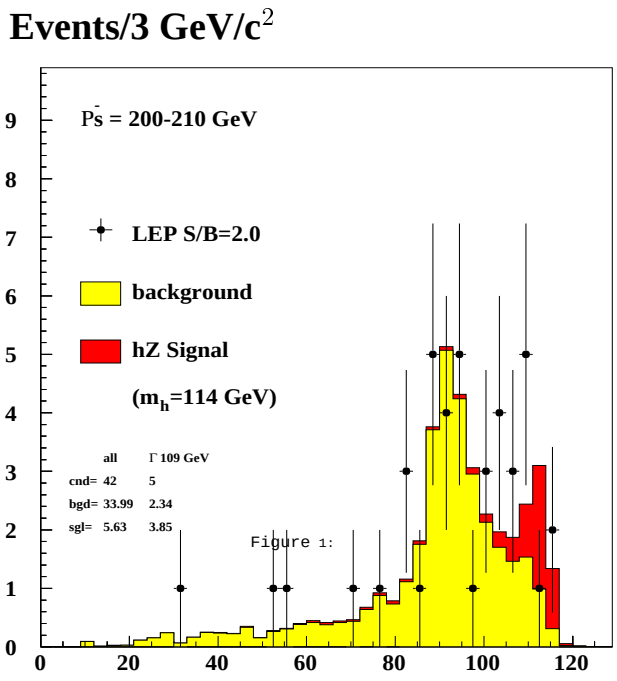


Fig. 45: Combined Confidence Level plots for the Higgs searches at LEP



	Data	Backg	Signal
All m_{rec}	42	34.0	5.6
$m_{rec} > 109$ GeV	5	2.3	3.9

Fig. 46: Higgs Candidate for 114 GeV/c^2

7.3 The Higgs boson mass in the MSSM

It has already been mentioned that in the MSSM the mass of the lightest Higgs boson is predicted to be less than the Z-boson mass. This is, however, the tree level result and the masses acquire the radiative corrections.

With account taken of the radiative corrections, the effective Higgs bosons potential is

$$V^{eff}_{Higgs} = V_{tree} + \Delta V, \tag{7.6}$$

where V_{tree} is given by eq.(5.16) and in the one-loop order

$$\Delta V_{1loop} = \sum_k \frac{1}{64\pi^2}(-1)^{J_k}(2J_k+1)c_k m_k^4 \left(\log\frac{m_k^2}{Q^2} - \frac{3}{2}\right). \tag{7.7}$$

Here the sum is taken over all the particles in the loop, J_k is the spin and m_k is the field dependent mass of a particle at the scale Q.

The main contribution comes from the diagrams shown in Fig.47. These radiative corrections van-

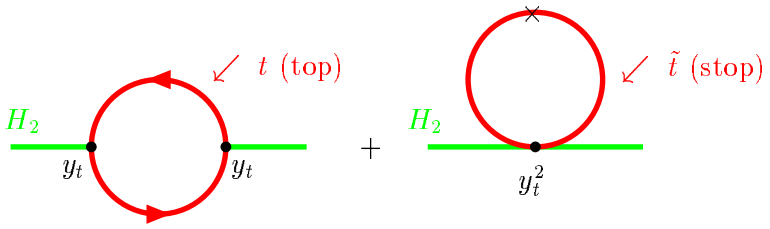


Fig. 47: Corrections to the Higgs boson self-energy from the top(stop) loops

ish when supersymmetry is not broken and are positive in the softly broken case. They are proportional to the mass squared of top (stop) quarks and depend on the values of the soft breaking parameters. Contributions from the other particles are much smaller [72, 73, 74]. The leading contribution comes from (s)top loops

$$\Delta V^{stop}_{1loop} = \frac{3}{32\pi^2}\left[\tilde{m}^4_{t_1}(\log\frac{\tilde{m}^2_{t_1}}{Q^2} - \frac{3}{2}) + \tilde{m}^4_{t_2}(\log\frac{\tilde{m}^2_{t_2}}{Q^2} - \frac{3}{2}) - 2m_t^4(\log\frac{m_t^2}{Q^2} - \frac{3}{2})\right]. \tag{7.8}$$

These corrections lead to the following modification of the tree-level relation for the lightest Higgs mass

$$m_h^2 \approx M_Z^2\cos^2 2\beta + \frac{3g^2 m_t^4}{16\pi^2 M_W^2}\log\frac{\tilde{m}^2_{t_1}\tilde{m}^2_{t_2}}{m_t^4}. \tag{7.9}$$

One finds that the one-loop correction is positive and increases the mass value. Two loop corrections have the opposite effect but are smaller and result in slightly lower value of the Higgs mass [75, 63, 76].

To find out numerical values of these corrections, one has to determine the masses of all superpartners. Within the Constrained MSSM, imposing various constraints, one can define the allowed region in the parameter space and calculate the spectrum of superpartners and, hence, the radiative corrections to the Higgs boson mass (see Figs.48, 49).

The Higgs mass depends mainly on the following parameters: the top mass, the squark masses, the mixing in the stop sector, the pseudoscalar Higgs mass and $\tan\beta$. As will be shown below, the maximum Higgs mass is obtained for large $\tan\beta$, for a maximum value of the top and squark masses and a minimum value of the stop mixing.

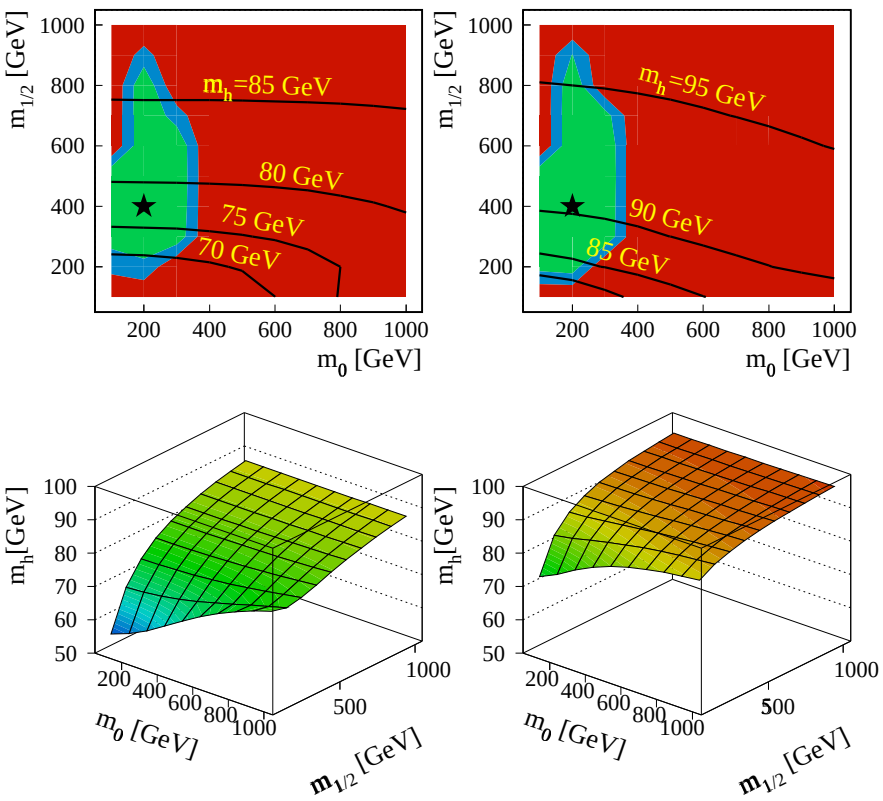


Fig. 48: The mass of the lightest Higgs boson for the low $\tan\beta$ solution as a function of m_0 and $m_{1/2}$. The contours at the upper plots correspond to fixed values of the Higgs mass. The lower plots demonstrate the saturation of the mass at high values of mass parameters.

Note that in the CMSSM the Higgs mixing parameter μ is determined by the requirement of EWSB, which yields large values for μ [39]. Given that the pseudoscalar Higgs mass increases rapidly with μ, this mass is always much larger than the lightest Higgs mass and thus decouples. This decoupling is effective for all regions of the CMSSM parameter space, i.e. the lightest Higgs has the couplings of the SM Higgs within a few per cent.

We present the value of the lightest Higgs mass in the whole $m_0, m_{1/2}$ plane for low and high $\tan\beta$ solutions, respectively [64] in Figs.48, 49. One can see that it is practically constant in the whole plane and is saturated for high values of m_0 and $m_{1/2}$.

The lightest Higgs boson mass m_h is shown as a function of $\tan\beta$ in Fig. 50 [64]. The shaded band corresponds to the uncertainty from the stop mass and stop mixing for $m_t = 175$ GeV. The upper and lower lines correspond to m_t=170 and 180 GeV, respectively.

The parameters used for the calculation of the upper limit are: $m_t = 180$ GeV, $A_0 = -3m_0$ and $m_0 = m_{1/2} = 1000$ GeV. The lowest line of the same figure gives the minimal values of m_h. For high $\tan\beta$ the values of m_h range from 105 GeV 125 GeV. At present, there is no preference for any of the values in this range but it can be seen that the 95% C.L. lower limit on the Higgs mass [66] of 113.3 GeV excludes $\tan\beta < 3.3$.

In order to better understand the Higgs mass uncertainties, the relevant parameters were varied one by one. The largest uncertainty on the light Higgs mass originates from the stop masses. The Higgs

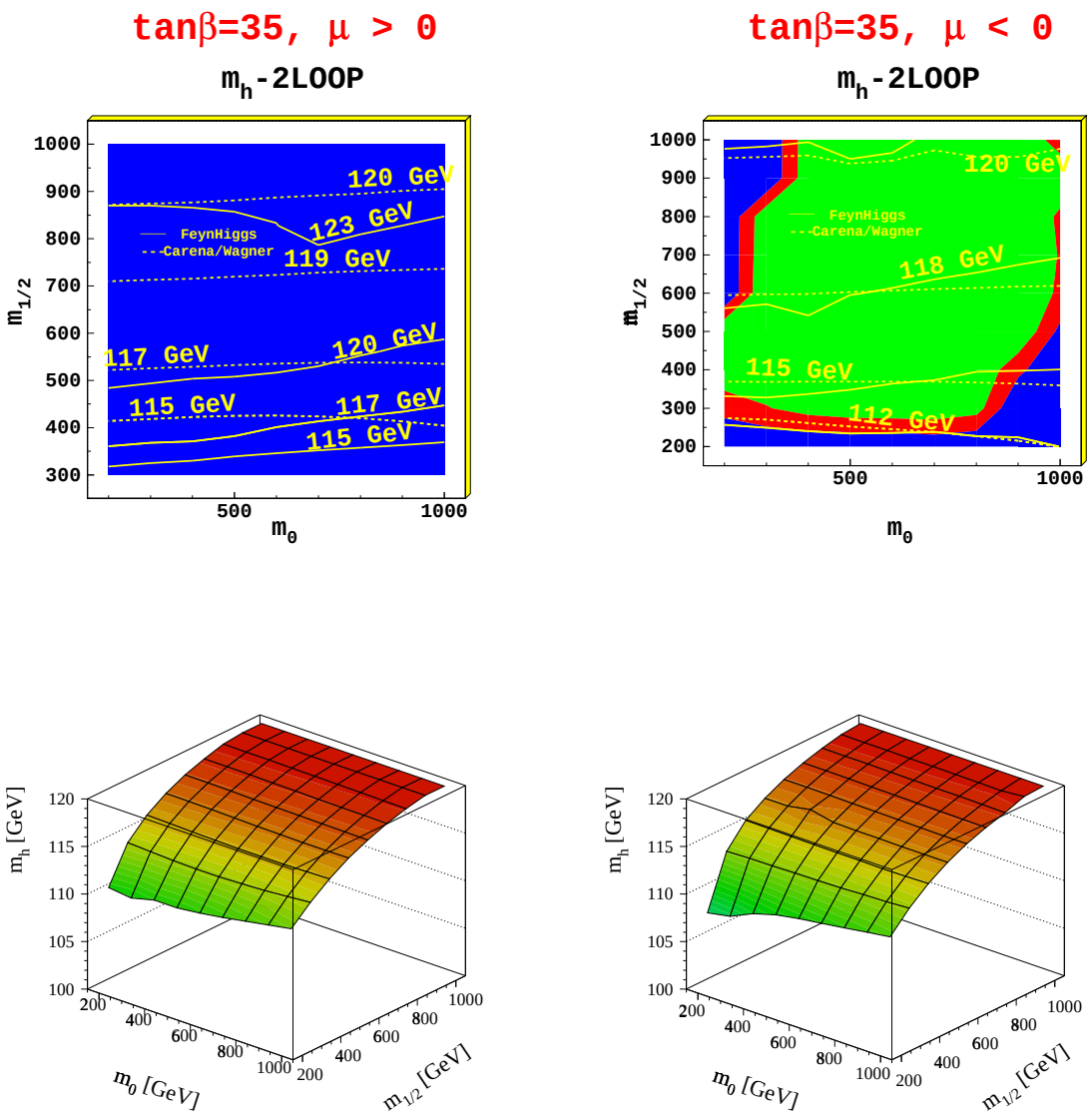


Fig. 49: The same as in Fig. 48 but for the high tan β solution $\tan\beta = 35$.

mass varies between 110 and 120 GeV, if m_0 and $m_{1/2}$ are varied between 200 and 1000 GeV, which implies stop masses varying between 400 and 2000 GeV. Since at present there is no preference for any of the values between 110 and 120 GeV, the variance for a flat probability distribution is $10/\sqrt{12}$=3 GeV, which we take as an error estimate.

The remaining uncertainty of the Higgs mass originates from the mixing in the stop sector when one leaves A_0 as a free parameter. The mixing is determined by the off-diagonal element in the stop mass matrix $X_t = A_t - \mu/\tan\beta$. Its influence on the Higgs mass is quite small in the CMSSM since the low energy value A_t tends to a fixed point so that the stop mixing parameter $X_t = A_t - \mu/\tan\beta$ is not strongly dependent on A_0. Furthermore, the μ term is not important at large $\tan\beta$. If we vary A_0 between $\pm 3m_0$, the error from the stop mixing in the Higgs boson mass is estimated to be ± 1.5 GeV. The values of $m_0 = m_{1/2} = 370$ GeV yield the central value of $m_h = 115$ GeV.

Given the uncertainty on the top mass of 5.2 GeV [43] leads to the uncertainty for the Higgs mass at large $\tan\beta$ of $\pm$ 5 GeV.

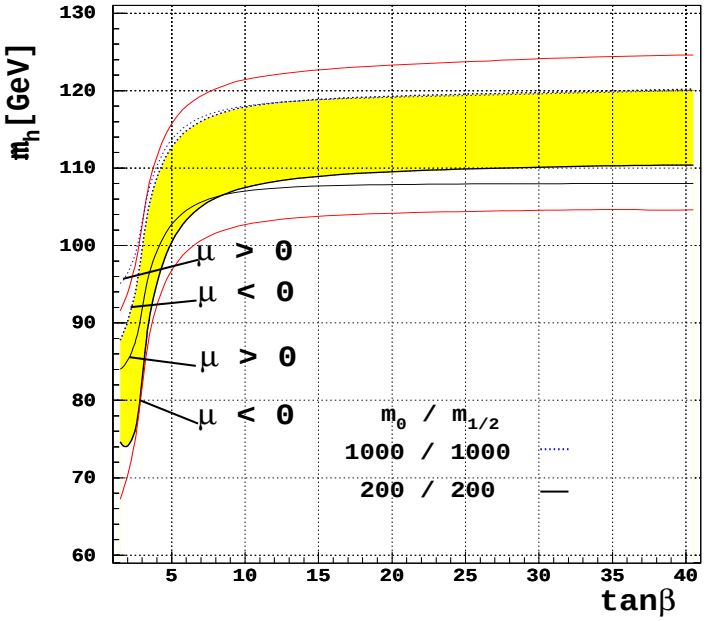


Fig. 50: The mass of the lightest Higgs boson as a function of $\tan\beta$

The uncertainties from the higher order calculations (HO) is estimated to be 2 GeV from a comparison of the full diagrammatic method [76] and the effective potential approach [75]. So combining all the uncertainties discussed before the results for the Higgs mass in the CMSSM can be summarized as follows:

- The low $\tan\beta$ scenario ($\tan\beta < 3.3$) of the CMSSM is excluded by the lower limit on the Higgs mass of 113.3 GeV [66].
- For the high $\tan\beta$ scenario the Higgs mass is found to be in the range from 110 to 120 GeV for $m_t = 175$ GeV. The central value is found to be [64]:

$$m_h = 115 \pm 3\,(\text{stopmass}) \pm 1.5\,(\text{stopmixing}) \pm 2\,(\text{theory}) \pm 5\,(\text{topmass})\ \text{GeV}, \quad (7.10)$$

  where the errors are the estimated standard deviations around the central value. This prediction is independent of $\tan\beta$ for $\tan\beta > 20$ and decreases for lower $\tan\beta$.

However, these SUSY limits on the Higgs mass may not be so restricting if non-minimal SUSY models are considered. In a SUSY model extended by a singlet, the so-called Next-to-Minimal model, eq.(5.26) is modified and at the tree level the upper bound looks like [77]

$$m_h^2 \simeq M_Z^2 \cos^2 2\beta + \lambda^2 v^2 \sin^2 2\beta, \quad (7.11)$$

where λ is an additional singlet Yukawa coupling. This coupling being unknown brings us back to the SM situation, though its influence is reduced by $\sin 2\beta$. As a result, for low $\tan\beta$ the upper bound on the Higgs mass is slightly modified (see Fig.51).

Even more dramatic changes are possible in models containing non-standard fields at intermediate scales. These fields appear in scenarios with gauge mediated supersymmetry breaking. In this case, the upper bound on the Higgs mass may increase up to 155 GeV [77] (the upper curve in Fig.51), though it is not necessarily saturated. One should notice, however, that these more sophisticated models do not change the generic feature of SUSY theories, the presence of the light Higgs boson.

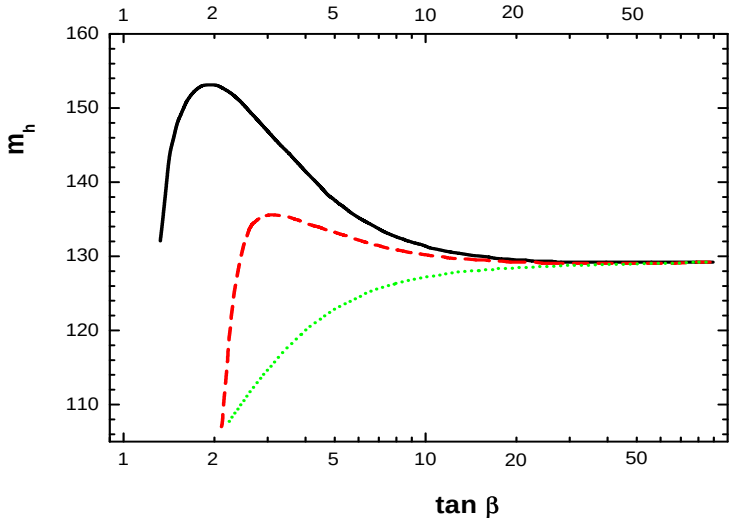


Fig. 51: Dependence of the upper bound on the lightest Higgs boson mass on $\tan\beta$ in MSSM (lower curve), NMSSM (middle curve) and extended SSM (upper curve)

7.4 Perspectives of observation

LEP

In the case of supersymmetry, contrary to the SM, there are two competing processes for neutral Higgs production. Besides the usual Higgsstrahlung diagram there is also the pair production one when two Higgs bosons (the usual one and the pseudoscalar boson A) are produced. The cross-sections of these two processes are complimentary and related to the SM one by a simple formula (see Fig.52). Thus, the

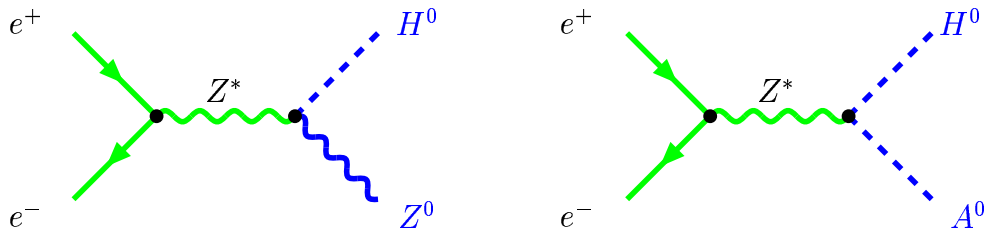
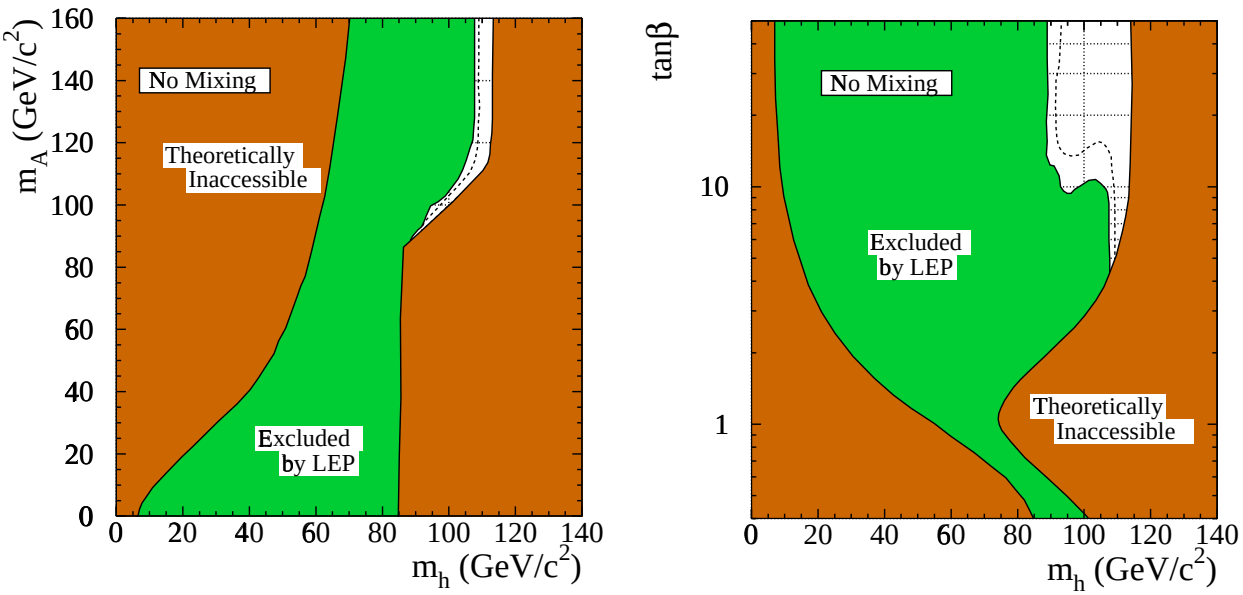


Fig. 52: MSSM Higgs production at LEP: complimentary diagrams

cross-section for Higgs production in the MSSM is usually lower than that of the SM. Therefore, searches for pair production are limited by a low cross-section rather than by a threshold.

Non-observation of the Higgs boson at LEP in general gives lower bound on the Higgs boson mass than that in the SM. Modern experimental limits on the MSSM Higgs bosons are [69]

$$m_h > 90.5\ \mathrm{GeV/c^2}, \quad m_A > 90.5\ \mathrm{GeV/c^2} \quad @\ 95\%\ C.L. \tag{7.12}$$



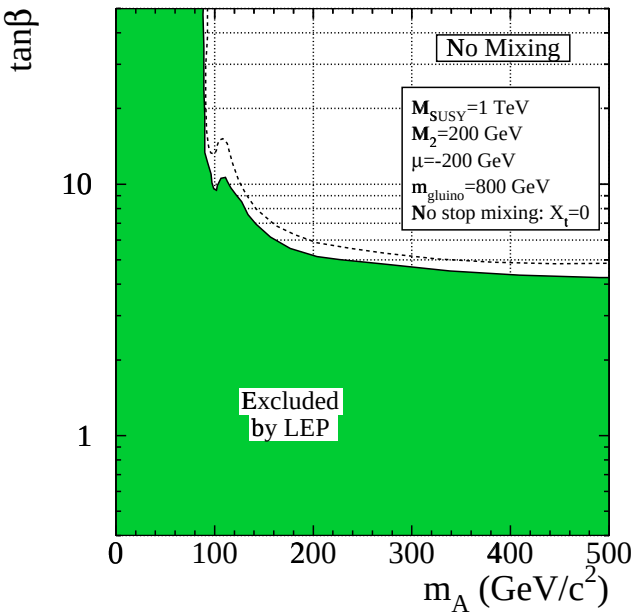


Fig. 53: Excluded regions for the neutral Higgs bosons search in the MSSM in the no-mixing case

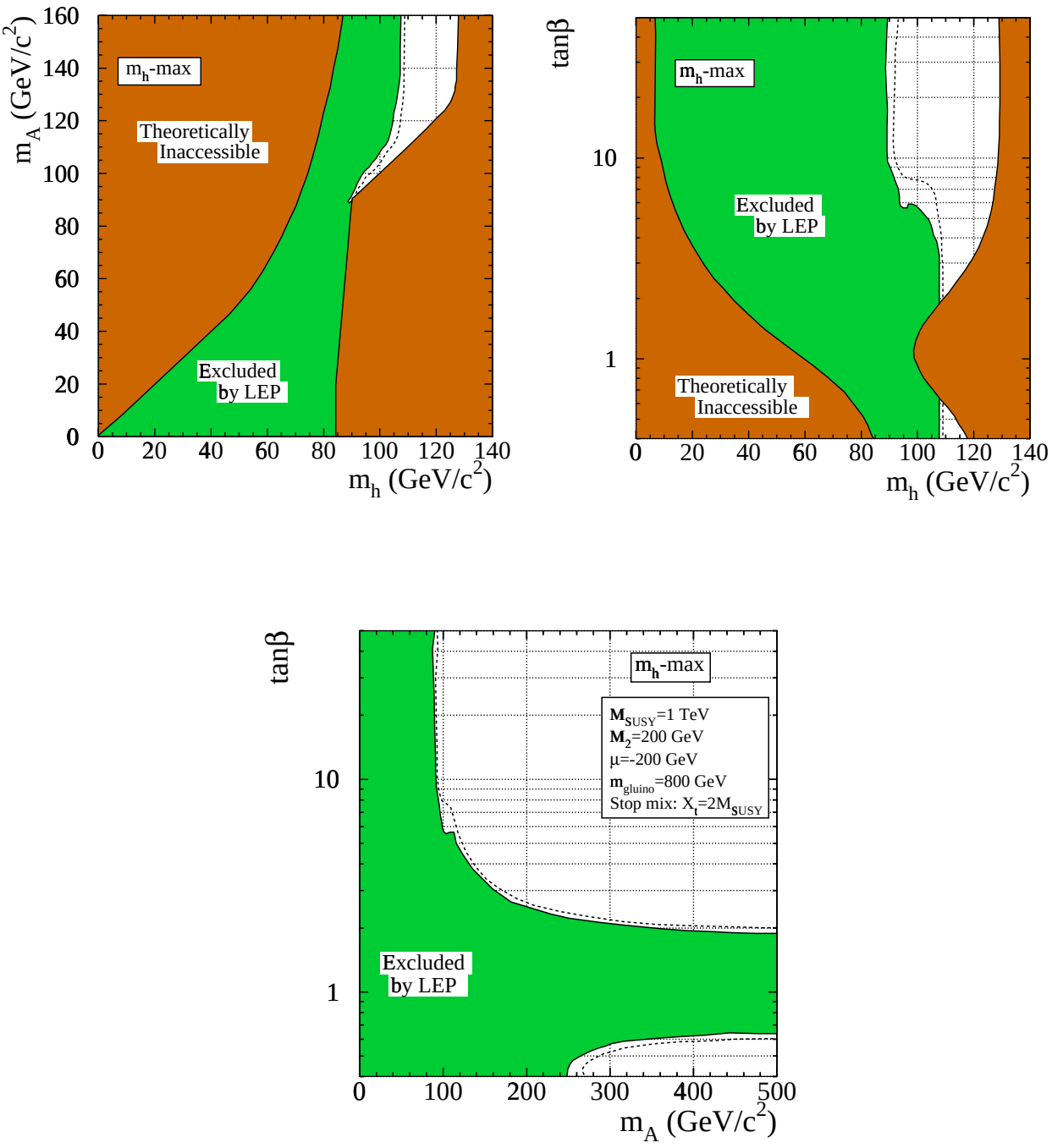


Fig. 54: Excluded regions for the neutral Higgs bosons search in the MSSM in the maximal mixing case

However, for a heavy pseudoscalar boson A the second process is decoupled and one basically has the same production rate as in the SM. Therefore, in this case the SM experimental limit is applicable also to the MSSM.

To present the result for the Higgs search in the MSSM, various variables can be used. The most popular ones are (m_h, m_A), $(m_h, \tan\beta)$ and $(m_A, \tan\beta)$ planes. They are shown below in Figs.53-54 for two particular cases: no-mixing and maximal mixing in the stop sector [66]. For comparison the theoretically allowed regions are shown. One can see that

a) low $\tan\beta$ solution ($0.5 < \tan\beta < 3.3$) is already excluded;

b) very small region for the lightest neutral Higgs boson mass is left (specially for the no-mixing case).

As it has been explained, in the MSSM one has also the charged Higgs bosons. The searches for the charged Higgs bosons are the attempts to look beyond the Standard Model. It is basically the same in the MSSM and in any two Higgs doublet model. The charged Higgs bosons are produced in pairs in an annihilation process like any charged particles. The couplings are the standard EW couplings and the only unknown quantity is the charged Higgs mass. However, the branching ratios for the decay channels depend on the mass and the model. A large background comes from the W-pair production. Nonobservation of charged Higgs bosons at LEP gives the lower limit on their masses. The combined exclusion plot for various channels is shown in Fig.55. This imposes the absolute lower limit on the charged Higgs boson mass [69]

$$m_{H^\pm} > 77.5\ \mathrm{GeV/c^2} \quad @\ 95\%\ C.L. \tag{7.13}$$

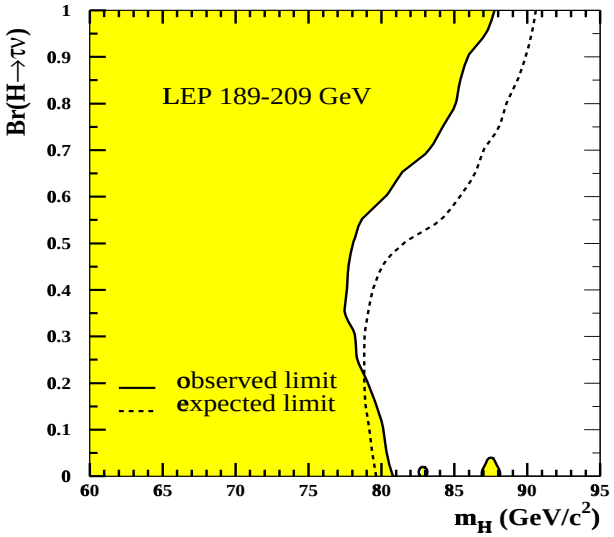


Fig. 55: Combined exclusion plot for the charged Higgs boson

Tevatron and LHC

With the LEP shut down, further attempts to discover the Higgs boson are connected with the Tevatron and LHC hadron colliders.

Tevatron will start the Run II next year and will reach the c.m. energy of 2 TeV with almost 10 times greater luminosity than in Run I. However, since it is a hadron collider, not the full energy goes into

collision taken away by those quarks in a proton that do not take part in the interaction. Having a very severe background, this collider needs a long time of running to reach the integrated luminosity required for the Higgs discovery. A combined CDF/D0 plot [78] shows the integrated luminosity at Tevatron as a function of the Higgs mass (see Fig.56). The three curves correspond to 2σ (95% confidence level), 3σ and 5σ signal necessary for exclusion, evidence and discovery of the Higgs boson, respectively. One can see that the integrated luminosity of 2 fb^{-1}, which is planned to be achieved at the end of 2001, will allow one to exclude the Higgs boson with the mass of an order of 115 GeV, i.e., just the limit reached by LEP. One will need RUN III to reach 10 fb^{-1} to cover the most interesting interval, even at the level of exclusion (2σ).

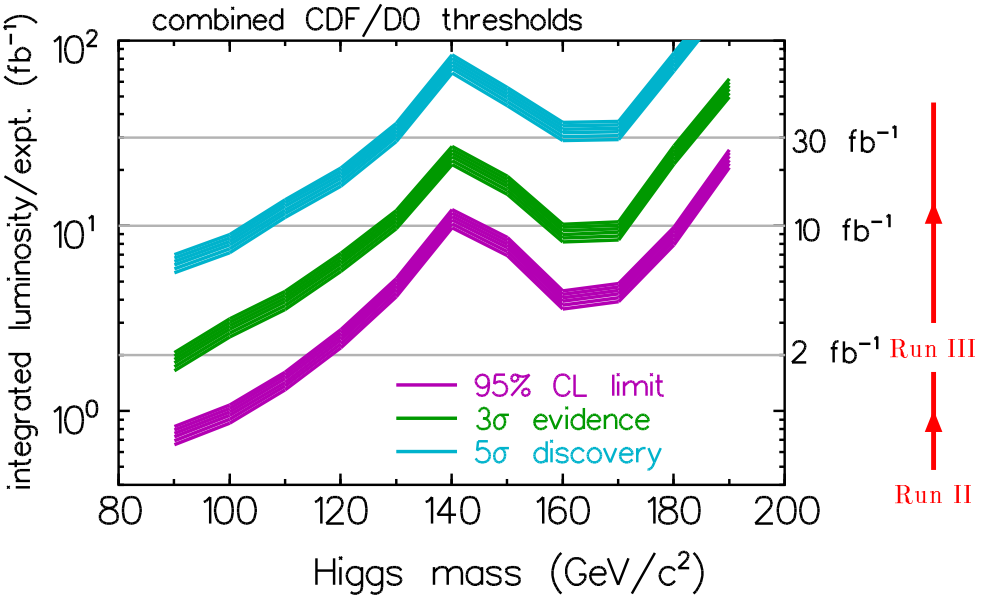


Fig. 56: Integrated luminosity needed for exclusion ((2σ), evidence (3σ) and discovery (5σ) of the Higgs boson at Tevatron

To find the Higgs boson, one will need still greater integrated luminosity. The signatures of the Higgs boson are related to the dominant decay modes which depend on the mass of the Higgs boson. In the Tevatron region they are

$$\begin{array}{ll} H \to b\bar{b}, & 100 < m_H < 140 \text{ GeV}, \\ H \to WW^*, & 140 < m_H < 175 \text{ GeV}, \\ H \to ZZ^*, & 175 < m_H < 190 \text{ GeV}. \end{array} \tag{7.14}$$

The LHC hadron collider is the ultimate machine for a new physics at the TeV scale. Its c.m. energy is planned to be 14 TeV with very high luminosity up to a few hundred fb^{-1}. It is supposed to start operating in 2006. In principle, LHC will be able to cover the whole interval of SUSY and Higgs masses up to a few TeV. It will either discover the SM or the MSSM Higgs boson, or prove their absence. In terms of exclusion plots shown in Figs.53,54 the LHC collider will cover the whole region [79]. Various decay modes allow one to probe different areas, as shown in Fig.57, though the background will be very essential.

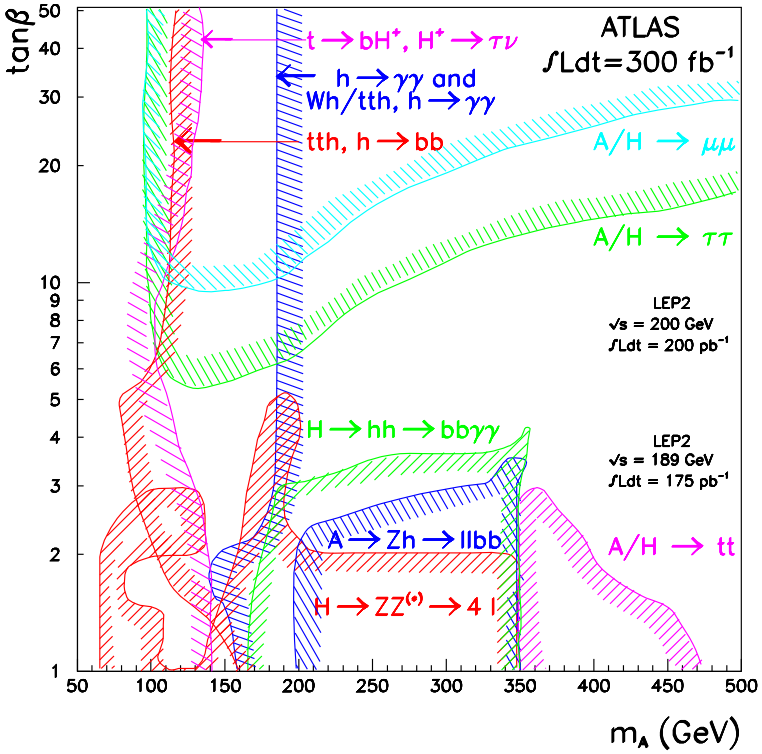


Fig. 57: Exclusion plots for LHC hadron collider for different Higgs decay modes

8 CONCLUSION

LEP II has neither discovered the new physics, nor has proven the existence of the Higgs boson. However, it gave us some indication that both of them exist. Supersymmetry is now the most popular extension of the Standard Model. It promises us that new physics is round the corner at a TeV scale to be exploited at new machines of this decade. If our expectations are correct, very soon we will face new discoveries, the whole world of supersymmetric particles will show up and the table of fundamental particles will be enlarged in increasing rate. If we are lucky, probably we will soon have the table of sparticles in new addition of Sparticle Data Group (see Fig.58) [80]. This would be a great step in understanding the microworld. If not, still new discoveries are in agenda.

Acknowledgements

I am grateful to the organizers of the School at Caramulo for providing very nice atmosphere during the school and to the students for their interest, patience and attention. I would like to thank V.Velizhanin for his help in preparation of this manuscript. Financial support from RFBR grants # 99-02-16650, # 98-02-17453 and # 00-15-96691 and Heienberg-Landau Programme is kindly acknowledged.

The SPDG is an international collaboration that reviews Sparticle Physics and related areas of Astrophysics, and compiles/analyzes data on particle properties. SPDG products are distributed to 130,000 physicists, teachers, and other interested people. **The Review of Sparticle Physics** is the most cited publication in particle physics during the last twenty years. Plots of SPDG statistics are available.

