

CERN 99-04
28 July 1999

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

1998 EUROPEAN SCHOOL OF HIGH-ENERGY PHYSICS

St Andrews, Scotland
23 August–5 September 1998

PROCEEDINGS

Editors: N. Ellis, J. March-Russell

GENEVA
1999

© Copyright CERN, Genève, 1999

Propriété littéraire et scientifique réservée pour tous les pays du monde. Ce document ne peut être reproduit ou traduit en tout ou en partie sans l'autorisation écrite du Directeur général du CERN, titulaire du droit d'auteur. Dans les cas appropriés, et s'il s'agit d'utiliser le document à des fins non commerciales, cette autorisation sera volontiers accordée.

Le CERN ne revendique pas la propriété des inventions brevetables et dessins ou modèles susceptibles de dépôt qui pourraient être décrits dans le présent document ; ceux-ci peuvent être librement utilisés par les instituts de recherche, les industriels et autres intéressés. Cependant, le CERN se réserve le droit de s'opposer à toute revendication qu'un usager pourrait faire de la propriété scientifique ou industrielle de toute invention et tout dessin ou modèle décrits dans le présent document.

Literary and scientific copyrights reserved in all countries of the world. This report, or any part of it, may not be reprinted or translated without written permission of the copyright holder, the Director-General of CERN. However, permission will be freely granted for appropriate non-commercial use.

If any patentable invention or registrable design is described in the report, CERN makes no claim to property rights in it but offers it for the free use of research institutions, manufacturers and others. CERN, however, may oppose any attempt by a user to claim any proprietary or patent rights in such inventions or designs as may be described in the present document.

ISSN 0531-4283

ISBN 92-9083-146-4

CERN 99-04
28 July 1999

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

1998 EUROPEAN SCHOOL OF HIGH-ENERGY PHYSICS

St Andrews, Scotland
23 August–5 September 1998

PROCEEDINGS

Editors: N. Ellis, J. March-Russell

GENEVA
1999

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, detection of gravitational waves, and lattice QCD. They also contain lectures on experimental techniques, the high-energy physics programme at JINR, the science behind Dolly the sheep, and malt whisky.

PREFACE

The 1998 European School of High-Energy Physics was held at the University of St Andrews, Scotland from 23 August to 5 September. The School was attended by 101 students, representing 27 different nationalities, mainly from CERN and JINR Member states. The 1998 School was the sixth in the new series organized on a yearly basis in collaboration between CERN and JINR, Dubna. The School was sponsored by CERN and JINR, and fifteen students from Eastern Europe and the former Soviet Union countries received financial support from the European Union and Unesco. Local support was obtained from the University of Edinburgh and from the Particle Physics and Astronomy Research Council.

Our sincere 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.

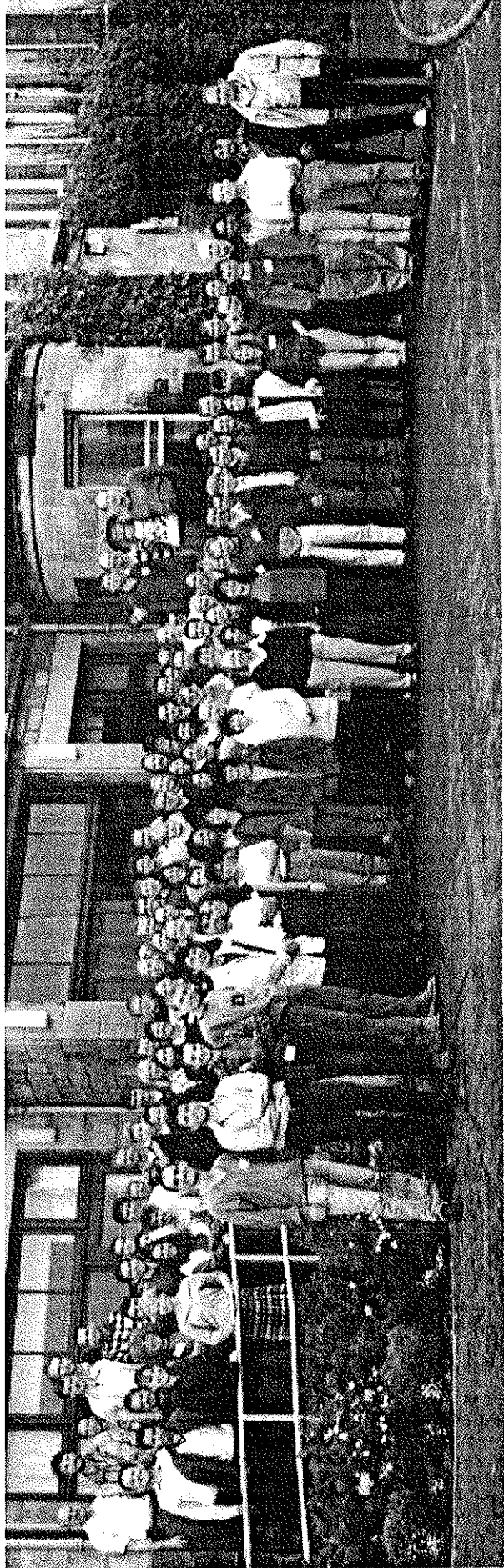
The major responsibility for organizing the 1998 School was with the University of Edinburgh. Our thanks are extended to Dr Alan Walker, who acted as Director of the School, and to the local organizing Committee and support staff from Edinburgh University who helped in setting up the School and taking care of the local arrangements. It has to be stressed that Alan took a very large share of the burden, and that the smooth running of the day-to-day organization and the social programme to a large extent was due to his devotion.

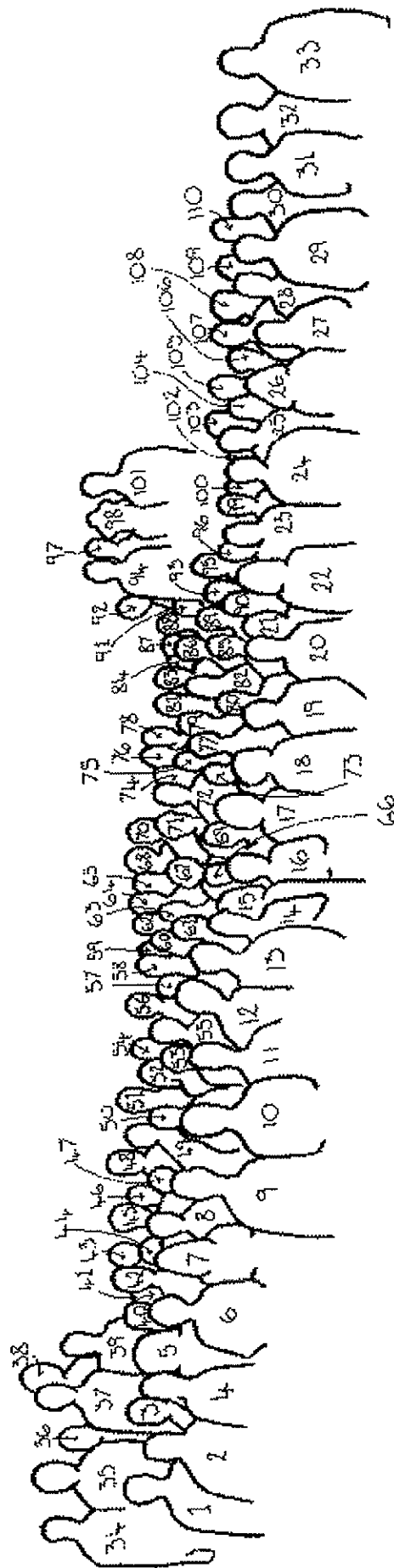
We are as always grateful to Susannah Tracy and Tatyana Donskova for their untiring efforts in the lengthy preparations for and the day-to-day care of the School. A special word of thanks goes to the Residence Managers, Jackie Smith and Ione Gilmour, and their staff for all their help and hospitality. Our thanks also go to the chef, Donald Smith, and his kitchen staff for providing an excellent selection of Scottish food.

It has become a tradition at the School to hear a special lecture outside the field of Particle Physics. And what is more appropriate in Scotland than an after dinner lecture on malt whisky? Dr P. M. Dryburgh gave an excellent overview of the history and traditions of whisky production followed by the tasting of ten specially selected 'malts'.

The students will remember playing golf on the most famous golf course in the world, a tough walk to the highest mountain in the area, and an excursion to Edinburgh. The success of the 1998 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





PEOPLE IN THE PHOTOGRAPH

- | | | | |
|-----|---------------------|-----|-------------------------|
| 70 | Aleksa Martin | 2 | Ghodbane Nabil |
| 75 | Alemaney Reyes | 44 | Giovannella Simona |
| 3 | Ambrosino Fabio | 74 | Gschwendtner Edda |
| 101 | Andress John | 63 | Guy Leanne |
| 66 | Arino Ignacio | 43 | Hansen Allan |
| 93 | Ball Patricia | 26 | Haroyan Lilith |
| 81 | Bastos Joao | 103 | Heising Stephan |
| 105 | Beckmann Marc | 39 | Hubaut Fabrice |
| 14 | Biguzzi Andrea | 57 | Huentemeyer Petra |
| 84 | Biscarat Catherine | 5 | Kaczmarska Anna |
| 95 | Borisov Vladimir | 109 | Kaestli Hans-Christian |
| 99 | Bouhali Othmane | 6 | Kiryluk Joanna |
| 82 | Buchalla Gerhard | 4 | Klimek Katarzyna |
| 42 | Buirra-Clark Daniel | 29 | Knowles Ian |
| 48 | Buis Ernst-Jan | 72 | Kokkonen Jukka |
| 79 | Cabibbo Giorgio | 56 | Krueger Henning |
| 17 | Cabrera Susana | 91 | Kupco Alexander |
| 36 | Chernyak Dmitry | 25 | Lebedev Igor |
| 80 | Coppola Nicola | 55 | Lehmann Giovanna |
| 67 | Coxe Robin | 87 | Lehti Sami |
| 11 | Cwiok Mikolaj | 104 | Leinonen Lena |
| 8 | Czermak Agnieszka | 97 | Lillestøl Egil |
| 69 | Da Silva Tatiana | 77 | Malina Roman |
| 30 | Del Duca Vittorio | 96 | Mangeol Dominique |
| 41 | Deschamps Olivier | 88 | Mannert Clemmens |
| 19 | Donskova Tatyana | 60 | Martin Victoria |
| 88 | Drage Lee | 73 | Mazzucato Federica |
| 76 | Dubbert Joerg | 13 | Mirea Adrian |
| 51 | Dudarev Andrey | 47 | Molokanova Natalia |
| 98 | Ellis John | 68 | Mommsen Remi |
| 78 | Ellis Malcolm | 15 | Moraes Danielle |
| 102 | Fagerstroem Peter | 64 | Moshous Basil |
| 27 | Farrington Sinead | 45 | Movilla Fernandez Pedro |
| 65 | Fox Harald | 52 | Muijs Sandra |
| 22 | Gavela Belén | 94 | Neufeld Niko |

107	Nir Yossi	59	Stojnev Stojan
33	Olchevski Sacha	61	Striegel Dieter
9	Oztuk Hakan	58	Tegenfeldt Fredrik
16	Paiva Raquel	28	Teryaev Oleg
18	Palomares Carmen	86	Thiergen Michael
53	Perez-Ochoa Rocio	20	Tracy Susannah
12	Pestotnik Rok	83	Vannerem Philippe
31	Piotto Enrico	21	Varanda Maria
49	Piskunov Oleg	106	Vasiljev Alexey
35	Popov Boris	7	Versille Sophie
50	Ruh Michael	37	Vishnevsky Dmitry
32	Rusu Vadim	46	Visser Jan
90	Schmitt Stefan	1	Voros Sandor
38	Shmatov Sergei	54	Voss Helge
40	Simard Laurent	34	Walker Alan
110	Simon Jürgen	62	Wegner Martin
24	Slepnev Roman	85	Ziegler Thomas
108	Somov Sasha		

CONTENTS

Preface	v
Field Theory and the Standard Model – Abstract	
<i>V. Novikov</i>	1
Lecture I: Quantum field theory. Birds-eye view	2
Lecture II: Symmetries	14
Lecture III: $SU(2) \times U(1)$ theory of electroweak interactions	25
Lecture IV: Higgs, W and Z	34
Lecture V: Radiative correction in SM	44
Introduction to QCD	
<i>M.L. Mangano</i>	53
1. Introduction	53
2. QCD Feynman rules	54
3. Renormalization, or ‘Theorists are not afraid of infinities!’	62
4. QCD in $e^+ e^-$ collisions	71
5. QCD and the proton structure at large Q^2	78
6. QCD in hadronic collisions	86
Flavor physics and CP Violation	
<i>Y. Nir</i>	99
1. Flavor physics	99
2. The mixing parameters	103
3. CP violation in meson decays: a model independent discussion	110
4. CP violation in the Standard Model	118
5. CP violation beyond the Standard Model	123
Beyond the Standard Model for Hillwalkers	
<i>J. Ellis</i>	133
1. Getting motivated	133
2. Introduction to supersymmetry	143
3. Phenomenology of supersymmetry	153
4. Grand unification	165
5. None of the above	178
Cosmology and Particle Physics	
<i>J.A. Peacock</i>	197
1. The isotropic universe	197
2. The hot big bang	208
3. Relics of the big bang	214
4. Inflationary cosmology	225
5. Evidence for vacuum energy at late times	240
6. Dynamics of structure formation	245
7. Cosmological density fields	257
8. Cosmic background fluctuations	279

The Detection of Gravitational Waves	
<i>S. Rowan and J. Hough</i>	301
1. Introduction	301
2. Detection of gravitational waves	302
3. Conclusion	309
An Introduction to Lattice QCD	
<i>R.D. Kenway</i>	313
1. Why do we need numerical simulations of QCD?	313
2. How is the computation done?	316
3. The determination of quark masses	321
4. The QCD spectrum	325
5. Conclusions	327
Heavy Quark Physics from Lattice QCD	
<i>C.T.H. Davies</i>	331
1. Introduction	331
2. Heavy quarks on the lattice	331
3. The heavy-light spectrum	334
4. Decay constants	337
5. Semi-leptonic matrix elements	341
6. Neutral B mixing	343
7. Conclusions	344
Experimental Techniques	
<i>T.S. Virdee</i>	347
1. Introduction	347
2. Interaction with matter	347
3. Experimental measurements: measurement of momentum	352
4. Measurement of energy	355
5. Energy resolution	361
6. Identification of particles	368
7. The experimental challenge at the LHC	381
8. The proton–proton experiments at the LHC	381
9. Inner tracking	386
10. Electronics noise	398
11. Inorganic scintillators	401
12. Calorimetry using noble liquids	407
13. Combined e.m. and hadronic calorimetry	412
14. Trigger and data acquisition system	413
15. Conclusion	418
JINR Programme in High Energy Physics	
<i>A. Sissakian</i>	421
1. General information about JINR	421
2. JINR is a major partner of world’s hep laboratories	425
3. JINR’s scientific potential	428
4. Dubna as an educational centre	434
5. Plans for future	435
6. The status of HEP in Russia and FSU countries	436
7. Conclusion	437

Malt Whisky – An Introduction to the Spirit of Scotland	
<i>P. M. Dryburgh</i>	439
Introduction	439
Lowland malts	440
Highland malts	440
Islay malts	441
Dolly: The Science Behind the World’s Most Famous Sheep	
<i>H.D. Griffin</i>	443
Dollymania	443
‘Science breakthrough of the year’	444
Cloning of humans	444
Nuclear transfer	445
Practical applications of cloning	446
Further information on cloning	449
Organizing Committee	451
Lecturers	451
Discussion Leaders	452
List of Students	452
List of Posters	455
Photographs	458

Field Theory and the Standard Model

V. Novikov
ITEP, Moscow

Abstract

The following 5 lectures are devoted to key ideas in field theory and in the Standard Model.

Lecture I. Quantum Field Theory. Bird's-eye-view.

Introductory remarks.

Quantum Field Theory (QFT) is the working language in the community of high energy physicists. It is not the esoteric theory accessible to the small group of experts, the basic ideas of QFT have to be familiar to all members of the community. There are number of excellent textbooks on QFT. The list of the recommended books can be found in ref [1]. As a rule they are rather lengthy, the average size of the standard textbook is of the order of 800 pages. The only way to understand QFT is to take one of these books and to spend one year or more to study it. It is hard way, but nobody knows the better one. So is life!

The goal of these lectures is not to provide any systematic introduction to the subject. It is impossible to do in five lectures. The goal is to remind the students what they actually had studied a few years ago. Just the basic concepts, notions and relations of QFT without long derivations and boring formalism.

1.1. Particles and Fields.

In the Classical Physics the particles and fields are very different dynamical systems. Particles are particles and fields are fields. There is no way to confuse these notions. This is everyday wisdom.

The system of particles has finite number of degrees of freedom N . To describe this dynamical system one have to introduce the general coordinates $q_i(t)$ ($i = 1, 2, ..N$) and their time derivatives $\dot{q}_i(t)$ or conjugated momenta $p_i(t)$. Either we study the bounded motion or the scattering processes at any time we can say how many degrees of freedom the system has. Even when we observe the decay of the system we are sure that outgoing particles were bounded inside the initial system before the decay. This is evident.

Field theory is a theory of the system with infinite number of degrees of freedom. To describe the electromagnetic fields we have to know four-potential A_μ at every space point x . Maxwell equations govern the evolution of the field in time.

The Quantum Mechanics (QM) of nonrelativistic particles was developed in 1925-26. In QM dynamical system with N degrees of freedom is described by wave function

$$\Psi(q, t) = \Psi(q_1, \dots, q_N; t) \quad (1.1)$$

that satisfies the wave equation

$$i \frac{\partial}{\partial t} \Psi(q, t) = H(p, q) \Psi(q, t) \quad (1.2)$$

where $H(p, q)$ is the Hamiltonian and p is the operator of momentum: $p = -i\partial/\partial q$. The number of degrees of freedom N was supposed to be fixed exactly like in Classical Mechanics.

The first quantization of electromagnetic fields had been done at the same time in 1926 by Born, Heisenberg and Jordan in their second paper on QM. They represented radiation electromagnetic field as an infinite set of harmonic oscillators and quantized these oscillators. They found that excitations of the oscillators behave like a free massless particles – photons. The number of photons was not fixed. They were created and destroyed by charged particles. Quantized theory of electromagnetic field became a theory of particles – photons. Photons were not "bounded" inside charged particles, they were created from "nothing" by scattered charged particles. The physical idea of photons was introduced by Einstein twenty years before

this paper, but the formal quantization of field showed that quantized field is equivalent to the system of particles that can be created and destroyed.

For some time physicists tried to find a relativistic version of wave equation (1.2) for the particles at high energy.

The first such equation was written in 1926 by Klein and Gordon for spin 0 relativistic particle

$$-\partial_\mu\partial_\mu\Phi(x) = m^2\Phi(x) \quad (1.3)$$

where $\partial_\mu = \partial/\partial x_\mu$ and $\Phi(x)$ is a complex function of $x = (t, \vec{x})$, m is a mass of particle.

Dirac pointed out that eq. (1.3) and the function $\Phi(x)$ can't be interpreted as a wave equation and wave function. In 1928 he suggested his own relativistic equation for spin 1/2 particles:

$$(i\gamma_\mu\partial_\mu - m)\Psi(x) = 0 \quad (1.4)$$

Here Ψ is a column with 4 complex components (4-spinor) and γ_μ are 4×4 matrices.

The troubles with interpretation of eq. (1.4) as the one-particle relativistic wave equation are not so evident as for the case of eq. (1.3). But the truth is that for any relativistic processes the single particle description breaks down. Any relativistic system has infinite numbers of degrees of freedom. More energy we pump into the system, more degrees of freedom can be excited. For example any scattering process in QED can be accompanied by creation of additional e^+e^- pairs. These pairs are not hidden inside initial particles, they are created during the scattering process. The natural description of relativistic physics is quantum field theory. So it is wrong to divide world on particles and fields. We have to use the quantum field theory for everything.

From this point of view both the Klein-Gordon and Dirac equations are not relativistic wave equations. They are field equations for scalar and spinor fields. These fields have to be quantize. The lowest excitations of these quantum fields behave like massive particles with spin 0 and 1/2 respectively.

The QFT is the right language for dealing with particle physics. This language is not unique. For example string theory also pretends to describe particles in low energy limit. We have also to note that one can construct "diagrammatica" (i.e. the set of rules for calculation of amplitudes in perturbation theory) without any reference to QFT.

1.2. Quantization and the Fock space.

Consider a field theory for a free scalar particle:

$$\Phi(x) = \Phi(\vec{x}, t) \quad (1.5)$$

In the "Classical Theory" $\Phi(x)$ is a real function of space-time point $x_\mu = (t, \vec{x})$ with the lagrangian density $\mathcal{L}(\Phi, \partial_\mu\Phi)$

$$\mathcal{L}(\Phi, \partial_\mu\Phi) = \frac{1}{2}\{\partial_\mu\Phi\partial_\mu\Phi - m^2\Phi^2\} \quad (1.6)$$

where $\partial_\mu\Phi = \frac{\partial}{\partial x_\mu}\Phi$ and m is the mass of the particle.