Mirror sites: USA (LBNL) Brazil CERN Italy (Genova) Japan (KEK) Russia (Novosibirsk) Russia (Protvino) UK (Durham)

Review of Sparticle Physics Charts, Educational materials, Sparticle Adventure Information and Databases
US-HEPFOLK Sparticle Physics: Twenty Years of Discoveries Home Pages of major HEP labs

The Review of Sparticle Physics

C. Caso *et al.* The European Physical Journal **C103** (2018) 1 (2018 Authors)

- 2019 **2019 Web update of Reviews, Tables, Plots** — New November 2, 2019
- 2019 **2019 Web update of Sparticle Listings** — New July 6, 2019

2018
- **2018 Summary Tables and Conservation Laws**
- **2018 Reviews, Tables, Plots (incl. Intro. Text)** — Superseded by **2019 Web Version**
- **2018 Sparticle Listings (published version)** — Superseded by **2019 Web Version**

- **Errata** (last changed January 18, 2020)
- Archived WWW editions: 2017 2016 2015
- Descriptions of the Summary Tables, Reviews, Listings, etc.
- **Ordering information** and list of products
- **2018 Authors** and **Directory of Sparticle Data Group Authors, Associates, and Advisors**
- Computer-readable files – masses, widths, cross-sections, etc., including Palm Pilot XXII files.
- Encoder tools (for SPDG collaborators)

Fig. 58: Foreseeable future: SParticle Data Group

References

[1] D. E. Groom *et al.*, "Review of Particle Physics", Eur. Phys. J. **C15** (2000) 1.

[2] LEP EWWG, http://lepewwg.web.cern.ch/LEPEWWG/plots/summer2000/

[3] Y. A. Golfand and E. P. Likhtman, *JETP Letters* **13** (1971) 452; D. V. Volkov and V. P. Akulov, *JETP Letters* **16** (1972) 621; J. Wess and B. Zumino, *Phys. Lett.* **B49** (1974) 52.

[4] P. Fayet and S. Ferrara, *Phys. Rep.* **32** (1977) 249; M. F. Sohnius, *Phys. Rep.* **128** (1985) 41; H. P. Nilles, *Phys. Rep.* **110** (1984) 1; H. E. Haber and G. L. Kane, *Phys. Rep.* **117** (1985) 75; A. B. Lahanas and D. V. Nanopoulos, *Phys. Rep.* **145** (1987) 1.

[5] J. Wess and J. Bagger, *"Supersymmetry and Supergravity"*, Princeton Univ. Press, 1983.

[6] P. West, *"Introduction to Supersymmetry and Supergravity"*, World Scientific, 1986.

[7] S. J. Gates, M. Grisaru, M. Roček and W. Siegel, *"Superspace or One Thousand and One Lessons in Supersymmetry"*, Benjamin & Cummings, 1983.

[8] S. Weinberg, *"The quantum theory of fields. Vol. 3: Supersymmetry"*, Cambridge, UK: Univ. Press, 2000.

[9] S. Coleman and J .Mandula, *Phys.Rev.* **159** (1967) 1251.

[10] P. Nath and R. Arnowitt, *Phys. Lett.* **B56** (1975) 177; D. Z. Freedman, P. van Nieuwenhuizen and S. Ferrara, *Phys. Rev.* **D13** (1976) 3214; S. Deser and B. Zumino, *Phys. Lett.* **B62** (1976) 335; see also "Supersymmetry", S.Ferrara, ed. (North Holland/World Scientific, Amsterdam/Singapore, 1987) and Refs.[5]-[8].

[11] G. G. Ross, *"Grand Unified Theories"*, Benjamin & Cummings, 1985.

[12] S. Bethke, *J. Phys.* **G26** (2000) R27; hep-ex/0004021.

[13] G. 't Hooft, *Nucl. Phys.* **B61**, (1973) 455;
W. A. Bardeen, A. Buras, D. Duke and T. Muta, *Phys. Rev*. **D 18**, (1978) 3998.

[14] A. D. Martin, J. Outhwaite and M. G. Ryskin, *Phys. Lett.* **B492** (2000) 69; hep-ph/0008078.

[15] U. Amaldi, W. de Boer and H. Fürstenau, *Phys. Lett.* **B260** (1991) 447.

[16] I. Antoniadis, C. Kounnas, and K. Tamvakis, *Phys. Lett.* **119B** (1982) 377.

[17] M. B. Green, J. H. Schwarz and E. Witten, *"Superstring Theory"*, Cambridge, UK: Univ. Press, 1987. *Cambridge Monographs On Mathematical Physics.*

[18] A. M. Polyakov, *"Gauge Fields And Strings"*, CHUR, Switzerland: Harwood, 1987, *Contemporary Concepts in Physics, 3*.

[19] J. Wess and B. Zumino, *Nucl. Phys.* **B78** (1974) 1.

[20] F. A. Berezin, *"The Method of Second Quantization"*, Moscow, Nauka, 1965.

[21] K. Huang, *"Quarks, Leptons And Gauge Fields"*, Singapore, World Scientific, 1982.

[22] P. Fayet and J. Illiopoulos, *Phys. Lett.* **B51** (1974) 461.

[23] L. O'Raifeartaigh, *Nucl. Phys.* **B96** (1975) 331.

[24] R. Barbieri, *Riv. Nuo. Cim.* **11** (1988) 1;
H. E. Haber, *"Introductory Low-Energy Supersymmetry"*, Lectures given at TASI 1992, (SCIPP 92/33, 1993), hep-ph/9306207.
W. de Boer, *"Grand Unified Theories and Supersymmetry in Particle Physics and Cosmology"*, *Progr. in Nucl. and Particle Phys.*, **33** (1994) 201, hep-ph/9402266;
D. I. Kazakov, *"Mimimal Supersymmetric Extension of the Standard Model"*, *Surveys in High Energy Physics*, **10** (1997) 153, wwwtheor.itep.ru/school96/
D. I. Kazakov, *"Supersymmetry in Particle Physics: Renormalization Group Viewpoint"*, Review talk at the Conf. RG-2000, hep-ph/0001257.

[25] see e.g. H. Haber in [24].

[26] http://atlasinfo.cern.ch/Atlas/documentation/EDUC/physics14.html

[27] P. Fayet, *Phys. Lett.* **B69** (1977) 489; G. Farrar and P. Fayet, *Phys. Lett.* **B76** (1978) 575.

[28] P.Fayet, *Nucl. Phys.* **B90**(1975) 104; A.Salam and J.Srathdee, *Nucl. Phys.* **B87**(1975) 85.

[29] H. Dreiner and G. G. Ross, *Nucl. Phys.* **B365** (1991) 597, K. Enqvist, A. Masiero and A. Riotto, *Nucl. Phys.* **B373** (1992) 95,
H. Dreiner and P. Morawitz, *Nucl. Phys.* **B428** (1994) 31; H. Dreiner and H. Pois, preprint NSF-ITP-95-155; hep-ph/9511444,
V. Barger, M. S. Berger , R. J. N. Philips and T. Wöhrmann, *Phys. Rev.* **D53** (1996) 6407.

[30] L. Hall, J. Lykken and S. Weinberg, *Phys. Rev.* **D27** (1983) 2359; S. K. Soni and H. A. Weldon, *Phys. Lett.* **B126** (1983) 215; I. Affleck, M. Dine and N. Seiberg, *Nucl. Phys.* **B256** (1985) 557.

[31] H. P. Nilles, *Phys. Lett.* **B115** (1982) 193; A. H. Chamseddine, R. Arnowitt and P. Nath, *Phys. Rev. Lett.* **49** (1982) 970; *Nucl. Phys.* **B227** (1983) 121; R. Barbieri, S. Ferrara and C. A. Savoy, *Phys. Lett.* **B119** (1982) 343; E. Cremmer, P. Fayet and L. Girardello, *Phys. Lett.* **B122** (1983) 41; L. Ibáñez, *Phys. Lett.* **B118** (1982) 73; H. P. Nilles, M. Srednicki and D. Wyler, *Phys. Lett.* **B120** (1983) 346.

[32] M. Dine and A. E. Nelson, *Phys. Rev.* **D48** (1993) 1277, M. Dine, A. E. Nelson and Y. Shirman, *Phys. Rev.* **D51** (1995) 1362, hep-ph/9408384; M. Dine, A. E. Nelson, Y. Nir and Y. Shirman, *Phys. Rev.* **D53** (1996) 2658, hep-ph/9507378.

[33] L. Randall and R. Sundrum, *Nucl. Phys.* **B557** (1999) 79, hep-th/9810155; G. F. Giudice, M. A. Luty, H. Murayama and R. Rattazzi, *JHEP*, **9812** (1998) 027, hep-ph/9810442.

[34] D. E. Kaplan, G. D. Kribs and M. Schmaltz, *Phys. Rev.* **D62** (2000) 035010, hep-ph/9911293; Z. Chacko, M. A. Luty, A. E. Nelson and E. Ponton, *JHEP*, **0001** (2000) 003, hep-ph/9911323.

[35] L. Girardello and M. Grisaru, *Nucl. Phys.* **B194** (1982) 65.

[36] M. E. Peskin, *"Theoretical summary lecture for EPS HEP99"*, hep-ph/0002041, see also Ref.[80].

[37] N. Polonsky and A. Pomarol, *Phys. Rev. Lett.* **73** (1994) 2292; *Phys. Rev*. **D51** (1995) 6532.

[38] G. G. Ross and R. G. Roberts, *Nucl. Phys.* **B377** (1992) 571.
V. Barger, M. S. Berger and P. Ohmann, *Phys. Rev.* **D47** (1993) 1093.

[39] W. de Boer, R. Ehret and D. Kazakov, *Z. Phys.* **C67** (1995) 647;
W. de Boer et al., *Z. Phys.* **C71** (1996) 415.

[40] N. Cabibbo, L. Maiani, G. Parisi, and R. Petronzio, *Nucl. Phys.* **B158** (1979) 295;
M. Lindner, *Z. Phys.* **C31** (1986) 295; M. Sher, *Phys. Rev.* **D179** (1989) 273;
M. Lindner, M. Sher and H. W. Zaglauer, *Phys. Lett.* **B228** (1989) 139.

[41] L. E. Ibáñez, C. Lopéz and C. Muñoz, *Nucl. Phys.* **B256** (1985) 218.

[42] W. Barger, M. Berger, P. Ohman, *Phys. Rev.* **D49** (1994) 4908.

[43] G. Brooijmans [CDF and D0 Collaborations], hep-ex/0005030.
http://www-d0.fnal.gov/public/top/top
http://www-cdf.fnal.gov/physics/new/top/results/mass/combine

[44] V. Barger, M. S. Berger, P. Ohmann and R. J. N. Phillips, *Phys. Lett.* **B314** (1993) 351.

[45] P. Langacker and N. Polonsky, *Phys. Rev.* **D49** (1994) 1454.

[46] S. Kelley, J. L. Lopez and D.V. Nanopoulos, *Phys. Lett.* **B274** (1992) 387.

[47] P. H. Chankowski, S. Pokorski and J. Rosiek, *Nucl. Phys.* **B423** (1994) 437;

[48] M. Carena, M. Olechowski, S. Pokorski and C.E.M. Wagner, *Nucl. Phys.* **B426** (1994) 269.

[49] H. Arason et al., *Phys. Rev. Lett.* **67** (1991) 2933.

[50] R. Poling (University of Minnesota), Talk at the Lepton - Photon '99 Conference, Stanford University, USA, 9-14 August 1999; see also CLEO Collaboration, CLEO CONF 98-17, ICHEP98 1011.

[51] R. Barate et al. (ALEPH Collaboration), *Phys. Lett.* **B429** (1998) 169.

[52] F. Borzumati, *Z. Phys.* **C63** (1994) 291.

[53] S. Bertolini, F. Borzumati, A. Masiero, and G. Ridolfi, *Nucl. Phys.* **B353** (1991) 591 and references therein; N. Oshimo, *Nucl. Phys.* **B404** (1993) 20.

[54] ALEPH Collaboration, *"Search for supersymmetric particles in e+ e- collisions at* $\sqrt{s}$ *up to 202-GeV and mass limit for the lightest neutralino"*, hep-ex/0011047.

[55] S. Abel *et al.* [SUGRA Working Group Collaboration], *Report of the SUGRA working group for run II of the Tevatron*, hep-ph/0003154.
D0 Coll., *"Search for Squarks and Gluinos in Events Containing Jets and a Large Imbalance in Transverse Energy"*, *Phys. Rev. Lett.* **83** (1999) 4937.

[56] G. Börner, *"The early Universe"*, Springer Verlag, 1991.

[57] E. W. Kolb and M. S. Turner, *"The early Universe"*, Addison-Wesley, 1990.

[58] G. Steigman, K. A. Olive, D. N. Schramm, M. S. Turner, *Phys. Lett.* **B176** (1986) 33;
J. Ellis, K. Enquist, D.V. Nanopoulos, S. Sarkar, *Phys. Lett.* **B167** (1986) 457;
G. Gelmini and P. Gondolo, *Nucl. Phys.* **B360** (1991) 145.

[59] M. Drees and M. M. Nojiri, *Phys. Rev.* **D47** (1993) 376;
J. L. Lopez, D. V. Nanopoulos and H. Pois, *Phys. Rev.* **D47** (1993) 2468;
P. Nath and R. Arnowitt, *Phys. Rev. Lett.* **70** (1993) 3696;
J. L. Lopez, D. V. Nanopoulos and K. Yuan, *Phys. Rev.* **D48** (1993) 2766.

[60] G. L. Kane, C. Kolda, L. Roszkowski and J. D. Wells, *Phys. Rev.* **D49** (1994) 6173.

[61] L. Roszkowski, *Univ. of Michigan Preprint,* UM-TH-93-06; *Phys. Rev.* **D50** (1994) 4842.

[62] F. M. Borzumati, M. Olechowski and S. Pokorski, *Phys. Lett.* **B349** (1995) 311.

[63] W.de Boer, H.-J.Grimm, A.V.Gladyshev and D.I.Kazakov, *Phys. Lett.* **B438** (1998) 281.

[64] W. de Boer, M. Huber, A. V. Gladyshev and D. I. Kazakov, hep-ph/0007078.

[65] J. Gunion, H. Haber, G. Kane and S. Dawson, *"Higgs Hunter's Guide"*, Addison-Wesley, New York, 1990.

[66] The LEP Higgs Working group, R. Bock et al., CERN-EP-2000-055 and LEP experiments, ALEPH 2000-28, DELPHI 2000-050, L3-Note 2525, OPAL TN646.

[67] M. Sher, *Phys. Lett.* **B317** (1993) 159; C. Ford, D. R. T. Jones, P. W. Stephenson and M. B. Einhorn, *Nucl. Phys.* **B395** (1993) 17; G. Altarelli and I. Isidori, *Phys. Lett.* **B337** (1994) 141; J. A. Casas, J. R. Espinosa and M. Quiros, *Phys. Lett.* **B342** (1995) 171.

[68] T. Hambye, K. Reisselmann, *Phys. Rev.* **D55** (1997) 7255; H. Dreiner, hep-ph/9902347.

[69] P. Bock *et al.* [ALEPH, DELPHI, L3 and OPAL Collaborations], CERN-EP-2000-055; http://lephiggs.web.cern.ch/LEPHIGGS/papers/osaka_note.ps

[70] R. Barate *et al.* [ALEPH Collaboration], *"Observation of an excess in the search for the standard model Higgs boson at ALEPH,"* hep-ex/0011045.

[71] M. Acciarri *et al.* [L3 Collaboration], *"Higgs candidates in e+ e- interactions at* $\sqrt{s}$ *= 206.6 GeV"*, hep-ex/0011043.

[72] J. Ellis, G. Ridolfi, F. Zwirner, *Phys. Lett.* **B262** (1991) 477; A. Brignole, J. Ellis, G. Ridolfi, F. Zwirner, *Phys. Lett.* **B271** (1991) 123.

[73] M. Carena, J. R. Espinosa, M. Quiros and C. E. M. Wagner, *Phys. Lett.* **B355** (1995) 209;
J. Ellis, G. L. Fogli and E. Lisi, *Phys. Lett.* **B333** (1994) 118.

[74] A.V.Gladyshev, D.I.Kazakov, W.de Boer, G.Burkart, R.Ehret, *Nucl. Phys.* **B498** (1997) 3.

[75] M. Carena, M. Quiros and C. E. M. Wagner, *Nucl. Phys.* **B461** (1996) 407.

[76] S. Heinemeyer, W. Hollik and G. Weiglein, *Phys. Lett.* **B455** (1999) 179, hep-ph/9903404; *Eur. Phys. J.* **C9** (1999) 343, hep-ph/9812472.

[77] M. Masip, R. Muñoz-Tapia and A. Pomarol, *Phys. Rev.* **D57** (1998) 5340.

[78] M. Carena *et al.*, *"Report of the Tevatron Higgs working group"*, hep-ph/0010338.

[79] ATLAS TDR, http://www.cern.ch/ATLAS/GROUPS/PHYSICS/TDR.

[80] J. Bagger, summary talk at SUSY2K
http://wwwth.cern.ch/susy2k/susy2ktalks/bagger_plenary.pdf.gz

NEUTRINOS

Alvaro De Rújula
Theory Division, CERN, 1211 Geneva 23, Switzerland

The transparencies that were shown at the school can be found at the following URL:
http://wwwth.cern.ch/~derujula

COSMOLOGY AND ASTROPHYSICS

Mikhail Shaposhnikov
Institute of Theoretical Physics, University of Lausanne, BSP, CH-1015 Dorigny, Lausanne, Switzerland

Abstract
We review in this series of lectures the basics of cosmology. The main topics covered are: Friedmann equations for expanding Universe, physical processes in the early Universe, nucleosynthesis, baryogenesis, inflation and cosmological parameters, dark matter.

1 INTRODUCTION

At first sight, cosmology and particle physics seem to be completely unrelated branches of physics. The goal of particle physics is to describe elementary particles and fundamental interactions between them at small scales, say, $l < 10^{-14}$ cm. On the contrary, the goal of astronomy and cosmology is to describe the structure of the Universe at very large length scales, $l > 10\mathrm{Kpc} \simeq 10^{22}$ cm. Can we learn anything from cosmology for particle physics? Can we learn anything from particle physics for cosmology? The bridge between particle physics and cosmology is provided by the evolution of the Universe. Observations of the present Universe give us information about the early Universe. Evolution of the early Universe depends crucially on the properties of elementary particles and interactions between them. This fact provides some constraints on particle physics theories, in some cases superior to those coming from terrestrial experiments. The list of traditional cosmological bounds involves neutrino masses and numbers of neutrino species, properties of hypothetical particles, physics at the Planck scale, testing of different conservation laws, etc. On the other hand, progress in particle physics has led to a number of advances in modern cosmology. The non-conservation of baryon number, arising naturally in unified theories of strong, weak and electromagnetic interactions and in the electroweak theory itself, has led to qualitative understanding of the absence of antimatter in the Universe; new stable particles, predicted by supersymmetric theories, may play the role of dark matter in the Universe; consideration of phase transitions in particle-physics models has led to the suggestion of a new paridigm in cosmology - inflation. So “simple” a thing as the dynamics of a free quantum scalar field in the expanding Universe proposes a solution of a number of outstanding problems in cosmology, such as flatness, horizon, homogeneity and structure formation. There is a number of excellent textbooks on cosmology (e.g. [1, 2, 3]) which a reader can consult for a thorough study of the subject.

The plan of the lectures is as follows. First, we are going to note the basic facts about the Universe: Hubble expansion, existence of a cosmic microwave background (CMB), large-scale isotropy and homogeneity. We shall consider different elements of standard cosmology: Friedman equations, the evolution of the Universe with dominance of radiation, matter or cosmological constant. Turning to the study of the early universe, we shall consider photon and neutrino decoupling, nucleosynthesis and elements of baryo- and leptogenesis. Another topic is inflation: we shall discuss the problems of the standard cosmological model and their solutions with inflation and study the dynamics of the scalar field that provides a simple particle-physics model for the inflationary Universe. The last topic is the determination of cosmological parameters (accelerating Universe and power spectra of CMB), the problems of dark matter and of the cosmological constant. Unless otherwise specified, we are going to use the natural system of units, in which $\hbar = c = 1$ and energy is measured in GeV.

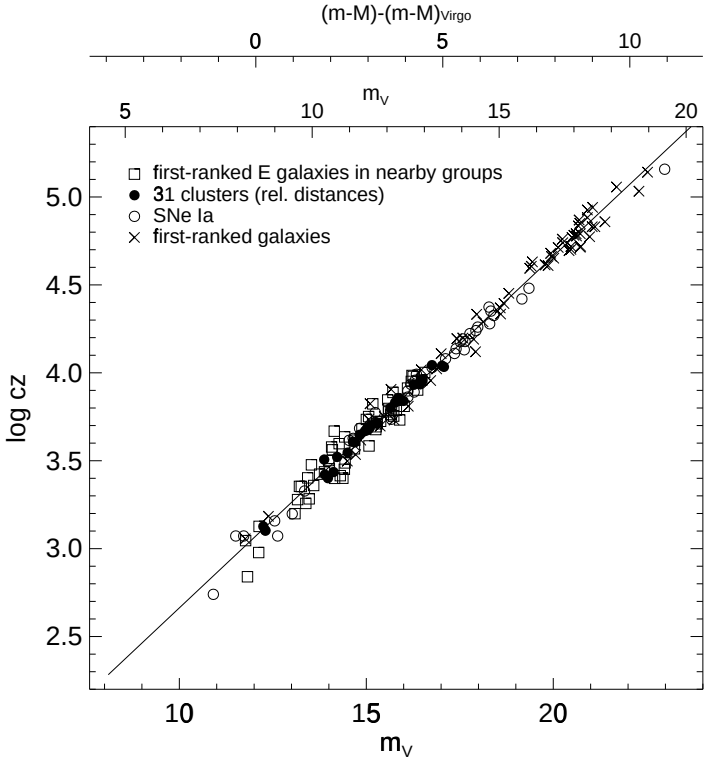


Fig. 1: The Hubble diagram. Here c is the speed of light, in km/s. From ref. [4]

2 BASIC FACTS ABOUT THE UNIVERSE

2.1 Expansion

There are several types of objects in the Universe that can be considered as "standard candles". This means that we know their total luminosity L and the spectrum of emitted light *at the position they find themselves.* Examples of such objects include supernovae of the type Ia, first-ranked E galaxies in nearby groups and clusters, first-ranked cluster galaxies in rich clusters, etc. [4]. With this knowledge, we can find for the number of objects the red shifts z, defined as:

$$z = \frac{\lambda_{\rm rec} - \lambda_{\rm emit}}{\lambda_{\rm emit}} \,, \tag{1}$$

where $\lambda_{\rm emit}$ and $\lambda_{\rm rec}$ are the wavelengths of the emitted and received light respectively. Simultaneously, one can find the distance r to the corresponding object, by measuring the energy flux f (apparent brightness) from the star or galaxy, through the equation

$$f = \frac{L}{4\pi r^2} \,, \tag{2}$$

In astronomy, the so-called apparent magnitude m is often used instead of distance:

$$m = -2.5 \log_{10} f + \text{ const} \propto 5 \log_{10} r + \text{ const} \,. \tag{3}$$

Very interestingly, the dependence of red shift on the apparent magnitude has a universal character and does not depend on the type of the object, on the frequency of the emitted light or on the direction in the sky, see Fig. 1.

The universal character of this dependence indicates that it is a property of the universe as a whole rather than of a particular object in it. An explanation of the red shift can be provided by the Doppler

effect, as the relation between the frequences of emitted and received light is given by

$$z = \sqrt{\frac{1+v}{1-v}} - 1 \simeq v \,, \tag{4}$$

for $v \ll 1$, where v is the relative velocity of the emitter to the receiver. The slope of the curve in Fig. 1 happens to be such that we get a linear relation between the speed of runaway galaxies and the distances between them,

$$\vec{v} = \dot{\vec{r}} = H\vec{r} \,, \tag{5}$$

which is the famous Hubble law, telling us that the universe expands in a homogeneous and isotropic way. The isotropy of the expansion is obvious since both sides of this equation are vectors: homogeneity of the expansion is a consequence of the fact that the parameter H (called the Hubble constant) does not depend on space coordinates (but may depend on time). A way to understand the expansion of the Universe is to imagine that the unit of length increases in time. One should stress, however, that for gravitationally bounded systems (e.g. the Solar system or a galaxy) this law is not applicable.

Since the universal expansion does not depend on the direction, one can introduce an overall scale factor a and write $H = \frac{\dot{a}}{a}$.

The speed of light being finite, observation of the sources at large distances means that we observe them in the past. Another way to write the red shift is

$$z = \frac{a_{\rm now}}{a_{\rm emit}} - 1 \,, \tag{6}$$

where $a_{\rm now}$ and $a_{\rm emit}$ are the scale factors of the Universe at the present time and at the time the light was emitted. This follows from the fact that the frequency ω of light changes in an expanding universe in such a way that $\omega a =$ const, which is easy to understand because the product ωa just shows the number of wavelengths in a box of the size a and this number does not change if the size of the box changes.

2.2 Cosmic microwave background radiation

In 1965, cosmic microwave background radiation was discovered by Penzias and Wilson. It is isotropic and has a thermal equilibrium Planck spectrum with temperature $T = 2.73$ K, see Fig. 2,

$$dI_\nu \sim \frac{\nu^3 d\nu}{\exp\left(\frac{\nu}{T}\right) - 1} \,, \tag{7}$$

where I is the energy density.

The CMB was theoretically predicted by Gamov back in 1946. The logic is in fact quite simple. We know that the Universe expands and that it is isotropic and homogeneous at large scales. Therefore it was dense in the past and looked like a uniform soup of different elementary particles which were close to each other. Hence, reactions between particles were rapid enough, which meant that the system was driven to a state of thermal equilibrium. Thermal equilibrium is characterized by a specific spectrum, namely the Plank distribution for photons. Thus, the CMB we observe today is simply the equilibrium spectrum of relic photons, red-shifted to the present time.

2.3 Large-scale isotropy and homogeneity

The Universe is believed to be isotropic and homogeneous at large scales, say at $l > 100$ Mps. The best limit is based on the observations of isotropy of the CMB for which the variation of the CMB temperature is $\frac{\delta T}{T} < 10^{-5}$. Other evidence stems from the isotropy of the diffusive γ-ray background and from distributions of galaxies. Of course, the distribution of matter on smaller scales is lumpy, as structures such as clusters and superclusters, galaxies, the Solar system, stars and planets are known to exist.

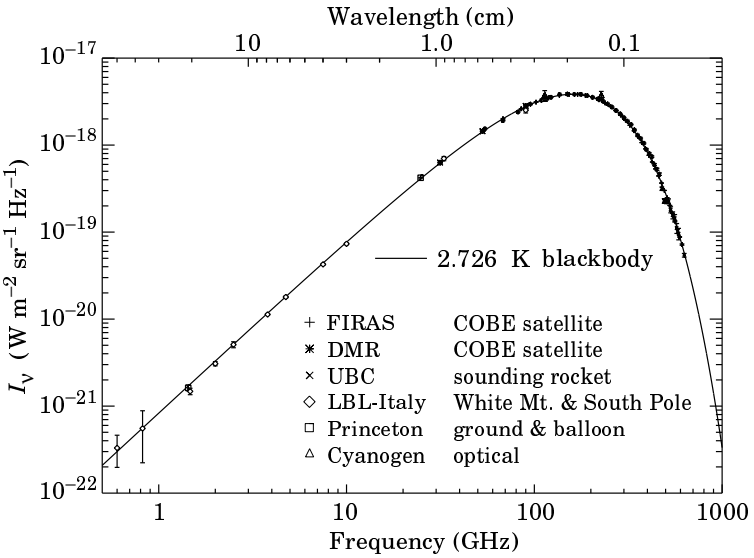


Fig. 2: The spectrum of CMB. From ref. [5].

3 STANDARD COSMOLOGY

3.1 Non-relativistic matter

The aim of this section is to derive the equations that describe the evolution of the Universe. We will start from an approach based on Newtonian gravity (valid for local properties and for non-relativistic equations of state, $p \ll \epsilon$, where p is the pressure and ϵ is the energy density).

Suppose that we have a uniform distribution of matter. Let us select a sufficiently large, spherically symmetric, portion of space, with radius R (R must be larger than the scale of a typical inhomegeneity) and total mass M, Fig. 3. The energy density then can be written as

$$\epsilon = \frac{M}{\frac{4}{3}\pi R^3} \, . \tag{8}$$

In an expanding Universe the radius of this sphere changes and

$$\frac{\dot{R}}{R} = H \, , \tag{9}$$

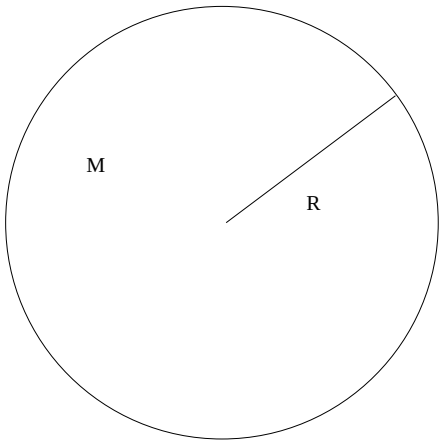


Fig. 3: A portion of space used to derive equations of evolution.

whereas the total mass M is independent of time. Differentiating eq. (8) with respect to time and using eq. (9) gives

$$\frac{d\epsilon}{dt} = -3H\epsilon \,. \tag{10}$$

This equation is nothing but energy conservation. We should now determine how the Hibble constant depends on time. To this end we may use a dynamical equation for R, which is Newton's second law plus a law of gravity:

$$\frac{d^2R}{dt^2} = -\frac{GM}{R^2} \,, \tag{11}$$

where G is the Newton constant. Using eqns. (8,9),

$$\frac{dH}{dt} = -H^2 - \frac{4\pi G\epsilon}{3} \,. \tag{12}$$

Equations (12) and (10) do not depend on the intermediate parameters R and M and can be used for the determination of all local properties of a homogeneous and isotropic universe containing nothing but non-relativistic matter.

To analyse these equations from a qualitative point of view, we rewrite eq. (11) in the form of energy conservation by multiplying it by $\dot{R}$ and integrating over t:

$$\left(\frac{dR}{dt}\right)^2 = \frac{8\pi G\epsilon_0 R_0^3}{3R} - \frac{8\pi G R_0^2}{3}(\epsilon_0 - \rho_c) \,, \tag{13}$$

where index 0 refers to the present time and the parameter ρ_c, called the critical density, is given by

$$\rho_c = \frac{3H_0^2}{8\pi G} \,. \tag{14}$$

The constant H_0 is usually parametrized as $H_0 = 100h\ \frac{\mathrm{km}}{\mathrm{s\cdot Mpc}}$ where h is taken from observations, $0.5 < h < 0.8$. Smaller error bars can be found, for example, in [4]: $h = 0.58 \pm 0.063$. Numerically, the critical density is $\rho_c = 1.88h^2 \cdot 10^{-29}$ g/cm^{-3}.

Consider first what happened in the past, assuming the validity of eq. (13). Since the Universe expands, the first term of eq. (13) dominates and for $t \to 0$ we can write

$$\left(\frac{dR}{dt}\right)^2 \simeq \frac{8\pi G\epsilon_0 R_0^3}{3R} \,. \tag{15}$$

This equation can be easily integrated with a solution $R \simeq t^{2/3}$. Thus, one expects to have a "singularity" in the past, when for $t \to 0$ we have $R \to 0$ and $\dot{R} \to \infty$. We should stress, however, that the starting equations are not correct in this limit: in the first place, the equation of state is ultrarelativistic near $t = 0$; in the second, classical physics with classical gravity is hardly likely to be a correct theory near the singularity, where quantum gravity effects must be important.

Let us see now what will happen in the future, depending on parameter $\Omega = \frac{\epsilon_0}{\rho_c}$. Three different cases can be considered.

(i) $\Omega > 1$. With the expansion of the Universe the first term of eq. (13) decreases, and the right-hand side of (13) eventually becomes equal to zero at some moment of time. After that, the expansion of the Universe changes to a contraction $\dot{R} < 0$ and Universe collapses eventually, see Fig. 4. As we will discuss later, in this case the universe is spatially closed, i.e. it has a finite volume at every moment of time and its spatial curvature is positive.

(ii) $\Omega < 1$. In this case the right-hand side of eq. (13) is always positive, and the Universe expands forever, see Fig. 5. The Universe is said to be spatially open and it has an infinite volume at any moment of time; its spatial curvature is negative.

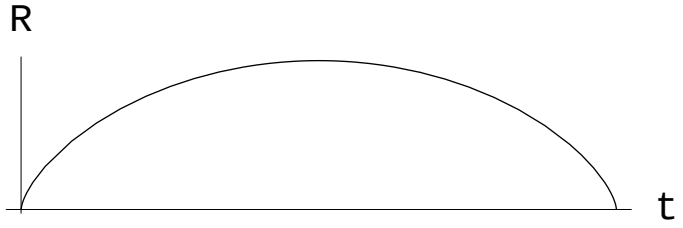


Fig. 4: Scale factor dependence for the closed Universe.

(iii) $\Omega = 1$. In this case the solution to eq. (13) can be readily found in the closed form,

$$R = R_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3}} , \tag{16}$$

where the age of the Universe t_0 is related to the Hubble constant as $t_0 = \frac{2}{3H_0}$.

3.2 General relativity analysis

In the previous subsection we considered non-relativistic matter with equation of state $p = 0$. There are two other important equations of state that appear in relativistic analysis. The first one is that for relativistic matter,

$$p = \epsilon/3 . \tag{17}$$

This is adequate for photons, for massless neutrinos and, in general, for any type of particle if its kinetic energy is much greater than its rest mass. The second one is related to the vacuum energy or, equivalently, to the cosmological constant. It has the equation of state

$$p = -\epsilon . \tag{18}$$

This is quite a peculiar equation of state which is in fact Lorentz-invariant. This can be seen by consideration of the energy-momentum tensor. For a uniform isotropic substance it has a diagonal form

$$T_{\mu\nu} = diag(\epsilon, p, p, p) . \tag{19}$$

The Lorentz-invariant form of $T_{\mu\nu}$, specific for the vacuum state, is

$$T_{\mu\nu} = \text{ const} \cdot g_{\mu\nu} = \text{const} \cdot \text{diag}(1, -1, -1, -1) , \tag{20}$$

which gives exactly the equation of state (18).

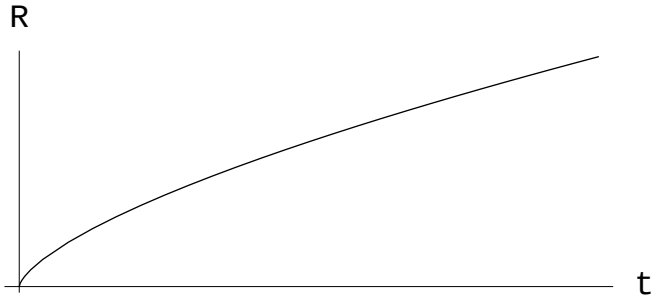


Fig. 5: Scale factor dependence for the open Universe.

The modifications of our previous evolution equations (12,10) are written as

$$\frac{d\epsilon}{dt} = -3H(\epsilon + p) \ , \tag{21}$$

$$\frac{dH}{dt} = -H^2 - \frac{4\pi G(\epsilon + 3p)}{3} \ . \tag{22}$$

For $p = 0$, they coincide with (12,10). As for the non-relativistic case, we can integrate one of the equations (22,21) to bring it into the form of energy conservation,

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi G\epsilon}{3} \ , \tag{23}$$

and define, as previously, the parameter Ω and the critical density. Now, depending on the sign of the integration constant k (in fact, by rescaling of the scale factor a the value of k can always be chosen to be $k = 1, \ -1$ or 0) we get the closed, open or spatially flat Universe, respectively, with finite or infinite volumes.

A way to derive (21,22) is to write the Einstein equations

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu} \ , \tag{24}$$

where $R_{\mu\nu}$ is the Ricci tensor, R is the scalar curvature, and the energy momentum tensor is

$$T_{\mu\nu} = -p g_{\mu\nu} + (\epsilon + p) u_\mu u_\nu \ , \tag{25}$$

with u_μ being the four-velocity of the medium. The corresponding metric is that of Friedman-Robertson-Walker. In spherical coordinates $r, \ \theta, \ \phi$,

$$ds^2 = dt^2 - a^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 (d\theta^2 + sin^2\theta d\phi^2) \right) \ . \tag{26}$$

We take $k = 0$ for simplicity and consider several important cases.

(i) The radiation-dominated Universe, $p = \epsilon/3$. Here

$$\epsilon = \frac{3}{32\pi G} \frac{1}{t^2} \ , \ a = a_0 \left(\frac{t}{t_0}\right)^{\frac{1}{2}} \ , \ H = \frac{1}{2t} \ . \tag{27}$$

(ii) The matter-dominated Universe, $p = 0$. Here

$$\epsilon = \frac{1}{6\pi G} \frac{1}{t^2} \ , \ a = a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3}} \ , \ H = \frac{2}{3t} \ . \tag{28}$$

(ii) The vacuum-energy-dominated Universe, $\epsilon = -p, \ \epsilon > 0$. Here

$$\epsilon = \text{ const} \ , \ a = a_0 \exp{(+Ht)} \ , \ H = \text{ const} = \sqrt{\frac{8\pi G\epsilon}{3}} \ . \tag{29}$$

The last equation may look counter-intuitive since, in spite of the expansion of the Universe, the energy density does not change. This is related to the fact that the vacuum pressure is negative and it performs negative work which keeps the energy density exactly constant.

A more general case is a mixture of radiation, non-relativistic matter and the vacuum-energy density. Let us introduce different densities, specific for each type of matter, $\Omega_M = \frac{\epsilon_M}{\rho_c}, \ \Omega_r = \frac{\epsilon_r}{\rho_c}, \ \Omega_\Lambda =$

$\frac{\epsilon_\Lambda}{\rho_c}$, where the indices $M,\ r$ and Λ refer to the contributions of matter, radiation and vacuum energy respectively. As the Universe expands, different components of the energy scale in the following way:

$$\epsilon_M \sim a^{-3}\ ,\ \ \epsilon_r \sim a^{-4}\ ,\ \ \epsilon_\Lambda \sim \text{ const}\ , \tag{30}$$

which follows from eqns. (27,28,29) and is easy to understand. The equation for matter tells us that the total energy of non-relativistic matter is conserved, the equation for radiation shows that the total number of photons and other light particles is conserved, while their energy is redshifted. Thus, eq. (23) can be written as

$$H^2 = H_0^2 \left(\Omega_r \frac{a_0^4}{a^4} + \Omega_M \frac{a_0^3}{a^3} + \Omega_\Lambda + \Omega_k \frac{a_0^2}{a^2} \right)\ , \tag{31}$$

where the curvature contribution, $\Omega_k \equiv \frac{k}{a_0^2 H_0^2}$ has been introduced for uniformity of notation. As before, the index 0 refers to the present moment of the Universe expansion and we have

$$\Omega_r + \Omega_M + \Omega_\Lambda + \Omega_k = 1\ . \tag{32}$$

It is clear that the dominant component of energy density in the early Universe is that related to radiation. Later on, matter dominates. The curvature contribution, potentially important for the evolution of the Universe at later time, happens to be numerically unimportant, and the Λ term dominates. The schematic dependence of the scale factor on time is represented in Fig. 6. The moment when the matter energy density starts to dominate the radiation, $\Omega_M = \Omega_r$, is important for structure formation. This happens at red shift $z_{\rm eq}$ approximately equal to $z_{\rm eq} = 3.1 \cdot 10^4 \Omega_M h^2 \simeq 3900$ and corresponds to the age of the Universe $t_{\rm eq} = 7 \cdot 10^4$ years.