The action S is given by

$$S = \int d^4x \mathcal{L}(\Phi, \partial_\mu\Phi) \quad (1.7)$$

The hamiltonian density is constructed according to the rules of hamiltonian dynamics

$$\mathcal{H} = \pi \frac{\delta\mathcal{L}}{\delta\dot{\Phi}} - \mathcal{L} = \frac{1}{2}\{\pi^2 + (\nabla\Phi)^2 + m^2\Phi^2\} \quad (1.8)$$

where

$$\pi = \frac{\delta \mathcal{L}}{\delta \dot{\Phi}} = \dot{\Phi} = \partial_0 \Phi$$

We use the natural units where $c \equiv 1$ and $\hbar \equiv 1$. So the action is dimensionless

$$[S] = m^0 \text{ ,}$$

and for other quantities we get

$$\begin{aligned} [E] &= [p] = m \\ [x] &= m^{-1} \\ [\mathcal{L}] &= [\mathcal{H}] = m^4 \\ [\phi] &= m \end{aligned} \tag{1.9}$$

Exercise: prove the dimension rule eq. (1.9)

Equations of motion are derived from Hamilton variational principle

$$\begin{cases} \delta S = 0 \\ \partial_\mu \left(\frac{\delta \mathcal{L}}{\delta \partial_\mu \Phi} \right) = \frac{\delta \mathcal{L}}{\delta \Phi} \end{cases} \tag{1.10}$$

Euler-Lagrange equations (1.10) for the density eq. (1.6) coincides with Klein-Gordon equation

$$(\partial^2 + m^2)\Phi = 0 \tag{1.11}$$

Consider the plane wave ansatz for the solution of eq. (1.11)

$$\Phi_{\vec{p}}(x, t) = a(t)e^{i\vec{p}\vec{x}} \tag{1.12}$$

The equation for the amplitude a

$$\ddot{a} + (\vec{p}^2 + m^2)a = 0 \tag{1.13}$$

is an equation for linear oscillator with frequency

$$\omega^2(p) = \vec{p}^2 + m^2 \text{ ,}$$

or

$$\omega(p) = \pm \sqrt{p^2 + m^2}$$

We get that dependence of frequency ω on \vec{p} and the dependence of particle energy on momentum \vec{p} are exactly the same (in the units $\hbar = c = 1$). This is why we can use free fields to describe free particles.

The general solution in the periodic box can be presented as a superposition of the solutions (1.12)

$$\Phi(x) = \sum_p [a(p)e^{-ipx} + a^+(p)e^{ipx}] \tag{1.15}$$

where

$$\begin{aligned} px &= p_\mu x_\mu = p_0 x_0 - \vec{p}\vec{x} \\ p_0 &= \sqrt{\vec{p}^2 + m^2} \\ \sum_p &= \int \frac{d^3 p}{(2\pi)^3 2p_0} \end{aligned}$$

In the Classical theory coefficient $a(\vec{p})$ are the arbitrary complex numbers. In terms of these variables the Hamiltonian is equal

$$H = \int d^3x \mathcal{H} = \sum_p \frac{1}{2} \omega(p) [aa^+ + a^+a] \quad (1.16)$$

$$\omega(p) = \sqrt{\vec{p}^2 + m^2}$$

This is the Hamiltonian for the set of decoupled linear oscillators.

In Quantum Field Theory we have to quantize these oscillators. The variables $a(p)$ become operators that satisfy commutation relations

$$[a(\vec{p}), a^+(\vec{p}')] = \delta_{\vec{p}\vec{p}'} \quad (1.17)$$

$$[a(\vec{p}), a(\vec{p}')] = [a^+(\vec{p}), a^+(\vec{p}')] = 0$$

Operators $a(p)$ and $a^+(p)$ are familiar from QM. They are the annihilation and creation operator for oscillator with frequency $\omega(\vec{p})$.

The Fock space is the Hilbert space of the states with definite values of the operator of particle number $N(p) = a^+(p)a(p)$:

$$\begin{aligned} & \underline{\text{vacuum}} \quad |0\rangle \\ & \left\{ \begin{array}{l} |0\rangle \\ a(p)|0\rangle \equiv 0 \end{array} \right. , \\ & \underline{\text{one-particle states}} \\ & |p\rangle = a^+(p)|0\rangle \quad (1.18) \\ & \underline{\text{two-particle states}} \\ & |p_1, p_2\rangle = a^+(p_1)a^+(p_2)|0\rangle \\ & |p; p\rangle = \sqrt{2}a^+(p)a^+(p)|0\rangle \end{aligned}$$

etc.

We get that commutation relation eq. (1.17) corresponds to Bose-Einstein statistic for spin 0 particle

$$|p_1, p_2\rangle = +|p_2, p_1\rangle$$

and to positively defined operator of energy

$$H = \sum \omega(p) [N(p) + \frac{1}{2}] \quad (1.20)$$

The vacuum energy is equal

$$E_{vac} = \sum \frac{1}{2} \omega(p) \quad (1.21)$$

We have constructed the space of free particles with given momenta. Now we have to describe the propagation of free particles. The operator

$$\Phi^{(+)} = \sum_{\vec{p}} a(\vec{p}) e^{-ipx}$$

with positive frequency is a combination of terms that annihilate 1 particle at point x . Operator

$$\Phi(x) = \sum_{\vec{p}} a^+(\vec{p}) e^{ipx} \quad (1.22)$$

creates the particle at point x .

Consider the time ordering product

$$T\{\Phi(x), \Phi(0)\} = \Theta(x_0)\Phi(x)\Phi(0) + \Theta(-x_0)\Phi(0)\Phi(x) \quad (1.23)$$

where the step function is equal

$$\Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

So the Feynman propagator

$$D_F(x, 0) = \langle 0 | T\{\Phi(x)\Phi(0)\} | 0 \rangle \quad (1.24)$$

is the amplitude for a particle to propagate from point 0 to point x . Time ordering implies that creation always comes before annihilation.

The theory of the complex scalar fields $\Phi(x) = \frac{1}{\sqrt{2}}(\Phi_1 + i\Phi_2)$ with lagrangian density

$$\mathcal{L} = (\partial_\mu \Phi^+ \partial_\mu \Phi) - m^2 \Phi^+ \Phi \quad (1.25)$$

is equal to the theory of two different scalar particles with degenerate masses. The general solution of the field equations can be presented in the form

$$\Phi(x) = \sum_p (a(p)e^{-ipx} + b(p)^+ e^{ipx}) \quad (1.26)$$

where the operator (a, a^+) and (b, b^+) are creation and annihilation operators for the particle with the same masses but with the opposite electric charges (see the next lecture). We consider these two particles as a particle and antiparticle.

The Feynman propagator of scalar particle in the momentum representation is equal to

$$\begin{array}{c} p \\ \longrightarrow \blacktriangleright \end{array} \quad D(p) = \frac{i}{p^2 - m^2 + i\epsilon} \quad (1.27)$$

For Dirac spinor field $\Psi(x)$ the lagrangian density

$$\mathcal{L} = \bar{\Psi}[i\gamma_\mu \partial_\mu - m]\Psi \quad (1.28)$$

The plane wave solutions of the Dirac equation look like

$$u(p, \lambda) e^{ipx}, \quad (\lambda = \pm 1/2)$$

(1.29)

$$v(p, \lambda)e^{-ipx}, \quad (\lambda = \pm 1/2)$$

where $u(p, \lambda)$, $v(p, \lambda)$ satisfy equations

$$(\gamma_\mu p_\mu - m)u(p, \lambda) = 0 \tag{1.30}$$

$$(\gamma_\mu p_\mu + m)v(p, \lambda) = 0$$

and $\lambda = \pm 1/2$ label the independent solution with different value of the spin projection on momenta \vec{p} . The general solution of Dirac equation can be presented in the form

$$\Psi(x) = \sum_{\vec{p}, \lambda} \left\{ a(p, \lambda)u(p, \lambda)e^{-ipx} + b^+(p, \lambda)v(p, \lambda)e^{ipx} \right\} \tag{1.31}$$

where $a(p, \lambda)$ and $b(p, \lambda)$ are annihilation operators for particles and antiparticles respectively.

The next step is the quantization. We have to consider $a(p, \lambda)$ and $b(p, \lambda)$ as operators in the Fock space. The great surprise is that to have positively defined energy the operators $a(p, \lambda)$ and $b(p, \lambda)$ should satisfy anticommutation relations

$$\{a(p, \lambda), a^+(p', \lambda')\} = \{b(p, \lambda), b^+(p', \lambda')\} = \delta_{pp'}\delta_{\lambda\lambda'} \tag{1.32}$$

$$\{a, a\} = \{a^+, a^+\} = \{b, b\} = \{b^+, b^+\} = 0$$

with $\{A, B\} = AB + BA$.

These imply Fermi-Dirac statistic for spin 1/2 particle. These two examples demonstrate the famous spin-statistic theorem.

Feynman propagator $S_F(x)$

$$S_F = \langle 0|T\{\Psi(x)\bar{\Psi}(y)\}|0 \rangle \tag{1.33}$$

in momentum representation looks like

$$\xrightarrow{p} \quad S(p) = \frac{i}{\hat{p} - m} \tag{1.34}$$

where $\hat{p} = \gamma_\mu p_\mu$.

Exercise. Calculate the dimension of field Ψ : $[\Psi] = m^{3/2}$.

Electromagnetic Field $A_\mu(x)$:

Lagrangian density is

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F_{\mu\nu} + e j_\mu^{ext} A_\mu \tag{1.35}$$

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$$

Because of gauge invariance the quantization of electromagnetic field is rather subtle matter. In Feynman gauge the propagator for photon

$$\text{~~~~~} \xrightarrow{p} \quad D_F^{\mu\nu} = \frac{(-i)g_{\mu\nu}}{p^2 + i\varepsilon} \tag{1.36}$$

Exercise. Show that

$$[A_\mu] = m$$

$$[j_\mu] = m^3$$

1.3. Feynman Rules and Feynman Amplitudes.

Tree approximation.

What we do understand well is the QFT in the framework of perturbation theory. Free field theory provides the asymptotic $|in\rangle$ and $|out\rangle$ states for free particles and the amplitudes for propagation of free particles from one space-time point to another point. The nonlinear interaction term \mathcal{L}_{int} in perturbation theory provides the vertices. Combining vertices and propagators one can construct in perturbation theory the transition amplitude from one asymptotic state to another one. Let us remind the main steps.

Transitions are described by means of unitary S -matrix: $S^\dagger S = I$

$$\langle f|S|i\rangle = \langle f|i\rangle + (2\pi)^4 i \delta^{(4)}(\Sigma p_f - \Sigma p_i) \langle f|T|i\rangle \quad (1.37)$$

where i and f refer to initial and final state.

In perturbation theory

$$S = T \exp\left\{i \int d^4x \mathcal{L}_{int}\right\} = \quad (1.38)$$

$$= I + i \int d^4x \mathcal{L}_{int}(x) + \frac{i^2}{2} T \left\{ \int d^4x_1 \mathcal{L}(x_1); \int d^4x_2 \mathcal{L}(x_2) \right\} + \dots$$

where the operators of fields are in the interaction representation.

Consider for example the case of QED. The interaction looks like a product of electromagnetic current and 4-potential

$$\mathcal{L}_{int} = j_\mu^{em}(x) A_\mu(x) \quad (1.39)$$

$$j_\mu^{em}(x) = (-ie) \left\{ \bar{e}(x) \gamma_\mu e(x) - \frac{2}{3} \bar{u}(x) \gamma_\mu u(x) + \dots \right\}$$

Feynman rules for this QED lagrangian are summarized in Fig. 1.

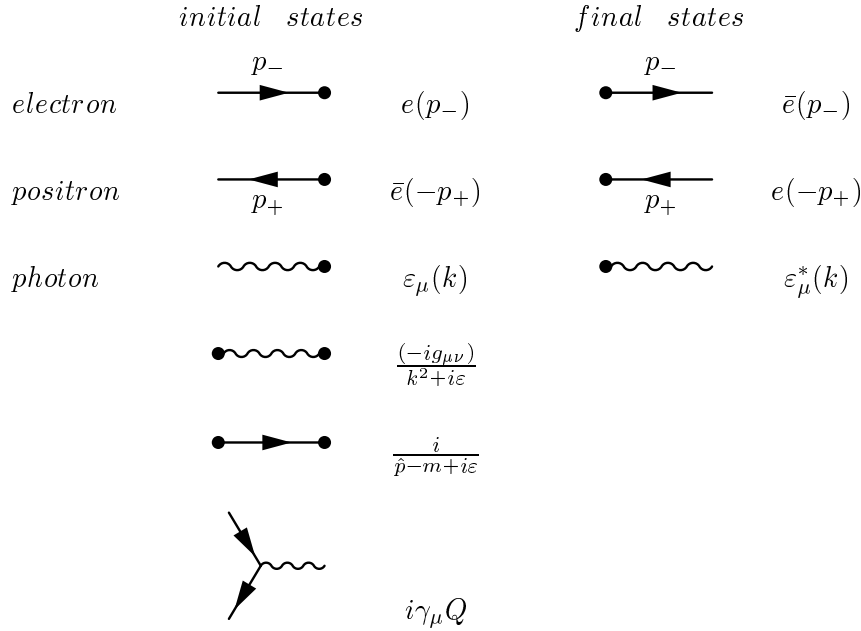
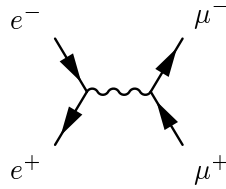


Figure 1: Feynman rules of QED.

Using these rules one can easily construct the transition amplitude. Consider as an example the process $e^+e^- \rightarrow \mu^+\mu^-$. There is one diagram for this process



The amplitude T is equal

$$iT(e^+e^- \rightarrow \mu^+\mu^-) = (-ie)^2 j_\alpha^{(e)} \frac{(-ig_{\alpha\beta})}{q^2} j_\beta^{(\mu)} \quad (1.40)$$

where

$$j_\alpha^{(e)} = \bar{e}(-p_+) \gamma_\alpha e(p_-)$$

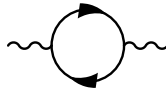
$$j_\beta^{(\mu)} = \bar{\mu}(k_-) \gamma_\beta \mu(-k_+)$$

This is the example of the amplitude in the lowest order in the coupling constant e . It contains no loops. There is special name for such diagrams – tree diagrams.

1.4. Loop corrections.

Propagator corrections in QED.

Consider one-loop correction to the photon propagator



$$\delta D_{\mu\nu} = \frac{(-ig_{\mu\alpha})}{q^2} (-i\Pi_{\alpha\beta}(q)) \frac{(-i)g_{\beta\nu}}{q^2} \quad (1.41)$$

where

$$(-i)\Pi_{\alpha\beta}(q) \equiv \alpha \begin{array}{c} \text{---} p \text{---} \\ \curvearrowright \\ \text{---} p - q \text{---} \\ \curvearrowleft \\ \beta \end{array} = \quad (1.42)$$

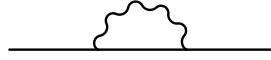
$$= e^2 \int \frac{d^4 p}{(2\pi)^4} (-1) S p \gamma_\alpha \frac{1}{\hat{p} - m + i\varepsilon} \gamma_\beta \frac{1}{\hat{p} - \hat{q} - m + i\varepsilon}$$

For large virtual momenta p the loop correction diverges quadratically

$$\Pi_{\alpha\beta} \simeq e^2 g_{\alpha\beta} \int_0^\Lambda \frac{d^4 p}{p^2} \simeq g_{\alpha\beta} e^2 \Lambda^2 \rightarrow \infty \quad (1.43)$$

for $\Lambda \rightarrow \infty$.

The one-loop correction to electron propagator

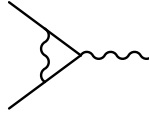


is proportional to

$$e^2 \frac{\hat{p}}{p^2 - m^2} \int \frac{d^4 q}{q^4} \sim e^2 \frac{\hat{p}}{p^2 - m^2} \ln \Lambda^2 \rightarrow \infty \quad (1.44)$$

for $\Lambda \rightarrow \infty$

The vertex



is proportional to

$$e^3 \gamma_\mu \ln \Lambda^2 \rightarrow \infty \quad (1.45)$$

for $\Lambda \rightarrow \infty$.

These corrections diverge logarithmically.

These are the simplest examples of the problem of divergences in QFT. It was a great success of theoretical physics when Dyson, Feynman, Schwinger and Tomonaga in the late 40th explained how to work with such theories.

1.5. Renormalizable Field Theories QED.

The general philosophy of renormalization can be formulated in the following way:

1) Suppose that we can split quantum fluctuations on the "fast" fluctuations (i.e. with virtual momenta $p > \Lambda$) and on the "slow" ones ($p < \Lambda$), where Λ is arbitrary large parameter.

2) Suppose that we can integrate over the "fast" fluctuations even though the physics at small distances ($p > \Lambda$) can be unknown.

3) For "slow" fluctuations we get "effective field theory" with $\mathcal{L}^{eff}(\Lambda)$ or $S^{eff}(\Lambda)$ and with parameters that depend on cut-off Λ .

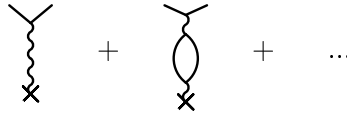
4) Physics of the low-energy processes does not depend on the value of the Λ .

5) For special class of renormalizable theories $S^{eff}(\Lambda)$ depends on finite number of parameters and interaction terms.

Consider how this program works in the case of QED. Suppose that effective lagrangian has the form

$$\mathcal{L}(\Lambda) = -\frac{1}{4}(F_{\mu\nu}^B)^2 + \bar{\Psi}_B(i\gamma_\mu\partial_\mu - m_B)\Psi_B - e_B\bar{\Psi}_B\gamma_\mu\Psi_B A_\mu^B \quad (1.46)$$

where all quantities with label B depend on Λ . Consider the scattering of heavy charged particle on the Coulomb center. In this case we have to sum up all corrections to the photon propagator.



As a result at low-energy the amplitude of Coulomb scattering is equal to

$$T = \frac{e_B^2(\Lambda)}{1 - \frac{e_B^2(\Lambda)}{12\pi^2} \ln \frac{\Lambda^2}{m_e^2}} \cdot \frac{1}{q^2} \quad (1.47)$$

The coefficient in front of $1/q^2$ is by definition the charge of particle ($1/q^2$ corresponds to $1/r$ dependence in the Coulomb law). So we claim that combination

$$e_{ph}^2 = \frac{e_B^2(\Lambda)}{1 - \frac{e_B^2(\Lambda)}{12\pi^2} \ln \frac{\Lambda^2}{m_e^2}} \quad (1.48)$$

is the physical charge. It does not depend on Λ and in this way we find $e_B^2(\Lambda)$ as a function of Λ .

In the similar way one can define the physical electron mass $m_{ph} = m_e$ as a pole in the exact propagator of the electron.

Now we are able to formulate the main theorem.

Theorem. If we rewrite the amplitudes of all QED processes that depend on e_B , m_B and Λ in terms of e_{ph} , m_{ph} the dependence on Λ in these amplitudes will disappear for large Λ !

1.6. Non-renormalizable Theories.

Fermi Theory (1934)

The first theory of weak interactions was formulated by Fermi. It was very similar to QED. The lagrangian of interaction was equal to a product of two vector currents. After the discovery of P and C parity violation this 4-fermion theory was modified so that

$$\mathcal{L}_W = \frac{G_F}{\sqrt{2}} j_\alpha j_\alpha \quad (1.49)$$

where $j_\alpha = \bar{\nu}_e \gamma_\alpha (1 + \gamma_5) e + \dots$

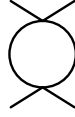
Remember that

$$[j] = m^3 ; [\mathcal{L}] = m^4$$

so the Fermi coupling constant has dimension -2 :

$$[G_F] = m^{-2}$$

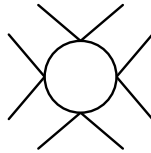
From the dimensional analysis it is clear that radiative corrections to 4-fermion interaction



should be of the order of

$$G_F (1 + \Sigma(G_F \Lambda^2)^n) (j)^2$$

where Λ is cut-off. It is also clear that 4-fermion interaction can generate multi-fermion interaction with diverge coupling constant, e.g. 8-fermion interaction



$$\Delta \mathcal{L}^{eff} = C G_F^4 [\ln \Lambda^2 + \Sigma(G_F \Lambda^2)^n] (j)^4$$

etc.

In this way we find that \mathcal{L}^{eff} should depend on infinite number of terms.

This is the example of non-renormalizable theory. In such theories we have to fix infinite number of terms in $\mathcal{L}^{eff}(\Lambda)$ at the scale Λ (i.e. at small distances $x \sim \Lambda^{-1}$) to reconstruct the amplitudes at low energy. Nobody knows how to work with nonrenormalizable theories.

1.7. From field theory to cross sections.

The steps from the formal QFT to the numerical predictions for cross sections and for the decay rates are very simple

$$\begin{aligned} & \text{Feynman diagrams} \\ & \quad \Downarrow \\ & \text{Amplitude } T \\ & \quad \Downarrow \\ & \text{Probability } \sim \Sigma |T|^2 \end{aligned}$$

For example the Cross Sections are calculated by the formula

$$d\sigma_{fi} = \frac{1}{2\sqrt{\lambda(s, m_1^2, m_2^2)}} |T_{fi}|^2 d\tau$$

where

$$d\tau = (2\pi)^4 \delta^4(p_1 + p_2 - \Sigma p_f) \prod_{j=1}^N \frac{d^3 p_j}{(2\pi)^3 2E_j}$$

is N-particle phase space and

$$\lambda(s, m_1^2, m_2^2) = 4[(p_1 p_2)^2 - m_1^2 m_2^2]$$

is relativistic flux.

The Decay Rate is given by formula

$$d\Gamma = \frac{1}{2E} |T|^2 d\tau$$

There exists a well developed routine technology of this kind of calculations. But sometime we can understand the basic properties of the answer without long calculations. Consider three examples of the order of magnitude estimates

1. Decay $\mu \rightarrow e \tilde{\nu}_e \nu_\mu$

In this case we know that

$$[\Gamma] = m \quad [G_F] = m^{-2};$$

and that in the limit $m_\mu \gg m_e$ nothing should depend on m_e . So on the ground of the dimensional analysis we conclude

$$\Rightarrow \Gamma(\mu \rightarrow e \tilde{\nu}_e \nu_\mu) = \frac{1}{192\pi^3} G_F^2 m_\mu^5$$

Certainly the numerical factor is beyond of these order of magnitude estimates but the dependence on mass is understood well.

2. Cross section $\nu e \rightarrow \nu e$

Cross sections have dimension

$$[\sigma] = m^{-2}$$

So for very large energy when we can forget about masses

$$\sigma(\nu_e e \rightarrow \nu_e e) = \frac{1}{3\pi} G_F^2 s$$

3. Cross section $e^+ e^- \rightarrow \mu^+ \mu^-$

Fine coupling constant is dimensionless. The cross section does not depend on masses at high energy. So

$$\sigma = \frac{4}{3} \frac{\pi \alpha^2}{s}, \quad s \gg m^2$$

It is interesting to note that after some training one can restore the powers of π that originate from the phase space. We have no time for such training.

Lecture II. Symmetries.

In the Standard Model the notion of global and local symmetries plays the very important role. In this lecture we shall study the different aspects of symmetries in QFT using a well known physical examples.

2.1. Global symmetries.

We start with the simplest $U(1)$ symmetry and consider as an example the theory of free electrons. Electrons are described by 4-component complex field ψ_i ($i = 1, 2, 3, 4$) named bispinor. The free Lagrangian density has the form

$$\mathcal{L} = \bar{\psi}(x)(i\gamma_\mu\partial_\mu - m)\psi(x) \quad (2.1)$$

where $\bar{\psi} = \psi^\dagger\gamma_0$ and γ_μ are 4×4 Dirac matrices, m is the electron mass.

It is quite evident that $U(1)$ global phase transformations

$$\begin{aligned} \psi(x) &\rightarrow \psi'(x) = e^{i\alpha}\psi(x) \\ \bar{\psi}(x) &\rightarrow \bar{\psi}'(x) = \bar{\psi}(x)e^{-i\alpha} \end{aligned} \quad (2.2)$$

leave Lagrangian (2.1) invariant. The global symmetry means that the phase of the transformation is the same for any of space-time points x .

A little bit less trivial example is the theory of two complex self-interacting scalar fields with the degenerate masses

$$\mathcal{L} = \partial_\mu\Phi^+\partial_\mu\Phi - m^2\Phi^+\Phi - \frac{\lambda}{4}(\Phi^+\Phi)^2, \quad (2.3)$$

where Φ is the two component column (doublet)

$$\Phi = \begin{pmatrix} \varphi^+(x) \\ \varphi^0(x) \end{pmatrix} \quad (2.4)$$

The Lagrangian (2.3) is invariant under global $SU(2)$ rotations of the complex doublet Φ

$$\begin{aligned} \Phi(x) &\rightarrow \Phi'(x) = S\Phi(x) \\ \Phi^+(x) &\rightarrow (\Phi'(x))^+ = \Phi^+(x)S^+ \end{aligned} \quad (2.5)$$

where S is unitary 2×2 matrix

$$S^+S = I$$

$$\text{with } \det S = 1 \quad (2.6)$$

This matrix S can be represented in the form

$$S = \exp\left[i\frac{\tau_a}{2}\alpha^a\right] \quad (2.7)$$

where $\tau_a = (\tau_1, \tau_2, \tau_3)$ are Pauli matrices

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (2.8)$$

that satisfy to $SU(2)$ Lie algebra commutation relations

$$[\tau_i, \tau_k] = 2ie_{ik\ell}\tau_\ell \quad (2.9)$$

where $e_{ik\ell}$ is totally antisymmetric tensor with $e_{123} = 1$. Three independent phases α^a ($a = 1, 2, 3$) do not depend on space-time. The Lagrangian density (2.3) describes Higgs scalars in the Standard Model.

Another useful example is the approximate $SU(2)_L \times SU(2)_R$ symmetry of strong interactions. This symmetry has been discovered long before the formulation of QCD in the framework of such general approach as the current algebra. It provided the first example of nonlinear realization of symmetry in QFT. In this lecture we shall formulate this symmetry using the Lagrangian of QCD.

The mass scale of QCD is defined by Λ_{QCD} :

$$\Lambda_{\text{QCD}} \simeq 0.5 \text{ GeV} \quad (2.10)$$

On the other hand according to particle data group booklet the masses of up and down quarks are of the order of a few MeV. So in a good approximation one can take

$$m_{u,d} = 0 \quad (2.11)$$

In this limit QCD Lagrangian for u and d quarks can be rewritten as

$$\mathcal{L} = \bar{\psi}_L(i\gamma_\mu \mathcal{D}_\mu)\psi_L + \bar{\psi}_R(i\gamma_\mu \mathcal{D}_\mu)\psi_R \quad (2.12)$$

where $\psi(x) = \begin{pmatrix} u(x) \\ d(x) \end{pmatrix}$ is $SU(2)$ doublet constructed from bispinor $u(x)$ and $d(x)$. The subscripts L and R means left and right field by definition

$$\begin{aligned} u_{L,R} &= \frac{1}{2}(1 \pm \gamma_5)u \\ d_{L,R} &= \frac{1}{2}(1 \pm \gamma_5)d \end{aligned} \quad (2.13)$$

where $\gamma_5 = -i\gamma_0\gamma_1\gamma_2\gamma_3$.

Any bispinor $\psi(x)$ can be represented as a sum of two spinors (left and right Weyl spinors)

$$\psi = \psi_L + \psi_R = \frac{1}{2}(1 + \gamma_5)\psi + \frac{1}{2}(1 - \gamma_5)\psi \quad (2.14)$$

For the case of massless particles the Lagrangian itself can be written as a sum of two independent Lagrangians : one for left spinors, another for right spinors. Each of them is $SU(2)$ invariant. So the total symmetry is $SU(2)_L \otimes SU(2)_R$.

2.2. Noether's Theorem. Conserved Currents.

So far we considered the examples of different Lagrangians that were invariant with respect to $U(1)$ and $SU(2)$ transformations. The transformations were global, i.e. they did not depend on space-time points x . As for invariance of the Lagrangian it looked like rather trivial property.

Noether's theorem states that for any continuous global symmetry of the Lagrangian one can construct the conserved vector currents. This is dynamical statement and it does not look like being trivial at all. I am going to prove this theorem in the classical field theory.

Let Lagrangian L depends on the set of fields φ^i and its first derivatives $\varphi^i_{,\mu} = \partial_\mu\varphi^i$. For infinitesimal global transformations the variations of fields are equal to

$$\begin{aligned}\delta\varphi^i &= i\epsilon^{(a)}T_{ij}^{(a)}\varphi^j \\ \delta\varphi_{,\mu}^i &= i\epsilon^{(a)}\partial_\mu(T_{ij}^a\varphi^j)\end{aligned}\tag{2.15}$$

The real infinitesimal parameters $\epsilon^{(a)}$ represent the independent symmetry transformations, matrices T_{ij}^a are the generators of the group of transformations in given representations. The invariance means that the actions S is not changed under transformation (2.15):

$$\delta S = \int d^4x \delta L \equiv 0\tag{2.16}$$

Let us calculate the variation of Lagrangian density directly

$$\begin{aligned}\delta L &= \frac{\partial L}{\partial\varphi^i}\delta\varphi^i + \frac{\partial L}{\partial\varphi_{,\mu}^i}\delta\varphi_{,\mu}^i = \\ &= +[\partial_\mu\frac{\partial L}{\partial\varphi_{,\mu}^i}]\delta\varphi^i + \frac{\partial L}{\partial\varphi_{,\mu}^i}\delta\varphi_{,\mu}^i\end{aligned}\tag{2.17}$$

where we have used the Lagrangian equation of motion for $\varphi^i(x)$

$$\frac{\partial L}{\partial\varphi^i} = \partial_\mu\frac{\partial L}{\partial\varphi_{,\mu}^i}\tag{2.18}$$

Substituting the variations for $\delta\varphi$ and $\delta\varphi_{,\mu}$ from eqs. (2.15) into (2.17) and (2.16) we get

$$\delta S = i\epsilon^{(a)}\int d^4x\partial_\mu j_\mu^a(x) = 0\ ,\tag{2.19}$$

where

$$j_\mu^{(a)} = \frac{\partial L}{\partial\varphi_{,\mu}^i}T_{ij}^{(a)}\varphi^j\tag{2.20}$$

Therefore we get the conservation of Noether currents

$$\partial_\mu j_\mu^{(a)}(x) = 0\tag{2.21}$$

and the conservation of the corresponding charges

$$\frac{d}{dt}Q^{(a)}(t) = 0\tag{2.22}$$

$$Q^{(a)}(t) = \int d^3x j_0^{(a)}(x)\tag{2.23}$$

(We assume that there are no fields at spatial infinity, i.e. $\int d^3x\partial_i j_i \equiv 0$)

The generalization of this proof to the Quantum Field Theory requires more advanced techniques such as operators algebra, commutators etc. But the final results, i.e. the expression for conserved Noethers currents, remain the same. So we are going to skip the proof of the theorem in QFT.

At the end of this subsection we present the conserved vector currents that correspond to the symmetries that we considered in the subsection 2.1 :

$$\begin{aligned}U(1) &: j_\mu = \bar{\psi}\gamma_\mu\psi\ , \\ SU(2) &: j_\mu^a = \Phi^+\tau^a\overleftrightarrow{\partial}_\mu\Phi\end{aligned}\tag{2.24}$$

$SU(2)_L \times SU(2)_R$:

$$(j_\mu^{L,R})^a = \bar{\psi}_{L,R}\gamma_\mu\tau^a\psi_{L,R}$$

where

$$\Phi^+ \overleftrightarrow{\partial}_\mu \Phi = \Phi^+ \partial_\mu \Phi - (\partial_\mu \Phi^+) \Phi$$

2.3. Spontaneous Violation of Global Symmetry. Goldstone Phenomenon.

The idea of spontaneous violation of symmetry was formulated first in the solid state physics. Consider, for example, the piece of some ferromagnetic material. The interaction of the elementary magnetic moments of electrons inside ferromagnetic is $O(3)$ invariant. On the other hand at low temperature $T < T_c$ the total magnetic moment \vec{M} of ferromagnetic piece is nonzero and has the definite direction, i.e. it violates $O(3)$ invariance of the system. Ground state is only $O(2)$ invariant for the rotations around \vec{M} and the “violated” symmetries are realized as a massless excitations.

In field theory analogous phenomenon is known as Nambu-Goldstone realization of symmetry. We are going to discuss this phenomenon using the example of $SU(2)_L \times SU(2)_R$ symmetry of strong interactions. Quark Lagrangian (2.12) is $SU(2)_L \times SU(2)_R$ invariant. There are 3 left ($V - A$) and 3 right ($V + A$) conserved currents of massless quarks. In other words there are 3 vector and 3 axial vector conserved currents. This is evident. On the other hand the quarks do not exist like a free particles. Instead we have a set of massive hadrons – baryons and mesons. The $SU(2)_V$ symmetry of hadrons was known for a long time. For example proton and neutron have practically degenerate mass and can be treated as an up and down members of $SU(2)_V$ doublet $N = \begin{pmatrix} p \\ n \end{pmatrix}$. (There are small corrections to the $SU(2)$ symmetric approximation of the order of $(\frac{m_{u,d}}{\Lambda})^2$ and of the order of electroweak coupling constant α .) The matrix elements of the conserved $SU(2)$ vector currents between nucleon states have the form

$$\langle N | \bar{\psi} \gamma_\mu \tau^a \psi | N \rangle \sim \bar{N} \gamma_\mu \tau^a N \quad (2.25)$$

In the limit of degenerate mass this m.e. is transversal

$$q_\mu \bar{N} \gamma_\mu \tau^a N = 0 \quad , \quad (2.26)$$

i.e. it corresponds to conserved currents.

For conservation it is crucial to have baryons with degenerate masses. The symmetry is realized in such a way that it transforms one-particle baryonic state into another one-particle state with the same mass.

If this realization of symmetry is unique we immediately get troubles with the $SU(2)_A$ symmetry. Indeed to construct the transversal matrix element of axial current we need degenerate baryons with opposite P -parity. The brief search for such baryons in the Table of Particle Properties shows that such baryons do not exist in Nature. So this way is prohibited for $SU(2)_A$.

One may try another possibility for matrix element of axial current between the same nucleon states

$$\langle N | \bar{\psi} \gamma_\mu \gamma_5 \tau^a \psi | N \rangle \sim (g_{\mu\nu} - \frac{q_\mu q_\nu}{q^2}) \bar{N} \gamma_\nu \gamma_5 \tau^a N \quad , \quad (2.27)$$

where q is momentum transfer from one nucleon to another. The transversality is now evident because

$$q_\mu (g_{\mu\nu} - \frac{q_\mu q_\nu}{q^2}) \equiv 0$$

So we have solved this problem easily. But now matrix element (2.27) is singular. It has a pole at $q^2 = 0$. The poles in the amplitudes correspond to one particle intermediate states and

pole at $q^2 = 0$ corresponds to massless particle. Fortunately there are π -mesons that contribute into m.e. (2.27) and that are almost massless. Now we are able to formulate the new way of realization of approximate $SU(2)_A$.

According to this new philosophy π -mesons should be massless in the limit $m_{u,d} = 0$ and there should be simple relation between the axial-vector part of matrix element

$$g_{\mu\nu} \bar{N} \gamma_\mu \gamma_5 N \quad (2.28)$$

and the pseudoscalar part

$$-\frac{q_\mu q_\nu}{q^2} \bar{N} \gamma_\mu \gamma_5 N = -\frac{q_\mu}{q^2} 2m_N (\bar{N} \gamma_5 N) \quad (2.29)$$

to have conserved current. This relation is known as Goldberger-Treiman relation. The matrix element of axial current between nucleon states can be measured in $n \rightarrow pe\tilde{\nu}$ decay. Experimental value of the ratio of axial coupling constant to the pseudoscalar coupling constant is very close to the theoretical prediction. So this realization of symmetry indeed works in the case of $SU(2)_A$.

Instead of one-particle degenerate states with opposite P -parity we have massless pseudoscalar Goldstone particles - pions. The symmetry transforms one-particle baryonic state into degenerate two-particle state (baryon plus massless pseudoscalar pion), then into 3-particle state with two pions etc. This is nonlinear realization of symmetry. One can proceed further and show that there are infinitely many states with the same lowest energy. The vacuum state is one of these states and it violates $SU(2)_A$ symmetry. This violation is due to nonzero vacuum expectation value of quark condensate

$$\begin{aligned} \langle 0 | \bar{u}u | 0 \rangle &= \langle 0 | \bar{d}d | 0 \rangle \neq 0 \\ \langle 0 | \bar{u} \gamma_5 u | 0 \rangle &= \langle 0 | \bar{d} \gamma_5 d | 0 \rangle = 0 \end{aligned} \quad (2.30)$$

It seems more instructive not to spend time proving eqs. (2.30) but to consider the same phenomenon using a very simple field model studied many years ago by Goldstone.

Let us start with the theory of complex scalar field $\varphi(x)$ with Lagrangian

$$\mathcal{L} = \partial_\mu \varphi^+ \partial_\mu \varphi - V(|\varphi|^2) \quad (2.31)$$

and with a special choice of potential (see fig.1)

$$V(|\varphi|^2) = \lambda \left(|\varphi|^2 - \frac{\eta^2}{2} \right)^2 \quad (2.32)$$

Lagrangian (2.31) is invariant under $U(1)$ transformations

$$\varphi(x) \rightarrow \varphi'(x) = e^{i\Lambda} \varphi(x) \quad (2.33)$$

and the Noether current is

$$j_\mu = i\varphi^+ \overleftrightarrow{\partial}_\mu \varphi \quad (2.34)$$

There are continuously many minima of the potential V (2.32)

$$\varphi = \frac{1}{\sqrt{2}} \eta e^{i\alpha} \quad (2.35)$$

The vacuum corresponds to one of these minima. This is spontaneous violation of symmetry : we have chosen one of state as a vacuum from the infinite set of minima. Let vacuum state corresponds to zero phase $\alpha = 0$:

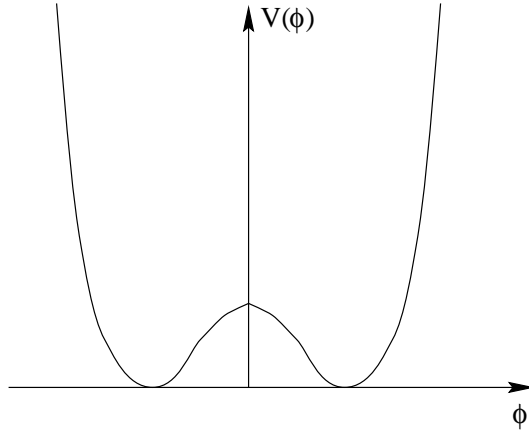


Fig. 1: Higgs Potential in Standard Model

$$\varphi_{cl} = \frac{1}{\sqrt{2}}\eta \quad (2.36)$$

Consider the small fluctuation of fields near vacuum configuration

$$\varphi = \frac{1}{\sqrt{2}}[\eta + \rho(x) + i\sigma(x)] \quad (2.37)$$

Potential can be rewritten as

$$V(\varphi) = V(\rho, \sigma) = \frac{\lambda}{2} \left\{ (\sigma^2 + \rho^2)^2 + 4\eta\rho(\rho^2 + \sigma^2) + 4\eta^2\rho^2 \right\} \quad (2.38)$$

The coefficients in front of bilinear terms determine the mass of the fields. So we get a theory of two particles with masses

$$\begin{aligned} M_\rho^2 &= 4\lambda\eta^2 \\ M_\sigma^2 &\equiv 0 \end{aligned} \quad (2.39)$$

Excitations that correspond to the motion along the valley of minima are massless! This is Goldstone phenomenon.

We can use more elegant and transparent representation for $\varphi(x)$ to demonstrate this phenomenon. Let us rewrite $\varphi(x)$ in terms of modulus and phase

$$\varphi(x) = \rho(x)e^{i\sigma(x)} \quad (2.40)$$

Then

$$\mathcal{L}(\rho, \sigma) = (\partial_\mu\rho)^2 - V(\rho^2) + \rho^2(\partial_\mu\sigma)^2 \quad (2.41)$$

There is no dependence on the field σ in the potential and therefore this field corresponds to massless particle.

In Quantum Field Theory we have two ways for realization of symmetry:

- 1) Vacuum state has the symmetry of the action S . Excitation states are degenerate.
- 2) Vacuum state has lower symmetry than action S . There are flat direction in configuration space of fields. The motions along these flat directions correspond to massless Goldstone particles.

Exercise

Consider the double well potential in Quantum Mechanics

$$V(x) = \lambda(x^2 - \eta^2)^2$$

and in the Quantum Field Theory for real scalar field $\varphi(x)$;

$$V(\varphi) = \lambda(\varphi^2 - \eta^2) .$$

Show that in QM there is only one lowest state and in QFT there are two degenerate and orthogonal lowest states.

2.4. Local U(1) gauge symmetry.

Now we are going to study the new type of symmetries : local gauge symmetries. Let us start with the theory of complex field $\varphi(x)$ described by the Lagrangian (eq.(2.31))

$$\mathcal{L} = \partial_\mu \phi^+ \partial_\mu \phi - V(\phi^+ \phi)$$

invariant under global U(1) transformation

$$\phi(x) \rightarrow \phi'(x) = e^{i\Lambda} \phi(x) .$$

Consider now the local U(1) transformation when we change the phase of the field independently for any point x

$$\phi(x) \rightarrow \phi'(x) = e^{i\Lambda(x)} \phi(x) \tag{2.42}$$

The potential $V(|\phi|^2) = V(|\phi'|^2)$ is invariant under this transformation but the kinetic term is not

$$\partial_\mu \phi^+ \partial_\mu \phi \rightarrow |(\partial_\mu + i(\partial_\mu \Lambda)\phi|^2 \tag{2.43}$$

To compensate this non-invariant change one can introduce new vector field $A_\mu(x)$ with the transformation law :

$$A_\mu(x) \rightarrow A'_\mu(x) = A_\mu(x) + \frac{1}{e} \partial_\mu \Lambda(x) \tag{2.44}$$

so that the new Lagrangian

$$\mathcal{L} = |(\partial_\mu - ieA_\mu)\phi|^2 - V(|\phi|^2) \tag{2.45}$$

is locally U(1) invariant or gauge invariant. The combination $\mathcal{D}_\mu = \partial_\mu - ieA_\mu$ has a name of covariant derivative (or long derivative). It has a simple transformations law

$$\begin{aligned} D_\mu &\rightarrow e^{i\Lambda} \mathcal{D}_\mu e^{-i\Lambda} \\ D_\mu \phi &\rightarrow e^{i\Lambda} (\mathcal{D}_\mu \phi) \end{aligned} \tag{2.46}$$

Up to now the fields $A_\mu(x)$ have no kinetic term in the Lagrangian and they are some kind of the auxiliary fields that do not propagate. To construct kinetic term we need gauge invariant combination of the derivatives of field A_μ . Notice that covariant derivatives and any

combinations of the covariant derivatives have a very simple transformation law eq. (2.46). Consider the commutator of two derivatives,

$$\begin{aligned} [\mathcal{D}_\mu \mathcal{D}_\nu] &\equiv -ieF_{\mu\nu} \\ F_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu \end{aligned} \quad (2.47)$$

We see that commutator is not the the differential operator but the function of x . According to (2.46) it is gauge invariant function. Now we are in position to write the totally gauge invariant Lagrangian

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F_{\mu\nu} + |\mathcal{D}_\mu\phi|^2 - V(|\phi|^2) \\ \phi(x) &\rightarrow \phi'(x) = e^{i\Lambda(x)}\phi(x) \\ A_\mu(x) &\rightarrow A'_\mu(x) = A_\mu(x) + \frac{1}{e}\partial_\mu\Lambda(x) \end{aligned} \quad (2.48)$$

The notion of gauge invariance was introduced by V. Fock in 1926 and by H. Weyl in 1929. (The very interesting history of this subject can be found in the lectures given by L.Okun at this school ten years ago).