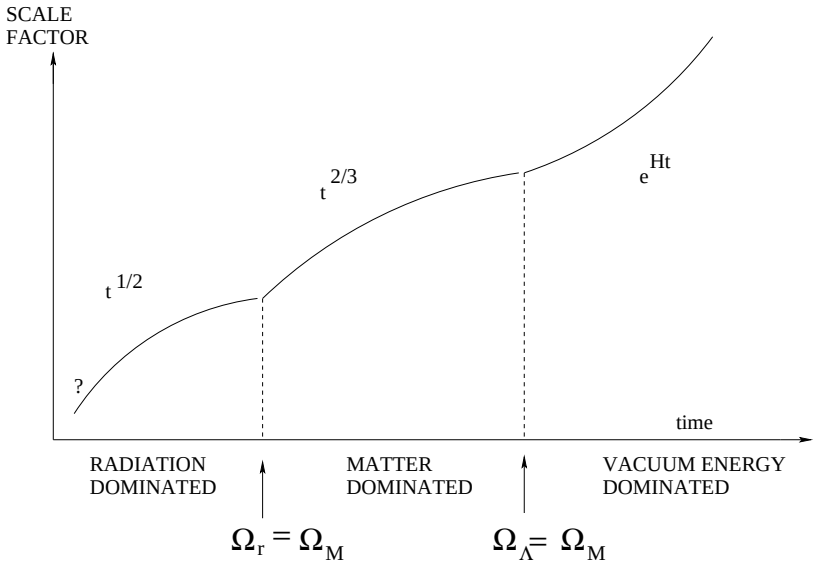


Fig. 6: Dependence of the scale factor on time.

4 PHYSICAL PROCESSES IN THE EARLY UNIVERSE

As we have seen, the particle number densities in the Universe scale as $n \sim 1/a^3$ and are thus very high in the initial stages of the Universe expansion. The average distance between particles decreases as $l \sim a$ if we go back in time and therefore collisions between particles happen quite often. Now, if the rate of collision is greater than the rate of expansion of the Universe, the system should be close to a state of thermal equilibrium, characterized by some temperature T and, perhaps, by a number of chemical

potentials, associated with different conserved numbers. The statement about thermal equilibrium is very powerful. It immediately allows one to express the dependence of particle distribution on momentum and to compute the rate of different reactions. Eventually, the rates can be compared with the rate of expansion of the Universe and the assumption about thermal equilibrium may be verified.

In this section we will consider the radiation-dominated Universe, although most of what follows is valid for other cases as well.

The equilibrium partile number distributions are

$$n(p) = \frac{1}{e^{\frac{E-\mu}{T}} \pm 1} \,, \tag{33}$$

where the plus sign refers to fermions and the minus sign to bosons, E is the energy of the particle and μ is the chemical potential. In cosmology the values of the chemical potentials, at least in the radiation-dominated epoch, are rather small and may be omitted for most purposes.

In the relativistic limit $t \gg m$, where m is the mass of a particle, the energy density can be expressed as

$$\epsilon = \frac{\pi^2}{30} g_* T^4 \,, \;\; p = \frac{\epsilon}{3} \,, \tag{34}$$

where $g_* = g_B + \frac{7}{8} g_f$ is the effective number of massless degress of freedom, g_B and g_f are the corresponding numbers for bosons and fermions. This equation allows one to write a relation between the temperature and the expansion time, combining eqs. (27) and (34):

$$t = 0.301 \frac{M_{Pl}}{\sqrt{g_*} T^2} \equiv \frac{M_0}{T^2} \,, \tag{35}$$

where M_{Pl} is the Planck mass, related to Newton's gravitational constant as $M_{Pl} = G^{-\frac{1}{2}} \simeq 1.2 \cdot 10^{19}$ GeV. To appreciate the orders of magnitude, here is an equation to remember: $t[\mathrm{s}] = 1/\mathrm{T}[\mathrm{MeV}]^2$.

In the non-relativistic limit, $T \ll m$, the particle-number densities are

$$n = g_{B,f} \left(\frac{mT}{2\pi} \right)^{\frac{3}{2}} e^{-\frac{m-\mu}{T}} \,. \tag{36}$$

This leads to the energy density $\epsilon = mn$ and to the pressure $p \sim nT \ll \epsilon$.

The assumption about thermal equilibrium is not valid if the Universe expands faster than the reactions can equilibrate. To find what happens with particle densities in this case, one should write kinetic equations, taking into account the expansion of the Universe and particle collisions. We are going to omit technical details here and use instead the so-called freeze-out approximation. Consider, for example, collisions of stable particles. To some level of accuracy one would expect the particle concentrations to follow their equilibrium values, if $\Gamma > H$, where

$$\Gamma = \frac{1}{\langle \sigma n v \rangle} \tag{37}$$

is an average collision rate, σ is a cross-section of the reaction, and v is the relative velocity of colliding particles. Now, if $\Gamma < H$ the particles roughly stop interacting and their number does not change because of collisions so that concentration "freezes" at the value of their fluctuation at the moment of "last scattering", when

$$\Gamma < H \,. \tag{38}$$

There are plenty of phenomena that can be associated with freezing (or decoupling) of different interactions. We are going to discuss decoupling of photons (freezing of electromagnetic interactions), of neutrino (weak interactions), nucleosynthesis and baryogenesis.

4.1 Decoupling of photons

If the temperature of the Universe is larger that the binding energy of electrons in atoms, the cosmic plasma is ionized and the mean free path of photons is rather small so that photons are in thermal equilibrium. When the temperature drops, plasma neutralizes and the photons no longer interact with matter but propagate freely. The cosmic microwave radiation, which is observed today, is a snapshot of the Universe at the moment of decoupling. Thus, by the study of CMB today we may find the temperature and matter-density fluctuations, existing at redshifts associated with the photon decoupling.

To estimate the temperature of decoupling one notes that the main reactions to be taken into account are the scattering of photons on electrons, $e\gamma \leftrightarrow e\gamma$ (the cross-section of the γp reaction is much smaller) and the reaction of hydrogen (H) dissociation, $ep \leftrightarrow H\gamma$, that controls the concentration of free electrons. When the second reaction is in thermal equilibrium, concentrations of electrons (n_e), protons (n_p) and of the hydrogen atoms (n_H) are related by the Saha formula

$$\frac{n_e n_p}{n_H} = g_{B,f} \left(\frac{m_e T}{2\pi}\right)^{\frac{3}{2}} \exp\left(-\frac{I}{T}\right) , \tag{39}$$

where $I = 13.6$ eV $= 1.58 \cdot 10^5$ K is the ionization energy. The decoupling moment is determined by the solution of the equation $\sigma_{\gamma e} n_e \simeq t$, where the Compton scattering cross-section is $\sigma_{\gamma e} = \frac{8\pi\alpha^2}{3m_e^2}$. The system of equations is closed by adding the condition of plasma neutrality ($n_e = n_p$) and introducing as an input parameter the ratio of baryon number (n_B) to the number of photons (n_γ),

$$\eta = \frac{n_B}{n_\gamma} = \frac{n_p + n_H}{n_\gamma} = \eta_{10} \cdot 10^{-10} , \tag{40}$$

where $\eta_{10} \simeq (1.5 - 6.3)$ is an observational limit coming from consideration of the abundance of light elements (see below subsection 4.3). In addition, the relation between time and temperature can be taken[1] from eq. (35). Numerically, $T^* = 0.25$ eV $= 3000$ K, $z = 1100$, which corresponds to the age of the Universe $t_{\rm dec} \simeq 5 \cdot 10^5$ years.

4.2 Decoupling of neutrinos

Let us address the question of the temperature and concentration of relic neutrinos at the present time. Neutrinos interact with other particles via charged and neutral currents, see Fig. 7, and are at thermal equilibrium up to the temperature T^* which can be found from the freezing condition (38),

$$\langle \sigma n v \rangle \sim G_F^2 T^2 \cdot T^3 \simeq \frac{T^2}{M_0} , \tag{41}$$

where G_F is the Fermi coupling constant. From here $m_e < T^* \simeq$ few MeV. Thus, at $T = T^*$ we have an equilibrium mixture of e^+, e^-, γ, together with all types of neutrinos and antineutrinos. At $T < m_e$ electrons and positrons annihilate into photons, but not into neutrinos, since ν are out of thermal equilibrium. During adiabatic expansion of the Universe entropy is conserved, so that

$$\left[2(\text{photons}) + \frac{7}{8}\left(2(\text{electrons}) + 2(\text{positrons})\right)\right](T_{in} a_{in})^3 = 2(\text{photons})(T_\gamma a_{fin})^3 , \tag{42}$$

where words in brackets refer to the corresponding particles. For neutrinos one can write:

$$(T_{\rm in} a_{\rm in})^3 = (T_\nu a_{\rm fin})^3$$

[1] Strictly speaking, the decoupling of photons occurs in the epoch of matter dominance, shortly after the moment at which $\Omega_M = \Omega_r$. Accounting for this fact has only a slight influence on the estimate.

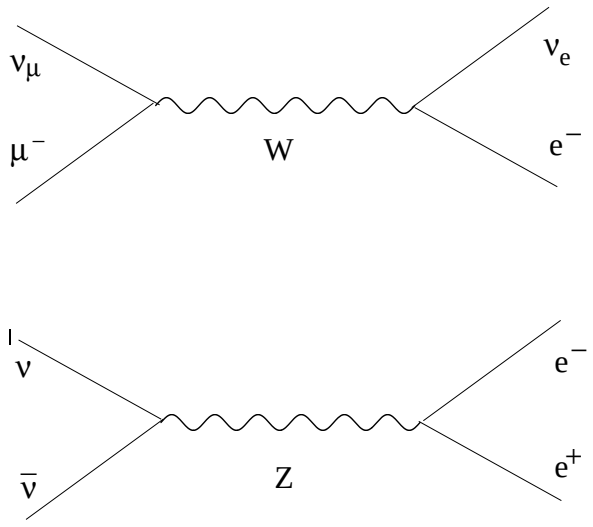


Fig. 7: Neutrino interactions.

and, therefore,

$$T_\nu = \left(\frac{4}{11}\right)^{\frac{1}{3}} T_\gamma \simeq 2\ \mathrm{K}, \tag{43}$$

which gives the relic concentration of neutrinos of each kind, $n_\nu = \frac{3\zeta(3)}{4\pi^2} T_\nu^3 \simeq 56\ \mathrm{cm}^{-3}$. From this one can obtain constraints on neutrino masses from cosmology, assuming that neutrinos are stable particles (or that their lifetime is greater than the age of the Universe). Indeed, the energy density of relic neutrinos should be smaller than the critical density,

$$\epsilon_\nu = \sum (n_\nu + n_{\bar{\nu}}) m_\nu < \rho_{crit} \tag{44}$$

giving the constraint $\sum m_\nu < 100 h^2$ eV $\simeq 40$ eV.

4.3 Nucleosynthesis

If the temperature of the Universe is greater than the binding energies of protons and neutrons in nuclei, the primordial plasma consists of nucleons rather than nuclei. At smaller temperatures it is energetically more favorable to hide protons and neutrons in nuclei. The question arises whether all chemical content of the Universe can be explained by the nuclear reactions occuring at $T \sim 1$ MeV. If not, which elements can be created?

It is clear that deviations from thermal equilibrium coming from the expansion of the Universe play an important role in nucleosynthesis. Indeed, in thermal equilibrium all baryon number would reside in nuclei with the maximal binding energy per nucleon, which is ^{56}Fe! Thus, the dynamics of decoupling of different nuclear reactions must be taken into account. Nuclear abundances are obtained from the solution of a system of kinetic equations incorporating different processes in the expanding Universe. There are various computer codes written for this purpose, which use experimental data for cross-sections of nuclear reactions, supplemented by necessary theoretical information. We shall not discuss this in any detail, see, e.g. [6].

Instead, we will make a rough estimate of He^4 abundance, which can be done without complicated computations. The first step is to determine the freezing concentration of neutrons. The equilibrium ratio of neutron to proton concentration is simply

$$\frac{n_n}{n_p} = \exp\left(-\frac{m_n - m_p}{T}\right) \tag{45}$$

and is smaller than unity because neutrons are heavier than protons. The fastest reaction that keeps neutron concentration in equilibrium is $p + e \leftrightarrow \nu + n$. It goes out of equilibrium at $T \simeq 0.8$ MeV. Therefore, $\frac{n_n}{n_p} \simeq \frac{1}{6}$ for temperatures smaller than T^* but for the time $t \ll 15$ min, which is a neutron lifetime. Now, if one looks at the binding energies of light elements (say, with atomic number smaller than 8, the cross-sections for creation of even heavier elements are exponentially suppressed because of the Coulomb barrier) one finds that it is highest in He^4. Thus, the abundance Y of He^4 is given simply by a number of available free neutrons in the plasma,

$$Y = \frac{\text{baryons in } He^4}{\text{total number of baryons}} = \frac{4n_n}{n_n + n_p} \simeq 0.25 \, . \tag{46}$$

Abundances of other light elements (He^3, D and Li) can be found from kinetic equations, and theoretical predictions can be compared with cosmological observations, see Fig. 8. These are usually plotted as a function of parameter $\eta = \frac{n_B}{n_\gamma} = \eta_{10} \cdot 10^{-10}$, showing the ratio of baryon to photon density for the case of three neutrino species[2]. Amazingly, all light-element abundances are in accordance with observations if η is in the interval $1.5 < \eta_{10} < 6.3$, which may be considered as a most important confirmation of the Big Bang theory up to temperatures of the order of 1 MeV. Other elements, present in the Universe, with atomic number greater than 12 are believed to be created in massive stars, while lighter elements, such as B, ^{6}Be, Li are created by a cosmic ray spallation process.

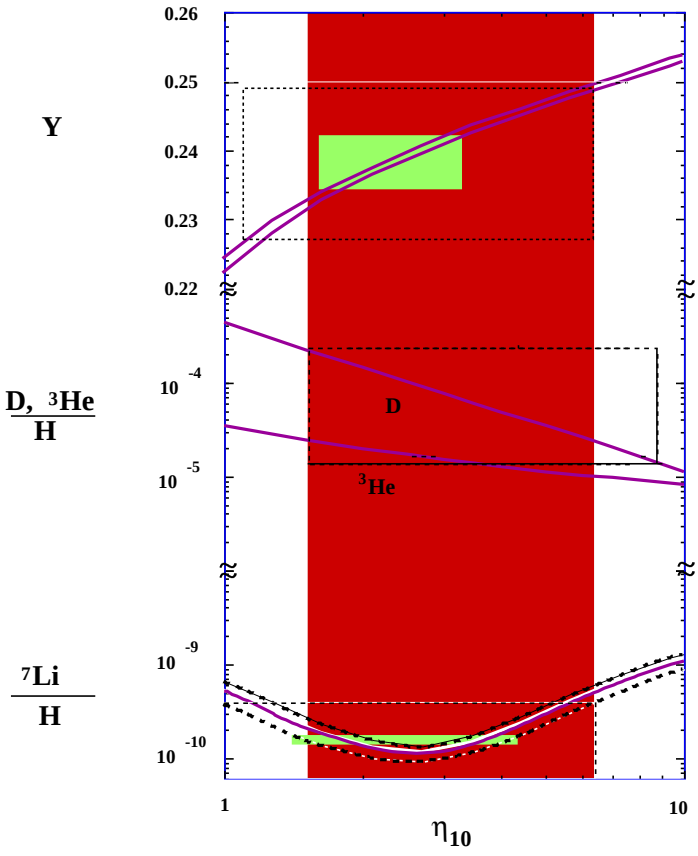


Fig. 8: Dependence of light element abundances on η_{10} and observational limits. From ref. [6].

[2]Changing the number of neutrino species changes the rate of the Universal expansion and thus predictions of Big Bang nucleosynthesis. One cannot admit more than four types of massless neutrinos in order not to spoil sucessful predictions of BBN, see, e.g. [6].

4.4 Baryogenesis

4.4.1 Evidence

As we discussed in the previous sections, the parameter $\eta = \frac{n_B}{n_\gamma}$, the inverse entropy per nucleon, plays an important role in cosmology. It determines the moment of matter-radiation equality and influences primordial abundances of light elements and structure formation. As we shall see, it is related to the fact that there is no antimatter in the Universe (at least, not in amounts comparable to matter).

Antimatter in the Universe can be detected by a number of different means. First, if antigalaxies exist, we should see antinuclei in cosmic rays, precisely in the same way as we see cosmic nuclei. However, no antinuclei have been observed in cosmic rays, the recent limit on the ratio of antihelium to helium nuclei in cosmic rays comes from the AMS experiment [7], $\bar{He}/He, 10^{-6}$.

Positrons, antiprotons and antineutrons are observed. These antiparticles can be produced in collisions of cosmic protons or nuclei with galactic gas and with particles in the Earth's atmosphere, and their amount is consistent with expectations [8]. On the contrary, the probability of forming a complicated nucleus, such as $\bar{He}$ by collision of particles (rather than antiparticles) is extremely small, and an observation of just one anitihelium nucleus would show that there must be antimatter in the Universe.

Secondly, in regions where matter and antimatter are mixed, annihilation of protons and antiprotons must take place. Annihilation will produce about 5-6 π^0 and $\pi^\pm$ mesons which, in turn, will decay into γ-quanta, electrons, positrons, neutrinos and antineutrinos. The spectrum of γ-quanta has a specific form, with energy peaked around $(2\text{ GeV})/(5-6)/2 \sim 150$ MeV. The form and normalization of annihilation $\gamma-$spectra at the present time depends on the size of antimatter clusters and on the amount of antimatter in them. Also, one should take into account the red shift, as γ-quanta we see today were created in the past. For the globally symmetric Universe one can put a strong constraint on the size l of antimatter clusters [9], $l > 1000$ Mps. This number may be compared with the visible size of the Universe, 3000 Mps $\sim 10^{10}$ light-years.

It is therefore very likely that the observable Universe is globally asymmetric and contains no antimatter. Even in the baryon-symmetric case with huge antimatter clusters, one should be able to understand why the Universe is asymmetric over cosmological distances and how this complicated structure with islands of matter and antimatter could emerge.

The parameter η in fact gives baryon asymmetry of the early Universe, at temperatures of the order of 1 GeV ($t \sim 10^{-6}$ s from the beginning of Big Bang):

$$\Delta(t) = \left.\frac{n_B - n_{\bar{B}}}{n_B + n_{\bar{B}}}\right|_{T\sim 1\text{ GeV}} . \tag{47}$$

Since the plasma was hot, with a temperature higher than the masses of light quarks, the number of quarks and antiquarks is the same as the number of photons, up to spin factors, and

$$\Delta(t) \simeq \left.\frac{n_B - n_{\bar{B}}}{n_\gamma}\right|_{T\sim\text{ GeV}} . \tag{48}$$

The ratio $\frac{n_B - n_{\bar{B}}}{n_\gamma}$ does not change much during the evolution of the Universe since then, because $(n_B - n_{\bar{B}})a^3$ gives (conserved) baryon number in comoving volume, and $n_\gamma a^3$ is constant up to entropy generation factors related to the annihilation of light particles, as we discussed in connection with neutrino temperature. Therefore, $\Delta(t \sim 10^{-6}) \simeq \eta \sim 10^{-10}$ and the baryon asymmetry at that time is tiny. When the Universe cools down from this state, the symmetric part of the baryon-antibaryon background annihilates into photons and neutrinos, but the nucleons that do not find a pair survive, see Fig. 9. These give rise to galaxies, stars and planets.

In 1967 Sakharov suggested [10] (see also the somewhat later paper by Kuzmin [11]) that the Universe is asymmetric because baryon number is in fact non-conserved. In this case, the Universe

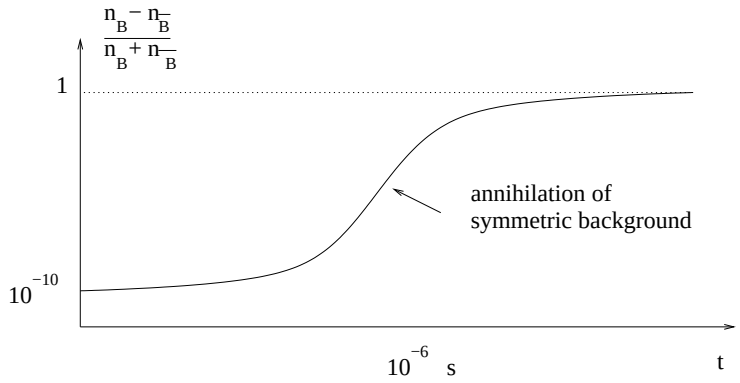


Fig. 9: Dependence of baryon asymmetry on time.

could start its expansion from a truly symmetric state, containing an equal number of particles and antiparticles. Then, in the course of the expansion, the particle physics reactions with B and CP non-conservation would produce an excess of particles over antiparticles.

He assumed that there exist some heavy particles X - maximons, with masses of the order of the Planck scale, 10^{19} GeV, which can decay with baryon number non-conservation and CP violation. If CP is broken, an equal number of X and $\bar{X}$ will, after their decay, leave a different number of baryons and antibaryons, precisely as the decays of K^0 and $\bar{K}^0$ mesons leave different numbers of electrons and positrons. It is sufficient to produce a small asymmetry, $\sim 10^{-9} - 10^{-10}$, which is then converted into a 100% asymmetry after the annihilation of matter and antimatter.

Besides B non-conservation and CP violation it is required that particle reactions occur in a non-equilibrium fashion, because in thermal equilibrium the baryon number of the system must be zero: the total rate of the processes which increase baryon number is exactly compensated by the rate of the processes that decrease it, as a consequence of the CPT-theorem.

From particle physics, baryogenesis requires baryon number non-conservation and CP-violation. Depending on the mechanism of B-violation, one can consider grand unified baryogenesis, electroweak baryogenesis and leptogenesis. We will briefly review different scenarios below.

4.4.2 GUT baryogenesis

The source of non-conservation of baryon number in GUT baryogenesis is associated with the unification of strong, weak and electromagnetic interactions. The leptoquarks X of grand unified theories can decay as $X \to q\ell,\ \bar{q}\bar{q}$ and $\bar{X} \to \bar{q}\bar{\ell}, qq$, see fig 10. If CP is broken, an equal number of X and $\bar{X}$ will, after

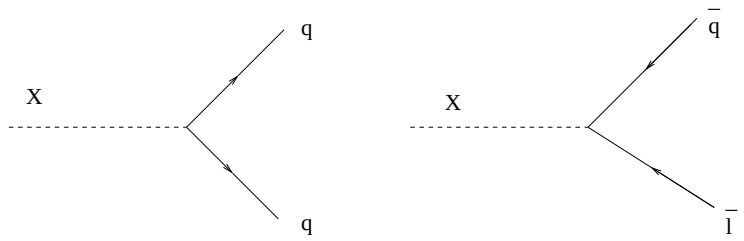


Fig. 10: Tree diagrams for leptoquark decay with baryon number non-conservation.

their decay, leave a different number of quarks and antiquarks. To find the magnitude of CP-violation one has to consider radiative corrections to the leptoquark decays, of the type shown on Fig. 11.

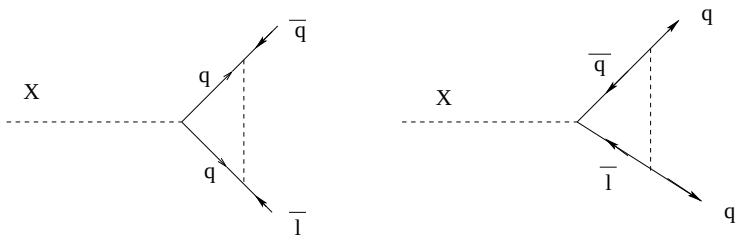


Fig. 11: Radiative corrections to leptoquark decay.

If the universe were as hot as the leptoquark mass M_X (typically, $M_X \sim 10^{15}$ GeV) and in a state close to thermal equilibrium then baryon asymmetry of the Universe resulting from leptoquark decays would be of the order

$$\Delta \sim \frac{1}{N_{\text{eff}}} \delta_{CP} \cdot S_{\text{macro}}, \tag{49}$$

where δ_{CP} is the asymmetry in leptoquark decays,

$$\delta_{CP} = \frac{\Gamma(X \to qq) - \Gamma(\bar{X} \to \bar{q}\bar{q})}{\Gamma_{\text{tot}}}, \tag{50}$$

Γ_{tot} is the total width, and S_{macro} is a factor taking into account the kinetics of the leptoquark decays: $S_{\text{macro}} \sim \frac{M_X^2}{\Gamma_{\text{tot}} M_{Pl}}$ for $M_X^2 < \Gamma_{\text{tot}} M_{Pl}$ and $S_{\text{macro}} \sim 1$ for the opposite case.

The factor S_{macro} can be understood in the following way. Besides the decay of X-particles, there are inverse processes (inverse decays) and four-fermion scattering of quarks and leptons. In a state of thermal equilibrium no baryon asymmetry is produced because of complete compensation of the rate of different reactions. Largest deviations are expected in concentrations of the heavy particles, in this case leptoquarks. Indeed, X-particles start to decay at a temperature $T_d = \sqrt{\Gamma M_0}$ determined from the condition $\frac{M_0}{T^2} = \Gamma^{-1}$ so, for $T > T_d$, the number of X-particles is effectively conserved and is equal to the number of photons, up to the spin factors, see Fig. 12. If $T_d < M_X$ then the number of X-particles is substantially greater than their equilibrium value, which is Boltzmann suppressed in this case, $\exp(-M_X/T)$. Thus, practically all leptoquarks decay in an out-of-equilibrium way, giving $S_{\text{macro}} \sim 1$. If, on the contrary, $T_d > M_X$, the processes equilibrating X-particles are in thermal equilibrium at $T < T_d$ and leptoquark concentration follows its equilibrium value, with small deviations of order $O(M_X^2/T_d^2)$, giving the suppression factor defined above.

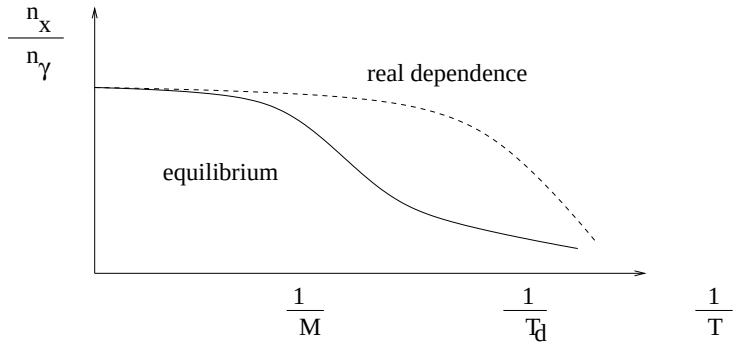


Fig. 12: Dependence of concentration of decaying particle on temperature in expanding Universe.

Under the assumption that the universe had temperatures of the requisite order of magnitude, many grand unified theories give rise to baryon asymmetry of the required order of magnitude. Some

care should be taken over the equilibrium character of anomalous electroweak reactions with B-non-conservation, see below. Basically, the requirement for sucessful GUT baryogenesis is that asymmetry in $B - L$ must be generated.

Perhaps the only drawback of GUT baryogenesis is that it is hardly compatible with inflation. In inflationary cosmology there are several constraints on the temperature of the Universe after reheating. It should not be larger than about 10^{10} GeV [12], otherwise a lot of gravitinos would be produced. However, the typical mass of leptoquarks in grand unified theories is of the order of 10^{15} GeV, which is substantially larger that the reheating temperature. Thus, there simply cannot be any leptoquarks to decay and produce the asymmetry.

There is a possible way out from this situation, related to the so-called pre-heating stage of the expansion of the universe [13]. At this time classical inflaton dynamics allows for a non-thermal production of heavy particles because of parametric resonance [14, 15]. Still, according to [16], the effective leptoquark concentration n_X/T^3 is typically quite small, $n_X/T^3 \sim 10^{-6}$, which would require very large CP asymmetry in the leptoquark decays, $\delta_{CP} \sim 1$.

4.4.3 Electroweak baryogenesis

Electroweak baryogenesis is based on the observation that the rate of B non-conservation in the electroweak theory is large at high temperatures [17].

On a perturbative level the electroweak theory has four conserved fermionic numbers: B - baryon number, and $L_e,\ L_\mu,\ L_\tau$ - leptonic numbers. Quantum anomaly

$$\partial_\mu J^B_\mu = \partial_\mu J^L_\mu = \frac{n_f}{32\pi^2}\mathrm{Tr}\left(F_{\mu\nu}\tilde{F}_{\mu\nu}\right) + U(1)\ \text{ part}$$

leads to anomalous processes with non-conservation of baryon and lepton number [18]:

$$\text{bosons} \leftrightarrow \text{bosons} + 9q + 3l.$$

Here J^B and J^L are baryon and lepton currents, $F_{\mu\nu}$ is the SU(2) field strength, n_f is the number of fermion generations, q and l are quarks and leptons.

The rate of anomalous fermion-number non-conservation at zero and non-zero temperatures is of the order of (see [18, 17, 19, 20] and reviews [21, 22]):

$$\Gamma \sim \begin{cases} \exp(-\frac{4\pi}{\alpha_W}) \sim 10^{-160}, & T = 0 \\ (\alpha_W T)^4 (\frac{M_{sph}}{T})^7 \exp\left(-\frac{M_{sph}}{T}\right), & T < T_c \\ (\alpha_W)^5 T^4, & T > T_c \end{cases} \tag{51}$$

where $M_{sph} \sim M_W/\alpha_W$ is the sphaleron mass and T_c is the temperature of the electroweak (EW) phase transition.

Other ingredients of baryogenesis, C and CP violations, are also present in the standard model or its extensions. The requirement of non-equilibrium happened to be quite non-trivial in the standard model. At the typical scale of electrowek theory, $T \sim 100$ GeV, all electroweak reactions are rapid enough to keep concentrations of different particles close to thermal equilibrium. Large deviations from thermal equilibrium may arise at the first-order phase transition with the breaking of SU(2)×U(1) group. For detailed discussion of the mechanism of electroweak baryogenesis see refs. [21, 22, 23]. Here we just mention the EW baryogenesis constraints on the particle spectrum of the SM and its SUSY extensions.

The nature of the electroweak phase transition depends crucially on the Higgs mass. If $M > M_{\rm crit}$ then there is no phase transition at all, so that during the cooling of the Universe the system gradually

changes from the so-called symmetric phase into the Higgs phase (in fact, there are no distinct phases but just one – the Higgs-confinement phase – which explains the possibility of the absence of phase transition). Lattice computations carried out in refs. [24, 25] have shown that the critical value of the Higgs mass in the MSM is

$$M_{\rm crit} = (72.3 \pm 0.7)\ {\rm GeV}\ . \tag{52}$$

In view of the experimental lower limit on the Higgs mass $m_H > 114$ GeV it is clear that the MSM does not have any phase transition at all, so that the deviations from thermal equilibrium at the electroweak scale are only associated with the expansion of the Universe. The baryon asymmetry that can appear in this situation is much smaller than the observed value, so that new physics is required for its explanation.

A popular extension of the standard model is the MSSM, in which the strength of the electroweak phase transition depends on a number of parameters, the most important being the mass of the Higgs and the mass of the right-handed stop (the scalar superpartner of the top quark). There can be a strong first-order phase transition, sufficient for electroweak baryogenesis, if the lightest Higgs mass is somewhere between 75 and 120 GeV and the right scalar top mass 100 GeV $< m_R < 160$ GeV [26], for a recent review see [27]. This set of parameters is constrained by existing experiments and this interval of Higgs masses has been partially covered at the electron–positron collider at CERN. The baryogenesis-carrying version of the SUSY extension of the standard model has a number of consequences for phenomenology as it requires a specific spectrum of SUSY particles and a particular pattern of CP-violation.

4.4.4 Leptogenesis

There is strong experimental evidence in favor of neutrino oscillations [28, 29]. If neutrino oscillates, it has a mass. Theoretically, a lowest-order SU(2)×U(1) gauge-invariant operator that can be added to the SM Lagrangian has the form:

$$\Delta L = f_{ab}\frac{(\bar{\nu^c_\alpha}\phi)(\phi^\dagger \nu_\beta)}{M}, \tag{53}$$

where ϕ is the Higgs doublet, M is some high-energy scale, and ν is left-handed neutrino. This term gives Majorana neutrino masses and a lepton number violation. The simplest way to obtain this effective interaction from renormalizable field theory is to have right-handed neutrino ν_R with a large Majorana mass M_R. Then (53) comes from the see-saw mechanism [30, 31]. A heavy right-handed neutrino can decay and produce lepton asymmetry in the early universe, in precisely the same way as leptoquarks produce baryon asymmetry in GUTs. There the lepton number is converted into baryon asymmetry by sphalerons [32] (for a recent review see [33]). The resulting baryon asymmetry is just a numerical factor of order one smaller than the lepton asymmetry.

This mechanism for baryogenesis requires sufficient concentration of right-handed neutrinos at the moment of their decay. If $m_R \sim 10^{10}$ GeV or less, right-handed neutrinos could be thermally produced at the end of the inflationary period; the reheating temperature is sufficiently low to prevent the overproduction of gravitino [12]. Right-handed neutrinos may also be produced non-thermally at preheating [34]. A detailed study of this mechanism can be found in [35].

4.4.5 Affleck-Dine baryogenesis

The Affleck-Dine mechanism [36] takes advantage of supersymmetry. Supersymmetric theories contain scalar fields that carry lepton or baryon numbers and the effective potential for squarks and sleptons has flat directions, i.e. the energy of the static scalar field configuration at large ϕ is much smaller than ϕ^4. In this scenario, a combination of squarks and sleptons, or some other fields carrying a baryon or lepton number, has a large expectation value along some flat direction of the potential at the end of inflation. At large VEV, the baryon number can be strongly violated by the high-scale physics. As a result of the baryon number non-conservation, along with the CP violation, the scalar condensate acquires a baryon number and the complex scalar field is characterized by the time-dependent phase,

$\phi = |\phi(t)| \exp(\Omega(t))$. The subsequent evolution leads it into the domain of conserved baryon number because the field amplitude ϕ decreases with time. Finally, squarks decay into ordinary quarks and release baryon number stored in the scalar condensate. A study of this scenario in different models was made in refs. [36, 37, 38] with the result that baryon asymmetry of the universe can be explained by this mechanism.

5 INFLATION

5.1 Problems of standard cosmology

To explain what kind of problems faced standard cosmology before the invention of inflation, we will introduce the notion of particle horizons. In a static Universe, if two events are separated by distance Δl and time Δt, they are causally independent, provided $\Delta l > \Delta t$. What is the analogue of this statement in an expanding Universe? To answer this, let us write the equation describing the propagation of light, taking into account the fact that the speed of light is 1 in the natural system of units:

$$\frac{dl}{dt} = 1 + \frac{\dot{a}}{a} l \, . \tag{54}$$

The solution of this equation is

$$l(t) = \int_{t_0}^{t} \frac{a(t)}{a(t')} dt' \, . \tag{55}$$

For both radiation- and matter-dominated Universes, with $a(t) \sim t^{1/2}$ and $a(t) \sim t^{2/3}$ respectively, the integral in eq. (55) converges even if $t_0 = 0$. The distance light travels since t_0 is called the particle horizon, $l_H(t)$. For different epochs:

$$l_H(t) = \begin{cases} 2t, & \text{radiation-dominated epoch,} \\ 3t, & \text{matter-dominated epoch.} \end{cases} \tag{56}$$

If the distance between two points is greater than $l_H(t)$, the points were not in causal contact in the past and thus we should expect that the parameters of the Universe may be different there.

5.1.1 The horizon and homogeneity problem

Assuming that there were only radiation- and matter-dominated epochs in the past, then, as already discussed, the photons decoupled from the plasma at some moment $t_d \simeq 5 \cdot 10^5$ y. Thus while looking at different points of the sky separated by some angle $\theta > \theta_H$ we observe CMB emitted from regions that were never in causal contact and so should have different temperatures. To estimate θ_H, one should find the present size L of the region that was the horizon at the decoupling time,

$$L \sim 3t_d \frac{a_{\rm now}}{a_d} \sim 3t_d \left(\frac{t_{\rm now}}{t_d} \right)^{\frac{2}{3}} \, , \tag{57}$$

where $3t_d$ is the horizon scale at t_d. The angle θ_H is simply the ratio of this scale to the present size of the horizon,

$$\theta_H \simeq \frac{3t_d}{3t_{\rm now}} \left(\frac{t_{\rm now}}{t_d} \right)^{\frac{2}{3}} \simeq \left(\frac{t_d}{t_{\rm now}} \right)^{\frac{1}{3}} \, , \tag{58}$$

which corresponds to $\theta_H \sim 1^\circ$. This means that the present horizon contains $O\left(\frac{t_{\rm now}}{t_d}\right) \sim 10^4$ domains that were not in causal contact before recombination, see Fig. 13. However, observations show that the cosmic microwave background is isotropic for these angles, with accuracy better than 10^{-4}. This is the essence of the horizon and homogeneity problem.

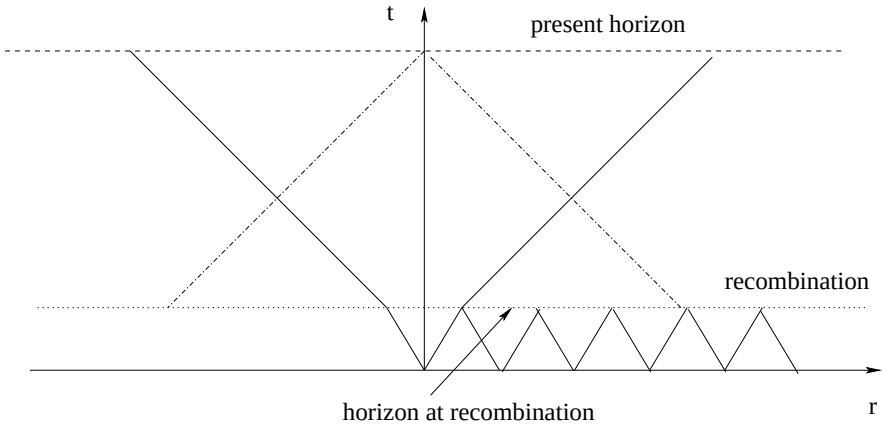


Fig. 13: An observer at point $r = 0$ sees many particle horizons corresponding to photon decoupling.

5.1.2 The flatness problem

Let us consider the relationship (23) in somewhat different form,

$$\Omega - 1 = \frac{k}{a^2 H^2} \,, \tag{59}$$

where Ω is the ratio of the total energy density to the critical density. For both matter- and radiation-dominated epochs, $H \sim \frac{1}{t}$ and $a \sim t^\alpha$, with $\alpha = \frac{1}{2}$ or $\alpha = \frac{2}{3}$. Thus, ω increases with t as $\Omega - 1 \sim t^{2(1-\alpha)}$. Therefore, to have $\Omega \simeq 1$ at the present time, as follows from observation, Ω must have been finely tuned to one with huge accuracy in the past. For example, at the nucleosynthesis time, $|\Omega - 1|$ must be of the order of 10^{-15}. It is unclear why the Universe should have been so flat in the past.

5.2 Inflation as a solution of cosmological problems

In fact, the two problems described above are related to each other. The inflationary paradigm, introduced in [39, 40, 41], provides a simultaneous solution to both of them.