2.5. Spontaneous Violation of local symmetry.

Higgs Phenomenon.

For the case when time derivative is zero $D_0\phi = 0$ and electric field is zero $F_{0i} = E_i = 0$ the Lagrangian (2.48) formally is equal to the free energy in the Ginzburg-Landau phenomenological theory of superconductivity, where $\varphi(x)$ plays a role of the order parameter. It is known that magnetic field does not penetrate into superconductor, it falls exponentially. Exponential fall in QFT corresponds to a massive particle. So one can expect that Lagrangian (2.48) at certain circumstances can describe the massive gauge field. This is the famous Higgs mechanism of spontaneous violation of local symmetry.

Consider Lagrangian (2.48) with the special choice of potential energy (2.32)

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^2 + |\mathcal{D}_\mu\phi|^2 - \lambda(|\phi|^2 - \frac{\eta^2}{2})^2 \quad (2.49)$$

Potential $V(\phi)$ has continuous valley of minima. Let us quantize the fields near the vacuum state (2.32)

$$\langle \varphi \rangle = \varphi_{cl} = \frac{1}{\sqrt{2}}\eta \quad (2.50)$$

As in the case of global symmetry it is convenient to use representation of $\phi(x)$ in term of modulus and phase

$$\phi(x) = \frac{1}{\sqrt{2}}(\eta + \rho(x))e^{i\sigma(x)} \quad (2.51)$$

The Lagrangian (2.48)-(2.49) is gauge invariant. So let us make gauge transformation with $\Lambda(x) \equiv -\sigma(x)$

$$\begin{aligned} \phi(x) &\rightarrow \phi' = e^{i\sigma}\phi \\ A_\mu(x) &\rightarrow A'_\mu = A_\mu - \frac{1}{e}\partial_\mu\sigma \end{aligned} \quad (2.52)$$

In this gauge (unitary gauge)

$$D_\mu\phi \rightarrow (\partial_\mu - ieA_\mu)\frac{1}{\sqrt{2}}(\eta + \rho(x)) \quad (2.53)$$

and the Lagrangian can be rewritten in the form

$$\mathcal{L} = \left[-\frac{1}{4}F_{\mu\nu}^2 + \frac{1}{2}e^2\eta^2 A_\mu^2 \right] + \frac{1}{2}(\partial_\mu\rho)^2 + (e^2\eta)\rho(x)A_\mu^2(x) + \frac{e^2}{2}A_\mu^2(x)\rho^2(x) \quad (2.54)$$

The term in bracket represents the free massive vector particle with mass

$$m_V = e\eta \quad (2.55)$$

Massless Goldstone mode $\sigma(x)$ has been eaten by massless vector field $A_\mu(x)$ (that had two polarization) and as a result we get massive vector field with three polarization. This is Higgs Phenomenon.

2.6. Local $SU(2)$. Yang-Mills theory of vector fields.

We have to make another nontrivial step to be ready for the construction of the Standard Model. We have to consider the general case of the local gauge groups.

Let us start with $SU(2)$ theory of massless fermion $\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}$

$$\mathcal{L} = \bar{\psi}[i\gamma_\mu\partial_\mu\psi] \quad (2.56)$$

and consider local $SU(2)$ transformations

$$\psi(x) \rightarrow \psi'(x) = S(x)\psi(x) \quad (2.57)$$

where

$$\begin{aligned} S(x) &= \exp i(T_j\Lambda_j(x)) ; \\ T_i &= \frac{1}{2}\tau_i \quad , \quad i = 1, 2, 3 ; \\ [T_i, T_j] &= ie_{ijk}T_k \end{aligned} \quad (2.58)$$

The Lagrangian (2.56) is not invariant under this transformation. To compensate the non-invariant piece in the Lagrangian we introduce the triplet of vector fields $A_\mu^i(x)$ so that:

$$\begin{aligned} \mathcal{L} &= \bar{\psi}i\gamma_\mu(\partial_\mu - igA_\mu(x))\psi \\ A_\mu(x) &= T^i A^i(x) \end{aligned} \quad (2.59)$$

with the transformation law

$$A_\mu(x) \rightarrow A'_\mu(x) = SA_\mu(x)S^+ - \frac{i}{g}(\partial_\mu S)S^+ \quad (2.60)$$

Again it is convenient to introduce covariant derivative

$$\mathcal{D}_\mu = \partial_\mu - igA_\mu \quad (2.61)$$

that transforms as a triplet under $SU(2)$ transformations:

$$\begin{aligned} D_\mu &\rightarrow SD_\mu S^+ \\ D_\mu\psi &\rightarrow S(D_\mu\psi) \end{aligned} \quad (2.62)$$

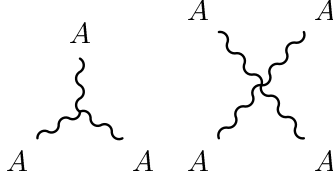
We can define the triplet of field-strength tensor $G_{\mu\nu}^i$:

$$\begin{aligned} G_{\mu\nu} &\equiv G_{\mu\nu}^i T^i = \frac{i}{g} [\mathcal{D}_\mu, \mathcal{D}_\nu] \\ &= \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu] \\ G_{\mu\nu} &\rightarrow G'_{\mu\nu} = SG_{\mu\nu}S^\dagger \end{aligned} \quad (2.63)$$

and construct the $SU(2)$ gauge invariant Lagrangian

$$\mathcal{L} = -\frac{1}{4} \text{Tr}[G_{\mu\nu}G_{\mu\nu}] + \bar{\psi}i\gamma_\mu \mathcal{D}_\mu \psi \quad (2.64)$$

This Lagrangian was invented by Yang and Mills in 1954. The very nontrivial part in this construction is that kinetic energy $\sim G_{\mu\nu}^2$ contains bilinear $\sim A^2$, trilinear $\sim A^3$ and quadrilinear $\sim A^4$ terms:



So we have a gauge theory of self-interacting vector fields.

2.7. Spontaneous Violation of Local $SU(2)$ Symmetry. Renormalizable theory of massive vector fields.

Consider the $SU(2)$ gauge theory of the couple of scalar fields $\phi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}$:

$$\mathcal{L} = -\frac{1}{4} \text{Tr}G_{\mu\nu}G_{\mu\nu} + |\mathcal{D}_\mu \phi|^2 - \lambda(|\varphi|^2 - \frac{\eta^2}{2})^2 \quad (2.65)$$

We expect that after spontaneous violation of $SU(2)$ symmetry three Goldstone bosons will be mixed with three massless vector fields and produce three massive vector field. To show this we repeat the steps that we had done in the case of local $U(1)$ symmetry.

Let us introduce a special representation for the doublet ϕ

$$\phi(x) = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} = e^{i\sigma^i(x)T^i} \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(\eta + \rho(x)) \end{pmatrix} \quad (2.66)$$

and consider gauge transformation with the parameter

$$\Lambda^i(x) = -\sigma^i(x) \quad (2.67)$$

In this gauge the fields $\sigma^i(x)$ disappear from the Lagrangian and vector part of \mathcal{L} gets the form

$$\begin{cases} \mathcal{L}_{\text{vect}} = -\frac{1}{4} \text{Tr}G_{\mu\nu}^2 - \frac{1}{2}m_V^2 A_\mu^2 \\ m_v = \frac{1}{2}g\eta \end{cases} \quad (2.68)$$

This is the theory of massive vector fields with the special choice of self-interactions.

The theory of massless Yang-Mills field was renormalizable theory. It seems that the property of the vacuum should not change the behavior of the amplitudes at high energy. So one can believe that Yang-Mills theory with spontaneous violation of gauge symmetry remains

renormalizable. The theory of massive vector fields with arbitrary interactions is nonrenormalizable in general. But if one takes the special case of interaction with quarks, with scalars and self-interaction that corresponds to the gauge-invariant Lagrangian (2.65) the nonrenormalizable divergences should disappear. Technically the rigorous proof of this statement is quite nontrivial business even now. This problem had been solved by t'Hooft and Veltman in 1971.

Lecture III. $SU(2) \times U(1)$ Theory of Electroweak Interactions.

In this lecture I am going to describe the fundamental Lagrangian of the Standard Model. It is important to understand that **a priori** there is no unique way to construct the model of electroweak interactions. There are plenty of them. In the review paper by B.Bjorken and Llewellyn-Smith in 1973 they discussed several dozens of models. We do not understand yet why the gauge group is $SU(2) \times U(1)$, why there are three generations of quarks and leptons etc. We have to deduce our theory from the experiment.

3.1. Minimal group.

It was well established in old four-fermionic theory of weak interactions that charged currents (responsible for β -decay of nucleons and other hadrons) have $V - A$ structure, i.e. they are constructed from the left-handed fermions.

The minimal group of gauge symmetry which includes charged vector currents is $SU(2)$ group. So any theory of weak interactions have to include $SU(2)_L$ symmetry as a subgroup. Photon interacts both with left- and right-handed fermions. So if we are going to unify weak and electromagnetic interactions the group of gauge symmetry should include $U(1)$ as well. The simplest choice of such group of symmetry is

$$G = SU(2)_L \times U(1)$$

3.2. Left and Right Fermions.

$SU(2)_L$ symmetry. Weak Mixing Angles.

Any Dirac 4-spinor Ψ can be presented as a sum of two Weyl spinors Ψ_L and Ψ_R :

$$\Psi = \Psi_L \oplus \Psi_R = \frac{1}{2}(1 + \gamma_5)\Psi + \frac{1}{2}(1 - \gamma_5)\Psi \quad (3.1)$$

These Weyl spinors are irreducible representation of Lorentz group, they depend on two complex parameters. For massless fermions

$$\Psi_L = \begin{pmatrix} (1 - \vec{\sigma}\vec{n})\varphi \\ -(1 + \vec{\sigma}\vec{n})\varphi \end{pmatrix}, \quad (3.2)$$

where φ is 2-spinor, $\vec{\sigma}$ are Pauli matrices, and $\vec{n} = \vec{p}/|p|$ is the direction of the motion of particle. So for left particle Ψ_L

$$\vec{\sigma}\vec{n} = -1 \quad (3.3)$$

and for right particles Ψ_R

$$\vec{\sigma}\vec{n} = +1 \quad (3.4)$$

Left leptons and quarks group into $SU(2)_L$ doublets. For the first generations they are

$$L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \quad \text{and} \quad Q = \begin{pmatrix} u \\ d \end{pmatrix}_L \quad (3.5)$$

To avoid $V + A$ charged current we have to put right fermions into singlet representation. So e_R , u_R and d_R are singlets. As for right-handed neutrino ν_R nobody has observed it so far. It is unknown whether such field exists. Just now we prefer not to introduce ν_R into the theory.

To include the electromagnetic interactions we have to define charge. For left-handed fermions the charge is different for up and down component so that

$$Q_L = T_3 + Y_L \quad (3.6)$$

where T_3 is the third component of $SU(2)_L$ and Y_L is left hypercharge.

From eq.(3.6) it follows that for leptonic doublet $Y_L = -1/2$ and for quark doublet $Y_Q = 1/6$.

For right fermions we identify Q and Y_R :

$$Q_R = Y_R \quad (3.7)$$

so that $Y_{u_R} = \frac{2}{3}$; $Y_{d_R} = -\frac{1}{3}$, $Y_{e_R} = -1$.

The minimal way to introduce $U(1)$ interactions is to consider gauge boson that interacts with

$$Y = Y_L + Y_R \quad (3.8)$$

This is the gauge group of Minimal Standard Model

$$SU(2)_L \times U(1)_Y \quad (3.9)$$

Let $A_\mu^i(x)$, $i = 1, 2, 3$ be gauge bosons of $SU(2)_L$ and $B_\mu(x)$ – the gauge boson of $U(1)$ group. The charged fields

$$A_\mu^\pm = \frac{1}{\sqrt{2}}(A_\mu^1 \pm iA_\mu^2) \quad (3.10)$$

can be identify with W_μ^\pm bosons.

Photon $A_\mu(x)$ is in general some combination of A_μ^3 and B_μ . Orthogonal combination represents another physical particle that we identify with Z boson. So

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos \theta_W & -\sin \theta_W \\ \sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} A_\mu^3 \\ B_\mu \end{pmatrix} \quad (3.11)$$

where θ_W is a weak mixing angle.

To violate spontaneously $SU(2)_L \times U(1)_Y$ group and to make masses to W^\pm and Z bosons we need three Goldstone fields. The $SU(2)$ doublet of Higgs particles

$$H = \begin{pmatrix} H^+ \\ H^0 \end{pmatrix}; \quad Y_H = \frac{1}{2} \quad (3.12)$$

can provide this number of Goldstone bosons after spontaneous violation. In the MSM we use only **one** Higgs doublet.

We have completed the construction of the MSM. Now we are ready to calculate the masses of vector bosons m_W , m_Z and phenomenological mixing angle θ_W in terms of coupling constants g_2 for $SU(2)$, g_1 for $U(1)$ and in terms of v.e.v. of Higgs field η .

In the unitary gauge Higgs doublet has the form

$$H(x) = e^{i\vec{T}\vec{\alpha}(x)} \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(\eta + \rho(x)) \end{pmatrix} \quad (3.13)$$

Covariant derivative

$$D_\mu \equiv \partial_\mu - ig_1 Y B_\mu(x) - ig_2 T^a A_\mu^a(x) \quad (3.14)$$

for the vacuum field H_{vac}

$$\begin{aligned}
D_\mu H_{vac} &= \left(-ig_1 \frac{1}{2} B_\mu - ig_2 \frac{1}{2} \tau^a A_\mu^a\right) \begin{pmatrix} 0 \\ \frac{\eta}{\sqrt{2}} \end{pmatrix} = \\
&= \frac{(-i)}{2\sqrt{2}} \eta \begin{pmatrix} \sqrt{2} g_2 W_\mu^- \\ -g_2 A_\mu^3 + g_1 B_\mu \end{pmatrix}
\end{aligned} \tag{3.15}$$

The mass term for vector fields originates from $(D_\mu H)^+ D_\mu H$ term in the Lagrangian. It looks like

$$\mathcal{L}_{mass} = \frac{1}{4} (g_2 \eta)^2 W_\mu^+ W_\mu^- + \frac{1}{8} \eta^2 (g_2 A_\mu^3 - g_1 B_\mu)^2 \tag{3.16}$$

From this expression we conclude that the massive combination of A_μ^3 and B_μ (i.e. Z -boson) is

$$Z_\mu = \frac{1}{\sqrt{g_1^2 + g_2^2}} (g_2 A_\mu^3 - g_1 B_\mu) \tag{3.17}$$

or that

$$tg\theta_W = g_1/g_2 \tag{3.18}$$

From eq. (3.16) it follows that

$$m_W = \frac{1}{2} g_2 \eta \tag{3.19}$$

and

$$m_W = m_Z \cos \theta_W \tag{3.20}$$

It is very interesting that Z boson should be heavier than W boson! After spontaneous violation there still remains unbroken $U(1)$ symmetry that corresponds to massless photon.

If we introduce electric charge e as a coupling constant of the photon we can relate $g_{1,2}$ with e and $\cos \theta_W$. Let us rewrite interaction of A_μ^3 and B_μ as an interaction of A_μ and Z_μ fields:

$$(-ig_2 T_3) A_\mu^3 - ig_1 Y B_\mu \equiv (-i) \frac{g_2}{\cos \theta_W} [T_3 - \sin^2 \theta_W Q] Z_\mu + (-i) (g_1 \cos \theta_W) Q A_\mu \tag{3.21}$$

This is identically rewritten universal expression for covariant derivative. So eq. (3.21) is applicable to the left and right fermions and to the Higgs doublet.

From eq. (3.21) it follows immediately that

$$e = g_1 \cos \theta_W = g_2 \sin \theta_W \tag{3.22}$$

We complete the description of bosonic sector of the SM.

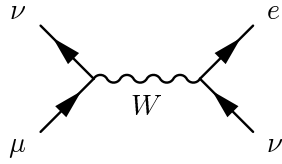
3.3. Weak interactions of leptons and quarks.

Neutral Current. Request for new particles.

Now we are ready to calculate the amplitude for the first physical process, for the decay of $\mu \rightarrow e \nu \tilde{\nu}$. Charged currents Lagrangian for leptons looks like

$$\Delta \mathcal{L}_{Charged} = \frac{g_2}{2\sqrt{2}} W_\mu^+ [\bar{\nu} \gamma_\mu (1 + \gamma_5) e + \dots] \tag{3.23}$$

where the dots are for the similar terms with μ and τ leptons. Feynman diagram for the μ -decay is presented in Fig.1



The amplitude for the decay can be read from this diagram and it is equal to

$$T(\mu \rightarrow e\nu\bar{\nu}) = \left[\frac{g_2}{2\sqrt{2}}\right]^2 \frac{1}{m_W^2 - q^2} (\bar{\nu}\gamma_\alpha(1 + \gamma_5)\mu)(\bar{e}\gamma_\alpha(1 + \gamma_5)\nu) \quad (3.24)$$

The momentum transfer q from muonic current to electronic current is of the order of muonic mass m_μ . So if $m_W \gg m_\mu$ the amplitude looks like a point-like interaction in Fermi theory.

$$T_{Fermi} = \frac{G_F}{\sqrt{2}} j_\alpha^e (j_\alpha^\mu)^+ \quad (3.25)$$

Comparing these two presentations for the same amplitude we conclude that

$$\frac{G_F}{\sqrt{2}} = \frac{g_2^2}{8m_W^2} \quad (3.26)$$

Taking into account eq. (3.19) for m_W we also get that v.e.v. η is directly connected with G_F :

$$\eta = [\sqrt{2}G_F]^{-1/2} = 246 \text{ GeV} \quad (3.27)$$

$$G_F \equiv G_\mu = 1.16639(2) \cdot 10^{-5} \text{ GeV}^{-2}$$

To fix remaining two fundamental parameters g_1 and g_2 we have to choose two other physical observables measured with the best accuracy. The choice is evident. They are the fine coupling constant α

$$\alpha^{-1} = \frac{4\pi}{e^2} = 137.035985(61) \quad (3.28)$$

and the mass of Z -boson

$$m_Z = 91.187(2) \text{ GeV} \quad (3.29)$$

To calculate g_1 and g_2 we first have to calculate the mixing angle θ_W in terms of G_F , α and m_Z . It is not difficult exercise to show that

$$\sin^2 \theta_W \cos^2 \theta_W = \frac{\pi\alpha}{\sqrt{2}(G_F m_Z^2)} \quad (3.30)$$

Exercise 1: Derive eq. (3.30).

Substituting the values of the parameters from eqs. (3.27), (3.28) and (3.30) we get

$$\begin{aligned} \sin^2 \theta_W &= 0.2120 \\ g_1 &= \frac{\sqrt{4\pi\alpha}}{\cos \theta_W} = 0.34 \\ g_2 &= \frac{\sqrt{4\pi\alpha}}{\sin \theta} = 0.66 \end{aligned} \quad (3.31)$$

So we are ready for the first prediction in SM: we can calculate m_W

$$(m_W)^{theor} = m_Z \cos \theta_W = 80.94 \text{ GeV} \quad (3.32)$$

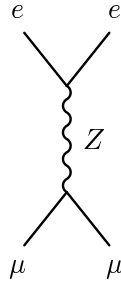
that has to be compared with the current experimental value

$$(m_W)^{exp} = 80.37(8) \text{ GeV} \quad (3.33)$$

The deviation from theoretical number is only 0.6%, but this tiny number is equal to 8σ deviation. To explain the huge discrepancy we have to take into account radiative correction that have the scale of the few per mill.

The old 4-fermionic point-like theory is the effective theory for momentum transfer much smaller than m_W . In this sense the SM is generalization of the old theory. But SM also predicts the new phenomena that were unknown in V-A theory. This is the neutral currents.

The effective 4-fermionic coupling of neutral currents is generated by Z boson exchange.



At small momentum transfer it is local interaction with the coupling constant equal to $G_F \cos^2 \theta_W$.

Exercise 2. Calculate the coupling constant for neutral currents.

Though this coupling is of the same order as G_F by some reasons the experimental search for neutral currents gave negative results for a long time and only in 1973 experimental groups at CERN observed neutral currents and provided the first experimental measurements of $\cos \theta_W$. This measurement gave the possibility to calculate m_W and m_Z theoretically (eqs. (3.19), (3.20)) with rather good accuracy. This estimate had been extremely helpful for the experimental discovery of W and Z bosons.