Supposing that, for some reason, the dependence of a on t before recombination were such that the integral in (55) is very large and the factor aH increases with time rather than decreases, both problems would be solved simultaneously. For example, if the dependence of a on time has a power-law behavior with exponent α such that $\alpha > 1$ during some time after an initial singularity, the integral in (55) diverges (i.e., formally, the size of the particle horizon is infinite at the recombination time), while the factor aH increases with time.

For a physical picture of how this might happen, we suppose that vacuum energy density ϵ_{vac} dominated the Universe expansion for some time $t_0 < t < t_1 < t_d$. The Universe therefore expands exponentially, $a \sim \exp(+H(t - t_0))$ with H defined in (29). Thus, the horizon at recombination is at least

$$l_H \simeq \frac{1}{H} \exp(+H(t_1 - t_0)) \,, \tag{60}$$

and is much greater than the horizon would be in a radiation- or matter-dominated epoch, provided $H(t_1 - t_0) \gg 1$. For example, if we take a typical Grand-Unified scale for $\epsilon_{vac} \sim (M_X)^4 \sim (10^{15}\ \text{GeV})^4$, it is sufficient to have $H(t_1 - t_0) > 65$ for the horizon problem to be solved, see Fig. 14. The flatness problem finds its solution as well, since after the period of exponential inflation $\Omega - 1 \sim \exp(-2H(t_1 - t_0))$, which gives an exponentially small deviation from the flat Universe.

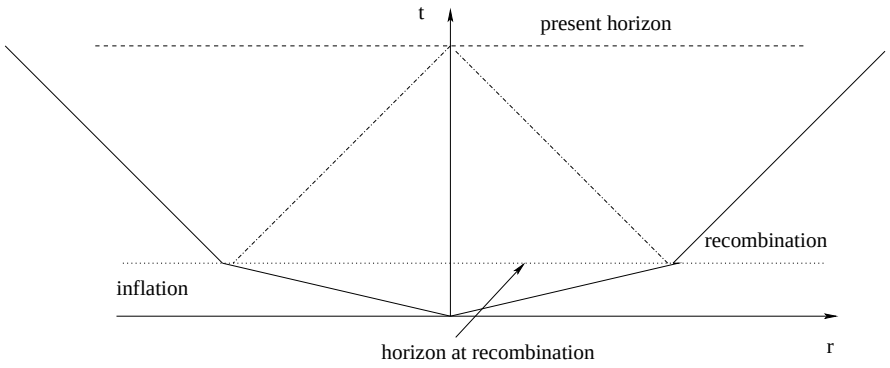


Fig. 14: An observer at point $r = 0$ sees just one particle horizon corresponding to photon decoupling.

5.2.1 A particle-physics model of inflation

There are many different particle-physics models of inflation. Practically all of them are associated with the dynamics of single or multiple scalar fields. We refer to a comprehensive review [42, 43] and describe here just one possibility which is called "chaotic inflation" [44].

Consider a theory of a single scalar field in a curved background with an action

$$S = \int d^4x\sqrt{-g}\left(\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - U(\phi)\right) , \tag{61}$$

where the potential is

$$U(\phi) = \frac{1}{2}\phi^2 + \frac{\lambda}{4}\phi^4 . \tag{62}$$

We will assume that $m^2 \ll M_{Pl}^2$ and $\lambda \ll 1$.

We do not know how to describe the state of the Universe at the Planck scale, since the classical theory of gravity is not applicable there. Nevertheless, it is natural to assume that at Planck time there were fluctuations in the scalar field with energy density

$$\epsilon \sim \frac{1}{2}\dot{\phi}^2 + \frac{1}{2}(\nabla\phi)^2 + U(\phi) \sim M_{Pl}^4 . \tag{63}$$

Suppose that there is a sufficiently large region of space where fluctuations of potential energy dominate, i.e.

$$U(\phi) \sim M_{Pl}^4 \gg (\nabla\phi)^2 \text{ and } \dot{\phi}^2 . \tag{64}$$

In these regions the value of ϕ is much larger than the Planck scale,

$$\phi \simeq \frac{M_{Pl}^2}{m} \gg M_{Pl} , \tag{65}$$

and the scalar field is nearly homogeneous, so that the equation representing its evolution is simply

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dU(\phi)}{d\phi} = 0 , \tag{66}$$

where

$$H^2 = \frac{8\pi G}{3}\left(\frac{\dot{\phi}^2}{2} + U(\phi)\right) . \tag{67}$$

This looks like an equation of motion of a non-relativistic particle with unit mass in the potential $U(\phi)$ with a friction term that depends on the position and velocity of the particle, see Fig. 15. For large values

of ϕ the regime is overdamped, with $H \gg \ddot{\phi}/\dot{\phi}$, where

$$H^2 \simeq \frac{4\pi m^2 \phi^2}{3M_{Pl}^2} \, . \tag{68}$$

Thus, eq. (66) has the form

$$\frac{\sqrt{12\pi} m \phi \dot{\phi}}{M_{Pl}} + m^2 \phi = 0 \tag{69}$$

and has a solution

$$\phi \simeq \phi_0 - \frac{m M_{Pl} t}{\sqrt{12\pi}} \simeq \phi_0 \left(1 - O(\frac{m^2 t}{M_{Pl}}) \right) \, . \tag{70}$$

The "slow-roll-down" approximation breaks down at $\dot{\phi}^2 \sim V(\phi)$, where $\phi \simeq M_{Pl}$, $t \simeq M_{Pl}/m^2$. Before this time the Universe expands exponentially and the scale factor changes by

$$\exp{(Ht)} \simeq \exp{(\frac{M_{Pl}^2}{m^2})} \gg 1 \, . \tag{71}$$

During exponential expansion the non-homogeneities are red-shifted away and the Universe becomes practically uniform at cosmological distances.

After time $t \simeq M_{Pl}/m^2$ inflaton oscillates near the origin, tranferring its energy to other particles. This process is usually called reheating: if the energy is converted through the parametric resonance, it is called preheating [13]. For $m \ll M_{Pl}$ this simple model of scalar field solves horizon, homegeneity

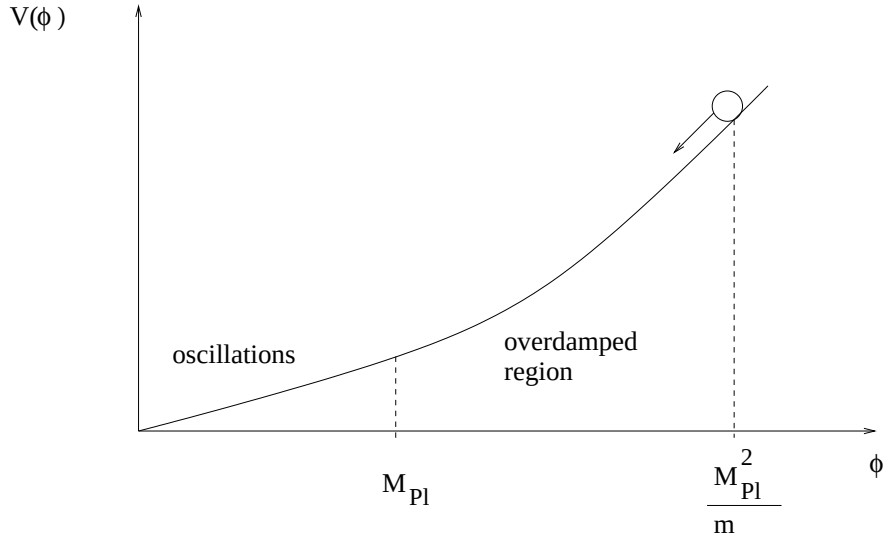


Fig. 15: Inflaton dynamics as the motion of a particle in potential $V(\phi)$ with friction term depending on ϕ.

and flatness problems. Moreover, quantum fluctuations of the scalar field, existing on the De-Sitter (exponential) part of the expansion of the Universe, give rise to the scale-invariant spectrum of primordial density perturbations, which eventually lead to structure formation in the Universe.

The most important and generic predictions of inflation are that the universe should be spatially flat, $\Omega_{tot} = 1 + O(10^{-5})$ and that the spectrum of density perturbations should be scale-invariant. Both predictions can be verified by cosmological observations.

6 CONTENT OF THE UNIVERSE OR COSMOLOGICAL OBSERVATIONS VERSUS THE STANDARD MODEL

6.1 Cosmological parameters

As we have seen, the amount of each of the different substances in the universe is measured as the ratio of the corresponding density ρ_A of the substance to the critical energy density ρ_c. The simplest cosmological model is obtained by considering the following contributions to the total energy density:

(i) matter: Ω_M with the equation of state $p = 0$ where p is the pressure. This counts all non-relativistic particles in the universe (protons, neutrons, electrons and, possibly, massive neutrinos).

(ii) radiation: Ω_r with $p = \epsilon/3$, where ϵ is the energy density. Radiation is associated with photons and very light or massless neutrinos.

(iii) The standard model also admits a cosmological constant with $p = -\epsilon$. The corresponding parameter is Ω_Λ.

In the Universe at the present time the contribution of radiation is not numerically essential, $\Omega_r \ll \Omega_M$. Thus, there are essentially two parameters Ω_M and Ω_Λ that determine the global structure of the Universe and its future. The "phase structure" of the Universe as a function of these parameters is shown in Fig. 16. It is more complicated than for the case $\Omega_\Lambda = 0$.

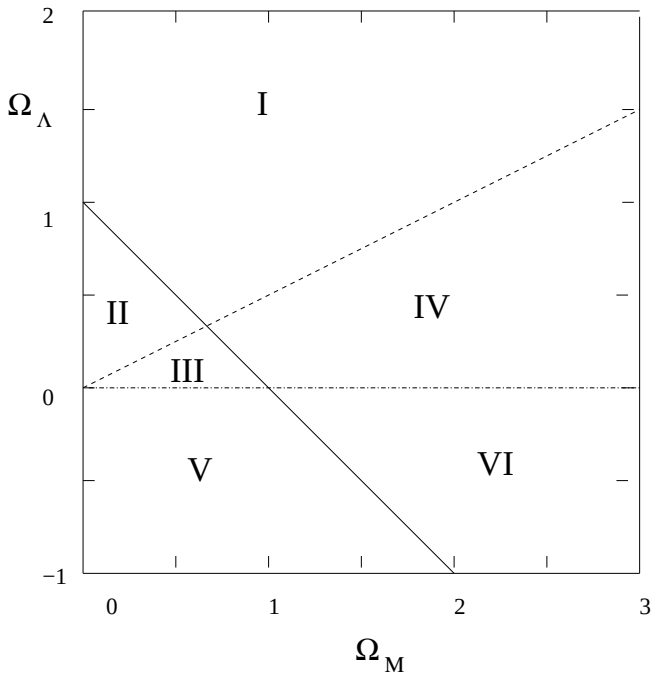


Fig. 16: Schematic phase structure of the Universe depending on parameters Ω_M and Ω_Λ. In regions I and II the Universe is accelerating now, and in regions III – VI the Universe is decelerating. In the regions I, IV, VI the Universe is closed, and in II, III, V it is open. In regions I – IV the Universe expands forever, while in regions V and VI it will collapse in the future.

The parameters Ω_M and Ω_Λ should be determined from observations. There are several independent ways to estimate them (for a review see [45, 46]).

6.1.1 Clusters and the dynamics of galaxies

The dynamics of clusters offers several ways of estimating the matter contribution. A cluster mass M_{cl} can be defined by the consideration of galaxy motion within the cluster and/or by gravitational lensing

by a cluster-gravitational potential. An estimate of the mass of matter in the Universe would then be

$$M_{\rm matter} = \frac{M_{cl}}{L_{cl}} L_U \ , \tag{72}$$

where L_{cl} and L_U are respectively the luminosities of a cluster and of the Universe as a whole. This gives the estimate [47, 48, 49, 50]

$$\Omega_M = 0.2 \pm 0.1 \ . \tag{73}$$

Incidentally, the fraction of energy in luminous matter is known to be much smaller,

$$\Omega_{\rm lum} \simeq 0.003 - 0.01 \ . \tag{74}$$

Another estimate comes from the baryon fraction in matter. Part of the baryonic matter in clusters is luminous, while another part corresponds to a gas whose mass can be estimated from X-ray emission. Assuming that the baryon fraction in clusters is the same as the average in the Universe, and taking $\Omega_B \simeq 0.045 \pm 0.0025$ from nucleosynthesis [51, 52], then [53, 54, 55]

$$\frac{M_{\rm baryons}}{M_{\rm total}} \simeq 0.15 \ , \tag{75}$$

which gives

$$\Omega_M = 0.3 \pm 0.1 \ . \tag{76}$$

Yet another estimate comes from consideration of the cluster abundance [56]

$$\Omega_M = 0.25^{+0.15}_{-0.10} \ . \tag{77}$$

These estimates show that the amount of matter in the Universe that can cluster is larger than that of baryonic matter. This indicates that the Universe contains dark matter of unknown nature. Another piece of evidence in favor of this conclusion comes from the consideration of rotational curves of spiral galaxies.

Orbital velocities of stars or of a gas far from the galactic center can be measured. The dependence of velocity on the distance from the center follows from the virial theorem, telling that for a gravitationally bounded system the kinetic energy of a body $\frac{1}{2}mv^2$ is proportional to its potential energy $\frac{MmG}{r}$, where m is the mass inside a sphere with radius r. Thus,

$$v \sim \frac{\sqrt{MG}}{\sqrt{r}} \ . \tag{78}$$

If it is assumed that only luminous matter gravitates, then $v \sim \frac{1}{\sqrt{r}}$ which is just Kepler's law. In reality, the rotational curves have plateaux, as shown on Fig. 17.

This type of dependence can be derived with the asssumption that the density of the halo of a galaxy changes with radius as $\rho_{\rm dark}(r) \sim 1/r^2$, giving $M_{\rm dark} \sim r$. Quite a good overall fit is given by

$$\rho_{dark}(r) = \frac{v_\infty^2}{4\pi G(r_c^2 + r^2)} \ , \tag{79}$$

where v_∞ is the limiting velocity and r_c is the typical size of density distribution.

Thus, the total amount of matter in galaxies is considerably larger than their visible part, so cosmological data indicates that the Universe contains non-baryonic dark matter which clusters but does not shine with $\Omega_M \simeq 0.3$ and $\frac{\Omega_B}{\Omega_M} \simeq \frac{1}{7}$.

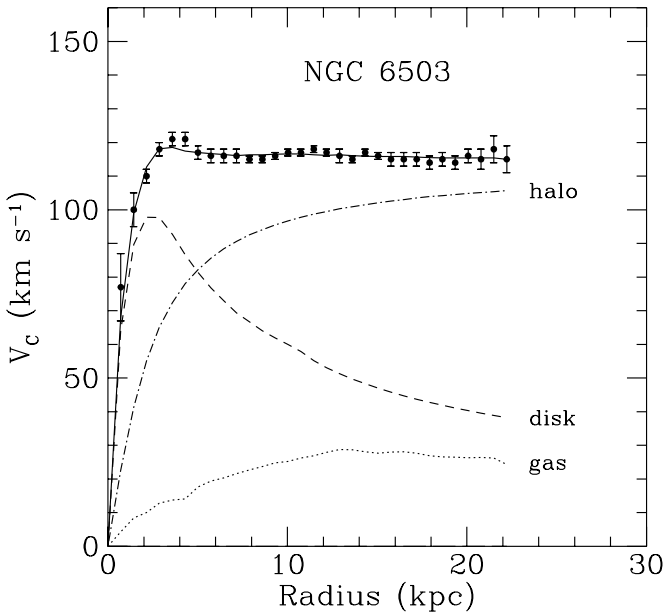


Fig. 17: Rotational curve of a typical spiral galaxy. The dashed curve shows expectations from luminous matter alone, the dotted line from gas and the dot-dashed line from dark matter distributed according to eq. (79). From ref. [57].

6.1.2 The accelerating universe

Further information about cosmological parameters comes from the observation of supernovae at cosmological distances. Two teams of observers have recently produced similar results: the Supernova Cosmology Project (SCP) and the High-Z Supernova Search (HZS). The idea is to take a number of "standard candles", i.e. supernovae of type Ia, and to find their distance from the earth by comparing the known luminosity with the observed luminosity. In addition, one can also define their red shift. At small distances there is a linear dependence between the red shift and the distance, given by the Hubble law, whereas at large distances cosmological evolution has some effect and information on the acceleration or deceleration of the universe can be derived. The main result of this study is that the Ia supernovae with a high red-shift are observed to be dimmer than would be expected in an empty Universe, $\Omega_M = 0$ with no cosmological constant, see Fig. 18

The result of fitting cosmological parameters, obtained by the SCP [58] (similar numbers from the HZS [59]), is

$$0.8\Omega_M - 0.6\Omega_\Lambda \simeq -0.2 \pm 0.1 \,, \tag{80}$$

see Fig. 19.

Assuming that the Universe is flat ($\Omega_M + \Omega_\Lambda = 1$), this gives [58]

$$\Omega_M = 0.28^{+0.09+0.05}_{-0.08-0.04} \,, \tag{81}$$

which is in reasonable agreement with the determination of Ω_M by other means.

6.1.3 Cosmic microwave background

Another important source of constraints on the cosmological parameters is related to cosmic microwave background radiation (CMB) and to the mass power spectrum. As we discussed, CMB anisotropy gives a snapshot of the Universe at redshifts of about 1000, corresponding to an age of $\sim 10^{12}$ s.

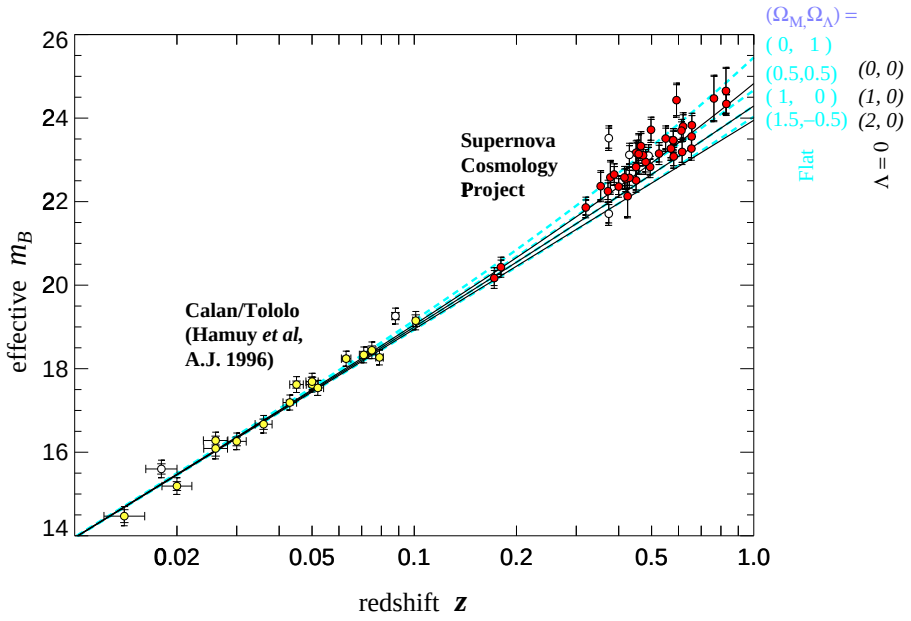


Fig. 18: Experimental data and fits to it for different cosmological parameters. From ref. [58].

In the first approximation CMB is isotropic: the temperature does not depend on the direction. However, exact isotropy must not be expected, as the Earth, together with the solar system, moves through CMB with some velocity v and the temperature in the forward direction must be larger than in backward direction. The form of these deviations is obtained from the exact Planck curve by replacing the photon energy in the Planck distribution (7) with its Lorentz transformed value. For non-relativistic velocities we have $\epsilon \to \epsilon(1+v\cos\theta)$, where θ is the angle between velocity v and the line of sight. Thus, the spectrum of deviations from the Planck curve is just the derivative of the equilibrium distribution with respect to frequency, see Fig. 20. Experimentally, this is indeed the case, $(T_{\rm forward} - T_{\rm backward})/T \simeq 1.23 \cdot 10^{-3}$ which corresponds to $v = 371 \pm 0.5$ km/s.

If dipole anisotropy is subtracted from the data on CMB, the fluctuations of the temperature on the level 10^{-5} are seen, as was first observed by COBE [60]. Nice colored pictures of COBE measuriments can be found on the home page of this experiment, http://space.gsfc.nasa.gov/astro/cobe/. A review of different CMB experiments can be found in [61].

Why are there fluctuations in CMB?

(i) *"Experimental" reason.* We do know that the Universe is clumpy at "small" scales of the order of supercluster size or below. There must therefore have been some fluctuations in matter distribution at decoupling. As matter and radiation were interacting at that time, there must be fluctuations in CMB as well. The theory of structure formation indicates that one should expect to have $\delta T/T \sim 10^{-5}$.

(ii) *Theoretical reason.* Quantum fluctuations of the scalar field-inflaton during inflation lead to adiabatic density perturbations that result in fluctuations in CMB.

Using theoretical machinery to study anisotropies of the microwave background, a first step consists of expanding the CMB temperature with the help of spherical harmonics $Y_l^m(\theta,\phi)$,

$$\frac{\Delta T(\theta,\phi)}{T} = \sum_{l,m} a_{lm} Y_l^m(\theta,\phi) \tag{82}$$

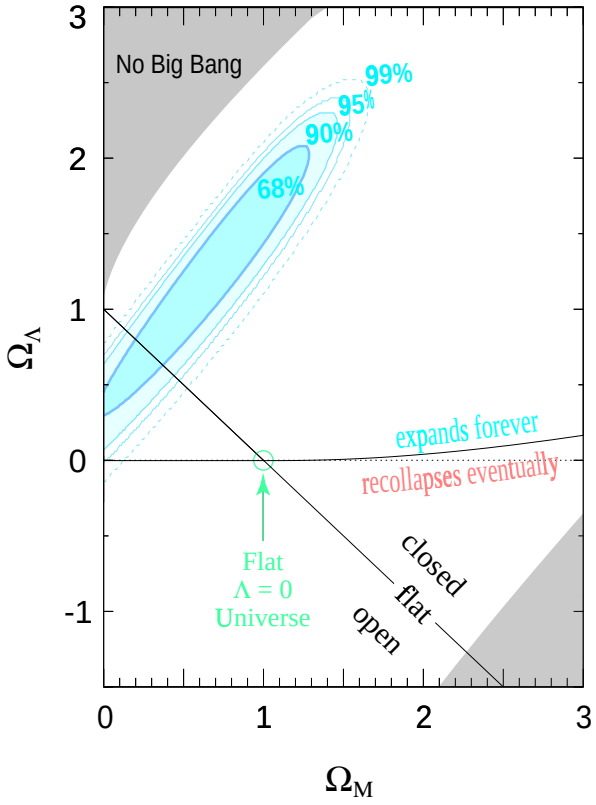


Fig. 19: Cosmological parameters from the Supernova Cosmology Project. From ref. [58].

and defining C_l as

$$C_l = \frac{1}{2l+1} \sum_m |a_{lm}|^2 \, . \tag{83}$$

For a particular cosmological model, values of C_l can be predicted. Schematically, one should take an initial spectrum of density perturbations (given, e.g., by inflation) and write kinetic equations for photons, taking into account various factors such as the interaction of photons with gravity potential induced by matter fluctuations (Sachs-Wolfe effect), the peculiar velocities of plasma (Doppler effect), damping of fluctuations, etc. Calculation then gives C_l as a function of cosmological parameters (Ω_Λ, Ω_M, baryon-to photon ratio) and other parameters essential for the model (for example, neutrino masses).

Recent experimental results, shown in Fig. 21, come from two different balloon experiments: Boomerang, [62] and Maxima [63]. In Fig. 22 the regions of cosmological parameters coming from different types of data are shown.

6.1.4 Global fit of cosmological data.

The results from different sources can be summarized as follows. A simple homogeneous cosmological model with two parameters $\Omega_M \simeq 1/3$ and $\Omega_\Lambda \simeq 2/3$ fits all the data well and the cosmological constant is nonzero at a confidence level of 99.7% (3σ) [58, 59]. With this set of parameters the Universe accelerates and will expand forever, provided the simplest cosmological model is correct. The total energy density in the Universe is close to the critical one [64]. This could be considered as an indication of the validity of the inflationary-universe scenario.

The age of the universe, according to [64], for $\Omega_{\rm tot} = 1$ is 11.8 ± 0.8 Gyr, and the value of the Hubble constant is $H = 75 \pm 10$ km s^{-1}Mps^{-1}.

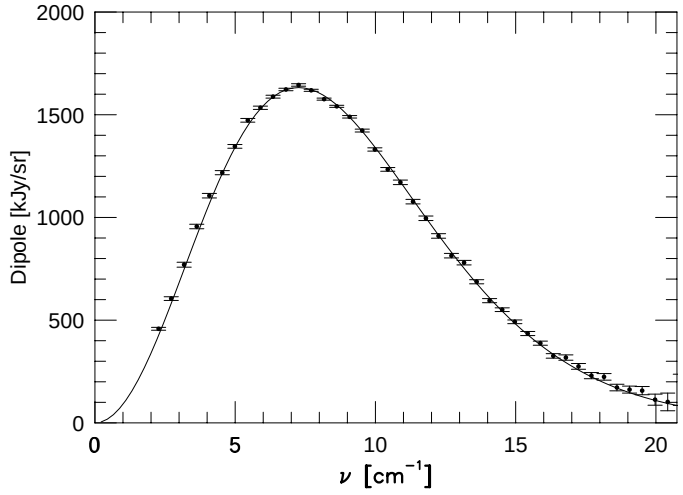


Fig. 20: The spectrum of the CMB dipole as measured by FIRAS. The solid line is the derivative of a T = 2.728 K Planck function. From ref. [5].

Thus, modern cosmological observations suggest that the Universe consists of baryonic matter, $\Omega_B \simeq 0.045$, non-baryonic dark matter with $\Omega_{DM} \simeq 0.3$,

$$\frac{\Omega_{DM}}{\Omega_B} \simeq 7\,, \tag{84}$$

and a cosmological constant with $\Omega_\Lambda \simeq 0.7$,

$$\frac{\Omega_\Lambda}{\Omega_M} \simeq 2\,. \tag{85}$$

In fact, all the components that we see in the Universe represent a challenge for the standard model of elementary particle interactions. We have already considered the problem of the origin of matter – baryon asymmetry of the Universe – which points towards physics beyond the standard model. The problems of dark matter and the cosmological constant remain to be discussed.

6.2 Origin of dark matter

Observations of cluster dynamics suggest that the amount of matter that can cluster in the Universe is about $\Omega_M \simeq 0.3$, while analysis of nucleosynthesis indicates that the amount of baryonic matter is much smaller, $\Omega_b \simeq 0.045$. The macho abundance found by the Eros collaboration [65] reveals that machos cannot contribute more than 20% of the galactic halo and thus cannot be used to explain the rotational curves of the galaxy. Thus, there should exist some non-baryonic dark matter.

The minimal standard model does not provide any candidate for the non-baryonic dark matter and, therefore, cosmological observations again point in the direction of physics beyond the standard model. A minimal extension of the standard model that gives neutrino mass through higher-order operators does not help for a number of reasons. Firstly, to explain the rotational curves of some dwarf galaxies, neutrino mass should be larger than 100 eV, which is in conflict with the cosmological upper bound on neutrino mass. Secondly, neutrinos form a so-called “hot” dark matter, and the theory of structure formation

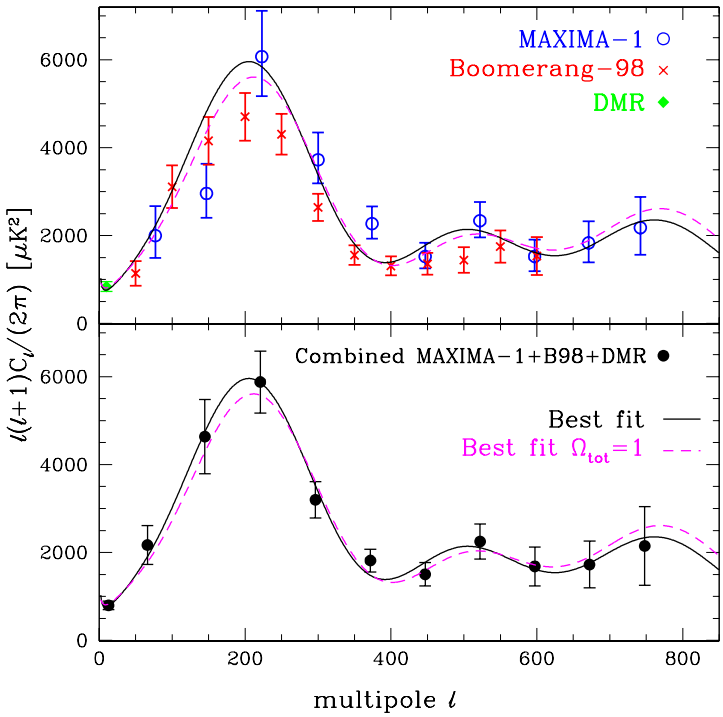


Fig. 21: The power spectrum seen by Boomerang, Maxima and COBE-DMR experiments. Best fit corresponds to $(\Omega_{tot}, \Omega_\Lambda, \Omega_b h^2, \Omega_{CDM} h^2) = (1.2, 0.5, 0.03, 0.12)$ and to $(1, 0.7, 0.03, 0.17)$ if Ω_{tot} is taken to be 1. The initial spectrum of perturbations is assumed to be scale-invariant. From ref. [64].

says that, with hot dark matter, small structures, such as galaxies, are not formed. Finally, if there is no extreme degeneracy between different neutrino flavors, the SuperK value of the neutrino mass splitting $\Delta m_\nu \sim 0.1$ eV is of the order of the neutrino mass, but then this value is too small to make a significant contribution to the dark matter. Thus, it looks as though more drastic modifications of the standard model are necessary in order to accomodate the cosmological data.

Particle physics provides a general answer to the question of dark matter by stating that there exist new stable objects (perhaps particles) which were produced in the course of the expansion of the universe. The non-observability of dark-matter particles is explained by their very weakly interacting character (if there are plenty of them) or by the fact that they are very rare, in which case they could be strongly interacting.

There are quite a number of particle-physics candidates for dark matter, for example axion, related to a solution of the strong CP problem (for a review of cosmological constraints on axion see, e.g, [66]) or the lightest supersymmetric particles, which may be stable due to R-parity conservation (neutralino, axino, gravitino). Neutralino is a Majorana fermion, which is a mixture of photino, zino and higgsino. It annihilates into lighter particles and its relic concentration can be computed in a particular supersymmetric extension of the standard model.

In a more general way, one can assume that dark matter consists of weakly interacting massive particles, characterized by their mass, concentration (mass and concentration are related to each other if the dark matter density is fixed from observation) and a cross-section of interaction with ordinary matter. One expects that their mass may be in the region of, say, $10 - 100$ GeV and that their average velocity is of the order of the gravitational galaxy escape speed 300 km/s. These particles may be searched for via their elastic scattering on ordinary mass particles, with a typical event rate of 1 event/kg/day, for existing detectors. There are several experiments looking for this type of event. I will just mention

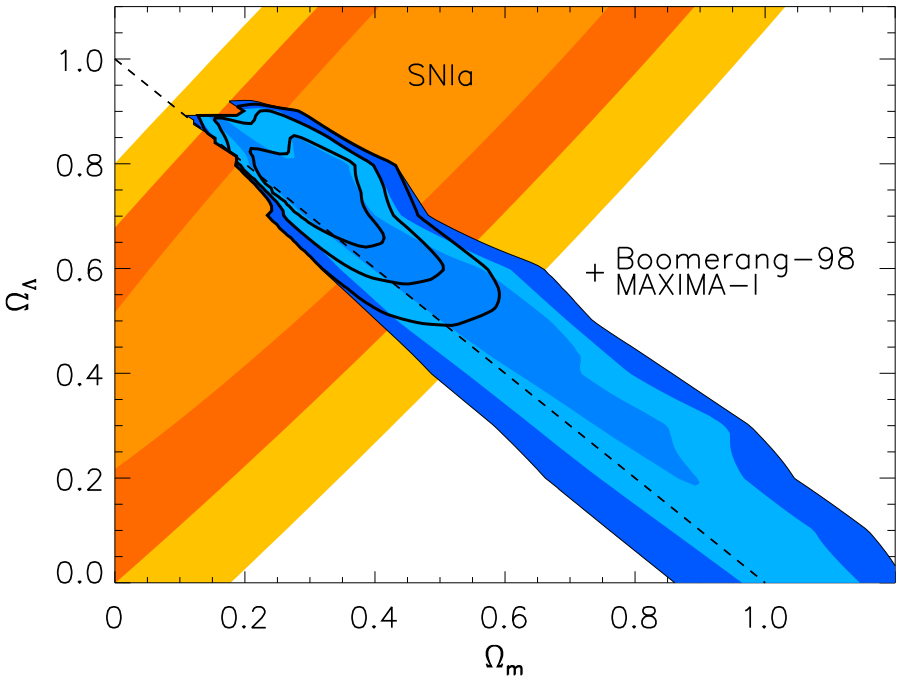


Fig. 22: The phase space of the Universe as determined from supernovae observations and from CMB. From ref. [64].

strong constraints on the SUSY relic coming from the Heidelberg-Moscow experiment [67] and unusual events seen by the DAMA collaboration [68].

There are also indirect methods for WIMP searches. Dark matter particles may be trapped by the Earth or by the Sun and concentrate in their centers. Annihilation of WIMPs will result in $\nu\bar{\nu}$ production in the centers of the Sun and Earth. These neutrinos can be detected by large neutrino telescopes, such as Amanda [69] or Antares [70]. To be compatible with the direct searches, the volume of neutrino telescope should be of the order of 1 km^3.

In addition to WIMPs, there are some "new" particle-physics candidates that have been suggested during the last three years. The first one is a superheavy particle relic. Usually it is assumed that dark-matter particles are relatively light. The reason is that if stable particles are in thermal equilibrium, their freezing concentration can be readily computed if the annihilation cross-section is known. Using the unitarity bound on the annihilation cross-section and requiring that these particles do not overclose the universe one can derive an upper bound on their mass, $M < 500$ TeV [71]. A possible loop-hole in this consideration is that the particles could never be in thermal equilibrium, so that their abundance may always be smaller than the freezing concentration in the equilibrium case. A hypothesis that superheavy particles play the role of dark matter requires that these particles be stable on the cosmological scale and that they be produced in sufficient amounts during the evolution of the Universe. In principle, these particles could be thermally created after inflation, provided their masses were not very much larger than the reheating temperature [72, 73]. Heavy particles might also be produced because of inflaton oscillations in parametric resonance [15]. Finally, they might be produced gravitationally [73, 74, 75].

Another example of a superheavy relic is provided by supersymmetric non-topological solitons. The existence of Q-balls is a generic feature of SUSY models [76]. In supersymmetric theories with low-energy SUSY breaking and flat directions in the effective potential ($V(\phi) \propto V_0$ as $\phi \to \infty$) there are stable states, Q-balls, that carry baryon number [77, 78]. The mass of a soliton with baryon charge B varies as [77] $M \sim V_0^{\frac{1}{4}} B^{\frac{3}{4}}$. Thus, non-topological solitons cannot decay into baryons for $B > V_0/m_p^4$

where m_p is the proton mass. In a number of gauge-mediated models, V_0 can be as small as $(1\ \mathrm{TeV})^4$ and Q-balls are stable with respect to decay into protons if their charge is greater than 10^{12}. Supersymmetric Q-balls might be produced in the early Universe through the decay of Affleck-Dine condensate together with ordinary baryons [78]. The search for Q-balls is possible [79, 80] in a number of working and projected experimental installations, such as Baikal, Macro, SuperK, Antares, etc., and some constraints are already available [81, 82].

6.3 The cosmological constant problem

The set of observations discussed above suggests that the cosmological constant is nonzero and positive, $\Omega_\Lambda \simeq 0.7$. This value of the cosmological constant corresponds to the energy density

$$V_0 = \left(\frac{\Lambda}{8\pi G}\right)^{\frac{1}{4}} \sim 10^{-3}\ \mathrm{eV} \sim 0.01\ \mathrm{cm}^{-1}\ . \tag{86}$$

Does this mean that a new scale in physics has been discovered?

The nonzero cosmological constant introduces several fine-tuning problems. The first comes from comparison between the magnitude of the scale (86) and other known scales, for example Λ_{QCD}, M_W, M_{GUT} and M_{Pl}. Why is the scale associated with Λ so small compared with the other scales?

Another problem arises from comparison of Ω_M and Ω_Λ. At the present stage of expansion of the Universe they are of the same order of magnitude. However, matter energy varies as $\rho_M \sim 1/a^3$, radiation energy varies as $\rho_\gamma \sim 1/a^4$, whereas vacuum energy does not change during the expansion of the universe. In other words, $\Omega_\Lambda \sim 10^{-120}$ on the Planck scale, and the question is whether this huge hierarchy could have any physical explanation.

Using a Λ term to fit the cosmological data is the simplest (but certainly not the only) possibility. It would also be reasonable to assume that extra matter in the Universe has a more general equation of state [83, 84],

$$p = \omega\epsilon\ , \tag{87}$$

where p is pressure, ϵ is an energy density and ω is a constant to be determined from observations. Equation (87) can be used to fit the cosmological parameters globally. The use of a set of cosmological observations (such as the Hubble constant, fraction of baryon mass, cluster abundance, age of the Universe, mass power spectrum, supernovae data, gravitational lensing and large scale structure) implies that ω is negative, $\omega < -0.5$ [85], see Fig. 23.

A model for a substance with a general equation of state can be provided by a uniform scalar field ϕ [86]-[95], for which

$$\omega = \frac{\dot\phi^2/2 - U(\phi)}{\dot\phi^2/2 + U(\phi)}\ . \tag{88}$$

Depending on the field evolution, one can set $-1 \le \omega \le 1$, where $\omega = -1$ and $\omega = +1$ correspond to the dominance of potential and kinetic energies, respectively.

At the present time an explanation of the first problem (why the cosmological constant is so small) is absent, whereas some solutions of the second problem (why energy density is roughly the same as the matter energy density) have been suggested [86]-[95]. The main idea is that the extra substance in the Universe is not a Λ-term but a time-dependent scalar field with an unusual type of potential (ground state at $\phi \to \infty$). The evolution of the scalar field is such that it adapts its energy to the energy contained in matter, so that its late time evolution practically does not depend on the initial conditions and is established dynamically (the so-called attractor solutions).