Another great achievement of the SM was the request for new particles needed for self-consistency of the theory. In 1970 the set of the known particles included 2 generations of leptons

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L; e_R, \mu_R \quad (3.34)$$

and three quarks u , d and s that belong to the following $SU(2)_L \times U(1)_Y$ representation

$$\begin{pmatrix} u \\ d' = d \cos \theta_c + s \sin \theta_c \end{pmatrix}_L, u_R, d_R, s_R \quad (3.35)$$

where θ_c is the Cabibbo angle. First of all there was no symmetry between quarks and leptons. No less important was the fact that for this set of quarks Z boson exchange produces flavour-changing $s \leftrightarrow d$ neutral currents

$$\begin{aligned} Z_\mu \bar{d}'_L \gamma_\mu d'_L &\sim Z_\mu \left[(\bar{d}d) \cos^2 \theta_c + (\bar{s}s) \sin^2 \theta_c + \right. \\ &\quad \left. + \sin \theta_c \cos \theta_c [\bar{d}s + \bar{s}d] \right] \end{aligned} \quad (3.36)$$

This was absolutely forbidden by experimental data. To save the SM Glashow, Illiopoulos and Maiani in 1970 introduced fourth c quark and the new $SU(2)_L$ doublet

$$\begin{pmatrix} c \\ -d \sin \theta_c + s \cos \theta_c \end{pmatrix} \quad (3.37)$$

As a result flavour-changing neutral currents disappear and all neutral currents become diagonal. This theoretical request for new particle was satisfied by experimental discovery of c -quark in 1974.

3.4. Quark masses. CKM matrix and CP-violation.

In the Standard Model the standard mass term for the electron violates $SU(2)_L$. Indeed this term

$$m_e \bar{e}e = m_e [\bar{e}_R e_L + \bar{e}_L e_R] \quad (3.38)$$

transforms like doublet instead of being invariant.

To preserve $SU(2)_L \times U(1)_Y$ symmetry we have to use Higgs mechanism to generate the masses for fermions. For example Yukawa coupling of L , e_R and H is $SU(2)_L \times U(1)_Y$ invariant

$$\begin{aligned} \Delta \mathcal{L} &= f_e (\bar{L} e_R) H + h.c. = \\ &= \frac{f_e}{\sqrt{2}} (\eta + \rho(x)) \bar{e}e = \\ &= m_e \bar{e}e + \frac{f_e}{\sqrt{2}} \rho(x) \bar{e}e \end{aligned} \quad (3.39)$$

where $\rho(x)$ is the field for physical Higgs in SM. From eq. (3.39) it follows that Yukawa coupling is proportional to m_e

$$f_e = \frac{\sqrt{2}}{\eta} m_e \simeq 3 \cdot 10^{-6} \quad (3.40)$$

Notice that before this step the fields $e_L(x)$ and $e_R(x)$ were absolutely different, i.e. they had different interaction with W and Z . Yukawa interaction unified this two Weyl spinors into one massive particle – electron. To give the mass to down quarks we can use the same type of Yukawa interaction

$$\Delta \mathcal{L}_{m_d} = f_d (\bar{Q}_L d_R) H \quad (3.41)$$

As for the mass of up quarks we need Higgs doublet with nonzero v.e.v. for up component of doublet. At that moment we can introduce new Higgs doublet. But in the case of $SU(2)$ group complex conjugated fields

$$\tilde{H} = (-i\sigma_2) H^* \quad (3.42)$$

also behave like a member of $SU(2)$ doublet. So we can use \tilde{H} to give mass to upper quark

$$\Delta \mathcal{L}_m = f_d (\bar{Q}_L d_R) H + f_u (\bar{Q}_L u_R) \tilde{H} \quad (3.43)$$

This is the solution of problem of fermion mass in the case of one generation. For more than one generation we have to take into account quark mixing.

Cabibbo-Kobayashi-Maskawa (CKM) matrix.

For three generations of quarks the general Yukawa couplings produce general non-diagonal mass 3×3 matrix

$$\Delta \mathcal{L} = (\bar{u}'_L)_i M_{ik}^{(u)} (u'_R)_k + (\bar{d}'_L)_i M_{ik}^{(d)} (d'_R)_k + h.c. \quad (3.44)$$

where u'_i, d'_i are quarks that belong to $SU(2)$ doublet of i -th generation $i = 1, 2, 3$.

The matrices $M^{(u)}$ and $M^{(d)}$ can be diagonalized by use of left and right unitary rotation, i.e.

$$M^{(u)} = U_L^+ M_{diag}^{(u)} U_R$$

(3.45)

$$M^{(d)} = D_L^+ M_{diag}^{(d)} D_R$$

where $M_{diag}^{(u,d)}$ are diagonal matrices. Substituting this expression into eq. (3.44) we get that diagonal massive quark fields are

$$u_R = U_R u'_R, \quad u_L = U_L u'_L; \tag{3.46}$$

$$d_R = D_R d'_R, \quad d_L = D_L d'_L;$$

It means that the members of the $SU(2)$ doublet are the mixture of different massive fields. We already met such doublet in the case of 2 generations (see eqs. (3.35) and (3.37)).

In terms of massive fields the charged currents look like

$$j_\mu^+ = \bar{u}_L \gamma_\mu d'_R = \bar{u}_L (U_L D_L^+) \gamma_\mu d_L \tag{3.47}$$

The unitary 3×3 matrix

$$V = U_L D_L^+ \tag{3.48}$$

is the famous CKM matrix.

Due to unitarity of U and D matrices the neutral currents remain diagonal

$$\bar{u}'_R u'_R = \bar{u}_R u_R, \quad \bar{u}'_L u'_L = \bar{u}_L u_L, \text{ etc.} \tag{3.49}$$

You will have a course of lectures on CKM matrix by Y.Nir. I do not want to interfere with his lectures but I'd like to mention one fundamental property of CKM. In general case $n \times n$ unitary matrix can be represented as the product of different orthogonal rotations dependent on n_a angles

$$n_a = \frac{1}{2} n(n-1) \tag{3.50}$$

and of the $U(1)$ factors with n_{ph} observable phases

$$n_{ph} = \frac{1}{2} (n-1)(n-2) \tag{3.51}$$

and of a number of nonobservable $U(1)$ factors that can be absorbed into redefinition of quark's fields. So for three generation there is one observable phase, i.e. charge currents contain complex couplings. This immediately gives violation of CP-invariance. Till now it is unknown whether this is the only source of CP-violation in the Nature. But in any case CKM mechanism of CP-violation does exist.

3.5. Subtle point. Triangle Anomaly.

To have renormalizable theory of electroweak interactions it was absolutely crucial to start from the gauge invariant theory where gauge bosons interact with conserved Noether currents. Spontaneous violation of symmetry does not spoil any symmetric relations between operators. They are exactly the same as in non-violent theory. The confusing notion of spontaneous violation means only nonlinear realization of the symmetry in the space of physical states.

In the SM we operate both with vector and axial currents. For any axial currents

$$j_\mu^A = \bar{\Psi} \gamma_\mu \gamma_5 \Psi$$

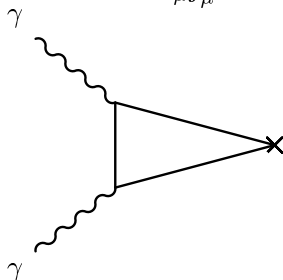
(3.52)

$$\partial_\mu j_\mu^A = 2im\bar{\Psi}\gamma_5\Psi$$

So naively for massless fermion axial current is conserved. But what is true in Classical Field Theory can be not true in Quantum Field Theory. Indeed one-loop calculation of the divergence of axial current for electrons gives instead of eq. (3.52)

$$\begin{cases} \partial_\mu j_\mu^5 = 2im\bar{\Psi}_e\gamma_5\Psi_e + \frac{\alpha}{2\pi}F_{\mu\nu}\tilde{F}_{\mu\nu} \\ \tilde{F}_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\alpha\beta}F_{\alpha\beta} \end{cases} \quad (3.53)$$

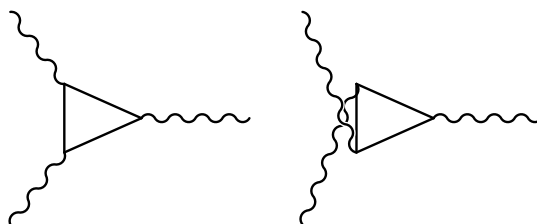
The term $F\tilde{F}$ originates from matrix element of $\partial_\mu j_\mu^5$ between vacuum and two-photon states.



So the axial current is not conserved even for $m \equiv 0$. Not any classical symmetry can survive in Quantum Mechanics. This very interesting theoretical phenomenon has special name – triangle anomaly.

In the SM there are plenty of axial currents that interact with gauge fields. Though fermions are massless (no mass terms in the Lagrangian) the anomaly can destroy the conservation of Noether currents and this will kill renormalizability. There is one possibility to save it. We see from eq. (3.53) that anomaly depends only on the "charge" of particle that is running inside loop. So if the total gauge current has different pieces it can happen that nonzero individual anomalies cancel each other for total current.

This cancellation imposes some restrictions on the charges of quarks and leptons. Let us check this possibility. We will calculate the triangle matrix elements between fields A_μ^i and B_μ . There are two crossing diagrams that contribute to anomalous interaction between 3 gauge fields.



Consider first the anomalous contribution of one generation of matter. It is easy to see that:

1) (A, A, A) and (A, B, B) anomalies are automatically disappeared for lepton doublet and for quark doublet separately.

2) (B, A, A) anomaly is disappeared if

$$Q_e + 2Q_u + Q_d \equiv 0 \quad , \quad (3.54)$$

i.e. quark contribution cancels lepton contribution only for this special relation between charges. This relation means that hydrogen atom has to be neutral!

It is very interesting that renormalizability of the SM takes place only if the charge of proton is opposite to the charge of electron.

We can proceed further and consider other anomalies. At that moment we have to make some statement about ν_R . Suppose first that it does not exist at all. In this case:

3) Cancellation of (B, B, B) anomaly takes place only if

$$Q_e = -1, \quad Q_\nu = 0; \quad Q_u = \frac{2}{3}, \quad Q_d = -\frac{1}{3} \quad (3.55)$$

(We suppose that QCD has $SU(3)_c$ symmetry.)

4) Cancellation of $(B \rightarrow \text{gluon} + \text{gluon})$ anomaly is automatic.

5) Cancellation of $(B \rightarrow \text{graviton} + \text{graviton})$ anomaly takes place only for the charge sample eq. (3.55). So we are able to fix the relative charges of leptons and quarks in this case.

If ν_R does exist anomalies 3) - 5) are disappeared automatically for any charge of neutrino.

Exercise IV. Prove 1) - 5).

If we suppose that the new generations are the exact replica of the old one (only masses are different, but the charges are the same) then we come to the same conclusion for each generation. If we allow to change the charges from generation to generation the restrictions on the choice of charges becomes weaker. But in any case it is very interesting that renormalizability impose restrictions on the property of matter fields.

Lecture IV. Higgs, W and Z .

Higgs boson H is the only missing particle in the Standard Model. The interaction of H with gauge bosons Z , W and photon γ is fixed by gauge invariance of the SM. There is no freedom in this interaction.

As for the Yukawa coupling constants of H with quarks and leptons they are free parameters of the theory. Yukawa coupling constants determine the masses of quarks and leptons, the mixing CKM angles and CP-odd phase. We do not understand yet why these parameters of SM have given value and not another one. So we have to take it from the experimental data.

A little can be said about Higgs particle. In this lecture we are going to discuss this isoteric subject.

4.1. Higgs Sector. Custodial symmetry of Higgs Potential.

The Higgs potential in the SM

$$V(H) = \frac{\lambda}{4} (H^+ H - \frac{\eta^2}{2})^2 \quad (4.1)$$

has been constructed to be $SU(2) \times U(1)$ invariant. Here $H = \begin{pmatrix} H^+ \\ H^0 \end{pmatrix}$ is $SU(2)$ doublet with $Y_L = 1/2$, $H^+ H$ is singlet. So $SU(2) \times U(1)$ symmetry is absolutely evident.

In the unitary representation

$$H(x) = \exp[i\alpha_a(x) \frac{\tau_a}{2}] \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(\eta + \rho(x)) \end{pmatrix} \quad (4.2)$$

$SU(2) \times U(1)$ transformations change three Goldstone fields $\alpha_a(x)$ and do not change the modulus $\rho(x)$. Potential $V(H)$ in this representation depends only on the modulus $\rho(x)$:

$$V(H) = V(\rho) = \frac{1}{2} (\frac{1}{2} \lambda \eta^2) \rho^2(x) + \frac{1}{4} \lambda \eta \rho^3 + \frac{1}{16} \lambda \rho^4 \quad (4.3)$$

Potential (4.1) has minima at nonzero value of H . Quantization near one of such minima

$$\langle H \rangle_{vac} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \eta \end{pmatrix} \quad (4.4)$$

spontaneously violates $SU(2) \times U(1)$ symmetry up to the $U(1)$ symmetry that remains unbroken.

In the SM the v.e.v. η is connected with Fermi coupling constant G_F (see previous lecture) and is equal numerically to

$$\eta = 246 \text{ GeV} .$$

In the unitary representation potential $V(\rho)$ eq. (4.3) has 3 terms: quadratic, cubic and quartic ones. The quadratic term determines the mass of Higgs boson

$$m_H^2 = \frac{1}{2} \lambda \eta^2 \quad (4.5)$$

and cubic and quartic terms determine self-interaction of Higgs bosons.

In general theory the mass and the self-interaction are independent parameters but not in the SM. Potential (4.1) depends only on two parameters η and λ . So the Higgs boson mass

completely fix the strength of self-interaction. For example the lower bound for m_H from LEP II experiment

$$m_H \gtrsim 90 \text{ GeV} \quad (4.6)$$

can be immediately transformed into lower bound for self-interaction coupling

$$\lambda = \frac{2m_H^2}{\eta^2} \gtrsim 0.27 \quad (4.7)$$

For large m_H the interaction of Higgs bosons becomes getting strong $\lambda \gtrsim 1$. In this case one should worry about large corrections to the low-energy observables, say to the ratio m_W/m_Z .

Naively one can expect that for $\lambda \sim 1$ the corrections are of the order of unity and that in this case the SM loses its predictive power. Fortunately this naive expectation is not correct and corrections to electroweak observables at low-energy (i.e. to LEP observables) are small even for heavy Higgs bosons. The reason is very subtle and beautiful. There is hidden symmetry of Higgs that preserves low-energy observables from large corrections. This is custodial $SU(2) \times SU(2)$ symmetry. Potential (4.1) that looks like $SU(2) \times U(1)$ invariant possesses larger $SU(2) \times SU(2)$ symmetry.

To find this symmetry it is convenient to rewrite (4.1) using new representation for doublet H :

$$H = \begin{pmatrix} H^+ \\ H^0 \end{pmatrix} = (\pi_0 + i\pi_a \tau_a) \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \pi_2 + i\pi_1 \\ \pi_0 - i\pi_3 \end{pmatrix} \quad (4.8)$$

where π_0 and π_a are real scalar fields. At first sight we have done nothing but changed a little bit the notation for H^+ and H^0 . It is true, but in this new notation $H^+ H$ looks like a scalar product of four-vector $\pi_\mu = (\pi_0, \pi_a)$. Indeed

$$H^+ H = (0, 1)[\pi_0 - i\vec{\pi}\vec{\tau}][\pi_0 + i\vec{\pi}\vec{\tau}] \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \pi_0^2 + \vec{\pi}^2 = \pi_\mu \pi_\mu \quad (4.9)$$

The quadratic form $\pi_\mu \pi_\mu = \pi_0^2 + \vec{\pi}^2$ is invariant with respect to $O(4)$ orthogonal transformation and $O(4)$ group can be presented as

$$O(4) = SU(2) \times SU(2) \quad (4.10)$$

So for one Higgs doublet H any $SU(2) \times U(1)$ invariant potential $V(H^+ H)$ is $SU(2) \times SU(2)$ invariant as well.

After spontaneous violation of symmetry the field π_0 gets nonzero v.e.v. and surviving symmetry is $O(3)$ rotational symmetry between different components of π_i or as a rotational symmetry between Goldstone fields α_i :

$$\alpha_i(x) \rightarrow \alpha'_i(x) = O_{ij} \alpha_j(x) \quad (4.11)$$

This $O(3)$ symmetry is custodial symmetry of Higgs potential.

The immediate consequence of the existence of the custodial symmetry is that the relation between m_W and m_Z

$$m_W/m_Z = \cos \theta \quad (4.12)$$

is correct to any order of λ . On other words there are no corrections of the order of λ^2, λ^3 etc. that could change the prediction (4.12) by order of magnitude for $\lambda \gtrsim 1$. Why?

If we start the calculation of the corrections of the order of λ^2, λ^3 etc. to the effective action we immediately find such large corrections to the v.e.v. η and to the wave function

renormalization Z of Goldstone fields α_i . (Wave function renormalization of gauge fields and gauge coupling constants are proportional to the square of gauge coupling constant and is considered as the small corrections.) Due to $O(3)$ custodial symmetry (4.11) of the selfinteraction the renormalization of the Goldstone fields should be the same for each of the components $\alpha_i(x)$, so that renormalized fields $\alpha^R(x)$ are transformed as a vector under $O(3)$ custodial symmetry

$$\vec{\alpha}^R(x) = Z^{1/2} \vec{\alpha}^B(x) \quad (4.13)$$

where $\vec{\alpha}^B$ are bare Goldstone fields. This is the central point. Indeed since $\vec{\alpha}^B(x)$ is a $O(3)$ vector it can be eaten by gauge transformation. In this gauge (unitary gauge) the renormalized effective action looks like the bare action that we considered in section 3.1. The only difference is that the Higgs vertices, the Higgs field $\rho(x)$ and v.e.v. η are renormalized by self-interaction in the order of $\lambda^1, \lambda^2 \dots$. So in this approximation we get that

$$\begin{aligned} m_W^R &= (\pi \alpha_W) \eta^R \\ m_Z^R &= (\pi \alpha_Z) \eta^R \end{aligned} \quad (4.14)$$

(see eqs. (3.19) and (3.20)) and that

$$m_W^R / m_Z^R = \alpha_W / \alpha_Z = \cos \theta_W$$

This relation is valid in any order of λ !

If we switch on the coupling constants g_2 and g_1 and consider the corrections of the order of $\alpha, \alpha\lambda, \alpha\lambda^2, \dots$ the custodial $SU(2) \times SU(2)$ symmetry will be broken and we can expect the correction of the same order $\alpha_1, \alpha\lambda, \alpha\lambda^2, \dots$ to the ratio m_W/m_Z and to other observables. These corrections do exist. They are of the order of two-loop electroweak corrections if the Higgs boson mass is not much larger than $m_{W,Z}$

$$\alpha_W \cdot \lambda \sim \alpha_W^2 \left(\frac{m_H}{m_Z} \right)^2$$

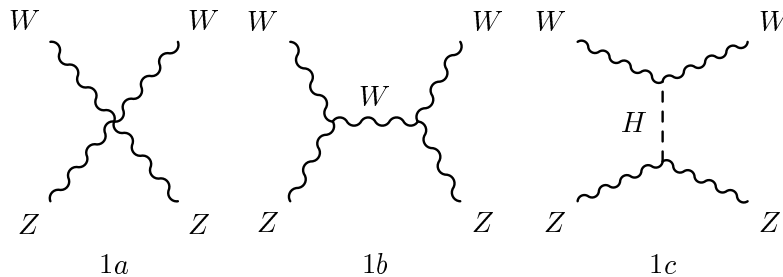
These small corrections will be studied in the next lecture.

4.2. Unitary Bound on Higgs mass.

Coupling constant λ of selfinteraction and the mass m_H^2 are independent parameters in the general theory but in the SM they are proportional to each other

$$m_H = \sqrt{\frac{\lambda}{2}} \cdot \eta \simeq \sqrt{\frac{\lambda}{2}} (246 \text{ GeV}) \quad (4.15)$$

If $m_H \gg m_{W,Z}$ the $SU(2) \times U(1)$ gauge theory looks like the old-fashioned theory of massive vector particles that is nonrenormalizable one. In nonrenormalizable theories scattering amplitudes grow with energy and violate unitary bounds at some value of energy. Consider for example the scattering of W boson on Z boson. There are three types of diagrams that contribute into this process.



It is clear that for $m_H \rightarrow \infty$ the diagrams (1c) goes to zero as $1/m_H^2$ and the remaining two diagrams (1a,b) are the same as in the theory of massive W and Z without Higgs mechanism.

We are going to demonstrate that these two amplitudes for longitudinally polarized W_L and Z_L linearly grow with s , where s is square of the energy in the c.m.f. Indeed any polarization vector of massive vector particle satisfies the condition

$$\varepsilon_\mu^V(p)p_\mu = 0 \quad (4.16)$$

For momenta $p_\mu = (\varepsilon, 0, 0, p)$ the solution of eq. (4.16) with longitudinal components is

$$\begin{aligned} \varepsilon_\mu^V(p) &= \frac{1}{m_V}(p, 0, 0, \varepsilon) \simeq \\ &\simeq \frac{P_\mu}{m_V} + 0\left(\frac{m_W}{\varepsilon}\right) , \end{aligned} \quad (4.17)$$

i.e. longitudinal component grows with energy. So any of the diagrams 1a,b is of the order of

$$Amp \sim \left(\frac{p_1}{m_W}\right)\left(\frac{p_2}{m_Z}\right)\left(\frac{p_3}{m_W}\right)\left(\frac{p_4}{m_Z}\right) \sim \frac{s^2}{m_W^2 m_Z^2} \quad (4.18)$$

So in general theory of massive vector particle the scattering amplitude $W_L Z_L \rightarrow W_L Z_L$ grows quadratically with s . In the loops this dependence transforms into divergence of the loop integrals and this is why the theory of massive particle in general is nonrenormalizable.

But SM is a very special theory because it is gauge invariant. The sum of the diagram 1a and 1b (and the diagram 1c itself) is gauge invariant. It means that gauge boson interacts with conserved Noether currents. Consider for example the interaction of incoming W :

$$T_{1a} + T_{1b} = g_W \varepsilon_\mu^W(p_1) \langle Z, W | J_\mu^W | Z \rangle \quad (4.19)$$

where $\langle Z, W |$, $\langle Z |$ represent the physical states of remaining particles. Since the J_μ^W is conserved current any matrix element of J_μ^W should be transversal

$$(p_1)_\mu \langle J_\mu^W(p_1) | \rangle \equiv 0 \quad (4.20)$$

It means that in the case of the theory of massive particles with gauge-invariant interaction the sum of two diagrams 1a and 1b eq. (4.19) should be transversal with respect to W -boson momenta $(p_1)_\mu$. If so the most dangerous term $\sim p_\mu^1/m_W$ in polarization vector ε_μ^W gives zero contribution and the remaining terms $\sim m_W/\varepsilon$ gives

$$(T_{1a} + T_{1b}) \sim \left(\frac{m_W}{E}\right) \langle Z | J_\mu | ZW \rangle \sim \left(\frac{m}{E}\right) \left(\frac{E}{m}\right)^3 \sim \frac{s}{m^2} \quad (4.21)$$

The natural idea now is to kill in the same way the dangerous terms for polarization vector of two W -bosons and to prove that amplitude does not grow with energy. Unfortunately our trick works only once. The interaction of the W -bosons with corresponding conserved currents look like

$$\begin{aligned} T_{1a} + T_{1b} &= g_W^2 \varepsilon_\mu^W(p_1) \varepsilon_\nu^W(p_3) \int d^4x_1 d^4x_2 e^{-ip_1x_1} e^{ip_3x_3} \times \\ &\times \langle Z | T \{ J_\mu^{W^+}(x_1) J_\nu^{W^-}(x_3) \} | Z \rangle \end{aligned} \quad (4.22)$$

where $T\{\}$ is the T -product of two currents. Just because currents in non-abelian theory do not commute the T -product does not allow to cancel $p_{1\mu}$ and $p_{3\nu}$ – dangerous terms simultaneously.

Exercise 1. Prove this technical statement.

So we conclude that amplitudes without Higgs exchange grows linearly with s . Unitarity of the scattering amplitudes says that the any partial amplitude is less than 1. So sooner or later this amplitude has to violate unitarity bound. The value of energy $\sqrt{s_0}$ at which one of the partial waves intersects unitary bounds varies both with angular momenta l and from one process to another one. We are not going to repeat the detailed calculations made in literature. The result of these calculations gives

$$\sqrt{s_0} \sim 700 \text{ GeV} \quad (4.23)$$

To stop the growth of the amplitude we have to switch on the diagrams with Higgs exchange. If we do it too late unitarity will be violated already. So to prevent the violation of unitarity bound

$$m_H \lesssim 700 \text{ GeV} . \quad (4.24)$$

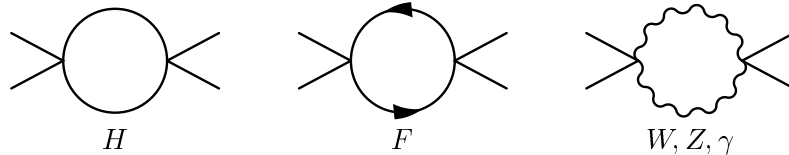
What is wrong with heavy Higgs? Actually the violation of unitarity that we found for tree diagrams means that the coupling constant becomes strong and that we can't restrict ourselves by calculation of tree diagrams, we have to include all loops and to calculate the whole infinite series in λ . We do not know how to deal with strong interaction. So the bounds (4.22), (4.23) and (4.24) say that for $m_H < 700 \text{ GeV}$ the theory can be treated perturbatively and for heavier Higgs we have a theory of strong interaction and we can't calculate $WZ \rightarrow WZ$ scattering theoretically. It is interesting to note that the attempts to work in strong coupling regime (e.g. using computer simulation for the Higgs theory on the lattice) demonstrate that the bound (4.23) and (4.24) is very stable, i.e. it takes place even for strong interacting Higgs.

4.3. Effective potential. Stability of the Universe and Bounds on m_H .

The Higgs potential in the SM

$$V_{\text{cl}}(H) = \frac{\lambda}{4} H^4 - \frac{\mu^2}{2} H^2 \quad (4.25)$$

has minima that corresponds to nonzero v.e.v. of field H : $\langle H \rangle_{\text{vac}} = \eta$. Loop corrections change self-interactions of Higgs particles.



The effective potential that takes into account loops corrections was calculated by Coleman and Weinberg in 1973. In one-loop approximation it looks like

$$V_{\text{eff}}(H) - V_{\text{cl}}(H) = \frac{1}{64\pi^2} \left\{ \frac{m_H^4 + 6m_W^4 + 3m_Z^4}{\eta^4} - \frac{12m_t^4}{\eta^4} \right\} H^4 \ln \frac{H^2}{M^2} \quad (4.26)$$

where we have neglected by small contributions from fermions other than t -quark. Note that due to Fermi-Dirac statistic the contribution of fermion loops has opposite sign in compared to the bosonic loops.

It is clear that corrections (4.26) become more important than the main classical potential (4.25) for very large field H when $\left(\frac{m}{\eta}\right)^4 \ln H \gtrsim \left(\frac{m_H}{\eta}\right)^2$.

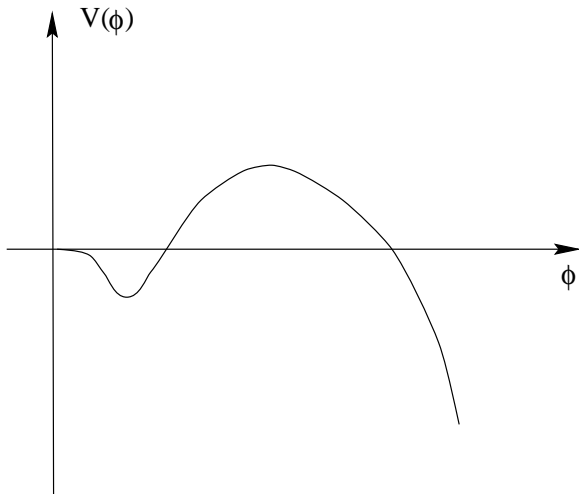


Fig. 2:

In one-loop approximation we get that the correction (4.26) has negative sign if $m_H < \sqrt[4]{12}m_t$. For this case the effective potential has no ground state(see fig. 2).

So even if our system was located first in the local minima at $\langle H \rangle = \eta$ it will decay at $t \rightarrow \infty$ and the average value of field H will run to infinity. We know that nothing like that has happened with our Universe that is near 10^{10} years old. So the stability of the Universe imposes strong constraints on the masses of top and Higgs particles.

We should improve a little our one-loop formula (eq. (4.26)). For large H one-loop logarithmic corrections $\lambda \ln H$ and $\alpha_W \ln H$ are of the same order as tree terms, two-loop double-logarithmic terms are of the order of one-loop terms etc. So all these logarithmic terms should be taken into account. Fortunately this technical problem is not very difficult- renormalization group techniques help to sum up such corrections. The result is

$$V^{\text{eff}}(H) = -\frac{1}{2}\mu^2(t)H^2(t) + \frac{1}{4}\lambda(t)H^4(t) \quad (4.27)$$

where $\lambda(t)$ and $\mu(t)$ are running parameters and $t = \ln H/\eta$. For small value of t (i.e. for small value of field H) the running parameters do not run far away from their classical values and the effective potential is equal to the classical one with the accuracy of small radiative corrections. For large H we can forget about $\mu^2 H^2$ and the whole dynamics at large H is governed by running coupling constant $\lambda(t)$. There are different contributions into differential equations for running $\lambda(t)$, coming from the loops with top quarks, vector bosons and Higgs boson itself. If the top quark contribution dominates, i.e. the higgs coupling to top (i.e. the top mass) is large, $\lambda(t)$ changes sign (see fig. 3) and the vacuum becomes unstable. This is reformulation of the phenomena that we had at one loop level.

If the Higgs selfinteraction dominates, i.e. the Higgs mass is large, then the evolution of $\lambda(t)$ is similar to the evolution of coupling in the H^4 theory without other fields. It is known that in this case the behavior of $\lambda(t)$ is

$$\lambda(t) = \frac{\lambda(t_0)}{1 - b\lambda(t_0) \ln \frac{t}{t_0}} \quad (4.28)$$

and running coupling goes to infinity at some finite value of H (see fig. 3).

$$H = \Lambda \ .$$

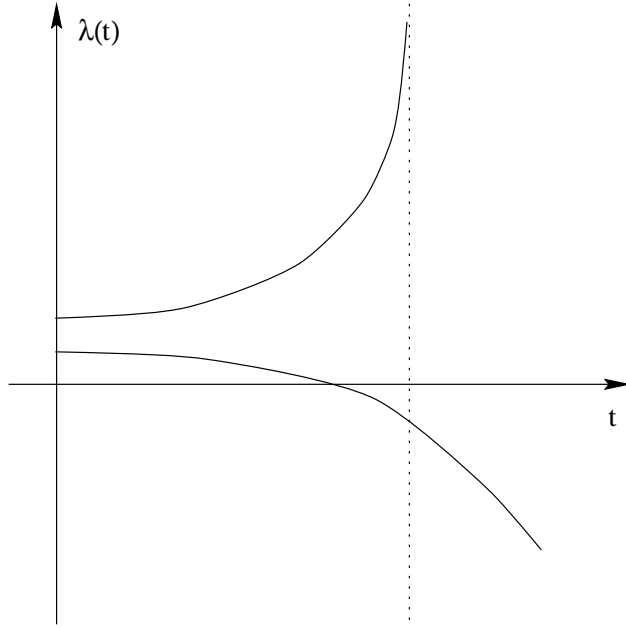


Fig. 3:

This is the Landau pole in the running coupling constant. When initial condition $\lambda(t_0)$ (i.e. the value of Higgs mass) increases the value of Landau pole goes done. If we substitute this running coupling constant into eq.(4.27) we get that the effective potential runs to infinity at this value of H as well. This is some new phenomena.

Such singular behavior of the coupling constant is unacceptable from the physical point of view. Indeed for any finite value of the bare coupling constant λ^B (λ^B is equal to the running coupling $\lambda(t)$ at the cut-off Λ) we get that renormalized coupling constant (i.e. $\lambda(t)$ at low value of t) is equal to zero. It means that at low energy we get trivial free theory. This pathological theory seems to be unphysical.

There are two possibilities to improve bad behavior of $\lambda(t)$. The first one is so to say dynamical. For $\lambda \sim 1$ the multiloop corrections and nonperturbative corrections change $\lambda(t)$ at large t so that Landau pole disappears. This is the possible solution of the problem in the strong coupling regime.

Another solution can be reached in perturbation theory if some new physics (i.e. new interactions and new particles) contribute into $\lambda(t)$ at scale near Λ so that pole disappears.

If we believe that there are no new physics up to some scale (or that the theory can be treated perturbatively up to this scale) we have to push the position of Landau pole Λ (calculated in one-loop approximation eq. (4.28)) to higher scale. This impose upper bound on the value of Higgs mass. So we have bounded m_H both from above and from below. This remarkable line of reasoning was invented by Cabibbo et al. in 1979.

There are different choices for the parameter Λ . For example Λ can be of the order of Planck scale

$$\Lambda \sim \Lambda_{Pl} = 10^{19} \text{ GeV} ,$$

or of the order of Grand Unification Scale

$$\Lambda \sim \Lambda_{GUT} = 10^{15} \text{ GeV} ,$$

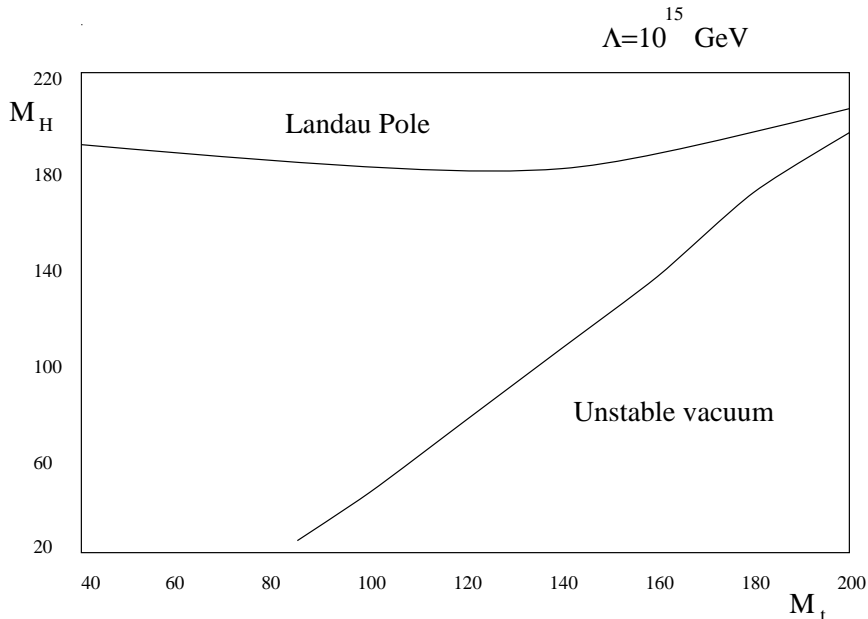


Fig. 4:

or of the scale of the energy of the accelerator of the next generation

$$\Lambda \sim 10^3 - 10^5 \text{ GeV} .$$

We have to keep in mind all these possibilities. It is evident that for the strongest assumption that new physics does not appear up to the Planck scale we should get the strongest upper bound for m_H .

To derive more quantitative results we have to solve differential equation for the running coupling constant $\lambda(t)$. The renormalization of $\lambda(t)$ depends on self-interaction coupling, on gauge coupling and on Yukawa coupling constants. So we have to solve the whole system of coupled differential equations. This can be done numerically with the help of computer. The result of calculation for $\Lambda = 10^{15}$ GeV is presented in the fig. .

This is so to say the phase diagram in the plane m_t and m_H . Allowed region is located between two curves, the lower region corresponds to unstable vacuum and for the parameters in the upper region Landau pole appears at the scale lower than $\Lambda = 10^{15}$ GeV. For experimental value of $m_t \simeq 175$ GeV the allowed region for m_H is very strong

$$170 \text{ GeV} < m_H < 190 \text{ GeV} \quad (\Lambda = 10^{15} \text{ GeV}).$$

For $\Lambda \simeq 10^5$ GeV the upper bound is much weaker.

4.4. Basic properties of W and Z bosons.

The interactions of gauge bosons with quarks, leptons, Higgs bosons and with gauge bosons itself are determined by the principle of gauge invariance. In this section we are going to calculate the partial widths of W and Z boson.

According to eq. (3.21) we write the amplitude of the Z -boson decay into fermion-antifermion pair $\bar{f}f$ in the form

$$T(Z \rightarrow \bar{f}f) = \frac{e}{2s_c} Z_\alpha \bar{\Psi}_f (g_V \gamma_\alpha + g_A \gamma_\alpha \gamma_5) \Psi_f , \quad (4.29)$$

where

$$g_V = T_3^f - 2s^2 Q^f$$

$$g_A = T_3^f$$

and

$$s = \sin \theta, \quad c = \cos \theta.$$

With good accuracy we can neglect the masses of fermions and derive the universal formula for any partial width of Z boson

$$\Gamma_{Z \rightarrow \bar{f}f} = \Gamma_0 [g_V^2 + g_A^2], \quad (4.30)$$

where Γ_0 is the standard width:

$$\Gamma_0 = \frac{G_F m_Z^3}{6\sqrt{2}\pi} = 332 \text{ MeV} \quad (4.31)$$

For neutrino we have

$$g_V = g_A = \frac{1}{2},$$

so that

$$\Gamma_{inv} = \sum_i \Gamma(Z \rightarrow \nu_i \bar{\nu}_i) = 3 \cdot \Gamma_0 \left(\frac{1}{4} + \frac{1}{4} \right) = 498 \text{ MeV}. \quad (4.32)$$

This is theoretical prediction for the decay of Z boson into invisible modes.

For the decays to any of the pairs of charged leptons we have

$$g_A = -\frac{1}{2}$$

$$g_V = \left(-\frac{1}{2}\right)(1 - 4s^2)$$

and

$$\Gamma_l = \Gamma(Z \rightarrow l\bar{l}) = \frac{1}{4} [1 + (1 - 4s^2)^2] \Gamma_0 \simeq 83.47 \text{ MeV} \quad (4.33)$$

Note that for $s^2 \simeq 0.2320$ the vector coupling g_V for charged lepton is a very small number!

For the decays into any of five pairs of quarks we have

$$\Gamma_q = \Gamma(Z \rightarrow \bar{q}q) = 3\Gamma_0 [g_A^2 R_A + g_V^2 R_V] \quad (4.34)$$

where the factor 3 is due to three color of each quarks and the factors $R_{A,V}$ are due to exchange of gluons between quarks in the final state. In the first order of strong coupling $\hat{\alpha}_s = \alpha(q^2 = M_Z^2)$ they are equal

$$R_A = R_V = 1 + \frac{\hat{\alpha}_s}{\pi} \quad (4.35)$$

As a result

$$\begin{aligned} \Gamma_h = \Gamma(Z \rightarrow \text{hadrons}) &= 3\Gamma_0 \left\{ 2 \left[\frac{1}{4} + \left(\frac{1}{2} - \frac{4}{3}s^2 \right)^2 \right] + \right. \\ &\left. + 3 \left[\frac{1}{4} + \left(\frac{1}{2} - \frac{2}{3}s^2 \right)^2 \right] \right\} \left(1 + \frac{\hat{\alpha}_s}{\pi} \right) = 1676 \left(1 + \frac{\hat{\alpha}_s}{\pi} \right) \text{ MeV} \end{aligned} \quad (4.36)$$

The numerical value of $\hat{\alpha}_s$ found from deep-inelastic processes is of the order of $\hat{\alpha}_s \simeq 0.119$. So the corrections due to strong interaction in the final state are of the order of 4%.

For the total width we have

$$\Gamma_Z^{th} = \Gamma_{inv} + 3\Gamma_l + \Gamma_h \simeq 2488.5 \text{ MeV} \quad (4.37)$$

that should be compared with the experimental LEP and SLC value

$$\Gamma_Z^{ex} = 2493.9 \pm 2.4 \text{ MeV} . \quad (4.38)$$

Though the agreement between Γ_Z^{th} , calculated in the tree approximation, and Γ_Z^{ex} is very good the experimental accuracy is better. So we have to improve the accuracy of the theoretical calculation and to take into account the radiative corrections (see lecture V).

Consider now the decays of W boson. According to eq. (3.23) the amplitude of W^+ decay into $\tilde{\nu}_l l$ can be written in the form

$$T(W \rightarrow \tilde{\nu}_l l) = \frac{e}{2\sqrt{2}s} W_\alpha (\bar{l} \gamma_\alpha (1 + \gamma_5) \nu) \quad (4.39)$$

Neglecting leptonic mass we get

$$\Gamma(W \rightarrow l \tilde{\nu}) = \frac{1}{6\sqrt{2}\pi} G_\mu m_W^3 \simeq 226 \text{ MeV} . \quad (4.40)$$

Consider now the hadronic decays of W -boson. The decay width into top quark is equal to zero since top quark is heavier than W -boson. With good accuracy we can neglect the masses of other quark. In this approximation

$$\Gamma(W \rightarrow \text{hadrons}) = 3 \cdot 2\Gamma(W \rightarrow e \tilde{\nu}) \quad (4.41)$$

where factor 3 is due to color of each quark and factor 2 is due to two quark channels of W decay.

Exercise: Prove that in massless approximation hadronic width eq. (4.41) does not depend on the CKM angles.

In this approximation the theoretical prediction for the total width is

$$\Gamma_W^{th} \simeq 9\Gamma(W \rightarrow \tilde{\nu}_l l) \simeq 2.03 \text{ GeV} \quad (4.42)$$

that has to be compared with the experimental value

$$\Gamma_W^{exp} \simeq 2.06 \pm 0.06 \text{ GeV} . \quad (4.43)$$

As it was expected the agreement is of the order of percent. The experimental accuracy is of the same order.

Lecture V. **Radiative Correction in SM.**
Hunting for virtual t -quark.

5.0. Z -physics at LEP and SLC.

To test the predictions of the SM the huge "factories" of Z -bosons (e^+e^- colliders) were constructed at CERN (LEP) and at SLAC (SLC). Electrons and positrons in these colliders collide at the centre of mass energy equal to the Z -boson mass. The reactions that are studied can be presented in the form

$$e^+e^- \rightarrow Z \rightarrow \bar{f}f$$

where

$$\bar{f}f = \begin{cases} \nu\bar{\nu} & \text{invisible modes} \\ \bar{l}l & \text{charged leptons} \\ q\bar{q} & \text{hadrons.} \end{cases}$$

The LEP I was terminated in the fall 1995 in order to give place to LEP II. Four groups — collaborations ALEPH, DELPHI, L3, OPAL — have collected near $2 \cdot 10^7$ Z -bosons. The SLC has worked from the fall 1989 till the fall 1998. The SLD detector recorded near $5 \cdot 10^5$ Z -bosons. The electrons in the SLC were longitudinally polarized so that the SLD group could provide very high precision data having relatively low statistics. More than two thousand experimentalists and engineers and hundreds of theorists participated in this unique project — one of the largest projects in the history of physics.