A simple example is provided by the exponential potential $V(\phi) = V_0 \exp(-\lambda\phi/M_{Pl})$ [86, 88, 91]. This potential does not have a ground state and $\dot\phi$ is nonzero at all times, see Fig. 24. The equations

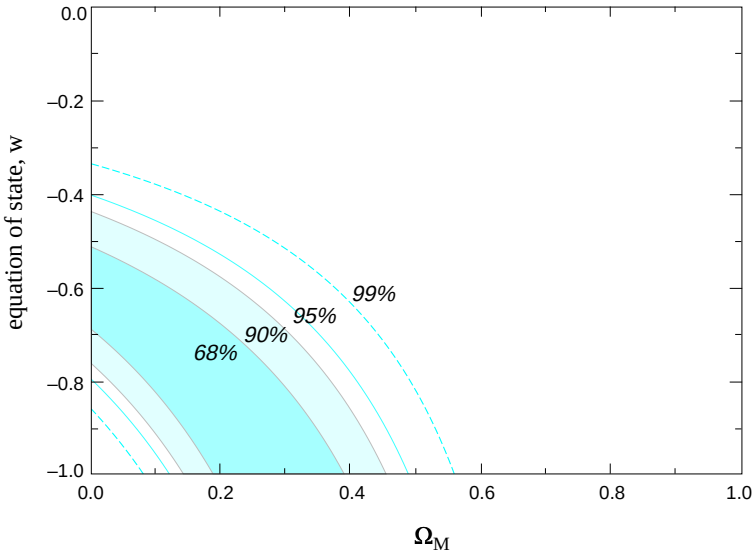


Fig. 23: Constraints on the equation of state from supernova data. From ref. [85].

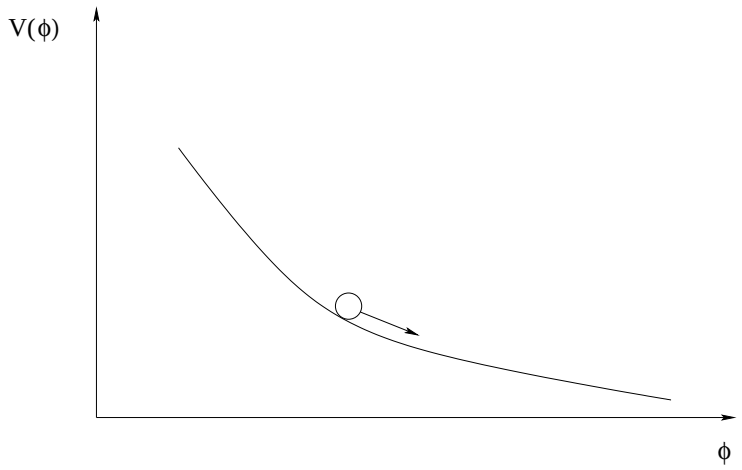


Fig. 24: The run-away potential used in quintessential models.

of motion (in conformal time, $ad\tau = dt$) are:

$$\frac{1}{a^2}\frac{\mathrm{d}}{\mathrm{d}\tau}(a^2\dot{\phi}) + a^2 V'(\phi) = 0\ , \tag{89}$$

$$H^2 = \frac{1}{3M_{Pl}^2}(\frac{1}{2}\dot{\phi}^2 + a^2 V(\phi) + a^2\rho_n)\ . \tag{90}$$

Here ρ_n is a matter density that obeys the equation of state:

$$\dot{\rho}_n + nH\rho_n, \quad \rho_n \sim 1/a^n\ , \tag{91}$$

with $n = 3$ for matter and $n = 4$ for radiation. This system of equations has the following attractor solutions: if $\rho_\phi \gg \rho_n$ then ρ_ϕ scales faster than $1/a^n$, but if $\rho_\phi \ll \rho_n$ then ρ_ϕ varies more slowly than $1/a^n$ (ρ_ϕ is an energy density of field ϕ). This means that after some time ρ_ϕ varies as a dominant component and

$$\Omega_\phi = \frac{\rho_\phi}{\rho_\phi + \rho_n} = \frac{n}{\lambda^2}\ . \tag{92}$$

In other words, the scalar field is self-tuned to the energy density of the dominant component, see Fig. 25. Thus, the value of Ω_ϕ is now not related to initial conditions (as was the case for the cosmological constant problem) but to parameters of potential. For this particular model the parameter ω in the equation of state is that of the dominant matter component, i.e. $\omega = 0$ at the present time, which is in disagreement with the analysis of [84]. Other types of potentials may give different types of behavior: in particular it is possible to obtain a negative value of ω at the present stage of expansion of the Universe, which is consistent with observations [93]. The main problem of this approach is that it does not give any solution to the cosmological constant problem, while the particle-physics origin of a scalar field with the necessary dynamics (existence of a correct attractor solution) remains to be found.

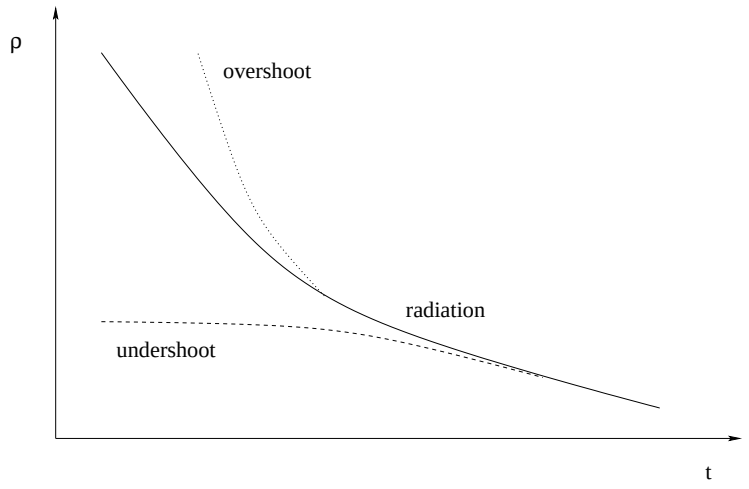


Fig. 25: Attractor solutions.

7 CONCLUSIONS

According to cosmological observations, most of the energy density of the Universe ($\sim 95\%$) cannot be described by the physics of the Standard Model. Roughly a third of this energy density can cluster and represents cold dark matter with unknown particle content, while almost two thirds of the energy appears to be uniform. There are quite a few theoretical proposals for cold dark matter particles and for the uniform component, but the existing cosmological data is not sufficient for a choice to be made. Future experiments, such as Planck and MAP in the area of cosmic microwave background, are extremely important as they will bring cosmology into a new era of precision and define constraints on drastically different cosmological and particle-physics models and even, perhaps, single one out.

References

[1] S. Weinberg, Gravitation and Cosmology : Principles and Applications of the General Theory of Relativity, John Wiley and Sons; 1972

[2] Ya. B. Zeldovich and I.D. Novikov, The Structure and Evolution of the Universe : Relativistic Astrophysics, University of Chicago Press, 1983

[3] E.W. Kolb and M.S. Turner, The Early Universe, Addison-Wesley, Reading, MA 1990

[4] G.A. Tammann, astro-ph/9805013

[5] G. F. Smoot, astro-ph/9705101.

[6] K. A. Olive, astro-ph/9901231.

[7] J. Alcaraz *et al.* [AMS Collaboration], Phys. Lett. **B461** (1999) 387

[8] G. Steigman, Annu. Rev. Astr. Ap. **14** (1976) 339.

[9] A.G. Cohen, A. De Rujula and S.L. Glashow, Astrophys. J. **495** (1988) 539.

[10] A. D. Sakharov. JETP Lett. **6** (1967) 24.

[11] V.A. Kuzmin. Pisma ZhETF **13** (1970) 335.

[12] S. Weinberg, Phys. Rev. Lett. **48** (1982) 1303;
J.Ellis, A.D. Linde and D.V. Nanopoulos, Phys. Lett. **118B** (1982) 59.

[13] L. Kofman, A. Linde and A. A. Starobinsky, Phys. Rev. Lett. **73** (1994) 3195;
Phys. Rev. **D56** (1997) 3258.

[14] E.W. Kolb, A. Linde and A. Riotto, Phys. Rev. Lett. **77** (1996) 4290

[15] E.W. Kolb, A. Riotto and I.I. Tkachev, Phys. Lett. **B423** (1998) 348

[16] S. Khlebnikov, L. Kofman, A. Linde, I. Tkachev. Phys. Rev. Lett. **81** (1998) 2012.

[17] V. A. Kuzmin, V. A. Rubakov, and M. E. Shaposhnikov. Phys. Lett., **155B** (1985) 36

[18] G. 't Hooft. Phys. Rev. Lett., **37** (1976) 8

[19] P. Arnold, D. T. Son and L. G. Yaffe, Phys. Rev. **D59** (1999) 105020

[20] P. Arnold, L.G. Yaffe. Phys. Rev. **D57** (1998) 1178.

[21] V.A. Rubakov and M.E. Shaposhnikov. Phys. Usp. **39** (1996) 461 (Usp. Fiz. Nauk **166** (1996) 493).

[22] A.G. Cohen, D.B. Kaplan, and A.E. Nelson. Annu. Rev. Nucl. Part. Sci. **43** (1993) 27.

[23] A. Riotto and M. Trodden, Ann. Rev. Nucl. Part. Sci. **49** (1999) 35

[24] K. Kajantie et al, Nucl. Phys. **B532** (1998) 283

[25] Y. Aoki, F. Csikor, Z. Fodor and A. Ukawa, Phys. Rev. **D60** (1999) 013001

[26] M. Carena, M. Quirós and C.E.M. Wagner. Phys. Lett. **B380** (1996) 81;
M. Laine. Nucl. Phys. **B481** (1996) 43;
J. M. Cline and K. Kainulainen. Nucl. Phys. **B482** (1996) 73;
M. Losada. Phys. Rev. **D56** (1997) 2893;
J.R. Espinosa. Nucl. Phys. **B475** (1996) 273;
D. Bödeker, P. John, M. Laine and M.G. Schmidt. Nucl. Phys. **B497** (1997) 387;
M. Laine , K. Rummukainen. Phys. Rev. Lett. **80** (1998) 5259;
M. Carena , M. Quiros, C.E.M. Wagner. Nucl.Phys. **B524** (1998)3.

[27] M. Laine, hep-ph/0010275.

[28] Y. Fukuda *et al.* [Super-Kamiokande Collaboration], Phys. Lett. **B436** (1998) 33

[29] Y. Fukuda *et al.* [Super-Kamiokande Collaboration], Phys. Rev. Lett. **81** (1998) 1158

[30] T. Yanagida, Report KEK-79-18 (1979)

[31] M. Gell-Mann, P. Ramond and R. Slansky, in “Supergravity” (North-Holland, Amsterdam, 1979).

[32] M. Fikugita and T. Yanagida, Phys. Lett. **B174** (1986) 45

[33] W. Buchmuller and M. Plümacher, hep-ph/9904310

[34] G.F. Giudice, M. Peloso, A. Riotto and I. Tkachev, JHEP **08** (1999) 014.

[35] W. Buchmuller and M. Plumacher, hep-ph/0007176.

[36] I. Affleck, M. Dine. Nucl. Phys., **B249** (1985) 361

[37] M. Dine, L. Randall and S. Thomas. Nucl. Phys., **B458** (1996) 291

[38] A. de Gouvêa, T. Moroi and H. Murayama, Phys. Rev., **D56** (1997) 1281

[39] A. H. Guth, Phys. Rev. **D23** (1981) 347.

[40] A. D. Linde, Phys. Lett. **B108** (1982) 389.

[41] A. Albrecht and P. J. Steinhardt, Phys. Rev. Lett. **48** (1982) 1220.

[42] A.D. Linde, “Particle physics and inflationary cosmology”, Harwood Academic Press (1980)

[43] A. R. Liddle and D. H. Lyth, “Cosmological inflation and large-scale structure,” Cambridge, UK: Univ. Pr. (2000).

[44] A. D. Linde, Phys. Lett. **B129** (1983) 177.

[45] W. L. Freedman, Phys. Scripta **T85** (2000) 37.

[46] N. Bahcall, J. P. Ostriker, S. Perlmutter and P. J. Steinhardt, Science **284** (1999) 1481.

[47] J. P. Ostriker, J.P.E. Peebles and A. Yahil, ApJ **193** (1974) L1

[48] N.A. Bahcall, L.M. Lubin and V. Dorman, ApJ **447** (1995) L81

[49] R.G. Carlberg et al., ApJ **462** (1996) 32

[50] N. Kaiser et al, astro-ph/9809268

[51] T.P. Walker et al, ApJ **376** (1991) 51

[52] D.N. Schramm and M.S. Turner, Rev. Mod. Phys. **70** (1998) 303

[53] S.D.M. White et al., Nature **366** (1993) 429

[54] R. Cen and J.P. Ostriker, ApJ **429** (1994) 4

[55] L.M. Lubin et al, ApJ **460** (1996) 10

[56] N.A. Bahcall and X. Fan, ApJ **504** (1998) 1

[57] K.G. Begeman, A.H. Broeils and R.H. Sanders, Mon. Not. R. Astr. Soc. **249** (1991) 523.

[58] S. Perlmutter *et al.* [Supernova Cosmology Project Collaboration], Astrophys. J. **517** (1999) 565.

[59] A.G. Riess et al, Astron.J. **116** (1998) 1009

[60] G. Smoot et al, ApJ, **396** (1992) L1

[61] G. F. Smoot, astro-ph/9705135.

[62] A. E. Lange *et al.*, astro-ph/0005004.

[63] A. Balbi *et al.*, astro-ph/0005124.

[64] A.H. Jaffe et al, astro-ph/0007333.

[65] F. Derue et al. astro-ph/9903209.

[66] H. Cheng, Phys. Rept. **158** (1988) 1.

[67] L. Baudis *et al.*, Phys. Rev. **D59** (1999) 022001

[68] R. Bernabei *et al.*, Phys. Lett. **B424** (1998) 195.

[69] R. Wischnewski *et al.*, Nucl. Phys. Proc. Suppl. **85** (2000) 141.

[70] J. J. Hernandez [ANTARES Collaboration], Nucl. Phys. Proc. Suppl. **81** (2000) 174.

[71] K. Griest and M. Kamionkowski, Phys. Rev. Lett. **64**, 615 (1990).

[72] D.J. Chung, E.W. Kolb and A. Riotto, Phys. Rev. **D59** (1999) 023501.

[73] D.J. Chung, E.W. Kolb and A. Riotto, Phys. Rev. Lett. **81** (1998) 4048

[74] V. Kuzmin and I. Tkachev, Phys. Rev. **D59** (1999) 123006

[75] V. Kuzmin and I. Tkachev, JETP Lett. **68** (1998) 271

[76] A. Kusenko, Phys. Lett. **B405** (1997) 108

[77] G. Dvali, A. Kusenko and M. Shaposhnikov, Phys. Lett. **B417** (1998) 99

[78] A. Kusenko and M. Shaposhnikov, Phys. Lett. **B418** (1998) 46

[79] A. Kusenko, V. Kuzmin, M. Shaposhnikov and P.G. Tinyakov, Phys. Rev. Lett. **80** (1998) 3185

[80] E. Aslanides *et al.* [ANTARES Collaboration], astro-ph/9907432.

[81] I.A. Belolaptikov *et al.*, astro-ph/9802223.

[82] M. Ambrosio *et al.* [MACRO Collaboration], hep-ex/9904031.

[83] R.R. Caldwell, R. Dave and P.J. Steinhardt, Phys. Rev. Lett. **80** (1998) 1582

[84] L. Wang, R.R. Caldwell, J.P. Ostriker and P.J. Steinhardt, astro-ph/9901388.

[85] S. Perlmutter, M. S. Turner and M. White, Phys. Rev. Lett. **83** (1999) 670 [astro-ph/9901052].

[86] B. Ratra and P.J. Peebles, Phys. Rev. **D37** (1988) 3406.

[87] P.J. Peebles and B. Ratra, Astrophys. J. **325** (1988) L17.

[88] C. Wetterich, Nucl. Phys. **B302** (1988) 645.

[89] C. Wetterich, Nucl. Phys. **B302** (1988) 668.

[90] P.G. Ferreira and M. Joyce, Phys. Rev. **D58** (1998) 023503

[91] P.G. Ferreira and M. Joyce, Phys. Rev. Lett. **79** (1997) 4740

[92] P.J. Steinhardt, L. Wang and I. Zlatev, Phys. Rev. **D59** (1999) 123504

[93] I. Zlatev and P.J. Steinhardt, Phys. Lett. **B459** (1999) 570

[94] I. Zlatev, L. Wang and P.J. Steinhardt, Phys. Rev. Lett. **82** (1999) 896

[95] G. Huey, L. Wang, R. Dave, R.R. Caldwell and P.J. Steinhardt, Phys. Rev. **D59** (1999) 063005

THE SEARCH FOR THE QUARK–GLUON PLASMA

Jorge Dias de Deus
Departamento de Física/CENTRA, IST, Lisboa, Portugal

Abstract
A short review is presented on the recent searches for the quark–gluo plasma (QGP), including theoretical motivation, the case for the Little Bang and the official QGP signatures. A discussion of the first RHIC results is also included.

1. INTRODUCTION

As far as strong interactions are concerned, we live in a world of hadrons; they appear in the atomic nuclei of normal matter, in galactic and extra-galactic cosmic rays from the sky, in radioactive decays of elements on earth; they are the nucleons and the pions of old nuclear physics and, in general, they are the baryons and the mesons registred in the particle physics booklet [1].

But we know that according to the theory, Quantum Chromo Dynamics, QCD, this world is, as well, a world of quarks and gluons. Our every day experience, however, only shows evidence for normal hadronic matter. Not quarks and gluons: just their bound states, i.e., hadrons.

Why is it so? Because freedom is only local and confinement dominates. In order to illustrate what the problem is, let us consider the conventional linear potential

$$V(r) = \sigma\ r\ , \tag{1}$$

σ being the string parameter and r the distance between the 3 and $\bar{3}$ colour triplets (for instance, quarks). The idea here is that the flux of the field lines, between 3 and $\bar{3}$, are limited to a constant area (the transverse area of the string) thus requiring (by Gauss theorem) the field σ to be constant. As the field σ does not decrease with distance the 3 – $\bar{3}$ system never ionizes and colour is not directly observed. This is the basic argument for confinement.

However, at a given finite temperature T_c, according to lattice QCD calculations [2], there is a phase transition from hadronic matter to quark–gluon matter, or a transition to the quark–gluon plasma (QGP). In terms of the potential (1) it amounts to induce an r dependence on σ such that the string tension vanishes,

$$\sigma \underset{T/T_c>1}{\longrightarrow} \frac{1}{r}\ . \tag{2}$$

The natural order paramenter for the transition is the Polyakov loop *L*(T), [2],

$$L(T) \sim\ e^{-V(r)/T}\ , \tag{3}$$

such tha

$$L(T) \underset{r\to\infty}{\sim} \mathbf{0}\ ,\ T/T_c < \mathbf{1} \qquad \mathbf{(4)}$$

and

$$L(T) \underset{r\to\infty}{\sim} const. > \mathbf{0}\ ,\ T/T_c > \mathbf{1}\ . \qquad \mathbf{(5)}$$

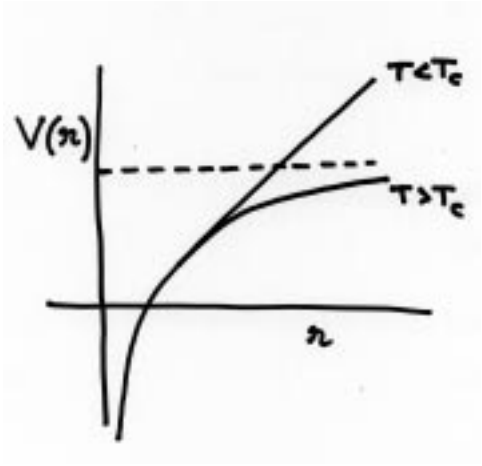


Fig. 1

In Fig. 1 we illustrate what happens to the state 3 – $\bar{3}$**, which is bound for** $T < T_c$**, and becomes free for** $T > T_c$**, the potential being no longer confining. In Fig. 2 we present the result of a lattice QCD calculation showing the abrupt transition of the order parameter** L**(T), for** $T = T_c$**. The transition for the derivative of** L**, the susceptibility, is also shown [3].**

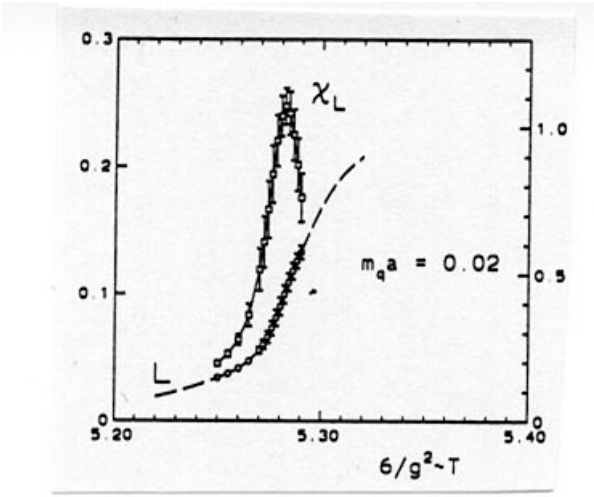


Fig. 2

Note that, again according to the theory, another transition, the chiral symmetry restoration, is taking place at a temperature T_χ (of the order of T_c), the order parameter being in this case the $\langle q\bar{q}\rangle$ condensate — measuring the value of the quark mass.

The deconfinement transition is a transition from an organized state (hadrons) to a less organized state (quarks and gluons). In general, to achieve this kind of transition, one has to heat or/and to compress the system. Heating means weakening the bounds between constituents, compressing meaning loosing the individuality of the structures. If in a given volume V one increases the size of hadrons (by increasing T) or increases the density (by increasing p) at some stage one reaches a situation similar to a percolation problem with the originally confined quarks and gluons starting to circulate over the whole of the interaction region (see Fig. 3). In a sense, deconfinement can also be seen as a transition from a colour insulator to a colour conductor.

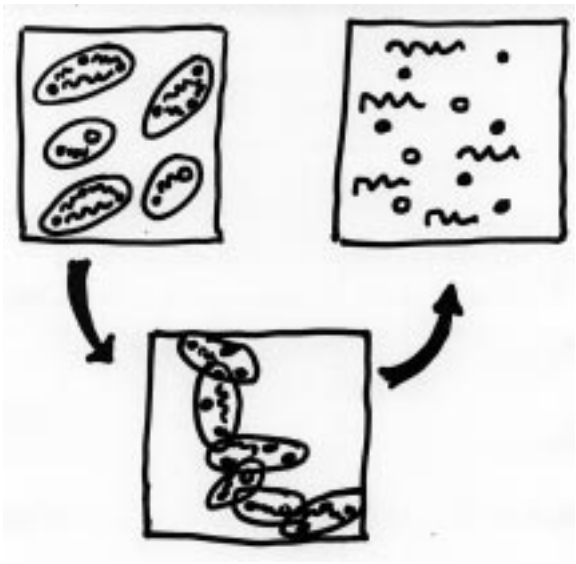


Fig. 3

In Fig. 4 we show the phase diagram, in the $T-\mu$ plane, where T is the temperature and μ the baryonic chemical potential, for the deconfinement transition. Theoretically, there is some room left for a colour superconductor phase with diquarks (quark–quark Cooper pairs).

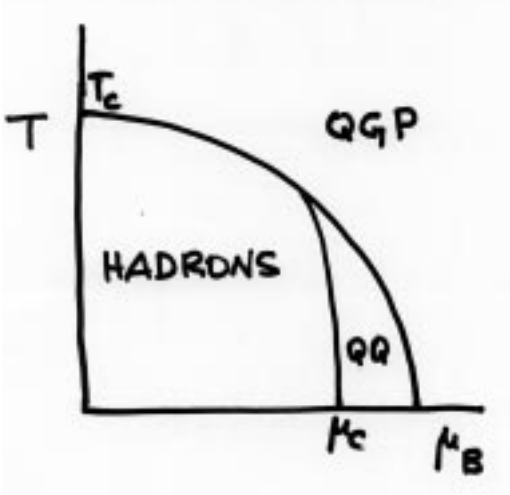


Fig. 4

Now we arrive at the main subject of the talk: is there any evidence for the transition from hadrons to quark–gluon plasma (or vice-versa)?

The natural laboratory to study the transition is the Universe as a whole: the transition should have ocurred in the first micro-seconds after the Big-Bang [4]. In fact, going back in time means heating and compressing and one should then reach situations with higher energy and mass densities. Do we have any evidence for the (de)confinement transition from the Big-Bang evolution? The answer seems to be no: additional transitions (like nucleosynthesis) and equilibrium (allowed by the slowness of the expansion rate) have erased most of past history. Primordial nucleosynthesis (the first 3 minutes) and the microwave background radiation (400 000 years later) are the best relics we can have today of that distant past.

However, in the laboratory we are now trying to recreate the conditions for the transition: the Little Bang.

2. THE LITTLE BANG

The idea is to produce concentrated nuclear matter in a small region of space and time, in order to achieve the high mass/energy densities (high temperature, if equilibrium exists), high enough to reach the deconfinement transition. The procedure is to shoot heavy ions against heavy ions making for that use of accelerators. The evolution, in space and time, of this concentrate of hadronic matter is expected to reveal the characteristic features of the QGP.

There has been a competition between Europe and USA, or between CERN and Brookhaven, with machine's energy growing fast with time. See the Table 1.

Table 1

MACHINE	$\sqrt{s}$ A GeV
AGS	~5
SPS	~20
RHIC (2000)	~200
LHC (2005)	~5000

One has to know, to start with, if enough incoming energy (forward/backward longitudinal energy) is transformed into transverse energy E_T (due mostly to small centre of mass rapidity production).

Several tests can be done:

i) Stopping power

One measures the net baryon number, $B-\overline{B}$, distribution in rapidity and sees how much of the incident nucleon energy was transfered into particle and E_T production. See Fig. 5.

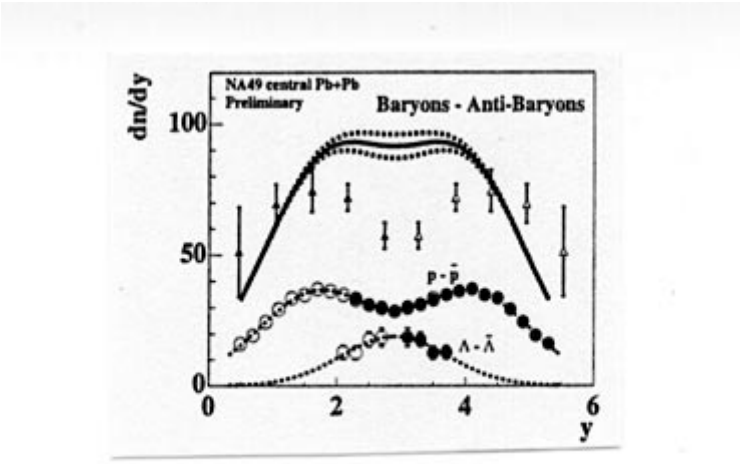


Fig. 5

Note that the incoming baryon distributions are at the centre of mass rapidity edges (forward and backward maximum rapidity). There is a clear strong stopping effect, in particular in heavy nuclei central collisions.

ii) Energy in forward–backward calorimeters

The ZDCs (Zero Degree Calorimeters) give information about the non-participating nucleons, $A-N_P$, in a collision. This allows the selection of highly central (small impact parameter) events. There is, as expected, a clear negative correlation between E_T and E_{ZDC}: central collisions, $E_{ZDC} \cong 0$, mean higher values of E_T (Fig. 6).

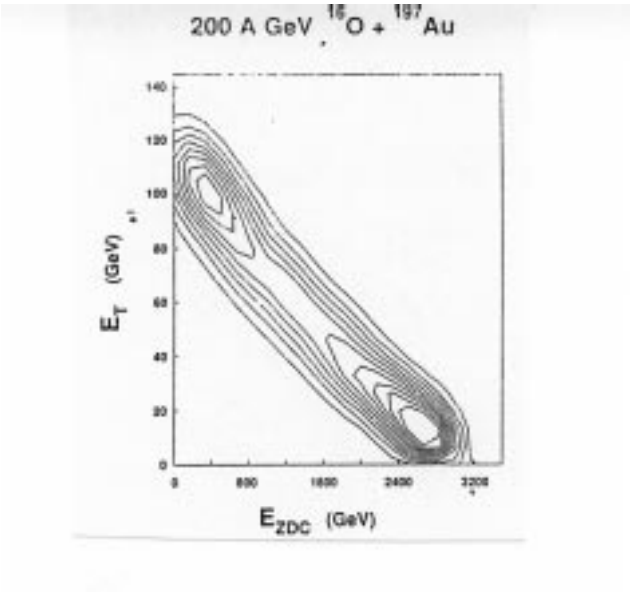


Fig. 6

iii) E_T **and multiplicity distributions**

E_T **and** *n* **distributions are very wide, in minimum bias events, and extend to very large values of** E_T **and** *n***. In general, the** *n* **and** E_T **distributions are quite similar, reflecting the statistics of nucleon collision distributions. For central collisions (** $E_{ZDC} \cong 0$ **) again the** E_T **and** *n* **distributions, in equivalent rapidity bins, are similar (Fig. 7). These results suggest that we are reaching a situation where high densities are created and general statistical fluctuations dominate.**

Let us suppose that the QGP is formed. What will happen next? How will we ever know that the plasma really existed?

Concerning the question about what will happen next, the answer is simple: the plasma will expand and cool with the deconfinement transition taking place at $T_c \cong 180$ **MeV, and the formation of an interacting gas of hadrons, with a chemical composition practically fixed, and finally reaching the interaction freezing temperature** $T_f \cong 120$ **MeV and the hadrons becoming free and flying away towards the detectors. See Fig. 8.**

The temperature *T* **is the parameter that controls chemical equilibrium and the average transverse mass** $m_T, m_T = \sqrt{m_h^2 + P_T^2}$ **, where** m_h **is the hadron mass and** P_T **the transverse momentum. In particular,** $T_f \sim \langle m_T \rangle$ **. However, as the particles approach the detectors, the temperature (like frequency) is blue shifted and the detectors measure not** T_f **but an effective temperature** T_e **with, in a non-relativistic Doppler approximation,**

$$T_e = T_f + m_h \langle v_T^2 \rangle \qquad \textbf{(6)}$$

where $\left\langle v_T^2 \right\rangle$ is the transverse expansion parameter (equivalent to the Hubble constant). Eq. (6) tells us that $T_e \to T_f$ when $m_h \to 0$ and that T_e increases linearly with the mass of the hadron. Data are not in detail consistent with (6) — there are also inconsistencies when comparing different experiments — but qualitatively, are quite suggestive (see Fig. 9). In Ref. [5] there is a much deeper discussion on the Little Bang evolution.

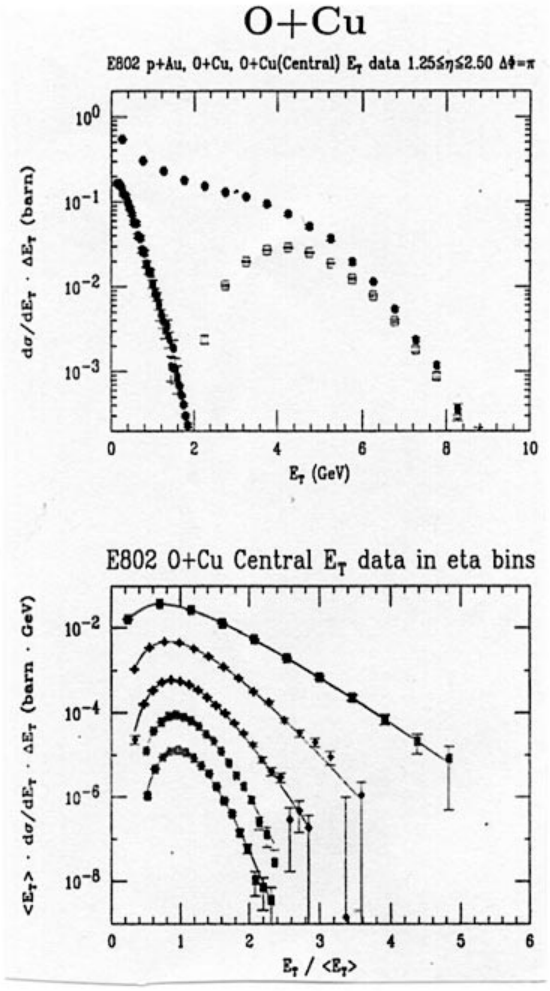


Fig. 7A

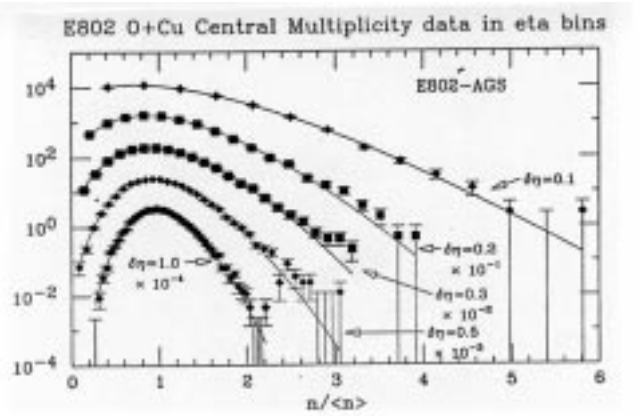

E802 O+Cu Central Multiplicity data in eta bins
E802-AGS
δη=0.1
δη=0.2
δη=0.3
δη=0.5
δη=1.0
n/<n>

Fig. 7B

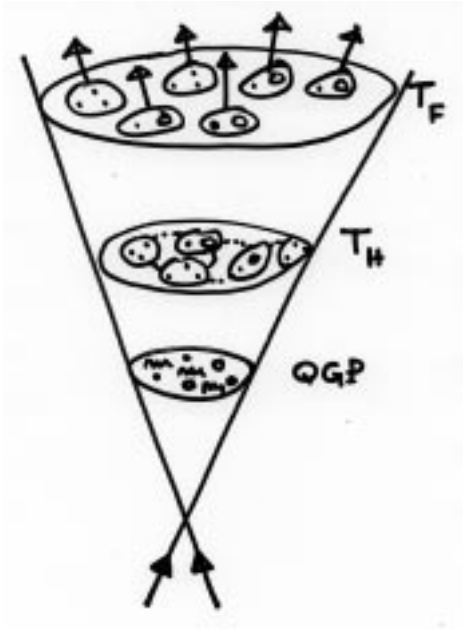

T_F
T_H
QGP

Fig.8

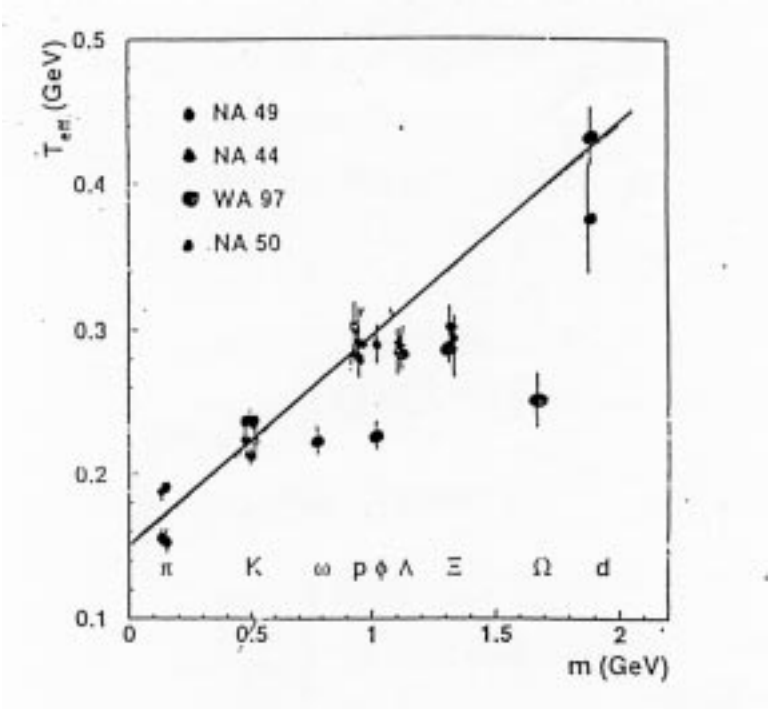


Fig. 9

3. THE SIGNATURES

According to the CERN press release of February 2000 the QGP (to be more precise: "a new state of matter") had been seen at the CERN Super Proton Synchroton (SPS) in seven different experiments.

The signatures that were considered as convincing ones are:

A – Strangeness enhancement.

B – Di-Lepton excess in low mass region.

C – J/ψ suppression

We shall next consider each one of these signatures.

A – Strangeness enhancement

A clear increase in the multiplicity of strange mesons per participant nucleon as the number of participants increases is observed in *Pb–Pb* collisions (Fig. 10). The effect is even more spectacular in baryon production, the amount of increase increasing with strangeness, 1, 2 or 3 (Fig. 11). Recentely, charm enhancement was also observed [6, 7, 8].

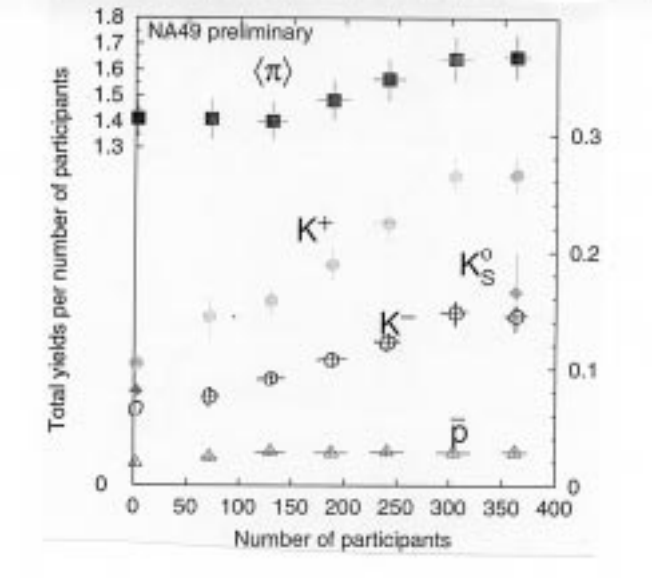


Fig.10

If QGP is formed in the early stages of the collision chemical equilibrium would produce a larger fraction of strange baryons than the one seen in ordinary hadronic collisions. The reason being that the masses of the strange and u and d quarks in the plasma are not so different. Additional hadron–hadron interactions in the following evolution would not substantially change the achieved equilibrium.

However, the argument is not fully convincing as strangeness enhancement also occurs in nucleon–nucleus collisions and even in nucleon–nucleon collisions when triggering on high multiplicities. See [9] for a discussion.

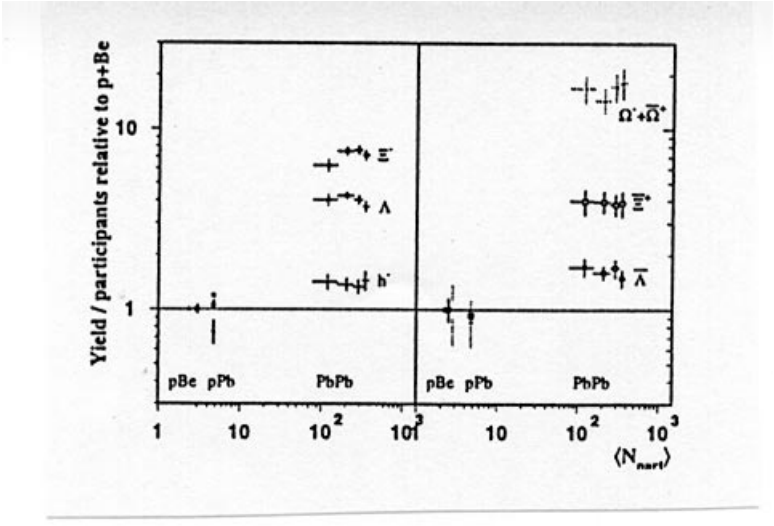


Fig.11

One can find other explanations for the strangeness enhancement. For instance, in the Dual String Model the role of sea quarks and gluons, contrary to valence quarks, increases with energy and density and strangeness is no longer disfavoured. On the other hand, string fusion, [10], leads to larger colour charges at the ends of the strings (ropes) with the consequence that, by Schwinger mechanism, heavier particle production is favoured.

B – Di-lepton excess in low mass region

In the QGP there will be a quantity of moving electric charges (quarks) in a small, transparent region, which can radiate real photons or photons detectable by their decay into di-leptons $(e^+e^-, \mu^+\mu^-)$**. This production competes with normal photon production by several hadronic processes:** $\eta \to e^+e^-\gamma, \omega \to e^+e^-\pi$ **,** $\rho/\omega \to e^+e^-$ **,** $\phi \to e^+e^-, \pi \to e^+e^-\gamma$ **,....**

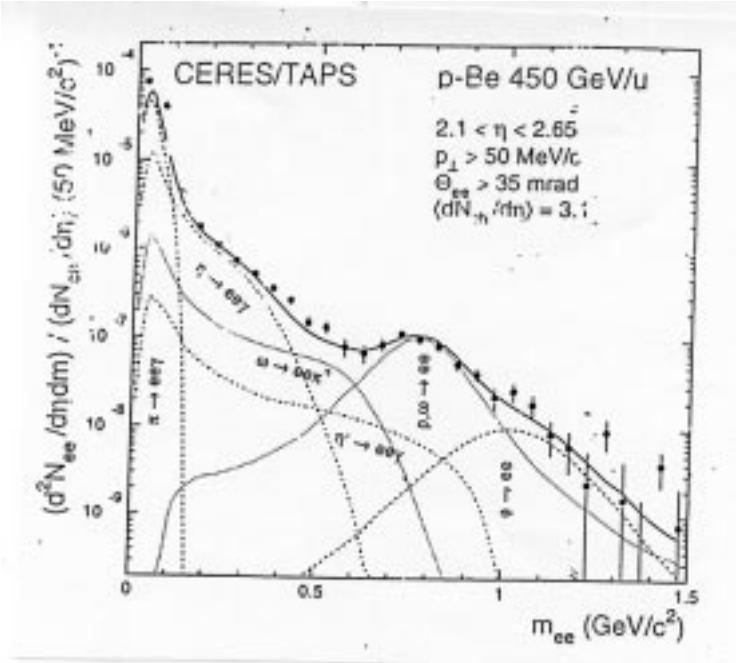


Fig. 12

The CERES collaboration [11] has measured the cross-section $dN/dy \;\; dm_{e^+e^-}$ **normalize to the central (pseudo) rapidity density i** *pBe* **and** *Pb-Au* **and, at the same time, studied the different hadronic processes. While in the** *p-Be* **case the hadronic processes account for the full distribution, in** *Pb-Au* **there is a large discrepancy of almost one order of magnitude (see Figs. 12 and 13). The same result was obtained in other experiments [12].**

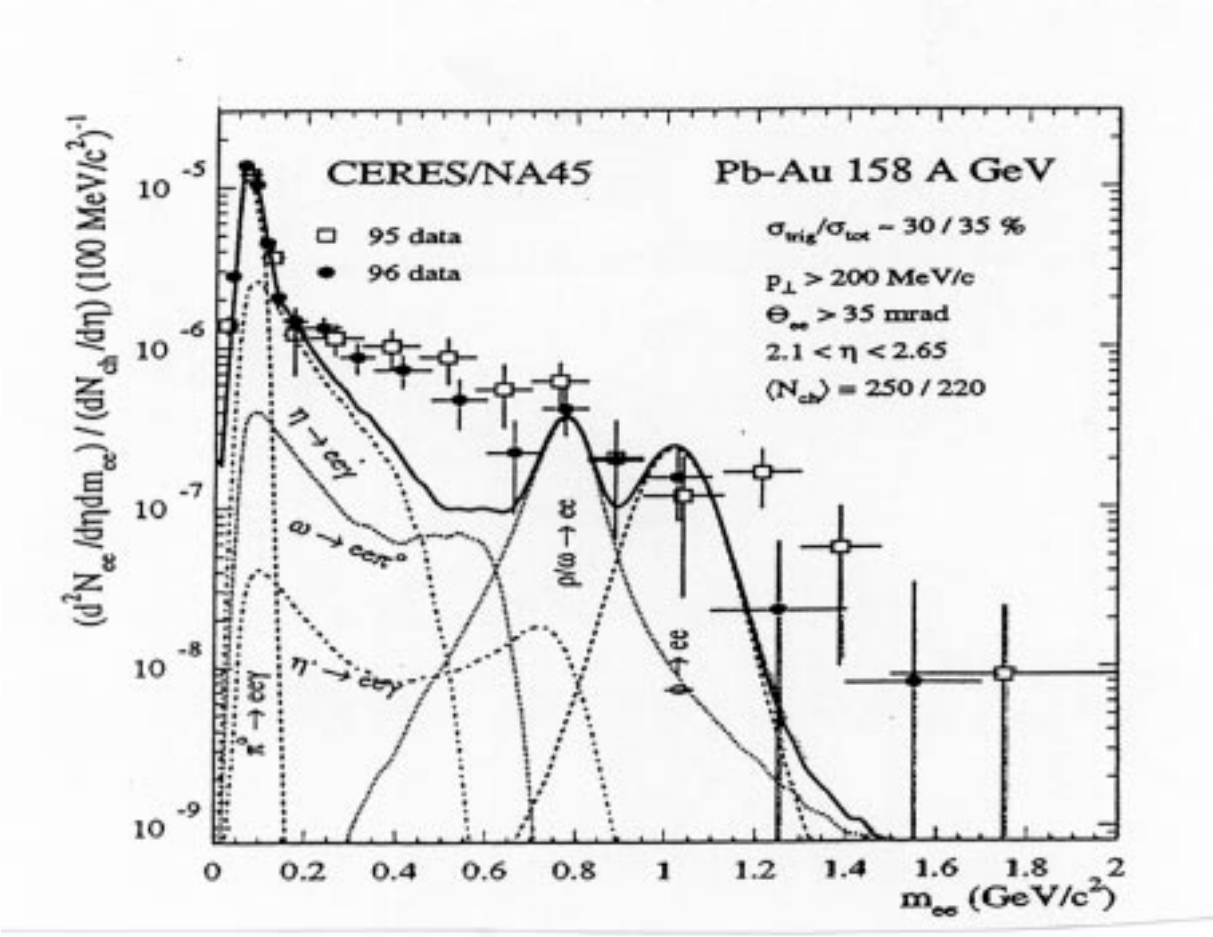


Fig. 13

Chiral symmetry restoration effects decreasing in medium the mass of the ρ and increasing its width were used in an attempt to explain the original CERES result [13]. However, such effects cannot explain the large disagreement found.

Note that in the present case — di-lepton excess — we have an effect which only occurs for heavy nucleus–nucleus collision and, to the extent that no other convincing explanation has appeared, it can be taken as direct evidence for the QGP. Do not forget, however, that thermal radiation from a hot hadronic gas may also be present.

C – J/ψ suppression

This is the most daring prediction and — to most of us — unexpected prediction for evidence of QGP: J/ψ suppression [14]. The prediction was made in 1985, before the first NA38 results [15].

Heavy quark physics, involving quark masses larger than the perturbative QCD scale parameter $\wedge$, can be described perturbatively and the plasma has features similar to electromagnetic plasma. In particular, there will be Debye screening, preventing heavy quarkonium, $c\bar{c}, b\bar{b}$,..., formation. At a given temperature T, T/T_c >1, the Debye radius r_D forbids bound states with radius $r_B > r_D$. As T/T_c increases, r_D decreases and more and more bound states (Υ , for instance) will not be formed (see Fig. 14).

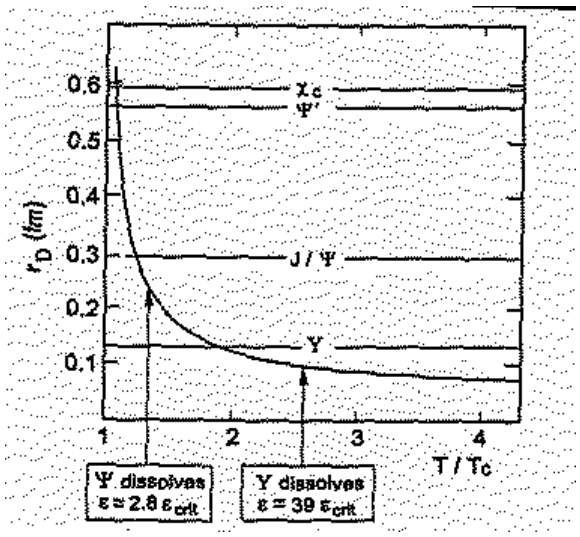


Fig. 14

From di-muon, $\mu^+\mu^-$, analysis the ratio of J/ψ production relative to Drell-Yan production does indeed decrease when going from nucleon–nucleon and nucleon–nucleus to nucleus–nucleus collisions, and it decreases with the increase of the associated transverse energy E_T. As large values of E_T mean centrality and high densities, the decrease of the J/ψ over Drell-Yan ratio is what to expect if plasma is formed.

However, the J/ψ is a strong interacting fragile object, it easily interacts inellastically with the medium and disappears, being absorbed.

The key difference, in the E_T dependence of the J/ψ over Drell-Yan ratio, between QGP (the J/ ψ is not formed) and absorption (the J/ ψ is formed but is killed after) is in the curvuture: QGP formation implies a drop in the E_T distribution (negative curvature) and absorption implies smooth decrease (positive curvature) [16]. See Fig. 15.

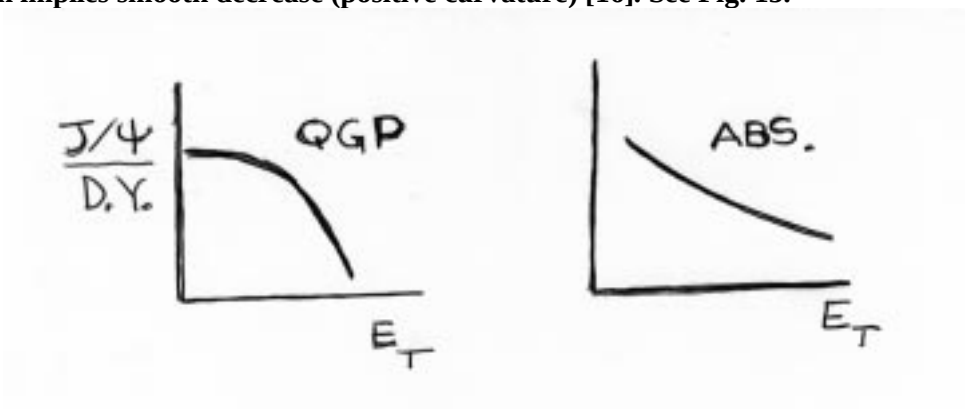


Fig. 15

At some stage, 1995 data, the large E_T data points were enhanced by rescattering in a thick target — thus favouring the absorption interpretation [17]. The situation has changed — the target was made thinner — and the final data [18] are more favourable to the QGP formation interpretation. See Fig.16. A typical absorption calculation, with the typical positive curvature, is also shown in Fig. 16. Sofistications can be introduced in absorption, but they do not substantially change the qualitative behaviour [19].

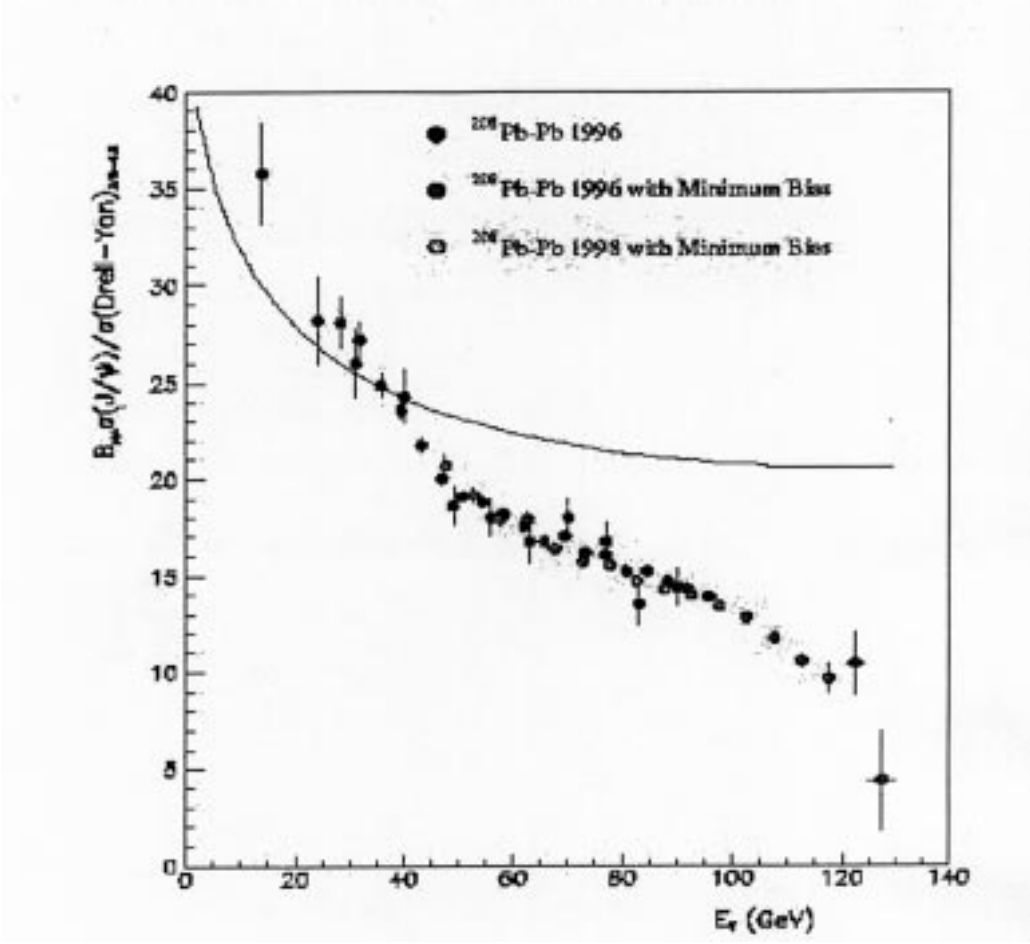


Fig. 16

4. CONCLUSION ABOUT THE QGP SEARCHES

The search for the QGP will continue with RHIC and LHC. The NA60 collaboration will extend the work at the SPS and it may help to fill in the gaps.

Where are we now — Summer 2000? There are positive indications, in particular from low mass di-leptons and J/ ψ' suppresion, that we may have reached the plasma. If it is so, at higher energy and temperature, photons radiated from the plasma will be shinning in an obvious manner. The strangeness enhancement, as mentioned before, remains an open question as a way to separate what is "normal" hadronic physics from QGP physics has not been invented. Finally, the J ψ' suppression is still involved in some ambiguities and it is better to wait the coming results before drawing final conclusions. The predictions of the absorption-percolation model of [20] are given in Fig. 17.

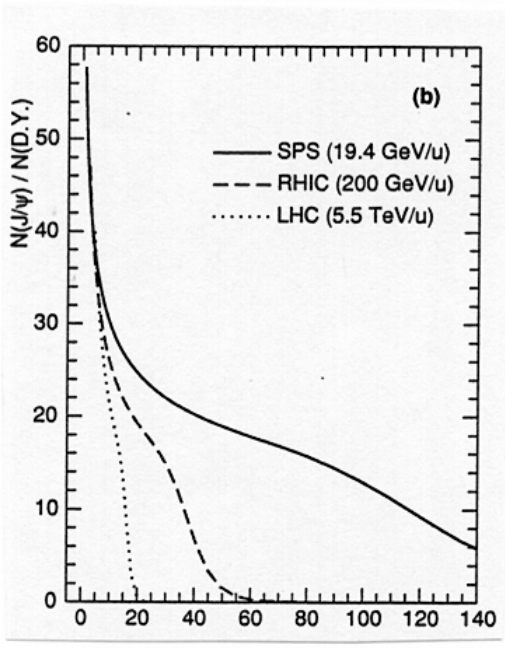


Fig.17

IN ANY CASE THE SEARCH WILL GO ON!

5. RECENT RESULTS FROM RHIC AND THE DUAL STRING MODEL

Recent results on charged particle pseudo-rapidity densities in central Au+Au collisions, at $\sqrt{s} = 56$ and $\sqrt{s} = 130$ AGeV, presented by the PHOBOS Collaboration, at RHIC, [21], give very interesting information that may help to clarify the way the expected Quark–Gluon Plasma (QGP) is approached as the energy increases (see Fig. 18).

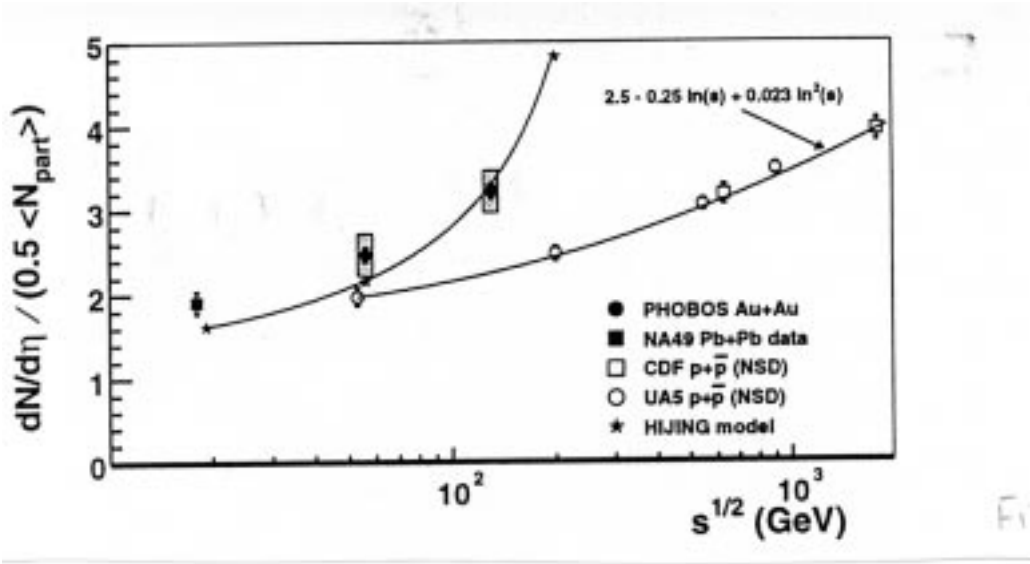


Fig.18

Those data also allow to select among different models of particle production. As in this experiment the charged particle densities and the average number of participating nucleons are simultaneously measured, that provides additional strong constraints to models.

As nuclei are made up of nucleons, it is natural to start by building nucleus–nucleus collisions as resulting from superposition of *nucleon–nucleon* **collisions, in the way it is done in the Glauber model approach and generalisations of it. In one (low energy) limit the nucleons are seen as structureless and emit particles only in their first collision: this is the wounded nucleon model [22]. The prediction for particle density, when** N_A **nucleons from each one of the nuclei in a AA collision participate, is**

$$\left.\frac{dN}{dy}\right|_{N_A N_A} = \left.\frac{dN}{dy}\right|_{pp} N_A \qquad \textbf{(7)}$$

where *dN/dy* **is the particle rapidity (or pseudo-rapidity) density (for** $N_A N_A$ **and nucleon–nucleon collisions). If the nucleon is seen as made up of quarks and gluons, with a growing number of participating sea quarks and gluons as the energy increases, one anticipates dominance of multi-collision processes [23] and the relation**

$$\left.\frac{dN}{dy}\right|_{N_A N_A} = \left.\frac{dN}{dy}\right|_{pp} \nu_{N_A} \qquad \textbf{(8)}$$

to hold, where ν_{N_A} **is the number of nucleon–nucleon collisions when** N_A **nucleons participate. Elementary multi-scattering arguments [24] give**

$$\nu_{N_A} = N_A^{4/3} \qquad \textbf{(9)}$$

In Fig. 19, together with the PHOBOS data, we have presented the quantity $\left.\frac{1}{N_A}\frac{dN}{dy}\right|_{N_A N_A}$ **as**

function of the c.m. energy $\sqrt{s}$ **for the bounds (7) — solid line — and (8) with (9) — dotted line. We used for** $dN/dy|_{pp}$ **the parametrisation** $0.957 + 0.0458\ln\left(\sqrt{s}\right) + 0.0494\ln^2\left(\sqrt{s}\right)$**, with** $\sqrt{s}$ **in GeV, which fits data from** pp **and** $p\overline{p}$ **non-single-diffractive collisions for c.m. energies** $\sqrt{s} \geq 22$ **GeV.**

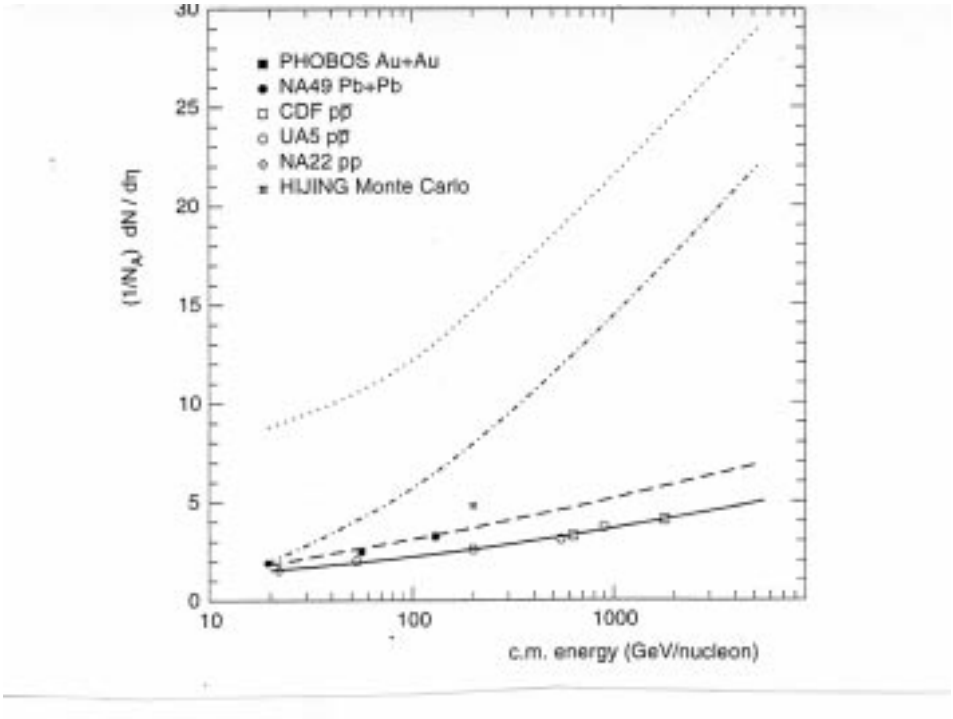


Fig.19

In the Dual String Model (DSM), i.e., the Dual Parton Model [25] with the inclusion of strings [26], the limits referred to above appear in a natural way. The valence quarks of the nucleon produce particles, via strings, only once — this is the wounded nucleon model case — and production is proportional to the number N_A of participant nucleons (Fig. 20a). As the energy and N_A increase the role of sea quarks and gluons increases, they interact and produce, again via strings, particles, and the number of collisions ν becomes the relevant parameter (Fig. 20b).

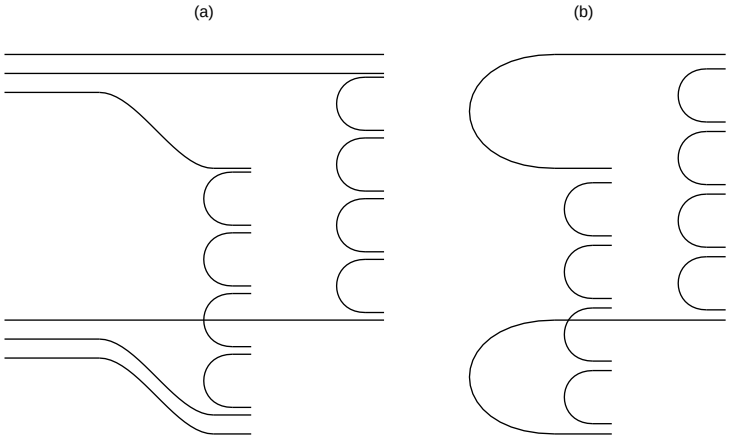


Fig.20

One should notice that the diagram of Fig. 20b may be interpreted as multiple inelastic scattering, either internally within a given nucleon–nucleon collision or externally involving

interactions with different nucleons. On the other hand, this diagram may appear repeated several times.

Following [26], and taking into account the basic diagrams of Fig. 20a and Fig. 20b, we now write an expression for the particle pseudo-rapidity density,

$$\left.\frac{dN}{dy}\right|_{N_A N_A} = N_A\left[2+2(k-1)\alpha\right]h + \left(\nu_{N_A} - N_A\right)2k\alpha h \,, \tag{10}$$

where h **is the height of the valence–valence rapidity plateau,** α **is the relative weight of the sea–sea (including gluons) plateau and** k **is the average number of string pairs per collision. The diagrams of Fig. 20a and Fig. 20b correspond to** $k = 1$**. However, as we mentioned above, the diagram of Fig. 20b can be iterated with** $k \geq 1$ **being, in general, a function of energy. The number of nucleon–nucleon collisions is, of course,**

$$N_A + \left(\nu_{N_A} - N_A\right) = \nu_{N_A} \,, \tag{11}$$

and the number N_s **of strings is**

$$N_s = N_A\left[2+2(k-1)\right] + \left(\nu_{N_A} - N_A\right)2k = 2k\nu_{N_A} \,. \tag{12}$$

The first term on the right-hand side of Eq. (4) is just a sum over nucleon–nucleo scattering contributions (including internal parton multiple scattering) and we can thus write

$$\left.\frac{dN}{dy}\right|_{N_A N_A} = \left.\frac{dN}{dy}\right|_{pp} N_A + \left(\nu_{N_A} - N_A\right)2k\alpha h \tag{13}$$

with

$$\left.\frac{dN}{dy}\right|_{pp} = \left[2+2(k-1)\alpha\right]h \,. \tag{14}$$

If external multiple scattering is absent, by putting $\nu_{N_A} = N_A$**, one obtains the wounded nucleon model limit, Eq. (7). If multiple scattering dominates,** $k >> 1$**, we obtain the limit of Eq. (8). We show the result of this model in Fig.19 (dash-dotted line). From comparison of Eq. (13) with** *pp* **data at low energy,** $k \cong 1$**, one obtains** $h \cong 0.75$**. The parameter** α **in Eq. (13) was put equal to 0.05.**

In the Dual String Model the strings interact, the simplest interaction being fusion due to overlap in the transverse plane ([26]. This is the mechanism that leads to percolation and to the Quark–Gluon Plasma formation [27,28,29]. When strings fuse, the strength of the colour field is reduced in comparison with the colour field generated by the same number of independent strings. This is essentially due to the random sum of colour vectors [30]: $Q_n^2 = \sum_{i=1}^{n} Q_i^2$ **and** $Q_n = \sqrt{n}Q$ **if all the** *n* **strings are of the same type.**

Introducing the dimensionless transverse density percolation parameter η**,**

$$\eta = \frac{r_s^2 N_s}{R_{N_A}^2} \ , \tag{15}$$

where r_s **is the string transverse radius (we shall take** $r_s = 0.2$ **fm, see [27,31]),** R_{N_A} **the radius of the interaction area** $\left(R_{N_A} \cong 1.14 N_A^{1/3}\right)$ **and** N_s **the number of strings, the effective reduction factor in particle production is [32],**

$$F(\eta) = \sqrt{\frac{1 - 2^{-\eta}}{\eta}} \ . \tag{16}$$

As $\eta \to 0, F(\eta) \to 1$ **(no fusion) and as** $\eta \to \infty, F(\eta) \to 1\sqrt{\eta} \approx 1/\sqrt{N_s}$ **, (all the strings fuse).**

We can consider the parameter η **in two situations. In nucleon–nucleon internal interactions, we have**

$$\eta_{pp} \equiv \frac{r_s^2}{R_{pp}^2}\left[2 + 2(k-1)\right] = \frac{r_s^2}{R_{pp}^2} 2k \ . \tag{17}$$

At present energies η_{pp} **is negligible,** $\eta_{pp} \approx 10^{-2} \div 10^{-1}$**. But we can also consider** η **in external interactions, with**

$$\eta_{N_A N_A} = \frac{r_s^2}{R_{N_A}^2} 2k\left(\nu_{N_A} - N_A\right) \cong \left(\frac{r_s}{1.14}\right)^2 2k\left(N_A^{1/3} - 1\right) N_A^{1/3} \ . \tag{18}$$

For $N_A \approx 10^2$, as in [21], $\eta_{N_AN_A} > 10\eta_{pp}$ and we shall then, in this context, only consider $\eta_{N_AN_A}$. Eq.(10) with string fusion becomes

$$\left.\frac{1}{N_A}\frac{dN}{dy}\right|_{N_AN_A} = \left.\frac{dN}{dy}\right|_{pp} + F\left(\eta_{N_AN_A}\right)\left(N_A^{1/3}-1\right)2k\alpha h \ . \qquad \textbf{(19)}$$

In Fig. 19 we have also shown the prediction of the DSM with string fusion (dashed line) again with $h = 0.75$ and $\alpha = 0.05$. The deviation from the wounded nucleon model limit becomes weaker and the agreement with PHOBOS data is quite satisfactory.

We would like now to make a few comments:

1. The predictions for particle densities in central Pb+Pb collisions of the DSM without fusion and of the DSM with fusion are very different at $\sqrt{s} = 200$ AGeV (RHIC) and at $\sqrt{s} = 5.5$ ATeV (LHC) as can be seen in Table 2, showing the average pseudo-rapidity density in the interval $[-1,1]$:

Table 2

c.m. energy	200 AGeV	5.5 ATeV
Without fusion	1500	4400
With fusion	700	1400

2. The models considered here are essentially soft models. The parameters of the elementary collision densities, h and α, were assumed constant, all the energy dependence being attributed to the parameter k, the average number of string pairs per elementary collision. If h and α are allowed to grow with energy, as a result, for instance, of semi-hard effects, the parameter k may then have a slower increase than the one obtained here.

3. The value found for $\alpha, \alpha \cong 0.05$, means that the height of the sea–sea plateau is much smaller than the height of the valence–valence plateau. By noticing that for valence–valence collisions the two strings stretch all over forward/backward rapidity without much overlap, while for sea–sea collisions the two strings do overlap, the value found for α means

$$\left.\frac{dN}{dy}\right|_{sea-sea} \cong 0.1\left.\frac{dN}{dy}\right|_{val-val} \ . \qquad \textbf{(20)}$$

4. In our Dual String Model with fusion, the parameter $\eta_{N_AN_A}$ at the CERN-SPS has the value $\eta_{N_AN_A} \approx 1.8$, larger than the critical density $\left(\eta_c \approx 1.12 \div 1.17\right)$ which means that

percolation transition is already taking place at $\sqrt{s} = 20$ AGeV, even allowing for non-uniform matter distribution in the nucleus $(\eta_c \approx 1.5)$ [33]; this result is valid even with $k = 1$. The observed anomalous J/ψ suppression may then be a signature of the percolation transition to the Quark-Gluon Plasma.

For a discussion of this problem see [34].

ACKNOWLEDGEMENTS

I would like to thank Roberto Ugoccioni for a very friendly and fruitfull collaboration.

REFERENCES

[1] **Particle Physics Booklet, Particle Data Group, July 2000, Springer, Available from LBNL and CERN.**

[2] **Lattice Gauge Theories: an introduction, H.J. Rothe, World Scientific, 1992.**

[3] **E. Laerman, Nucl. Phys. A610 (1996) 1, and references therein.**

[4] **J.Cleymans, R. Gavai and E. Suhonen, Phys. Ref. 130 (1986) 217, K. Kajantie, Second Eso – CERN Symposium (1986).**

[5] **U. Heinz, in Proceedings of Strong and Electroweak Matter 98, Ed. By J. Ambjorn, P. Damgaard, K. Kainulainen and K. Rummukainen, World Scientific, Singapore, 1999.**

[6] **E. Andersen et al (WA97 Coll.), Phys. Lett. B449 (1999) 401; J. Phys. G25 (1999) 171.**

[7] **F. Sikler and I. Králik, Talks at Quark Matter'99, Torino (1999),**

[8] **M.C. Abreu et al, CERN-EP-2000-012, Euro Phys. J.C., to appear, (2000).**

[9] **A Satz, pre-print BI-TP 2000/31, hep-ph/0009099.**

[10] **N.S. Amelin, M.A. Braun and C. Pajares, Phys. Lett. B3 06 (1993) 312; Z. Phys. C63 (1994) 507.**

[11] **G.Agakichiev et al. (CERES Coll.), Phys. Rev. Lett. 75 (1996) 1272.**

[12] **E. Scomparim (NA38 Coll.) Nucl. Phys. A610 (1996) 331c.; A.L.S. Angelis et al. HELIOS3 Coll.), Eur. Phys. J. C13 (2000) 433.**

[13] **C.M. Ko, C. Q. Li, G. E. Brown and H. Sorge, in Proc. QM96; W. Cassing, W. Ehehalt and I. Kralik, Phys. Lett. B300 (1996) 5.**

[14] **T. Matsui and H. Staz, Phys. Lett. B178 (1986) 416.**

[15] **S. Ramos et al, Nucl. Phys. A590 (1995) 117c and ref. therein.**

[16] **J. Dias de Deus, C. Pajares, C.A. Salgado, Phys. Lett B409 (1997) 474.**

[17] **J. Dias de Deus and J. Seixas, Phys. Lett. B430 (1998) 363.**

[18] **C. Cicalò (NA50 Collaboration), talk presented at "QM99", Torino, May 1999.**

[19] **S. Gavin and R. Vogt, Phys. Rev. Lett. 78 (1997) 1006; A. Capella, A. Kaidalov, A. Kovider Akil and G. Gerschel, Phys. Lett. B393 (1997) 431.**

[20] **J. Dias de Deus, R. Ugoccioni and A. Rodrigues, Eur. Phys. J. C16 (2000) 537.**

[21] **B.B. Back et al., PHOBOS Collaboration, preprint hep-ex/0007036, PHOBOS / RHIC.**

[22] **A. Bialas, B. Bleszynski and W. Czyz, Nucl. Phys. B111, 461 (1976).**

[23] **J. Dias de Deus, C. Pajares and C.A. Salgado, Phys. Lett. B409, 474 (1997).**

[24] **N. Armesto and C. Pajares, Int. J. Mod. Phys. A15, 2019 (2000).**

[25] **A. Capella, U.P. Sukhatme, C.I. Tan and J. Trân Thanh Vân, Physics Reports 236, 225 (1994).**

[26] **N.S. Amelin, M.A. Braun and C. Pajares}, Phys. Lett. B306, 312 (1993); N.S. Amelin, M.A. Braun and C. Pajares, Z. Phys. C63, 507 (1994).**

[27] **N. Armesto, M.A. Braun, E.G. Ferreiro and C. Pajares, Phys. Rev. Lett. 77, 3736 (1996).**

[28] **M. Nardi and H. Satz, Phys. Lett. B442, 14 (1998).**

[29] **J. Dias de Deus, R. Ugoccioni and A. Rodrigues, Eur. Phys. J. C16, 537 (2000).**

[30] **T.S. Biro, H.B. Nielsen and J. Knoll, Nucl. Phys. B245, 449 (1984).**

[31] **H. Satz, Nucl. Phys. A642, 130c (1998).**

[32] **M.A. Braun and C. Pajares, Eur. Phys. J. C16, 349 (2000).**

[33] **A. Rodrigues, R. Ugoccioni and J. Dias de Deus, Phys. Lett. B458, 402 (1999).**

[34] **J. Dias de Deus, R. Ugoccioni, Phys. Lett. B (2000), to appear.**

THE JINR SCIENTIFIC PROGRAMME

A.N. Sissakian
Joint Institute for Nuclear Research, Dubna, Russia

1. INTRODUCTION

1.1 General information about the Joint Institute for Nuclear Research (JINR)

The Joint Institute for Nuclear Research (JINR) in Dubna was established on the basis of the convention signed by the Plenipotentiaries of the governments of the Member States of the JINR in March 1956 in Moscow. The JINR was created in order to unify the intellectual and material potential of the Member States in order to study the fundamental properties of matter.

Dubna as a town of science was founded immediately after the end of World War II. In 1947 a group of scientists led by Academician I.V. Kurchatov initiated construction of the then largest accelerator of charged particles – the synchrocyclotron. The accelerator was commissioned already in 1949. Extensive fundamental and applied investigations into the properties of nuclear matter immediately started at the newly established Institute for Nuclear Problems (INP) with its operating 680 MeV synchrocyclotron, headed by the young physicists M.G. Meshcheryakov and V.P. Dzhelepov, later world-known scientists.

After the INP, the Electrophysical Laboratory of the USSR Academy of Sciences (EFLAN), headed by Academician V.I. Veksler, was set up in Dubna. A new accelerator, namely a synchrophasotron with record parameters for that time, was constructed at EFLAN.

In 1954 the European Organization for Nuclear Research (CERN) was established near Geneva to unite the efforts of Western European countries for studying the fundamental properties of matter.

About the same time, under the stimulus of the USSR Government, the countries then belonging to the socialist world took the decision to establish the Joint Institute for Nuclear Research in Dubna from the INP and EFLAN laboratories. The same year, specialists from 12 countries (Albania, Bulgaria, China, Czechoslovakia, East Germany, Hungary, Mongolia, N. Korea, Poland, Romania, USSR, and Vietnam) came to Dubna. The town became international, and investigations into many fields of nuclear physics of interest for research centres of the JINR Member States were launched there.

Many scientists and engineers from the Member States have been trained in the JINR scientific schools established by N.N. Bogoliubov, D.I. Blokhintsev, G.N. Flerov, I.M. Frank, B.M. Pontecorvo, V.I. Veksler, and other outstanding physicists. The development of different scientific directions at JINR is connected with the names of L. Infeld and H. Niewodniczanski (Poland), G. Nadjakov (Bulgaria), H. Hulubei (Romania), L. Janossy (Hungary), N. Sodnom (Mongolia), Wang Gangchang (China), Nguyen Van Hieu (Vietnam), V. Votruba and J. Kozesnik (Czechoslovakia), H. Pose and K. Lanius (Germany), and others.

The Charter of the JINR was adopted in 1956, the text of which was revised in 1992 and more recently in 1999. In accordance with the Charter, the activities of the Institute are achieved on the

basis of its openness, and the mutual and equal co-operation of all the interested parties to participate in research.