Near the dozen of independent observables were measured with fantastic precision of the order of 10^{-3} (10^{-5} for the case of Z -boson mass). The scale of the radiative corrections in the SM is of the order of $\alpha_{W,Z}/\pi \sim 10^{-2} - 10^{-3}$. Therefore LEP-I and SLD data provide a precision test of the SM as a renormalizable field theory, i.e. with loops included.

The theoretical study of electroweak corrections in the SM started in the 1970's and was elaborated by a number of theoretical groups. Near 10^2 theorists had published near $2 \cdot 10^3$ papers on radiative corrections (see e.g. ref. [2]). This study has been summarized in Yellow CERN Reports of Working Groups on precision calculations for the Z resonance in 1995 [3]. The deviations in theoretical calculations of different groups are by the order of magnitude smaller than the experimental uncertainties.

By comparing the $\Gamma_{invisible}$ with theoretical predictions for neutrino decays the result of fundamental importance was established — the sequence of the generations with light neutrinos is completed with number of generations $N_f = 3$.

The best fit of the most recent data submitted to the Vancouver (1998) conference is presented in Table I.

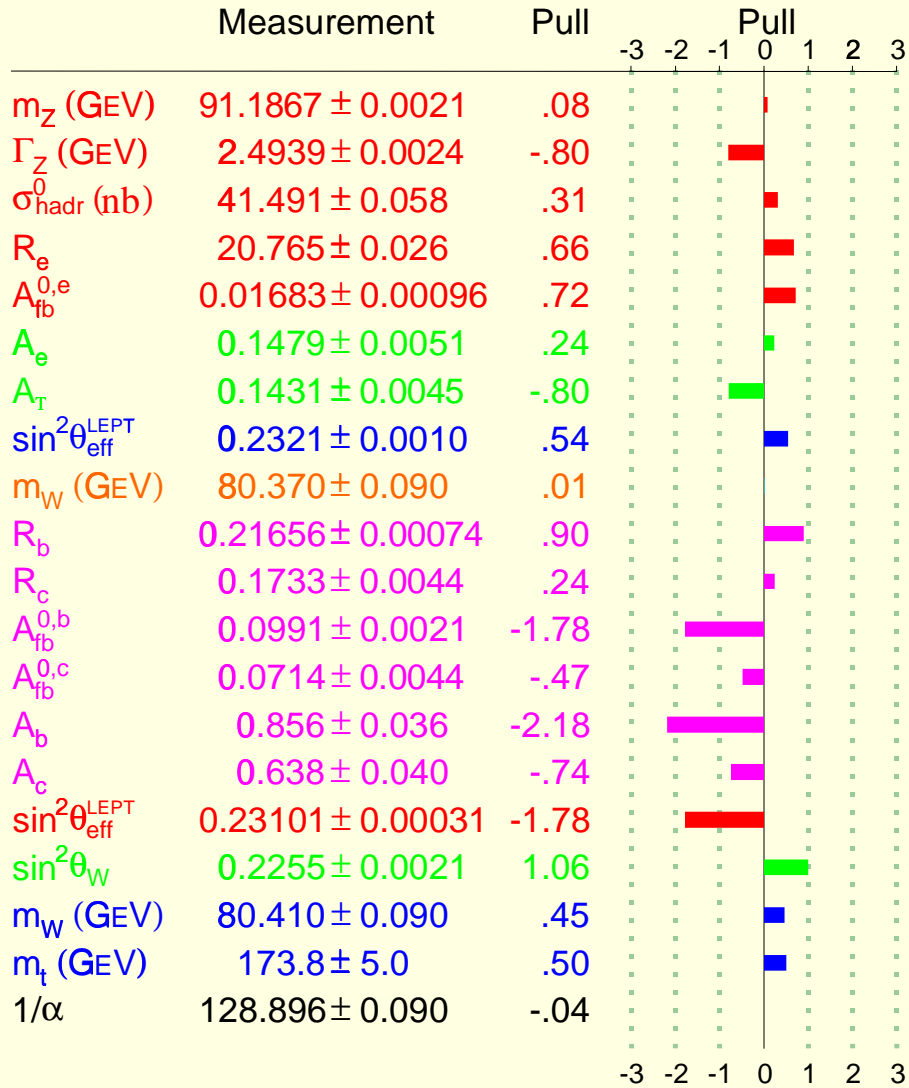
This fit gives

$$m_t = 171.6 \pm 4.9 \text{ GeV}$$

$$m_H = 159.1^{+144}_{-83} \text{ GeV}$$

$$\hat{\alpha}_s(m_Z) = 0.119 \pm 0.003$$

Pulls from SM fit to all data



$\chi^2 / dof = 16.4 / 15$

$$\chi^2/n.d.f. = 16.4/15$$

The quality of this fit is very good. We conclude that the SM gives the perfect description of Z physics. New physics can not improve the fit of LEP and SLC data. Thus the Standard Model has been confirmed up to the loop corrections. What is more important is that the loop corrections can be used to gather data on the not yet discovered particle. For instance, even before t -quark was discovered at Tevatron, its mass was predicted by analyzing the loops and LEP-SLC data. The hunting for virtual top quark is a very bright example of the collaboration of the theory with the experiment.

5.1. Decoupling of heavy flavors from Low-Energy Physics in QED and QCD.

It is interesting to understand why in 1950's nobody worried about the contribution of top quark (and other heavy flavours) into magnetic moment of the electron known with very high accuracy. The answer to this question is that for $q \sim m_e$ the corrections due to top quark are suppressed as a power of (m_e^2/m_t^2) i.e. the contribution was negligible. In QED we have decoupling of heavy (unknown) particles from the low-energy observables. Why? Consider the contribution of t -quark into QED observables. The only diagram with t -quarks in loop is the self-energy of the photon

$$\Pi_{\mu\nu}(q) = i\langle 0 | \{j_\mu(q), j_\nu(-q)\} | 0 \rangle \quad (5.1)$$

where $j_\mu(q)$ is the electromagnetic current of t -quark. Self energy has dimension 2: $[\Pi_{\mu\nu}] = m^2$. So one can expect that there exit terms of the order of

$$\Pi_{\mu\nu} \sim \alpha m_t^2 g_{\mu\nu}$$

This expectation is wrong in the case of conserved currents

$$q_\mu j_\mu(q) = 0 \quad (5.2)$$

Indeed for conserved current the self-energy operator should be transversal $q_\mu \Pi_{\mu\nu} = 0$. So

$$\Pi_{\mu\nu}(q) = (q_{\mu\nu}q^2 - q_\mu q_\nu) \Pi(q^2), \quad (5.3)$$

Equation (5.3) implies that the photon remains massless. The scalar function $\Pi(q^2)$ has dimension zero and the only possible contribution of t -quark into $\Pi(q^2)$ can be written in the form

$$\Pi(q^2) \sim \alpha \ln \frac{\Lambda^2}{m_t^2 + q^2}$$

where Λ is cut-off. The self-energy keeps the memory of heavy flavours!

The crucial step is renormalization. For instance consider the Coulomb scattering. If we take into account the infinite chain of self-energy contribution into the photon propagator we get for Coulomb amplitude

$$T_{Coulomb} = \frac{e_0^2(\Lambda)}{q^2(1 + \Pi(q^2))} \quad (5.4)$$

At low q^2 we reproduce the Coulomb-low

$$T = \frac{e_{phys}^2}{q^2} \quad (5.5)$$

with

$$e_{phys}^2 = \frac{e_0^2(\Lambda)}{1 + \Pi(0)} \quad (5.6)$$

When we rewrite the amplitude (5.4) in terms of e_{phys}^2 we get

$$T \simeq \frac{e_{phys}^2}{q^2[1 + \Pi(q^2) - \Pi(0)]} \quad (5.7).$$

As a result:

1) the dependence on cut-off Λ disappears

$$\Delta\Pi = \Pi(q^2) - \Pi(0) \sim \alpha \ln \frac{m_t^2}{m_t^2 + Q^2};$$

2) the contribution of heavy flavour is suppressed as a power (q^2/m_t^2):

$$\Delta\Pi \sim -\alpha \left(\frac{q^2}{m_t^2} \right) \rightarrow 0.$$

This is the sample of the famous decoupling theorem. It works for the theories with conserved vector currents.

5.2. Non-decoupling of chiral matter.

Heavy Flavor contribution into electroweak observables.

In the Standard Model the left components of t - and b -quarks belong to $SU(2)_W$ doublet representation: $Q_L = \begin{pmatrix} t_L \\ b_L \end{pmatrix}$. Therefore for the case when $m_t \gg m_b$ and for small energies $E \leq m_t$ we have effectively the explicit violation of $SU(2)_W$ symmetry. For the virtual momenta $q \sim \Lambda \sim m_t$ theory looks like the old nonrenormalizable theory. It mean that one-loop corrections diverge quadratically $\delta_1 \sim \alpha\Lambda^2/m_Z^2$, two-loop corrections diverge quartically $\delta_2 \sim (\alpha\Lambda^2/m_Z^2)^2$. So we expect that the corrections to the low-energy observables due to top contribution are of the order of

$$\delta_1 \sim \alpha_W t$$

$$\delta_2 \sim \alpha_W^2 t^2$$

where $t = m_t^2/m_Z^2$, i.e. corrections are not suppressed, they grow with top mass m_t . Heavy flavours are not decoupled from the low-energy observables for the chiral matter. As a result the radiative corrections in the SM are sensitive to the top contribution.

Consider for example the ratio of g_V and g_A for the decay $Z \rightarrow \bar{l}l$. It is rather simple exercise to calculate this ratio

$$R_l = \left(\frac{g_V}{g_A} \right)_l = 1 - 4s^2 - \frac{3\bar{\alpha}}{4\pi(c^2 - s^2)}(t + \delta V_R) \quad (5.8)$$

The linear term $\sim t$ originates from different self-energies with top quark in the loops and δV_R is the contribution of the rest of the diagrams (a few dozens ones) that do not depend on t .

If we compare experimental value of t

$$t^{\text{exp}} \simeq 3.7$$

with experimental value of $(t + \delta V_R)$

$$(t + \delta V_R)^{\text{exp}} \simeq -0.49 \pm 0.32$$

we find that the constant terms are as much important as a linear term.

So we have to perform the accurate calculation of the whole set of one-loop diagrams.

5.3. Radiative corrections in SM.

a) Strategy.

There are 3 steps in the calculations of the radiative corrections.

1) Calculate all observables in terms of base parameters g_1^B, g_2^B, η^B and cut-off Λ .

2) Find 3 basic observables and express base parameters in terms of these 3 physical observables and cut-off Λ .

3) Substitute these expressions into formulas for other observables. Dependence on cut-off disappears.

b) Basic parameters.

It is reasonable to calculate observables in terms of quantities that are known to the highest degree of precision.

$$\alpha^{-1} = 137.035985(61)$$

$$G_\mu = 1.16639(2) \cdot 10^{-5} \text{Gev}^{-2} \quad (5.9)$$

$$m_Z = 91.1867(21) \text{Gev}$$

c) The Choice of Born approximation

In contrast to $\alpha(q^2)$ that is running coupling constant, the two electroweak coupling constants $\alpha_Z(q^2)$ and $\alpha_W(q^2)$ are not "running" but "crawling" for $q^2 \leq m_Z^2$.

The natural scale for electroweak physics is $q^2 = m_Z^2$. Therefore it is evident that $\bar{\alpha} \equiv \alpha(m_Z^2)$, not $\alpha \equiv \alpha(0)$ is relevant parameter in electroweak physics. The value of $\bar{\alpha}$ is less accurate.

$$\bar{\alpha} = \alpha(m_Z^2) = [128.896 \pm 0.090]^{-1} \quad (5.10)$$

With this parameterization it is convenient to introduce the weak angle θ by the relation

$$G_\mu = \frac{\pi \bar{\alpha}}{\sqrt{2} m_Z^2 s^2 c^2} \quad (5.11)$$

where $s^2 = \sin^2 \theta$, $c^2 = \cos^2 \theta$. Numerically

$$s^2 = 0.2311(22) \quad (5.12)$$

d) Basic relations for electroweak observables.

A simple calculation gives the following result for one-loop electroweak corrections in the case gluon free observables:

$$\frac{m_W}{m_Z} = c \left[1 + \frac{3\bar{\alpha}}{32\pi s^2(c^2 - s^2)}(t + \delta V_m) \right]$$

$$g_{Al} = -\frac{1}{2} - \frac{3\bar{\alpha}}{64\pi s^2 c^2}(t + \delta V_A) \quad (5.13)$$

$$R_l = (g_V/g_A)_l = 1 - 4s^2 + \frac{3\bar{\alpha}}{4\pi(c^2 - s^2)}(t + \delta V_R)$$

There are a few diagrams that contribute into leading term $\sim t$ and there are dozens of them for nonleading δV .

The hadronic decays of Z are more sensitive to the value of the gluonic coupling constant α_s .

$$\Gamma(Z \rightarrow q\bar{q}) = \frac{G_\mu m_Z^3}{2\sqrt{2}\pi} [g_A^2 R_A + g_V^2 R_V] \quad (5.14)$$

The "radiator" $R_{A,V}$ contain QCD and QED corrections caused by the final state emission and exchange of gluons and photons. There are hundreds of diagrams that contribute into $R_{A,V}$.

5.4. Hunting for virtual top. Great success.

A comparison of the experimental data with the result of theoretical calculation eq.(5.13) led to the prediction of the t -quark mass m_t .

$$m_t = 180(5)_{-20}^{+17} \text{ Gev}$$

where the number in parentheses is the uncertainty due to the uncertainties of the data. The center value corresponds to the assumption that $m_H = 300$ Gev, the upper and lower "shifts" correspond to $m_H = 1000$ Gev and 60 Gev, respectively.

The best fit of all observables gives

$$(m_t) \simeq 173.8 \pm 5.3 \text{ Gev}$$

These numbers are in perfect agreement with the recent direct measurement of the top-quark mass by two collaborations at FNAL

$$m_t = 173.8 \pm 5.0 \text{ Gev.}$$

5.5. Hopeless hunting for virtual Higgs.

Consider the limit of very large Higgs boson mass m_H . For $E \ll m_H$ we have $SU(2)$ symmetric theory of massive gauge bosons, i.e. effectively nonrenormalizable theory. As I

explained in the Lecture II due to the gauge symmetry the leading divergence of the loop disappears. So the one-loop corrections diverge logarithmically

$$\delta_1 \sim \alpha_W \ln \frac{\Lambda^2}{m_Z^2} \sim \alpha_W \ln \frac{m_H^2}{m_Z^2} = \alpha_W \ln h$$

two-loop corrections diverge quadratically

$$\delta_2 \sim \alpha_W^2 \left(\frac{\Lambda^2}{m_Z^2} \right) \sim \alpha_W^2 h.$$

Here $h = m_H^2/m_Z^2$.

This is the famous Veltman screening theorem. The weak dependence of radiative corrections on h results in a rather poor accuracy for m_H derived from the precision data. The central value of m_H from the fit should not be taken seriously. It is very unstable. Any tiny corrections or any change of the parameter can shift it by the order of magnitude.

The one sigma upper bound is more reliable. According to the recent fit

$$m_H < 280 \text{Gev} \quad 95\% \text{ c.l.}$$

It seems that the fit of the precision data is not the best way for hunting for Higgs boson.

REFERENCES

- 1 . Steven Weinberg, "The Quantum Theory of Fields:Foundations", Vol.1, Cambridge Univ.Pr.,1995 and "Quantum Theory of Fields:Modern Applications",Vol.2, Cambridge Univ.Pr.,1996 .
 Michael E.Peskin, Daniel V.Schroeder, "An Introduction to Quantum Field Theory", Addison-Wesley Pub.Co.,1995
 Martinus Veltman, "Diagrammatica:The Path to Feynman Rules", Cambridge Univ.Pr.,1994
 Lev B.Okun, "Leptons and Quarks" ,Amsterdam,North-Holland,1982
- 2 . 't Hooft G., Nucl. Phys. **B33**, 173 (1971), **35**, 167 (1971).
 Veltman M., Nucl. Phys. **B123**, 89 (1977); Acta Phys. Pol. **B8**, 475 (1977);

 Passarino G., Veltman M., Nucl. Phys. **B160**, 151 (1979).

 Berman S.M., Sirlin A., Ann. Phys. (N.Y.) **20**, 20 (1962);

 Sirlin A., Rev. Mod. Phys. **50**, 573 (1978).
 Sirlin A., Phys. Rev. **D22**, 971 (1980);

 Marchiano W.J. and Sirlin A., Phys. Rev. **D22**, 2695 (1980);

 Degrassi G., Fanchiotti S. and Sirlin A., Nucl. Phys. **B351**, 49 (1991).
 Fleischer J. and Jegerlehner F., Phys. Rev. **D23**, 2001 (1981).
 Aoki K.I., Hioki Z., Kawabe R., Konuma M. and Muta T., Suppl. Prog. Theor. Phys. **73**, 1 (1982).
 Consoli M., LoPresti S. and Maiani L., Nucl. Phys. **B223**, 474 (1983);

 Barbieri R., Maiani L., Nucl. Phys. **B224**, 32 (1983).
 Lynn B.W., Peskin M.E., Report SLAC-PUB-3724 (1985) (unpublished);

 Lynn B.W., Peskin M.E., Stuart R.G., *in* Physics at LEP (Report CERN 86-02) (CERN, Geneva, 1986), p. 90.
 Kennedy D.C. and Lynn B.W., Nucl. Phys. **B322**, 1 (1989).
 Bardin D.Yu., Christova P.Ch. and Fedorenko O.M., Nucl. Phys. **B175**, 235 (1980);

 Bardin D.Yu., Christova P.Ch. and Fedorenko O.M., Nucl. Phys. **B197**, 1 (1982);

 Bardin D.Yu., Bilenky M.S., Mitselmakher G.V., Riemann T. and Sachwitz M., Z. Phys. **C44**, 493 (1989).
 Bardin D. et al., Program ZFITTER 4.9, Nucl. Phys. **B351**, 1 (1991); Z. Phys. **C44**, 493 (1989); Phys. Lett. **B255**, 290 (1991); Preprint CERN-TH.6443-92 (1992).
 Altarelli G., Barbieri R., Phys. Lett. **B253**, 161 (1991);

 Altarelli G., Barbieri R. and Jadach S., Nucl. Phys. **B369**, 3 (1992), Erratum-ibid **B376**, 446 (1992);

 Altarelli G., Barbieri R., Caravaglios F., Nucl.Phys. **B405**, 3 (1993), Phys. Lett. **B314**, 357 (1993), Phys. Lett. **B349**, 145 (1995).
 Ellis J., Fogli G., Phys. Lett. **B213**, 189, 526 (1988); **232**, 139 (1989); **249**, 543 (1990).
 Hollik W., Fortschr. Phys. **38**, 3, 165 (1990);

 Consoli M., Hollik W., Jegerlehner F., Proc. of the Workshop on Z physics at LEPI

(CERN Report 89-08) Vol. I, p. 7;

Burgers G. et al., *ibid.* p. 55.

Montagna G. et al., *Nucl. Phys.* **B401**, 3 (1993);

Montagna G. et al., Program TOPAZO, *Comput. Phys. Commun.* **76**, 328 (1993).

Chetyrkin K.G., Kühn J.H., *Phys. Lett.* **B248**, 359 (1990);

Chetyrkin K.G., Kühn J.H., Kwiatkowski A., *Phys. Lett.* **282**, 221 (1992).

Gorishny S.G., Kataev A.L., Larin S.A., *Phys. Lett.* **B259**, 144 (1991);

Surguladze L.R., Samuel M.A., *Phys. Rev. Lett.* **66**, 560 (1991).

Novikov V., Okun L., Vysotsky M., *Nucl. Phys.* **B397**, 35 (1993);

Vysotsky M.I., Novikov V.A., Okun L.B., Rozanov A.N., *Uspekhi Fiz. Nauk*, v. 166, 539 (1996) (Russian), *Physics-Uspekhi* **39**(5), 503 (1996) (English);

Novikov V.A., Okun L.B. and Vysotsky M.I., *Mod. Phys. Lett.* **A8**, 2529 (1993); 3301 (E).

- 3** . Reports on the working groups on precision calculations for the Z resonance (Report CERN 95-03), (CERN, Geneva, 1995), p. 7-163.

INTRODUCTION TO QCD

Michelangelo L. Mangano

CERN, TH Division, Geneva, Switzerland

Abstract

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

1 Introduction

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

If you go through lecture series on QCD (e.g., the lectures given in previous years at the CERN Summer School, refs. [9,10,11]), you will hardly ever find the same item twice. This is because QCD today covers a huge set of subjects and each of us has his own concept of what to do with QCD and of what are the “fundamental” notions of QCD and its “fundamental” applications. As a result, you will find lecture series centred around non-perturbative applications, (lattice QCD, sum rules, chiral perturbation theory, heavy quark effective theory), around formal properties of the perturbative expansion (asymptotic behaviour, renormalons), techniques to evaluate complex classes of Feynman diagrams, or phenomenological applications of QCD to possibly very different sets of experimental data (structure functions, deep-inelastic scattering (DIS) sum rules, polarized DIS, small x physics (including hard pomerons, diffraction), LEP physics, $p\bar{p}$ collisions, etc.