The aim of the Institute is:

- to carry out theoretical and experimental investigations on adopted scientific topics;
- to organize the exchange of experience when carrying out research and the exchange of information obtained as a result of these investigations through publication of scientific papers, holding of conferences, symposia etc.;
- to promote the development of the intellectual and professional capabilities of the scientific personnel;
- to establish and maintain contacts with other national and international scientific organizations and institutes to ensure the stable and mutual co-operation;
- to use the results of the investigations of an applied nature to provide supplementary financial resources for fundamental research by implementing them into industrial, medical, and technological developments.

The results of investigations carried out at the Institute can be used solely for peaceful purposes for the benefit of mankind. So until the late 1980s, Dubna was a centre that unified the efforts of leading research groups of nuclear sciences from socialist countries and the Soviet Union.

After the disintegration of the USSR, membership of the JINR underwent the following changes: the majority of Eastern European countries, such as Poland, the Czech and Slovak Republics, Bulgaria, Romania, and others continue to be Member States and contribute to the budget. Germany remains as an observer and makes a substantial financial contribution. Most of the former Soviet Union republics, which became independent states, entered the JINR as new members.

I would like to recall the words of the great Russian writer A. Chekhov, who said:
"Science cannot be national, in the same way that a multiplication table cannot be national. If a science becomes national, it ceases to be a science".

JINR is a perfect illustration of this idea.

There are different ways to participate in the activities of the Institute: on the basis of membership, or bilateral and multilateral agreements in order to perform particular scientific programmes. JINR Member States contribute financially to the Institute's activities and have equal rights in its management.

At present the JINR has 18 Member States: Armenia, Azerbaijan, Belarus, Bulgaria, Cuba, Czech Republic, Georgia, Kazakhstan, D.P. Republic of Korea, Moldova, Mongolia, Poland, Romania, Russian Federation, Slovak Republic, Ukraine, Uzbekistan, Vietnam.

The JINR has special co-operation agreements concluded on the governmental level with:

- Germany (in the field of theoretical physics, heavy-ion physics, condensed matter physics, and high-energy physics);
- Hungary (in the field of condensed matter physics);
- Italy (in the field of intermediate and low energy physics).

Among the major partners with whom JINR has long-term co-operation agreements are:

- CERN, in the field of high-energy physics;
- IN2P3 (France), in the field of nuclear and particle physics;
- INFN (Italy), in the field of nuclear and particle physics;
- FNAL, BNL, SLAC and other research centres in the USA.

Recently, special agreements were signed with UNESCO and CLAF (Latin American Centre for Physics). The latest political changes in Eastern Europe and especially in Russia have made the JINR more and more open. New collaborating countries are welcomed to join the JINR.

Particularly, successful negotiations have been lately conducted between the JINR Directorate and Governmental bodies of Greece on the subject of Greece entering the JINR as Associate Member.

THE AGREEMENT BETWEEN THE GOVERNMENT OF THE RUSSIAN FEDERATION AND JINR HAS BEEN RATIFIED

On the 3rd December 1999 the State Duma of the Russian Federation passed a Federal Law "On Ratification of the Agreement between the Government of the Russian Federation and the Joint Institute for Nuclear Research on the Location and Terms of Activity of the Joint Institute for Nuclear Research in the Russian Federation". The Federal Law was approved by the Federation Council on 22 December 1999 and signed by the Acting President of the Russian Federation V. Putin on 2 January 2000. The Federal Law came into force from the date of its official publication – 6 January 2000.

The Agreement between the Russian Government and the JINR was signed in Dubna in 1995. Now it is supported by the law and confirms the international personality and legal capacity of JINR. It also grants facilities, privileges and immunities in correspondence with the established practice for international intergovernmental organizations.

GOVERNING AND ADVISORY BODIES OF THE JINR

- Committee of Plenipotentiaries of the JINR Member States
- Finance Committee (one delegate from each Member State)
- Scientific Council
- Programme Advisory Committee for Particle Physics
- Programme Advisory Committee for Nuclear Physics
- Programme Advisory Committee for Condensed Matter Physics

The main fields of the Institute's activities are theoretical physics, elementary particle physics, relativistic nuclear physics, physics of low and intermediate energies, heavy-ion physics, nuclear physics with neutrons, condensed matter physics, radiobiology and nuclear medicine, experimental instruments and methods, computing technologies.

The basic facilities of the Institute for experimental investigations are the Nuclotron, U400 and U400M cyclotrons, and IBR-2 neutron reactor. The other JINR facilities such as the synchrophasotron, phasotron, cyclotron U-200, and the IBR-30 reactor are mainly supported by resources which are outside the JINR budget.

The internal organization of the JINR is determined by scientific specialization. There are seven Laboratories in the Institute:

- Bogoliubov Laboratory of Theoretical Physics (BLTP)
- Laboratory of High Energies (LHE)
- Laboratory of Particle Physics (LPP)
- Dzhelepov Laboratory of Nuclear Problems (DLNP)
- Flerov Laboratory of Nuclear Reactions (FLNR)

- Frank Laboratory of Neutron Physics (FLNP)
- Laboratory of Information Technologies (LIT)

There are three more all-Institute subdivisions in the JINR's structure:

- Division of Condensed Matter Physics
- Division of Radiation and Radiobiological Research
- University Centre of JINR

The total number of JINR personnel is about 6,000 (including service divisions). Approximately 1,100 scientists work at the Institute. The scientific policy of the JINR is established by the Scientific Council, whose members are prominent scientists from the Member States (A.A. Logunov, V.A. Matveev, A.N. Skrinsky, R. Sosnowski, A.N. Tavkhelidze, etc.) and well-known physicists from CERN, France, Germany, Italy, the USA and other countries. M. Della Negra, C. Detraz, F. Dydak, F. Lehar, B. Peyaud, H. Schopper, P. Spillantini, G. Trilling, and others are among the members of the Scientific Council. JINR's Director, Academician V.G. Kadyshevsky, is the Chairman of the Scientific Council.

Several associate experimental physics workshops are also part of the Institute. The personnel of the Central Workshop totals about 400. It is equipped with everything necessary for manufacturing large-size non-standard facilities, electronics, and has technological lines for constructing detectors for physics. Here the main units of JINR's heavy-ion cyclotrons — U400 and U400M — were constructed in recent years, as well as the Nuclotron — a new superconducting accelerator for relativistic nuclear physics. It is an excellent result especially in view of the difficult economic situation in Russia in recent years.

2. JINR's SCIENTIFIC ACTIVITY AND BASIC FACILITIES

Let us consider the research programmes at the JINR's main facilities and a few examples of the recent results.

2.1 Nuclotron

The new superconducting accelerator, the Nuclotron (see Fig. 1) was put into operation five years ago (A.M. Baldin et al.). It will enable an extensive programme of research in relativistic nuclear physics to be performed both in asymptotic mode (an accelerated nuclei energy higher than 4 GeV/n) and transmission regime (less then 4 GeV/n). In the asymptotic mode the nucleons cannot be considered as quasiparticles of nuclear matter and the influence of quark–gluon degrees of freedom in interactions of hadrons and/or nuclei should be observed.

The Laboratory of High Energies (LHE) of the JINR was a pioneer in designing and constructing the first low-cost accelerator based on low-field iron dominated superconducting magnets. The Nuclotron was built over a period of five years. The main magnetic equipment and many other systems were fabricated by JINR workshops.

The Nuclotron ring of 251 m in perimeter is installed in the Synchrophasotron technological tunnel. The total 'cold mass' of the magnetic system is about 80 t. Cooling of the system down to 4.5 K takes about 90 h. The cooling system was designed taking into account the fast cycling mode of the Nuclotron (up to 0.5–1.0 Hz), which is a specific feature of this superconducting accelerator.

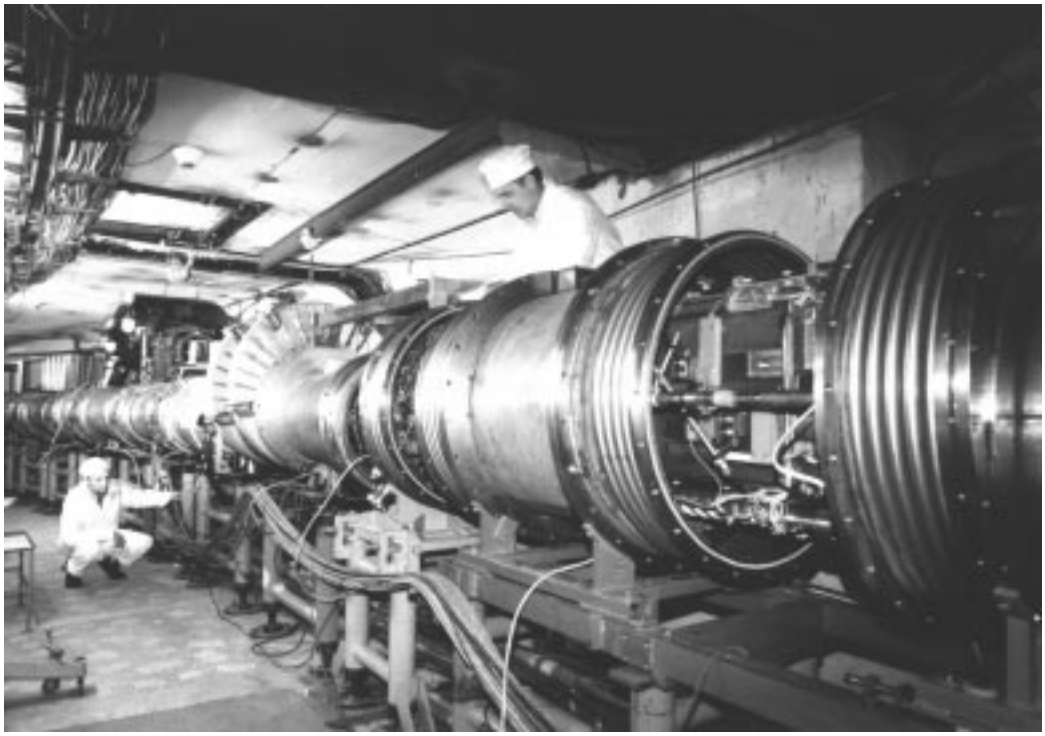


Fig. 1: The Nuclotron.
A new superconducting accelerator which permits an extensive programme of research in relativistic nuclear physics to be performed.

In 1999, a special system for extraction of a nuclear beam from the Nuclotron to external set-ups was completed. The system is a complex structure based on superconducting magnetic elements, which does not have an analogue in the world. In December 1999, the system was successfully tested and the first extraction of the proton beam from the Nuclotron was carried out. In March 2000, an experiment was conducted to optimize the modes of the slow extraction beam and research activity was started at the extracted beam in the LHE experimental hall.

Two experiments at the STRELA and SCAN-2 set-ups have been started this year with the extracted beam. The first test measurements with the slow extracted deuteron beam of the Nuclotron were carried out within the framework of research on the charge-exchange process in deuteron–proton interactions: $\mathbf{dp \rightarrow (pp)n}$. The goal of the measurements is to register two protons moving forward in a narrow cone at the large background of single protons from the deuteron fragmentation process. The two-proton events are distinguished by measuring the Cherenkov radiation amplitude (see Fig. 2).

The SCAN-2 set-up is constructed to study the proton-formation length in
$\mathbf{d + A_T \rightarrow p_1(\sim 0^o) + p_2\ (\sim 0^o) + \ldots}$

The injection complex under development will consist of a booster, linac, and ion sources. This complex will allow one to accelerate nuclei from hydrogen to uranium with an intensity of 10^{13} to 10^8 particles per pulse, respectively (see Table 1), in the energy range of 6–7 GeV per nucleon. Polarized deuteron beams are foreseen.

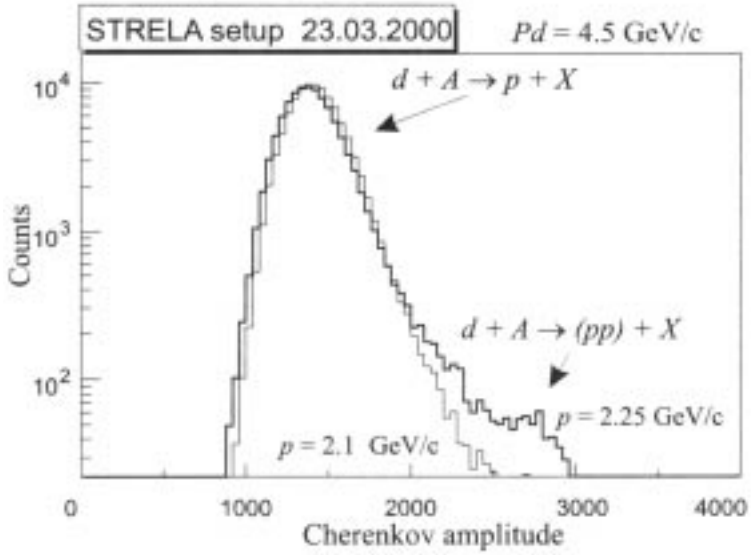


Fig. 2: The amplitude distribution of signals from the Cherenkov counter at the zero-angle scattering of deuterons with a momentum of 4.5 GeV/c from a polyethelene target

Table 1: **The NUCLOTRON beams**

Beam	INTENSITY (Particles per cycle)		
	Nuclotron *(available)*	Nuclotron + Ion sources development	Nuclotron + Booster
p	$2 \cdot 10^{10}$	$1 \cdot 10^{11}$	$1 \cdot 10^{13}$
d	$2 \cdot 10^{10}$	$5 \cdot 10^{10}$	$1 \cdot 10^{13}$
^{3}He		$2 \cdot 10^{9}$	$\sim 10^{12}$
^{4}He	$8 \cdot 10^{8}$	$5 \cdot 10^{9}$	$2 \cdot 10^{12}$
^{7}Li		$2 \cdot 10^{10}$	$5 \cdot 10^{12}$
^{12}C	$1 \cdot 10^{8}$	$7 \cdot 10^{9}$	$2 \cdot 10^{12}$
^{20}Ne		$1 \cdot 10^{8}$	$5 \cdot 10^{9}$
^{24}Mg		$3 \cdot 10^{8}$	$5 \cdot 10^{11}$
^{40}Ar		$3 \cdot 10^{7}$	$2 \cdot 10^{9}$
^{56}Fe		$5 \cdot 10^{7}$	$1 \cdot 10^{11}$
^{84}Kr	$1 \cdot 10^{3}$	$2 \cdot 10^{7}$	$5 \cdot 10^{8}$
^{96}Mo			$1 \cdot 10^{10}$
^{131}Xe		$1 \cdot 10^{7}$	$2 \cdot 10^{8}$
^{181}Ta			$1 \cdot 10^{8}$
^{209}Bi		$3 \cdot 10^{6}$	$1 \cdot 10^{8}$
^{238}U			$1 \cdot 10^{8}$

2.2 Synchrophasotron

The Synchrophasotron (see Fig. 3) is an accelerator of 10 GeV protons commissioned in 1957 (V.I. Veksler, A.L. Mints et al.). In the 1970s the acceleration of nuclei heavier than hydrogen was accomplished in the broad energy spectrum from a few hundred MeV to 4.5 GeV per nucleon. The average densities of the beams range from 10^4 to 10^{11} ions per cm^2s depending on the atomic number of the accelerated nuclei and the experimental requirements.

Synchrophasotron beams attract physicists from many laboratories in the world. The collaboration SPHERE (JINR+Nagoya University) performed a study of nuclear matter at short distances.

Fig. 3: Synchrophasotron

In order to clarify the mechanism of the reactions and the structure of the non-nucleon degrees of freedom, an experiment with polarized deuteron fragmentation into cumulative hadrons was performed. The observed difference between the model and the experimental data is especially large for K = 0.2 GeV/c, where it would be natural to expect the manifestation of non-nucleon degrees of freedom in the deuteron wave function (see Fig. 4).

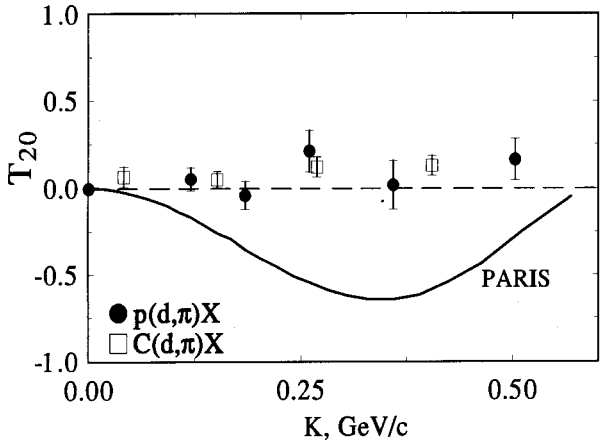


Fig. 4: SPHERE Collaboration. Measurement of the tensor analyzing power T_{20} in inclusive polarized deuteron fragmentation into pion at zero angle.

2.3 Phasotron

The Phasotron is an accelerator of 680 MeV protons (M.G. Mesheryakov, V.P. Dzhelepov et al.). It commenced operation in 1949, was reconstructed in 1984, and is the oldest basic facility of the JINR (see Fig. 5). In December 1999 the staff of the JINR's Laboratory of Nuclear Problems has celebrated the 50th anniversary of commissioning the Dubna Synchrocyclotron (from 1984 – Phasotron). The person who immensely contributed to the creation, development, and subsequent upgrading of the Synchrocyclotron for applied nuclear physics and medico-biological research was Professor V.P. Dzhelepov, an outstanding scientist and one of the founders of the JINR. At its 87 session (January 2000), the JINR Scientific Council welcomed the decision of the JINR Directorate to name the Laboratory of Nuclear Problems after Professor V.P. Dzhelepov, in recognition of his great contribution to the activities of this Laboratory and of the whole Institute.

It is owing to his enthusiasm that an onco-radiological department on the basis of the Phasotron was opened in Dubna on 1 December 1999. Twenty-four patients having serious forms of cancer (4th stage) have already undergone a course of treatment in the onco-radiological department; 18 of them received a gamma–proton treatment, and 6 were exposed to gamma-rays. The results of the treatment were positive.

Fig. 5: Phasotron with space variation of the magnetic field, proton energy (660+3.1)MeV, extracted beam intensity 3x10^{13} p/s, duty cycle - 85%

Ten beam channels are available at this machine and are used to carry out experiments with pions, muons, neutrons and protons. Five secondary beams are designed to carry out medical investigations. The intensity of the extracted proton beam is 4 μA.

The scientific activities of the Laboratory cover experimental research in particle physics (at high, low and intermediate energies); investigation of nuclear structure (including relativistic nuclear physics and nuclear spectroscopy); study of condensed matter properties; biological and medico-biological investigations; applied research including cancer therapy.

There is a proposal to construct the external injection of the beam into the accelerator's centre in order to increase the intensity of the Phasotron accelerated proton beam 10–20 times.

2.4 U400 and U400M

At present the complex of two heavy ion cyclotrons — U400 and U400M (G.N. Flerov, Yu.Ts. Oganessian et al.) — is the main experimental base of the Flerov Laboratory of Nuclear Reactions. As you know, element 105 discovered at this laboratory was named Dubnium in honour of Dubna.

The Laboratory's research programme involves investigations in a wide region of low- and intermediate-energy heavy-ion physics. Intense heavy ion beams open up new vistas for the synthesis and study of radioactive and chemical properties of heavy elements, studies of nuclei far from the line of beta-stability. Research is under way in the field of reactions with exotic nuclei and the mechanism of fusion and fission processes.

The isochronous U400 cyclotron was constructed in 1978 and produces ion beams of atomic masses from 4 to 100 and maximum energy up to 25 MeV/nucleon. The maximum ion beam intensity is presently $2x10^{14}$ ion/s for Ne, and drops to $2x10^{13}$ ion/s for Ca. The experimental set-ups are located at 12 extracted beam channels.

The U400M is an isochronous cyclotron (see Fig. 6) which commenced operation in 1991-92 to accelerate heavy ions. At the U400M cyclotron, internal ion beams of light elements from He to Ar with energies up to 50 MeV/nucleon (the maximum energy is 100 MeV/nucleon) have been obtained. A system of channels for extracted beam transportation is being developed.

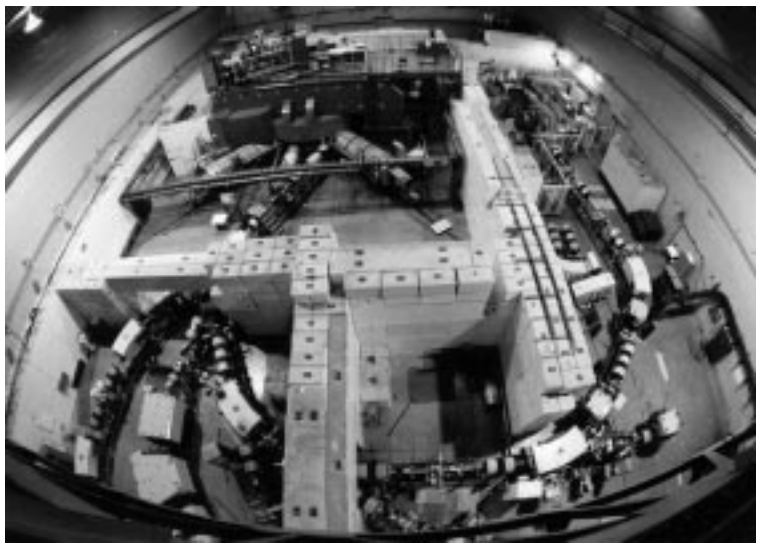


Fig. 6: U400M

To widen the range of accelerated ions at the U400M and to increase the energy, the ECR-source DECRIS has been constructed at the Laboratory. Also, a more powerful ECR-source is being developed in collaboration with the GANIL Centre (France). The use of the ECR-type heavy-ion sources will allow the Laboratory to use less expensive separated isotopes in research on the synthesis of heavy elements or exotic nuclei. In order to obtain and transport secondary beams of radioactive nuclei, the COMBAS fragment separator with a large acceptance and high resolution is being created. The project of the future development of the Laboratory accelerator complex is linked with the production of radioactive nuclear beams in the tandem mode of operation of the U400M and U400 accelerators, and with the creation of a storage ring, acting as the third stage.

The most spectacular result obtained at U400 in late 1998 was the synthesis of a new long-lived (30s) superheavy element with atomic number 114 and mass 289 and thus the discovery of an island of nuclear stability for superheavy elements. This discovery has crowned the 35-year efforts of physicists from JINR, USA and Germany in search for this stability island.

During 1999, three isotopes of the new element, 114, having masses A=287, 288 and 289 were synthesized. Among their decay products there were identified the heaviest isotopes of the elements: 112 with A=283, 284 and 285, element 110 with A=280 and 281, and element 108 with A=277.

Fig. 7 shows the decay chains of the new isotopes, including the energy values, the alpha-decay periods, and spontaneous fission.

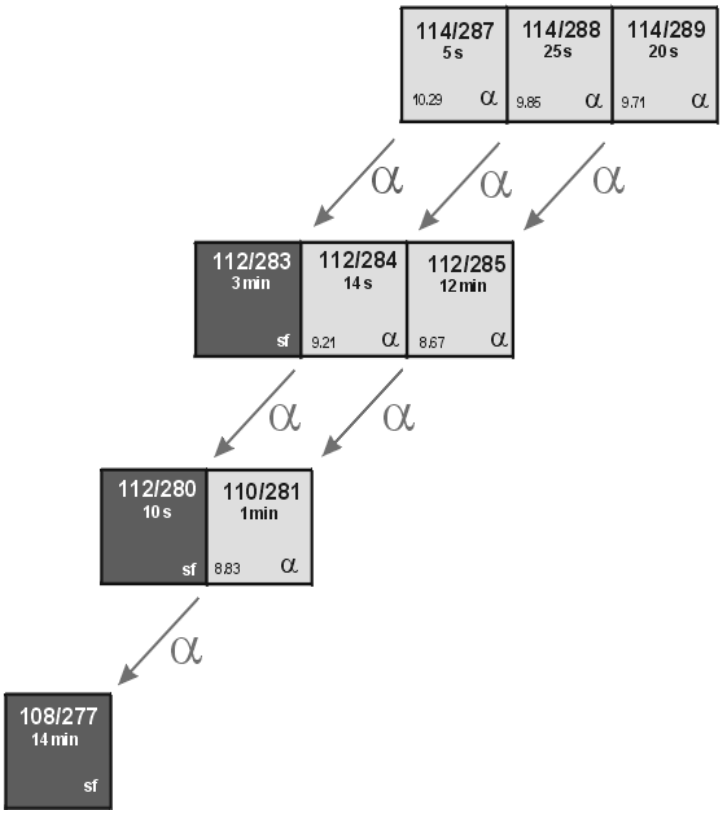


Fig. 7: Decay chains of isotopes of element 114

On 24 May 2000 an experiment on element 116 synthesis was started. The unique curium target is created with the support of the Ministry for Atomic Energy of the Russian Federation.

The other intensively developing direction at FLNR is a research programme with radioactive nuclear beams, namely the Dubna Radioactive Ion Beams (DRIBs) project. Realization of the DRIBs project assumes some modernization of U400, U400M and Microtron accelerators and construction of new beam channels between these three accelerators. The new complex will provide generation and acceleration of secondary beams of lightest nuclei ^{4}He – ^{14}O with the energy from 4 to 20 MeV per nucleon and radioactive ion beams of heavier elements generated from the products of fission fragments. It will be the first source of intense radioactive ions with unique parameters in the JINR Member States for experiments with exotic nuclei.

2.5 The IBR-2 fast neutron reactor

The main basic facility of the Frank Laboratory of Neutron Physics (FLNP) is the fast neutron pulsed reactor IBR-2 used for condensed matter research (D.I. Blochintsev, I.M. Frank et al.), see Fig. 8. Neutron scattering investigations in the field of condensed matter physics are conducted at IBR-2 using four main experimental techniques: diffraction, small-angle scattering, inelastic scattering, and polarized neutron optics.

IBR-2 is a pulsed reactor with an average thermal power 2 MW, peak power in pulse of 1,500 MW, half width of the pulse 215 μs, pulse repetition rate 5 Hz, thermal neutron flux density in moderator at maximum of the pulse – $2x10^{16}$ n/cm^{2}s.

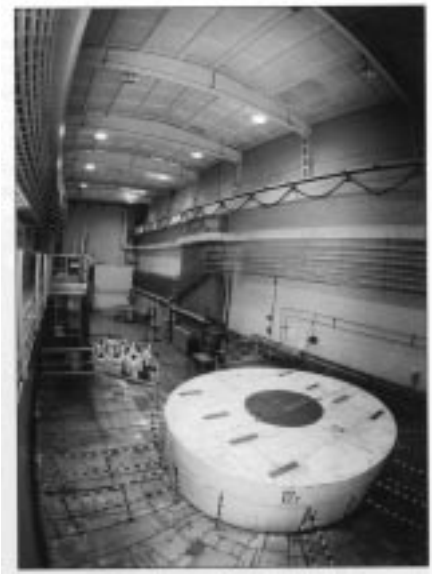


Fig. 8: IBR-2

In December 1999 the staff of FLNP completed the tests of the unique cryogenic moderator on solid methane at the IBR-2 reactor. Thus, the cold neutron flux turned out to be 3–4 times higher than the flux from the best for the moment cold neutron source ISIS in Britain. Now there are new opportunities for experimental research of complex structures in biology, physics of polymers, materials science, pharmacology, etc.

By 2002, the principal parts of the reactor IBR-2 will have their radiation resource exhausted and will have to be replaced. The programme for upgrading the IBR-2 reactor will be accomplished over a period of 10 years (1996–2005). This programme includes:

- improvement of the reactor parameters,
- increase in nuclear safety and reliability of the reactor,
- updating of the reactor systems.

2.6 IBR-30 + LUE-40

The IBR-30+LUE-40 is a pulsed neutron source consisting of the old pulsed reactor IBR-30 and the electron 40 MeV linac LUE-40 (D.I.Blochintsev, I.M.Frank et al.). The average heat power of the reactor is 10 kW, and the instant pulse power is 150 MW. It generates neutron pulses with a frequency of about or less than 100 Hz and a duration of 4.5μs. The total neutron yield is $5x10^{14}$ n/s, the flux of fast neutrons is about 10^{12} n/cm^{2}s.

A developed set of time-of-flight neutron spectrometers allow a wide spectrum of investigations to be carried out including the study of P- and T-symmetry violation of fundamental interactions in nuclei, the electromagnetic structure of the neutron, and some problems related to fundamental nuclear physics.

2.7 IREN

IREN (Intense Resonance Neutron Source) is a project aimed at constructing a high-flux pulsed neutron source to carry out investigations with resonance neutrons (Fig. 9).

The facility will comprise a modern 200 MeV electron linac and subcritical plutonium booster with a neutron multiplication coefficient 30. The pulse rate is 150 Hz, duration 0.4 μs, and the total neutron yield ~ 10^{15} n/s. This project was started in 1994 and the IREN facility will commence operation in 2002.

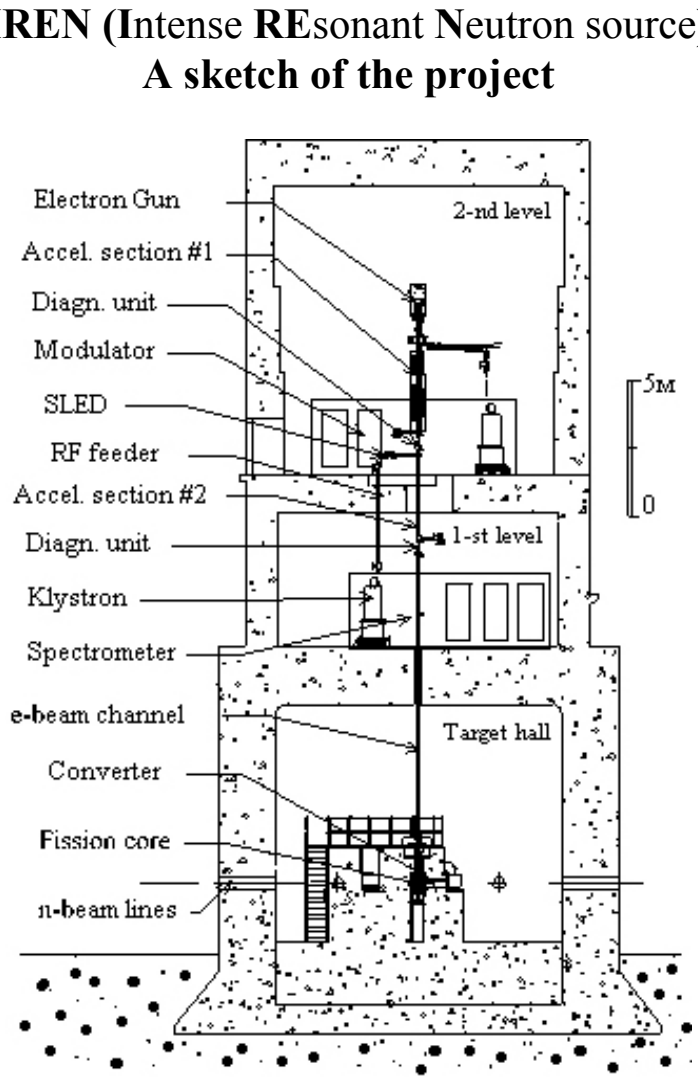


Fig. 9: The research programme for the IREN includes investigation of P- and CP - violations in slow neutron interactions with nuclei and other fundamental nuclear physics topics.

2.8 The JINR computing and networking infrastructure

According to the recommendation made by the JINR Scientific Council at its 85th session in January 1999, the Institute's telecommunication links, computing and networking infrastructure are considered to be a basic facility, too.

In 2000 the JINR Directorate took a decision to reorganize the Laboratory of Computing Techniques and Automation, which was responsible for this infrastructure, into the Laboratory of Information Technologies (LIT).

The efforts of the Laboratory of Information Technologies are directed at providing adequate networking, information, computing and software support of the research under way at JINR. The main activities of this laboratory are shown in Fig. 10.

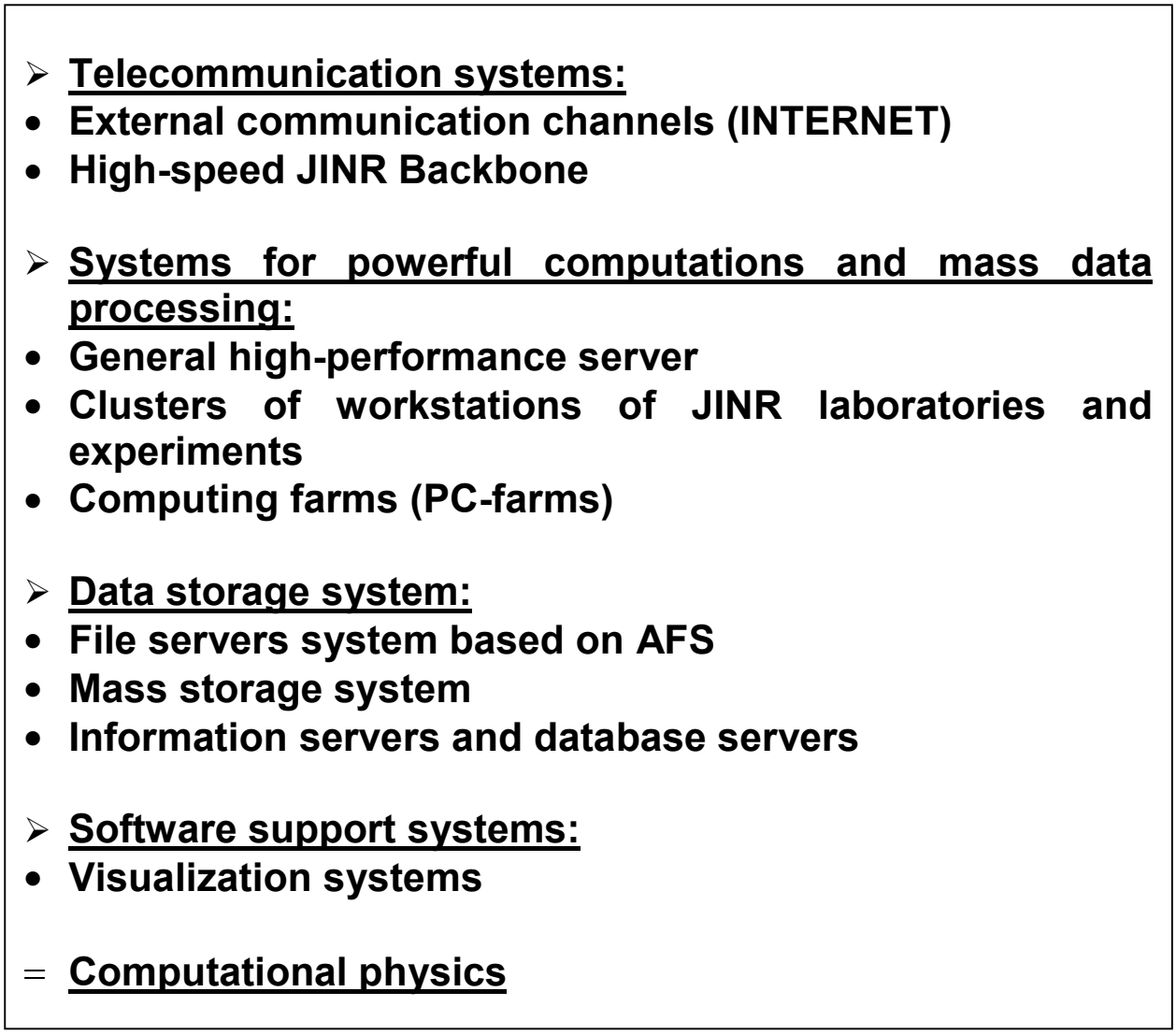


Fig. 10: The main activities of the Laboratory of Information Technologies

3. THE JINR: A MAJOR PARTNER OF WORLD HEP LABORATORIES

3.1 International collaboration

Broad international co-operation is one of the most important principles of the JINR's activities. Almost all investigations are carried out in close collaboration with the JINR member-state scientific centres, as well as international and national institutions and laboratories of the world. Effective co-operation is achieved with institutes in Russia such as IHEP (Protvino), Kurchatov Institute (Moscow), Institute of Nuclear Physics (Gatchina near St. Petersburg), ITEP (Moscow), INR (Troitsk), Lebedev Institute of Physics (Moscow), Moscow State University, Budker Institute of Nuclear Physics (Novosibirsk), and others.

Very fruitful scientific co-operation has been achieved with CERN, especially in recent years, as well as with many physics laboratories in the USA, France, Germany, Italy, and other countries.

In 1999, approximately 1,200 JINR specialists participated in joint experiments and international conferences and symposia, and more than 1000 scientists from collaborating laboratories and centres visited Dubna. The JINR also organizes about 50 conferences, workshops and other meetings annually. Scientists from Dubna participated in more than 150 international conferences in 1999.

3.2. Co-operation with CERN

For more than 40 years the co-operation between JINR and CERN has been very fruitful and mutually beneficial. A general agreement between the JINR and CERN was signed in 1992, but co-operation between the two international organizations has a very long and rich history. The JINR Directorate attends much attention to the CERN–JINR collaboration in research programmes, including theoretical studies and education of a new generation of young scientists, particularly JINR–CERN Schools.

Dubna physicists are involved in a large part of the CERN experimental programme (see Table 2). Dubna has made progress in fulfilling its obligations with respect to the LHC Programme (ATLAS, CMS and ALICE experiments, as well as the LHC machine).

ATLAS project

A great amount of work for the ATLAS Barrel Hadron Tile Calorimeter has been performed: in co-operation with Belarusian and Slovak Institutes and industry about 300,000 units of steel absorber plates with very high mechanical tolerance (50 microns) were manufactured. All of them have been delivered to calorimeter submodule producers (Dubna, Prague, Protvino, Pisa).

CMS project

The JINR is participating in the CMS project in the framework of the Russia and Dubna Member States Collaboration (RDMS). The involvement of the Member States in this activity through RDMS has given them an opportunity to play leading roles and to contribute significantly to the preparation of the hadron calorimeter, electromagnetic calorimeter and the muon detector.

Belarusian scientists and industry have developed and manufactured a special electronics scheme (for proportional chambers) which successfully passed radiation tests at CERN. Belarus and Bulgaria are responsible for CMS end-cap calorimetry. They assembled the very first full-scale prototype of h-cal sector (2x2x1 m^3 in dimensions and weighing 30 t).

ALICE project

The JINR contributes to the warm dipole magnet, the production of large-scale Pestov counters, detector assembly, data taking runs, and data processing & analysis.

Table 2: Experiments at CERN in which Dubna is involved

Project	Location	a) main goals b) JINR contribution
NA45	SPS	a) Study of electron–positron pair production in relativistic nuclear collisions. b) Design & manufacture of a new magnet system for precise spectroscopy. Design & construction of the trigger system. Data analysis.
NA48	SPS	a) Highest precision direct CP violation searching in neutral kaons decays. b) Subsystems design & construction, data taking runs. Data analysis.
NA49	SPS	a) Search for the predicted phase transition from hadrons to deconfined quarks and gluons in Pb+Pb collisions at SPS. b) 900-channel time-of-flight detector for identification of $h^{\pm}$, $K^{\pm}$, p, p, d and d. Data analysis.
NOMAD	SPS	a) Search for $\nu_\mu \rightarrow \nu_\tau$ and $\nu_\mu \rightarrow \nu_e$ oscillations. b) Data taking & analysis, new proposal preparation.
COMPASS	SPS	a) Hadron structure and hadron spectrometry on high rate hadron and muon beams; q&g contribution to nucleon spin; polarization of nucleon sea q's etc. Glueballs search for exotics. b) Hadron Cal, muon detector, large area track chambers.
DIRAC	PS	a) 5% accuracy test of low energy QCD by 10% precision ($\pi^+\pi^-$) atom life time measurement. b) Experiment proposed by JINR, drift chambers, secondary particles channel, trigger development, MC simulation & software. Data taking runs, data processing & analysis.
DELPHI	LEP	a) Precision measurements of m (W), search for new particles, etc. b) Maintenance of Hadron Calorimeter and Surround Muon Chambers, physics analysis.
ATLAS	LHC	a) General purpose *pp*-experiment. b) Subsystems: calorimeters, muon, transition rad. det., rad. Hard. tests, phys. Software & simulation, trigger and data acquisition
CMS	LHC	a) General purpose *pp*-experiment. b) Subsystems: forward μ-station, hadron end cap cal., e/m cal preshower; simulation.
ALICE	LHC	a) Heavy-ion relativistic beams. Study of q-g plasma and phase transition. b) Warm dipole magnet, large-scale Pestov counters production. Detector assembly. Data taking runs. Data processing & analysis.
R&D for LHC Accel. Complex elements	LHC	a) Development & construction of LHC beams formation & control system elements. b) Design & construction of transverse oscillation damping system. Simulation & prototypes study.

LHC damper

One of the main goals of the JINR's participation in the LHC Project is to manufacture a powerful amplifier and a kicker for the Transverse Feed-back System for the LHC.

At the JINR the design of a special power device was proposed. It allows different types of damping to be used and investigated (including non-linear "bang-bang" and "logical" ones). Joint CERN–JINR experiments at the SPS (SL RF) have shown the effectiveness of this solution.

I would like also to comment some new scientific results obtained in CERN experiments in co-operation with JINR scientists. You already know about an impressive result connected with the discovery of a new state of matter – the so-called 'quark–gluon plasma'. CERN Director General Luciano Maiani mentioned that this is a result of combined data coming from the seven experiments ongoing in the frame of CERN's Heavy-Ion Programme. It is remarkable that physicists from the JINR Laboratory of High Energies participate in three of them, namely NA49, NA45 and WA98, and have significantly contributed to obtaining of this result that pertains to relativistic nuclear physics. Let me remind you that this field of research was first suggested and developed at the JINR by Academician A. Baldin – the author of the idea of Dubna's accelerator Nuclotron created to study phenomena in which particles are moving with relativistic velocities. The interpretation of the CERN experiment is still not completely clear and definite. The future experiments at RHIC (BNL, USA), in which Dubna physicists also participate, might clarify this phenomena.

One more example of our fruitful co-operation with CERN is connected with the NA48 experiment. On 29 February 2000, the NA48 collaboration presented a new result on direct CP violation. You know that this experiment is carried out at the SPS beam-line, which takes data to study direct CP violation in the neutral kaon system. The Spokesman of this experiment is Prof. V. Kekelidze, Director of the JINR Laboratory of Particle Physics. The measured parameter $(14{,}0 \pm 4{,}3)\cdot 10^{-4}$ has been obtained by analysing a big amount of experimental data accumulated in 1997–1998.

Scientists from the JINR made a significant contribution also to other CERN experiments performed at LEP, the SPS, and the PS. Prof. L. Nemenov from the JINR was elected by the collaboration as a Spokesman of the DIRAC experiment, which successfully started its first run in late 1998. The main goal of this experiment is a precise measurement of $\pi^+\pi^-$ atom lifetime for the non-model check of the Quantum Chromodynamics (QCD).

3.3 Co-operation with IHEP (Protvino, Russia)

JINR scientists are carrying out experiments at the IHEP U-70 proton synchrotron using such set-ups as the EXCHARM, HYPERON, and the Neutrino Detector (Table 3).

Table 3: JINR's participation in research at U-70

EXCHARM	**Search for exotic states with strange quarks, study of processes of production and decay of particles containing heavy quarks**
HYPERON	**Investigations of rare *K*-meson decays**
NEUTRINO DETECTOR	**Investigations of neutrino oscillations and neutrino-nucleon interactions, measurement of nucleon structure functions, coupling constant of QCD at low energy**

4. DUBNA AS AN EDUCATIONAL CENTRE

The educational programme plays an important role in the JINR's activities. It should be stressed that the concept of the JINR's development is the integration of fundamental science, technological studies, and education.

To achieve this task, in 1991 we established the JINR University Centre and in 1994, together with the Russian Academy of Natural Sciences, the International University "Dubna".

The University Centre of the JINR offers graduate programmes in the fields of:

- Nuclear Physics
- Elementary Particle Physics
- Condensed Matter Physics
- Theoretical Physics
- Technical Physics
- Radiobiology

Since 1995 the University Centre of JINR has been offering post-graduate training.
Fig. 11 shows the educational contacts of the JINR University Centre.

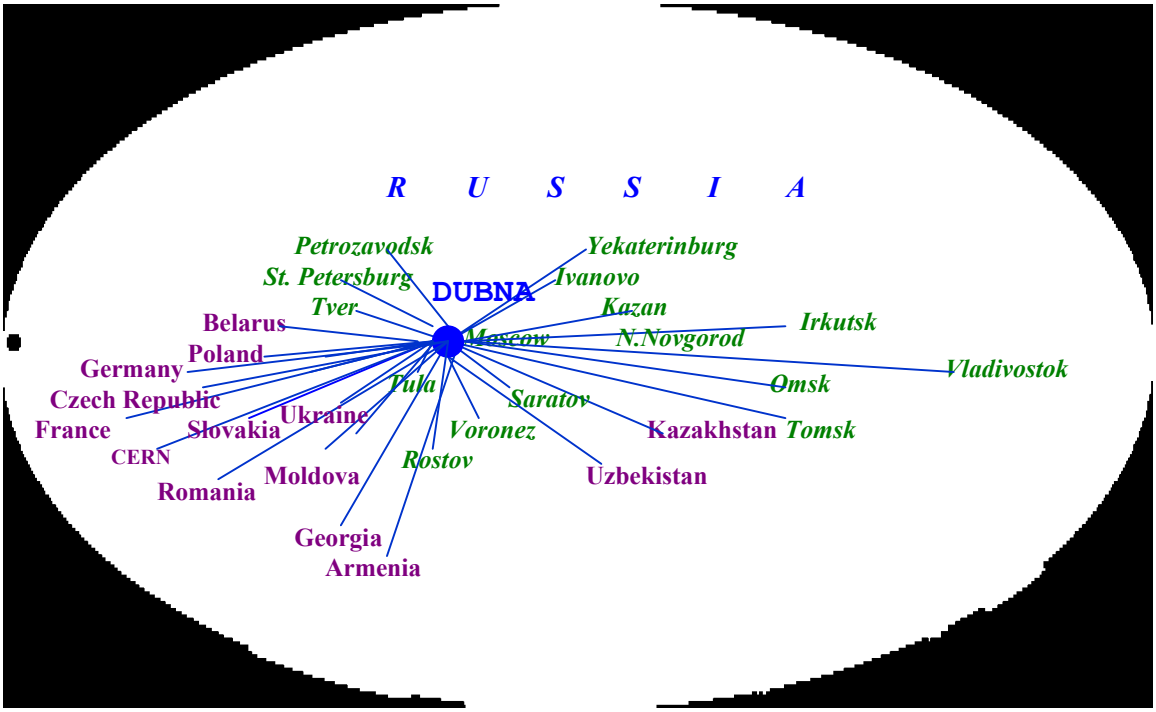


Fig. 11: Educational contacts of the JINR University Centre

5. PLANS FOR THE NEAR FUTURE

The JINR has the following plans for the near future:

- Significant broadening of the international usage of the JINR's unique research machines such as Nuclotron, cyclotron complex U400-U400M, and IBR-2.
- Development of methodical and computing possibilities for participation in experimental programmes of the world's largest HEP laboratories (CERN, FNAL, DESY, IHEP and others).
- Construction of the IREN facility.
- Development of the Nuclotron's injecton complex.
- Further development of the JINR University Centre.
- The use of the JINR's advanced infrastructure for holding international conferences, meetings and schools.
- Development of the project DRIBS dedicated to produce intense beams of unstable nuclei.
- Development of the DELSY facility — a third-generation synchrotron light source.

The JINR is continuing its programme of reforms to create better conditions for using its facilities and infrastructure.

6. CONCLUSION

This short review only presents some general information about the research centre in Dubna. It should be noted that over its 44 years of existence, the JINR has become a well-known international scientific centre, which incorporates the fundamental research of the structure of matter, the development and application of high technologies, and university education in the relevant fields of knowledge. The scientific policy pursued by the Directorate of the JINR has been developed in the context of the world's scientific trends. At the same time, recent years have been marked by a struggle for survival and the preservation of the Institute as a unified scientific centre in the time of radical political changes and serious economic difficulties in Russia and most of the Member States. Nevertheless, thanks to the joint efforts, the Institute has survived, and continues to contribute significantly to world science in the fields of particle physics, nuclear physics, and condensed matter physics.

The JINR Directorate is convinced in the further development of the Institute after the recent ratification of the Agreement between the Government of the Russian Federation and JINR.

In conclusion, I would like to stress the following: in order to preserve and multiply the achievements of JINR one should take care of its three pillars, namely:

– The international character of the Institute;
– The traditions of the Dubna scientific schools;
– The attractive experimental facilities, including the computing and networking infrastructure.

I would like to express my gratitude to Dr G. Arzumanyan and Mr O. Kronshtadtov for their assistance in the preparation of these lecture notes.

HISTORY OF THE PORTUGUESE MUSIC: AN OVERVIEW

Rui César Vilão
Physics Department, University of Coimbra, PORTUGAL

Abstract
A brief overview of the history of the Portuguese music is given.

In memoriam

José Carlos Travassos Cortez

1. PREVIOUS REMARKS

Presenting a communication on such a vast field as the subject of this work is usually a task for specialists. Only these are able to smoothly draw a global and coherent picture of their field of study. So, if you are not a specialist (even if you have a deep interest for the field) and you occupy your days working in an entirely different subject (say the physics of muonium states in semiconductors), you are surely in trouble!

There are two possible solutions for this dilemma: a) ask a specialist and b) do it *à la* Newton (even if you're not one), climbing in the shoulders of giants. The mountaineering option was adopted here, and four giants helped me in viewing higher: Rui Vieira Nery and Paulo Ferreira de Castro with their *História da Música* [1], João de Freitas Branco with his *História da Música Portuguesa* [2] and José Carlos Travassos Cortez with his passioned and illuminating lessons on the history of Portuguese music that I was lucky to attend in the Conservatório de Musica de Coimbra [3]. I am very pleased to dedicate this work to his memory.

Apart from Refs. [1-3], which are available in Portuguese only, all the other references are relative to introductory notes of compact-discs and are generally available in English as well. Of course these discographic indications are not extensive, but I believe they constitute a basis (surely not the only one!) for a good start in Portuguese-specialized melomania.

2. MEDIEVAL MUSIC

2.1 Liturgical repertoire

In the first times of the Catholic church, several local liturgies devellop (the galican in the franc realm, the sarum in England, the antique roman in Rome, the Ambrosian in Milan, etc). In the visigothic kingdom established in the Iberian Peninsula, the Council of Toledo organizes in 633 the hispanic rite. This liturgy is also known as visigothic rite or mozarab rite.

The main source of study of this particular hispanic liturgy is the *León Antifonary* (X^{th} century), which was most probably copied from an original colected at Beja (now in southern Portugal). It is also from the Beja region (Mértola) that comes the most ancient reference in Europe to a musician: there we find reference to the activity of *Andre princeps cantorum* (Andrew, prince of the singers) in 489-525.

The oldest manuscript (XI^{th} century) keeping liturgical music in hispanic (toledan) notation is kept at the General Library of the University of Coimbra. Most of the existing documents have

aquitan notation. From middle XIIIth century on, the notation presents tipically portuguese variations; this portuguese notation is used until the XVth century, when modern notation in staves is adopted.

However, the church would start worrying soon about the proliferation of liturgies. From the mixture of the galican liturgy with the antique roman one would result, traditionally under pope Gregory I (540-604), the modern roman liturgy, also known as gregorian liturgy or gregorian chant. This would become the official liturgy of the Catholic church (until today!) and substituted gradually the local ones. In the Iberian Peninsula, the Council of Burgos decreed the substitution of the hispanic rite by the modern roman one in 1080. This measure was easened by the fact that, during the christian reconquest, most part of the bishops were french (Gérard, Maurice Bourdin, Jean Péculier, Bernard, Hughes).

2.2 Profane music

In Portugal, it was cultivated at least since the independence (1143) an aristocratic poetical-musical genre whose texts are kept in three main collections (Cancioneiros):

- Cancioneiro da Ajuda (XIIIth century)
- Cancioneiro da Biblioteca Nacional (XVIth, on originais from the XIVth)
- Cancioneiro da Vaticana (XVIth, on originais from the XIVth)

The 1680 poems kept in the Cancioneiros are divided in three forms: *cantigas de amigo*, *cantigas de amor* and *cantigas de escárnio e maldizer*. The intrinsic link to music is well expressed in the *Cancioneiro da Ajuda*, where the staves have been drawn, but no melodies have been writen...

The only known musical source is due to a bookseller in Madrid, who found in the beginning of the XXth century a parchment with the 7 *Cantigas de Amigo* by Martin Codax, six of them with the respective melodies. Codax was a (galecian? portuguese?) troubadour from the court of King Dinis of Portugal (1261-1325) [4].

3. THE DEVELOPMENT OF POLYPHONY

We know nearly nothing about the introduction of polyphony in Portugal. Of course the Portuguese musicians were aware of the new polyphonic practices: for example, polyphony was practiced since early times in Santiago de Compostela, in the near Galicia, a famous pilgrimage centre in the medieval times. The earliest notice we have reflects already a developed stage of the polyphonic practice in Portugal. It is related to the presence in the court of king D. Fernando I (1345-1383) of the soon-to-be famous composer Jehan Simon de Haspre (*L'Hasprois*), who was a defender of the extremely sophisticated *Ars Subtilior* (a complex form of the already complex *Ars Nova*).

The new polyphonic practices developed in the main centres of musical activity in Portugal during the middle ages, Renaissance and Mannerist periods, which were the royal chapel, the royal court, the main cathedrals (particularly Évora, but also Coimbra and Braga),the main monasteries (Santa Cruz at Coimbra and Alcobaça) and the University.

3.1 The royal chapel

The royal chapel was founded by D. Dinis in 1299. D. Duarte (1391-1438) elaborated a Regiment (*Ordenaçam*) of the Chapel, which indicates that the standard practice was a three-voice

singing. His son, Afonso V (1432-1481), sent the *Mestre de Capela* (Master of the Chapel), Álvaro Afonso, to the court of Henry VI of England (1421-1471) in order to get a copy of the statutes, regiment and liturgy practiced in the English Royal Chapel. The detailed description written by William Say is still kept at Évora.

3.2 The Court

As with the trovadoresque poetry, we keep important collections of texts of the XVth and XVIth century (e.g. *Cancioneiro Geral*, compiled by Garcia de Resende), but the musical documents are fewer. The main sources of the court music in the Renaissance and Mannerist Periods are:

- C*ancioneiro de Elvas* (Públia Hortênsia Library, at Elvas)
- *Cancioneiro da Biblioteca Nacional* (National Library, Lisbon)
- *Cancioneiro de Paris* (*École Nationale Supérieure des Beaux Arts*, Paris)
- *Cancioneiro de Belém* (*Museu Nacional de Arqueologia e Etnologia*, Lisbon) [5]

The poetical forms are the *vilancete* (or *vilancico*), the *cantiga* and the *romance*. The first two, similar to the French *virelai* and to the Italian *ballata* (all probably descendants of the Arab *zejel*), are generally dedicated to the love thematic, though satire and social criticism are not excluded. They share a refrain and stanzas structure. The *romance* is dedicated to celebrate historical events, applying the same musical text to all the stanzas of the poem.

3.3 The Cathedrals

Cardinal-Princes D. Afonso (1509-1540) and D. Henrique (1512-1580), sons of D. Manuel I (1469-1521) administrated the main Portuguese dioceses through the XVIth century. Afonso administrated the Évora and Lisboa dioceses until his death. Henrique was successively Archbishop of Braga, Lisboa and Évora, as well as head of the Portuguese Inquisition. He became King of Portugal when his grand-nephew Sebastião I (1554-1578) died at Alcácer-Quibir (1578). As princes, they had their personal chapels and imposed a magnificent liturgy in the cathedrals they administrated.

In Évora, D. Afonso attracts high-quality musicians (like Mateus de Aranda, *Mestre de Capela* from 1528 to 1544) for the cathedral by establishing significant wages; Pedro do Porto (also known as Pedro Escobar, *El Portugués*), *Cantor* of the chapel of Isabel, *the Catholic Queen* and Master of the choir boys at Sevilla, comes as *Mestre de Capela* to Évora. He is the author of the most ancient polyphonic piece by a Portuguese author (a three-voice *Magnificat*), as well as the most ancient polyphonic treatment of the *Requiem* in the Iberian Peninsula. D. Afonso also founds a school for the choir boys, allowing them to study after the voice-change; many of these boys became professional musicians. This Évora school formed high-standard musicians for more than 150 years.

Besides Évora, Braga and Coimbra show a particular care in the liturgy. The most ancient version from a Mass by a Portuguese author is from a *Cantor* of the Coimbra cathedral, Fernão Gomes Correia (active 1505-32).

3.4 The Main Monasteries

The most important monasteries kept a solemn liturgy. From these, the Monastery of Santa Cruz, in Coimbra, had a particular importance. Founded in the XIIth century by D. Afonso Henriques, it was the first school of superior studies in Portugal (St. Anthony of Lisbon - or of Padova - studied

there). In the XVIth century, several monks distinguished by their musical gifts, as D. Heliodoro de Paiva and D. Francisco de Santa Maria. The musical performances at Santa Cruz competed with those at *El Escorial*, and were praised for their conciliation between polyphony and the respect for the sacred texts.

3.5 The University

The Portuguese University was founded in Lisbon by D. Dinis in 1290 and had a Music teacher as early as 1323. After several transfers between Coimbra and Lisbon, King João III (1502-1557) established it definitively at Coimbra in 1537. The move to Coimbra was followed by a reorganization in 1544, in which the King himself proposed Mateus de Aranda (*Mestre de Capela* at Évora after Pedro do Porto) as music teacher. The music teacher was also *Mestre de Capela* of the University.

4. THE MANNERIST PERIOD (2ND HALF XVITH AND XVIITH CENTURIES)

4.1 Historical context

In the end of the XVIth century, the circumstances lead to the disappearing of profane music in Portugal and a taking over by religious music. There are economical and political factors, like the troubles to keep the Portuguese conquests in Morocco and the competition lead by Venetians and Turks (later by Dutch and English) to the spice trade, which leads to the closing of the Portuguese *feitoria* (which was a kind of "spice supermarket") in Antwerp. In cultural terms, the influence of the Counter-Reformation in Portugal is enormous:

- João III introduces the Inquisition in Portugal in 1536; his brother Henrique will be the first General Inquisitor;
- The Jesuits come to Portugal in 1540 and soon start teaching in their own colleges in Coimbra and Lisbon. In 1555, they are in charge of the Arts College in Coimbra (the superior school in Portugal with most *prestige*), after the expulsion by the Inquisition of its most reputed teachers (like André de Gouveia);
- the Portuguese church participated actively in the Trento Council and, in 1564, Portugal becomes the only catholic country where the council decisions (namely those concerning the musical practice in the church) are integrally published as laws.

In this context, the profane music declines in the courts of João III and his grand-son Sebastião I. In 1578, with the death of Sebastião I, Cardinal Henrique becomes king of Portugal. After his death in 1580, Portugal loses its independence, as the throne is inherited by Filipe II, king of Spain (1527-1598). With the disappearance of the court in Lisbon, the aristocracy retires to their homes in the countryside, and the profane music nearly vanishes. The development of the Portuguese music in the end of XVIth is thus mainly in the sacred polyphony.

4.2 Climax of the Évora school

In 1575, Cardinal Henrique brings Manuel Mendes (?-1605), *Mestre de Capela* at Portalegre, to Évora, where he takes the *Mestre de Claustra* position. Besides his qualities as a composer, Manuel Mendes is remarkable as a teacher. He formed the most part of the extremely competent professional musicians who would have the most reputated musical positions in Portugal in the next decades. Between his students at Évora, we have the most noted poliphonists of the next

generation: Fr. Manuel Cardoso (1566-1650), Filipe de Magalhães (- 1652) and Duarte Lobo (1564/69-1646) [6]. These continued the pedagogical action of their teacher, worthing him references as «*mestre de toda boa musica deste reino*» («teacher of every good music in this kingdom») and «*el Mendes Sonoroso que de Musicos llena toda a Europa*» («the sound Mendes who replenishes Europe with musicians»).

4.3 Other centres of musical activity in the XVIIth century

4.3.1 Santa Cruz at Coimbra

The main composers in the XVIIth century are D. Pedro de Cristo, D. Pedro da Esperança and D. Gabriel de S. João. The manuscripts kept at the General Library of the University of Coimbra reveal innovative polyphonic practices, such as policorality, accompanied monody and instrument *obligato*.

4.3.2 The Royal Chapel

In spite of the absence of the King, it remains an important centre, with *Mestres de Capela* such as Francisco Garro, Filipe de Magalhães and Marcos Soares Pereira.

4.3.3 Chapel of the Dukes of Bragança at Vila Viçosa

During the Spanish domination, the dukes of Bragança retired to their palace in Vila Viçosa. The ducal chapel maintained a magnificent liturgy and, in 1609, Teodósio II founded the Santos Reis Magos College, working in a similar way to the Évora school. Roberto Tornar, *mestre de Capela* at Vila Viçosa, would become the musical instructor of the young Duke of Barcelos, D. João (later D. João IV, king of Portugal). D. João would become an melomaniac and, after inheriting the ducal title and even after becoming King of Portugal, would enlarge immensely the musical library of his father, transforming it in the biggest musical library of the time in Europe. D. João IV was a composer and a theorist himself, though a limited one. As a patron of the music, D. João paid a special friendship to his music-classes mate at Vila Viçosa, João Lourenço Rebelo (1610-1661), whose works he would send for printing at Rome. João Lourenço Rebelo, as the Santa Cruz friars, composed in an innovative style, making use of an opulent polichoral writing *à la Gabrieli* [7].

4.4 Instrumental music

It is in the organ domain, the liturgical instrument *par excellence*, that is placed the essential of the instrumental music in the mannerist period. Portuguese organs, as well as the Spanish ones, characterized for the existence of only one manual, without pedals. The iberian organ has original characteristics as the *meio-registo* ("half-stop", dividing the keybord in two distinct parts and allowing accentuated timbrical contrasts between the two halfs) and the horizontal placing (*"em chamada"*) of particularly strident pallet stops.

The first printed volume of Portuguese instrumental music is the *"Flores de Música para o instrumento de tecla e harpa"* (*"Music flowers for the keyboard instrument and harp"*), by Manuel Rodrigues Coelho (1620), which contains only sacred compositions. Coelho worked as an organist in the Badajoz, Elvas and Lisbon cathedrals and finally in the Royal Chapel [8].

In Braga, appears to have developed a flourishing organ school in the XVIIth century, dominated by Gaspar dos Reis, *Mestre de Capela* of the cathedral. Other relevant composers are Pedro de Araújo and Fr. Diogo da Conceição.

In the tipically iberian repertoire, we count the *Tento de Meio-Registo* (Half-Stop Tento) and the *Batalha* (Battle). This last form goes back to one of the most famous pieces by Clément Jannequin - *La bataille de Marignan ou La Guerre*, in which are imitated the characteristic sounds of a battle. The iberian composers tried to use the same effects in sacred works, in an allusion to the mystical battle between good and evil [9].

5. THE BAROQUE PERIOD AND THE ITALIAN INVASION

5.1 João V, *the Magnificent*

In the end of the XVIIth, Portuguese composers gradually evolve towards the new musical language that would result in the modern tonalism. The government of D. João V (1689-1750) marks a profound transition in the Portuguese society and culture. After the definitive peace with Spain, the monarch will try to modernize the Portuguese economy and to drive the country to a development scheme similar to the French Absolutism of Louis XIV. The main originality on D. João V's absolutism is that he managed, using his influence with the Pope, to face the huge political, economical and cultural power of the Church, by reorganizing it in order to strength its unity and discipline and then putting it under the royal authority. In a very clever process, João V got for his chapel the dignity of Patriarchal Basilica, by dividing the Lisbon archdioceses. The chaplain became a Cardinal. Then he got the reunification of the dioceses under the command of the royal chaplain. So, the Cardinal-Patriarch, Archbishop of Lisbon, was merely the chaplain of the King of Portugal...

João V took a special care with the liturgy in his chapel, with he wanted as monumental as the Papal chapel in Rome. He got it repeating somehow the formula of Cardinal D. Afonso two hundred years before: contracting high-standard professional musicians and creating structures for the adequate formation of Portuguese musicians. As such, he contracted the brilliant Master of the *Capella Giulia*, in Rome, Domenico Scarlatti, as *Mestre da Capela Real* and music teacher of princess D. Maria Bárbara and founded in 1713 a specialized school annex to the Patriarchal Basilica: the Patriarchal Seminary, which would become the major music school in Portugal and form generations of professional musicians of remarkable quality until the foundation of the National Conservatory in 1835. The most gifted students of the Patriarcal Seminary were sent to Rome at the King's expenses. Those were the cases, namely, of António Teixeira, João Rodrigues Esteves and Francisco António de Almeida, who were hence formed in the Roman ecclesiastic baroque school and had the chance of getting acquainted with the roman operatic tradition.

5.2 The incipient Opera

It is in 1728 that takes place in the Ribeira Royal Palace at Lisbon the first performance of *Il Don Chisciotte della Mancia*, with music by Scarlatti. This was the first operatic-style performance in Portugal and was followed by other *opera buffa* performances in the Royal Palace in the next years. However, they had little impact in the music life, not only because the public was extremely restricted, but also because the King did not pay them much attention. It was in the Trindade Theatre, in 1735, that the Alessandro Paghetti company had permission to perform the first *opera seria* for a wider (aristocratic) audience. The success was enormous and the company kept performing until 1742, now in the Rua dos Condes Theatre. At the same time, had begun (1733)

in the Bairro Alto Theatre a set of performances in Portuguese of the António José da Silva's (*o Judeu*) plays, with music by António Teixeira. The audience of these plays was even wider. However, D. João V got ill in 1742, and the mysticism in which he fell in the his last years implied that every theatrical performance was forbidden until his death.

5.3 Instrumental music

Undoubtedly, the most important Portuguese keyboard composer of the time is José António Carlos de Seixas (1704-1742). Son of Francisco Vaz, organist of the Coimbra Cathedral, Carlos Seixas comes only with sixteen, but already very famous to Lisbon, where he is appointed as organist of the Patriarchal Cathedral. There, he would soon be appointed as *Vice-Mestre de Capela* (the *Mestre de Capela* was Scarlatti himself and Seixas was, at the time, the only Portuguese member of the Royal Chapel). Seixas has left us 105 two-part baroque *Sonatas* (or *Tocatas*) for keyboard. He also wrote religious and orchestral music. However, his most original contribution is a Concert for harpsichord and strings, one of the first examples of this form in Europe [10].

5.4 Opera and Sacred music under D. José I and D. Maria I

With D. José I (1714-1777), the operatic activity is taken again. The neapolitan David Perez (1711-1778), one of the most reputed Italian opera composers, is hired in 1752. The climax of Perez activity would be the inauguration of the monumental Tejo's Opera, in March 1755, with the opera *Alessandro nell'Indie*. But the earthquake of the 1st November 1755 destroyed the new building, together with Lisbon downtown. The royal palace also disappeared, and with it the musical Library of D. João IV.

After the earthquake, the public theatres like the Rua dos Condes Theatre and the Bairro Alto Theatre are rebuilt (but not the *Tejo's Opera*). Already under D. Maria I, would be built the S. Carlos Theatre, in Lisbon (1792) and the S. João Theatre in Oporto (1798). The neapolitan influence is enormous and, under D. José and D. Maria, the gifted music students of the Patriarchal are sent to Santo Onofre Conservatory in Naples. Afterwards, these students distinguished in the Neapolitan operatic style, as well as in the sacred music. Between these we have João de Sousa Carvalho(1745-1798), a Vila Viçosa school student and perhaps the most prominent composer of the 2nd half of the XVIIIth century. Besides his operatic and sacred music production, he may also be considered the most remarkable keyboard composer of the time.

Other relevant portuguese composers of the time are Jerónimo Francisco de Lima, Luciano Xavier dos Santos, José Joaquim dos Santos, José dos Santos Maurício, António Leal Moreira and, particularly, Marcos Portugal, perhaps the Portuguese composer with the most international career ever [11].

6. THE XIXTH CENTURY

With the Napoleonic invasions, the Royal family goes to Brazil and the court establishes in the Rio de Janeiro. This presence would conduce to the independence of this colony (1822) and would be benefic as well to the development of Brazilian music (the first significative Brazilian composer is José Maurício Nunes Garcia, member of the royal chapel at the Rio de Janeiro). Meanwhile, the constitutional *régime* is proclaimed (1820) and King D. João VI (1767-1826) is forced to come back. The activity of the Royal Chamber Orchestra (founded by D. João V), which had been in the previous century one of the most important chamber orchestras in Europe, declines

irreversibly. However, in the turn of the century, generalizes the tradition of amateur academies performing the contemporary instrumental music. The generalization of public concerts is due to João Domingos Bomtempo (1775-1842), the most prominent musical figure of the first half of the XIXth century.

Bomtempo, son of an Italian musician of the court Orchestra, studied with the Patriarchal masters. Unlike most of his contemporaries, he was not interested in opera and, in 1801, instead of going to Italy, he travels to Paris, starting a *virtuoso* pianist career. He moves to London in 1810 and gets acquainted with the liberal circles. In 1822 he is back to Lisbon, and founds a Philharmonic Society to promote public concerts of the contemporary music. After the civil war between liberals and absolutists, Bomtempo becomes music teacher of Queen D. Maria II (1819-1853) and first Director of the National Conservatory, created in 1835 and which replaced the old Patriarchal Seminary, extinct by the liberal *régime*. As a composer, Bomtempo produced a vast amount of *concerti*, sonatas, variations and fantasias for the piano. His two known symphonies are the first to be produced by a Portuguese composer. The master piece of Bomtempo is his *Requiem to the memory of Camões* [12].

7. ANTECEDENTS OF THE ACTUALITY

7.1 The turn of the century

All over the XIXth century, there is a proliferation of the concert societies. Bernardo Moreira de Sá (1853-1924), in Oporto, is the director, among others, of the Quartet Society and forms the Moreira de Sá Quartet, which will have an international career. He will have a decisive influence in the Formation of the Oporto Conservatory (1917). However, opera remained as the favourite activity of Portuguese composers, though the creative activity moved slowly towards the symphonic and chamber music fields. The two most significative lyric composers are Alfredo Keil (1850-1907) and Augusto Machado (1845-1924).

José Vianna da Motta (1868-1948) and Luiz de Freitas Branco (1890-1955) have a special place in the Portuguese musical life in the turn of the century.

7.1.1 Vianna da Motta

Vianna da Motta went to Scharwenka Conservatory in Berlin in 1882 at King Fernando II's expenses. He also attended Liszt classes at Weimar in 1885, as well as Hans van Bülow's. In Germany, he started a career as a concertist and exceptional interpret of Bach, Beethoven and Liszt. During the First World War, he taught at Geneva Conservatory. In 1917, he came back to Portugal, becoming director of the National Conservatory. As a composer, he is very close to the German Romanticism, and dedicates himself to the production of a national style, by including and recreating the national folklore. His most emblematic work is the *A Major* Symphony *"À Pátria"* (1895) [13].

7.1.2 Luiz de Freitas Branco

Luiz de Freitas Branco (1890-1955) is usually appointed as the «introducer of modernism in Portugal», by his decisive role in the approximation of Portuguese music to the most innovative European aesthetics, namely the Schönberg atonalism and the French impressionism. Pupil of Augusto Machado and Tomás Borba, he studied with the Belgian organist and composer Désiré Pâque and, in 1910, went to Berlin to study with Humperdinck. There, he attended to a

performance of Debussy's *Pélleas et Mélisande*, which was determinant in his aesthetic orientation. In his early work we count the symphonic poems *"Váthek"* and *"Paraísos Artificiais"* and several piano pieces [14]. His prolific production includes five symphonies, a violin concert and inumerous vocal works.

7.1.3 Other composers

In the turn of the century, other relevant composers are Francisco de Lacerda (1869-1934) [15], Óscar da Silva (1870-1958), Luiz Costa (1879-1960) and António Fragoso (1897-1918). Lacerda was as well a famous director specialist in the French and Russian repertoire. He became assistant of d'Indy at the *Schola Cantorum* in Paris. His musical language is very close to that of Fauré and Debussy.

7.2 The Estado Novo régime

The military *coup* of 1926 installed in Portugal a dictatorship (self-called *Estado Novo*, "the new state") which would condition the Portuguese life for near half century. The concept of culture is substituted, in the main stream of European fascisms, by the concept of *propaganda*. This propaganda had its maximum height at the Nationality Centenary in 1940; the S. Carlos Theatre was then reopened after a restoration with an opera by the regime semi-official composer Ruy Coelho. Curiously, the most important figure of Portuguese musical life in that period is a composer who openly contested the *régime* and its aesthetic orientations and who, consequently, was forced to do his entire activity outside the institutional circuits: Fernando Lopes Graça.

7.3 Lopes Graça

Fernando Lopes Graça (1906-1995) was student of Tomás Borba, Luiz de Freitas Branco and Vianna da Motta at the National Conservatory and finished the Superior Course on Composition in 1931. He tried to get a position at that institution, but was arrested by political reasons and the place was not conceded to him.

He taught for some time in the Music Academy in Coimbra and, in 1937, went to Paris at his expenses, where he studied musicology. There he composed the first works of his musical maturity (2nd Piano Sonata, Quartet for Violin, Viola, Cello and Piano). After coming back to Portugal in 1939, Lopes Graça taught at the *Academia de Amadores de Música* at Lisbon. Of his production, it worth mentioning the numerous harmonizations or adaptations of popular Portuguese songs for choir or soloist, the songs for voice and piano over the poems of the most important Portuguese poets, the inumerous political songs, as well as the symphonic music, chamber music and piano music production. Lopes Graça undertook, with the Corsican ethnologist Michael Giacometti, a systematic study of Portuguese folk music, which he assimilated and used thoroughly in his musical speech. His view from the folklore is far from the regime bucolic or picturesque view, rather strengthening the hard dimensions of rural life [16,17].

The contemporaries of Lopes Graça generally choose a more pacific conservative "neo-classic" style: these were the cases of Armando José Fernandes (1906-1983), Jorge Croner de Vasconcelos (1910-1974), Frederico de Freitas (1902-1980), Joly Braga Santos (1924-1988) and Cláudio Carneyro (1895-1963).

8. WHAT ABOUT NOW?

The *coup* of April, 25th of 1974 restored the democracy in Portugal. In these last twenty five years the country knew a great development, particularly since the adhesion to the European Economic Community (now European Union) in 1986. The intellectual and cultural life had particular improvements (these seem not to be very cherished areas under dictatorships...). The music has also benefited from the increasing number of Conservatories and specialized superior schools, in a freedom context, as well as from the generalization of Music Festivals. The role of Foundation Calouste Gulbenkian (founded in 1953) has been of outstanding importance in every aspect of the cultural life, particularly the musical one (the Foundation has been the Portuguese "Ministry of Culture" for decades...). If, however, Portuguese Physics has already got a solid place in the international community (this CERN school is an experimental evidence!), Portuguese music still has a much more fragile situation. There is still much to do... This can be a challenge for Portuguese musicians and politicians, as they have in their hands the possibility and the responsibility for a fruitful continuation of this history.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Rui Marques for his kind challenge, as well as to my colleagues at the Grupo de Física Nuclear e Óptica Aplicadas for their encouragement and support.

REFERENCES

[1] Nery, Rui Vieira and Castro, Paulo Ferreira de, *História da Música*, 2nd edition, «Sínteses da cultura portuguesa», Lisbon: Imprensa Nacional Casa da Moeda (1999)

[2] Branco, João de Freitas, *História da Música Portuguesa*, 3rd edition, Lisbon: Publicações Europa-América (1995)

[3] Cortez, José Carlos Travassos, Lessons on *História da Música III*, Coimbra: Conservatório de Música (1995/96)

[4] Pickett, Phillip, introductory note to *The Pilgrimage to Santiago*, London: Editions de l'oyseau-lyre, Decca (1991)

[5] Nery, Rui Vieira, introductory note to *Portuguese Mannerist Music: Cancioneiro Musical de Belém*, Lisbon: Movieplay (1988)

[6] Mallavibarrena, Raúl, introductory note to *Duarte Lobo - Requiem and Mass Vox Clamantis: the Tallis Schollars*, Barcelona: Editiones Altaya (1998)

[7] Nevel, Paul van, introductory note to *João Lourenço Rebelo: Vespers, Lamentations*, s.l.: Crépuscule (1994)

[8] Doderer, Gerhard, introductory note to *Manuel Rodrigues Coelho: Flores de Música I*, Lisbon: Strauss (1998)

[9] Nery, Rui Vieira, introductory note to *Batalhas & Meios Registos: Iberian Organ Music of the seventeenth century*, Lisbon: Movieplay (1994)

[10] d'Alvarenga, João Pedro, introductory note to *Carlos Seixas: Harpsichord Concerto, Sinfonia, Harpsichord Sonatas*, London: Virgin Classics (1995)

[11] d'Alvarenga, João Pedro, introductory note to *João de Sousa Carvalho, Luciano Xavier dos Santos, José Joaquim dos Santos: Sacred Vocal Works*, Lisbon: Movieplay (1995)

[12] Graça, Fernando Lopes, introductory note to *Bomtempo: Requiem Op. 23 (in the memory of Camões)*, Lisbon: Strauss (1997)

[13] Branco, João de Freitas, introductory note to *Vianna da Motta: Symphony Op. 13 "À Pátria"*, 3rd edition, Lisbon: Strauss (1996)

[14] Maissa, Nella, introductory note to *Música do Século XX*, Lisbon: Jorsom (1994)

[15] Sousa, Filipe de, introductory note to *Francisco de Lacerda: Four Orchestral Pieces*, 2nd edition, Lisbon: Strauss (1995)

[16] Giacometti, Michel and Graça, Fernando Lopes, introductory note to *Portuguese Folk Music*, 5 volumes, Lisbon: Strauss (1998)

[17] Leitão, Eugénio, introductory note to *Fernando Lopes Graça: Portuguese Folk Songs*, 2nd edition, Lisbon: Strauss (1995)