I can anticipate that I will not be able to cover or to simply mention all of this. After introducing some basic material, I will focus on some elementary applications of QCD in high-energy e^+e^- , ep and $p\bar{p}$ collisions. The outline of these lectures is the following:

1. Gauge invariance and Feynman rules for QCD.
2. Renormalization, running coupling, renormalization group invariance.
3. QCD in e^+e^- collisions: from quarks and gluons to hadrons, jets, shape variables.
4. QCD in lepton-hadron collisions: DIS, parton densities, parton evolution.
5. QCD in hadron-hadron collisions: formalism, W/Z production, jet production.

Given the large number of papers which contributed to the development of the field, it is impossible to provide a fair bibliography. I therefore limit my list of references to some excellent

books and review articles covering the material presented here, and more. Papers on specific items can be easily found by consulting the standard `hep-th` and `hep-ph` preprint archives.

2 QCD Feynman rules

There is no free lunch, so before starting with the applications, we need to spend some time developing the formalism and the necessary theoretical ideas. I will dedicate to this purpose the first two lectures. Today, I concentrate on Feynman rules. I will use an approach which is not canonical, namely it does not follow the standard path of the construction of a gauge invariant Lagrangian and the derivation of Feynman rules from it. I will rather start from QED, and empirically construct the extension to a non-Abelian theory by enforcing the desired symmetries directly on some specific scattering amplitudes. Hopefully, this will lead to a better insight into the relation between gauge invariance and Feynman rules. It will also provide you with a way of easily recalling or checking your rules when books are not around!

2.1 Summary of QED Feynman rules

We start by summarizing the familiar Feynman rules for Quantum Electrodynamics (QED). They are obtained from the Lagrangian:

$$\mathcal{L} = \bar{\psi}(i\partial\!\!\!/ - m)\psi - e\bar{\psi}\mathbf{A}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \quad (1)$$

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

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (2)$$

The resulting Feynman rules are summarized in the following table:

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

$$\begin{array}{c} \mu \\ \text{~~~~~} \\ \text{~~~~~} \end{array} \xrightarrow{p} \begin{array}{c} \nu \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = -i\frac{g_{\mu\nu}}{p^2 + i\epsilon} \text{ (Feynman gauge)} \quad (4)$$

$$\begin{array}{c} \diagdown \\ \text{~~~~~} \\ \diagup \end{array} = -ie\gamma_\mu Q \quad (Q = -1 \text{ for the electron, } Q = 2/3 \text{ for the u-quark, etc.}) \quad (5)$$

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

$$\begin{array}{c} \text{q} \xrightarrow{\quad} \\ \quad \quad \quad \diagup \quad \quad \quad \diagdown \\ \quad \quad \quad \text{k}_1, \mu \\ \quad \quad \quad \diagdown \quad \quad \quad \diagup \\ \quad \quad \quad \text{k}_2, \nu \\ \text{q} \xleftarrow{\quad} \end{array} \quad \begin{array}{c} \text{q} \xrightarrow{\quad} \\ \quad \quad \quad \diagdown \quad \quad \quad \diagup \\ \quad \quad \quad \text{k}_1, \mu \\ \quad \quad \quad \diagup \quad \quad \quad \diagdown \\ \quad \quad \quad \text{k}_2, \nu \\ \text{q} \xleftarrow{\quad} \end{array} = D_1 + D_2 \quad (6)$$

The total amplitude M_γ is given by:

$$\frac{i}{e^2} M_\gamma \equiv D_1 + D_2 = \bar{v}(\bar{q}) \not{\epsilon}_2 \frac{1}{\not{q} - \not{k}_1} \not{\epsilon}_1 u(q) + \bar{v}(\bar{q}) \not{\epsilon}_1 \frac{1}{\not{q} - \not{k}_2} \not{\epsilon}_2 n(q) \equiv M_{\mu\nu} \epsilon_1^\mu \epsilon_2^\nu \quad (7)$$

Gauge invariance demands that

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

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

$$\begin{aligned}\partial_\mu M^\mu = 0 &\Rightarrow \frac{d}{dt} \int M^0 d^3x = \int \partial_0 M^0 d^3x \\ &= \int \vec{\nabla} \cdot \vec{M} d^3x = \int_{S \rightarrow \infty} \vec{M} \cdot d\vec{\Sigma} = 0\end{aligned}\quad (9)$$

In momentum space, this means

$$k_1^\mu M_\mu = 0. \quad (10)$$

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

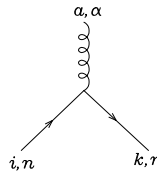
$$A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu f(x) \quad (11)$$

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

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

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

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



$$= ig\lambda_{ki}^a \gamma_{mn}^\mu \quad (13)$$

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

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

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

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

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

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

$$M_g = (\lambda^a \lambda^b)_{ij} M_\gamma - g^2 f^{abc} \lambda_{ij}^c D_1. \quad (16)$$

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

$$k_1^\mu \epsilon_2^\nu M_g^{\mu\nu} = \epsilon_1^\mu k_2^\nu M_g^{\mu\nu} = 0. \quad (17)$$

Substituting $\epsilon_1^\mu \rightarrow k_1^\mu$ in (16) we get instead, using (12):

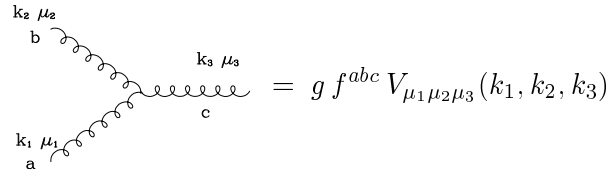
$$k_{1\mu} M_g^\mu = -g^2 f^{abc} \lambda_{ij}^c \bar{v}_i(\bar{q}) \not{\epsilon}_2 u_i(q) \quad (18)$$

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

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

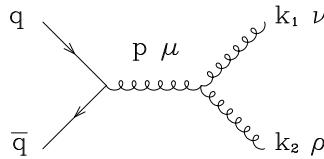
$$k_{1\mu} M_g^\mu = i (f^{abc} g \epsilon_2^\mu) \times (i g \lambda_{ij}^c \bar{v}_i(\bar{q}) \gamma_\mu u_i(q)) \quad (19)$$

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



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

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



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

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

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

Lorentz invariance requires then that V be built out of terms of the form $g_{\mu_1 \mu_2} p_{\mu_3}$. Bose symmetry requires V to be fully antisymmetric under the exchange of any pair $(\mu_i, p_i) \leftrightarrow (\mu_j, p_j)$

since the colour structure f^{abc} is totally antisymmetric. As a result, for example, a term like $g_{\mu_1\mu_2}p_3^{\mu_3}$ vanishes under antisymmetrization, while $g_{\mu_1\mu_2}p_1^{\mu_3}$ doesn't. Starting from this last term, we can easily add the pieces required to obtain the full antisymmetry in all three indices. The result is unique, up to an overall factor:

$$V_{\mu_1\mu_2\mu_3} = V_0 [(k_1 - k_2)^{\mu_3} g_{\mu_1\mu_2} + (k_2 - k_3)^{\mu_1} g_{\mu_2\mu_3} + (k_3 - k_1)^{\mu_2} g_{\mu_1\mu_3}] \quad (22)$$

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

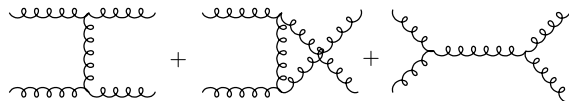
$$k_1^{\mu_1} \epsilon_2^{\mu_2} V_{\mu_1\mu_2\mu} (k_1, k_2, k_3) = V_0 \{ -(k_1 + k_2)^\mu (k_1 \cdot \epsilon_2) + 2(k_1 \cdot k_2) \epsilon_2^\mu - (k_2 \cdot \epsilon_2) k_1^\mu \} . \quad (23)$$

The gauge variation is therefore:

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

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

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



$$\quad (25)$$

Once more it is necessary to verify the gauge invariance of this amplitude. It turns out that one more diagram is required, induced by a four-gluon vertex. Lorentz invariance, Bose symmetry and dimensional analysis uniquely determine once again the structure of this vertex. The overall factor is fixed by gauge invariance. The resulting Feynman rule for the 4-gluon vertex is given in fig. 1.

You can verify that the 3- and 4-gluon vertices we introduced above are exactly those which arise from the Yang-Mills Lagrangian:

$$\mathcal{L}_{YM} = -\frac{1}{4} \sum_a F_{\mu\nu}^a F^{a\mu\nu} \quad \text{with} \quad F_{\mu\nu}^a = \partial_{[\mu} A_{\nu]}^a - g f^{abc} A_{[\mu}^b A_{\nu]}^c \quad (26)$$

It can be shown that the 3 and 4-gluon vertices we generated are all is needed to guarantee gauge invariance even for processes more complicated than those studied in the previous simple examples. In other words, no extra 5- or more gluon vertices have to be introduced to achieve the gauge invariance of higher-order amplitudes. At the tree level this is the consequence of dimensional analysis and of the locality of the couplings (no inverse powers of the momenta can appear in the Lagrangian). At the loop level, these conditions are supplemented by the renormalizability of the theory [3,7].

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

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

$$\begin{aligned}
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ b, \beta \end{array} \begin{array}{c} p \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = \delta^{ab} \frac{-i g^{\alpha\beta}}{p^2 + i\epsilon} \quad (\text{Feynman gauge}) \\
& \begin{array}{c} a \\ \text{-----} \\ \text{-----} \\ \text{-----} \\ b \end{array} \begin{array}{c} p \\ \text{-----} \\ \text{-----} \\ \text{-----} \end{array} = \delta^{ab} \frac{i}{p^2 + i\epsilon} \\
& \begin{array}{c} i, n \\ \text{-----} \\ \text{-----} \\ k, m \end{array} \begin{array}{c} p \\ \text{-----} \\ \text{-----} \\ \text{-----} \end{array} = \delta^{ik} \frac{i}{\not{p} - m + i\epsilon} \Big|_{mn} \\
& \begin{array}{c} b, \beta \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ a, \alpha \end{array} \begin{array}{c} q \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = g f^{abc} \left[g^{\alpha\beta} (p - q)^\gamma + g^{\beta\gamma} (q - r)^\alpha + g^{\gamma\alpha} (r - p)^\beta \right] \\
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ c, \gamma \end{array} \begin{array}{c} p \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} \begin{array}{c} r \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} \begin{array}{c} c, \gamma \end{array} \\
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ c, \gamma \end{array} \begin{array}{c} b, \beta \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = -i g^2 f^{xac} f^{xbd} \left(g^{\alpha\beta} g^{\gamma\delta} - g^{\alpha\delta} g^{\beta\gamma} \right) \\
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ c, \gamma \end{array} \begin{array}{c} b, \beta \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = -i g^2 f^{xad} f^{xbc} \left(g^{\alpha\beta} g^{\gamma\delta} - g^{\alpha\gamma} g^{\beta\delta} \right) \\
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ c, \gamma \end{array} \begin{array}{c} d, \delta \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = -i g^2 f^{xab} f^{xcd} \left(g^{\alpha\gamma} g^{\beta\delta} - g^{\alpha\delta} g^{\beta\gamma} \right) \\
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ b \end{array} \begin{array}{c} q \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \end{array} = -g f^{abc} q^\alpha \\
& \begin{array}{c} a, \alpha \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ \text{~~~~~} \\ i, n \end{array} \begin{array}{c} k, m \end{array} = i g \lambda_{ki}^a \gamma_{mn}^\alpha
\end{aligned}$$

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

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

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

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

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

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

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

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

$$k_1^{\mu_1} k_2^{\mu_2} M_{\mu_1 \mu_2} = 0 \quad (30)$$

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

Exercise: show that

$$\sum_{\text{non-physical}} |\epsilon_1^\mu \epsilon_2^\nu M_{\mu\nu}|^2 = \left| i g^2 f^{abc} \lambda^c \frac{1}{2k_1 k_2} \bar{v}(\bar{q}) k_\lambda u(q) \right|^2 \quad (31)$$

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

$$\left| \begin{array}{c} \text{Diagram: } q\bar{q} \text{ annihilation into } gg \text{ via a ghost loop} \end{array} \right|^2 = - \left| i g^2 f^{abc} \lambda^c \frac{1}{2k_1 k_2} \bar{v}(\bar{q}) k_\lambda u(q) \right|^2, \quad (32)$$

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

The detailed derivation of the need for and properties of ghosts (including their Feynman rules and the “−1” prescription for loops) can be found in the suggested textbooks. I will not derive these results here since we will not need them for our applications (we will use physical gauges or will consider processes not involving the $3g$ vertex). The full set of Feynman rules for the QCD Lagrangian is given in fig. 1.

2.2 Some Useful Results in Colour Algebra

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

$$[\lambda^a, \lambda^b] = if^{abd}\lambda^c \quad (33)$$

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

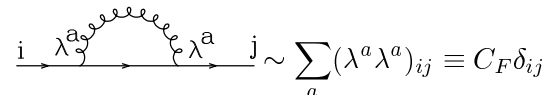
$$\text{tr}(\lambda^a \lambda^b) = T_F \delta_{ab} \quad (34)$$

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

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

$$\delta \lambda^a = \sum_{b,c} \epsilon_b f^{abc} \lambda^c \quad (35)$$

The definition of T_F allows to evaluate the colour factor for an interesting diagram, i.e. the quark self-energy:



$$\text{Diagram} \sim \sum_a (\lambda^a \lambda^a)_{ij} \equiv C_F \delta_{ij} \quad (36)$$

The value of C_F can be obtained by tracing the relation above:

$$C_F N = \text{tr} \sum_a \lambda^a \lambda^a = \delta^{ab} T_F \delta_{ab} = \frac{N^2 - 1}{2} \quad (37)$$

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

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

$$\longrightarrow \text{fermion} \quad (38)$$

$$\begin{array}{c} \longrightarrow \\ \longleftarrow \end{array} \text{gluon} \quad (39)$$

$$\frac{1}{\sqrt{2}} \left(\begin{array}{c} \downarrow \\ \uparrow \\ \longrightarrow \end{array} - \frac{1}{N} \begin{array}{c} \downarrow \\ \downarrow \\ \longrightarrow \end{array} \right) \text{Fermion-Gluon Vertex } (t^a) \quad (40)$$

$$\frac{1}{\sqrt{2}} \left(\begin{array}{c} \downarrow \\ \downarrow \\ \downarrow \end{array} - \begin{array}{c} \downarrow \\ \downarrow \\ \downarrow \end{array} \right) \text{3-Gluon Vertex } (f^{abc}) \quad (41)$$

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

As a first example of applications, let us reevaluate C_F :

$$\begin{aligned} i \xrightarrow{\lambda^a} \text{---} \xrightarrow{\lambda^a} j &= \frac{1}{\sqrt{2}} \left(\begin{array}{c} \downarrow \\ \uparrow \\ \longrightarrow \end{array} - \frac{1}{N} \begin{array}{c} \downarrow \\ \downarrow \\ \longrightarrow \end{array} \right) \times \frac{1}{\sqrt{2}} \left(\begin{array}{c} \downarrow \\ \uparrow \\ \longrightarrow \end{array} - \frac{1}{N} \begin{array}{c} \downarrow \\ \downarrow \\ \longrightarrow \end{array} \right) \\ &= \frac{1}{2} \left(\begin{array}{c} \text{---} \xrightarrow{N} \text{---} \\ \text{---} \xrightarrow{N} \text{---} \end{array} - \frac{1}{N} \begin{array}{c} \text{---} \xrightarrow{N} \text{---} \\ \text{---} \xrightarrow{N} \text{---} \end{array} - \frac{1}{N} \begin{array}{c} \text{---} \xrightarrow{N} \text{---} \\ \text{---} \xrightarrow{N} \text{---} \end{array} + \right. \\ &\quad \left. + \frac{1}{N^2} \begin{array}{c} \text{---} \xrightarrow{N} \text{---} \\ \text{---} \xrightarrow{N} \text{---} \end{array} \right) = \delta^{ij} \frac{N^2 - 1}{2N} \quad (42) \end{aligned}$$

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

$$\sum_a (\lambda^a)_{ij} (\lambda^a)_{lk} = \begin{array}{c} j \\ \uparrow \\ \text{---} \\ \downarrow \\ k \\ i \end{array} = \frac{1}{2} \left(\begin{array}{c} \downarrow \\ \uparrow \\ \longrightarrow \\ \downarrow \\ \uparrow \\ \longrightarrow \end{array} - \frac{1}{N} \begin{array}{c} \downarrow \\ \downarrow \\ \longrightarrow \\ \downarrow \\ \downarrow \\ \longrightarrow \end{array} \right) = \frac{1}{2} \left(\delta_{ik} \delta_{lj} - \frac{1}{N} \delta_{ij} \delta_{lk} \right) \quad (43)$$

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

$$\begin{array}{c} \text{---} \\ \uparrow \\ \text{---} \\ \downarrow \\ \text{---} \end{array} = \frac{1}{2} \left(\begin{array}{c} \text{---} \\ \uparrow \\ \text{---} \\ \downarrow \\ \text{---} \end{array} - \frac{1}{N} \begin{array}{c} \text{---} \\ \uparrow \\ \text{---} \\ \downarrow \\ \text{---} \end{array} \right) = \frac{1}{2} \frac{N^2 - 1}{N} \delta_{ij} = C_F \delta_{ij} \quad (44)$$

For the interaction with a gluon we have instead:

$$\begin{aligned}
 & \text{Diagram: } \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} \\
 &= \frac{1}{\sqrt{2}} \left(\begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} - \frac{1}{N} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} \right) \times \frac{1}{2} \left(\begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} - \frac{1}{N} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} \right) \\
 &= \frac{1}{2\sqrt{2}} \left(\begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} - \frac{1}{N} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} - \frac{1}{N} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} + \frac{1}{N^2} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} \right) \\
 &= -\frac{1}{2N} \frac{1}{\sqrt{2}} \left(\begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} - \frac{1}{N} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} \right) = -\frac{1}{2N} \begin{array}{c} \text{gluon} \\ | \\ \text{triangle} \\ | \\ \text{gluon} \end{array} \quad (45)
 \end{aligned}$$

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

The remaining important relation that one needs is the following:

$$\sum_{a,b} f^{abc} f^{abd} = C_A \delta^{cd} \quad \text{with } C_A = N. \quad (46)$$

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

3 Renormalization, or: “THEORISTS ARE NOT AFRAID OF INFINITIES!”

QCD calculations are extremely demanding. Although perturbative, the size of the coupling constant even at rather large values of the exchanged momentum, Q^2 , is such that the convergence of the perturbative expansion is slow. Several orders of perturbation theory (PT) are required in order to obtain a good accuracy. The complexity of the calculations grows dramatically with the order of the approximation. As an additional complication, the evaluation of a large class of higher-order diagrams gives rise to results which are a priori ill-defined, namely to infinities. A typical example of what is known as an *ultraviolet* divergence, appears when considering the corrections to the quark self-energy. Using the Feynman rules presented in the previous lecture, one can obtain:

$$\begin{array}{c} \text{quark} \\ | \\ \text{gluon} \\ | \\ \text{quark} \end{array} = (-ig)^2 C_F \int \frac{d^4 \ell}{(2\pi)^4} \gamma_\mu \frac{i}{\not{p} + \not{\ell}} \gamma_\nu \left(-\frac{ig^{\mu\nu}}{\ell^2} \right) \equiv i \not{p} \Sigma(p), \quad (47)$$

where simple manipulations lead to the following expression for $\Sigma(p)$:

$$\Sigma(p) = iC_F \int \frac{d^4\ell}{(2\pi)^4} \frac{1}{\ell^2(p+\ell)^2}, \quad (48)$$

which is logarithmically divergent in the ultraviolet ($|\ell| \rightarrow \infty$) region. In this lecture we will discuss how to deal with these infinities. To start with, we study a simple example taken from standard electrostatics.

3.1 The potential of an infinite line of charge

Let us consider a wire of infinite length, carrying a constant charge density λ . By definition, the dimensions of λ are $[\text{length}]^{-1}$. Our goal is to evaluate the electric potential, and eventually the electric field, in a point P at distance R from the wire. There is no need to do any calculation to anticipate that the evaluation of the electric potential will cause some problem. Using the fact that the potential should be linear in the charge density λ , we write $V(R) = \lambda f(R)$. Since the potential itself has the dimensions of $[\text{length}]^{-1}$, we clearly see that there is no room for $f(R)$ to have any non-trivial functional dependence on R . The problem is made explicit if we try to evaluate $V(R)$ using the standard EM formulas:

$$V(R) = \int \frac{\lambda(r)}{r} dx = \lambda \int_{-\infty}^{+\infty} \frac{dx}{\sqrt{R^2 + x^2}} \quad (49)$$

where the integral runs over the position x on the wire. This integral is logarithmically divergent, and the potential is ill-defined. We know however that this is not a serious issue, since the potential itself is not a physical observable, only the electric field is measurable. Since the electric field is obtained by taking the gradient of the scalar potential, it will be proportional to

$$V'(R) \sim \lambda \int_{-\infty}^{+\infty} \frac{dx}{(R^2 + x^2)^{3/2}}, \quad (50)$$

which is perfectly convergent. It is however interesting to explore the possibility of providing a useful operative meaning to the definition of the scalar potential. To do that, we start by *regularizing* the integral in eq. (49). This can be done by introducing the regularized $V(R)$ defined as:

$$V_\Lambda(R) = \int_{-\Lambda}^{\Lambda} \lambda \frac{dx}{\sqrt{R^2 + x^2}} = \lambda \log \left[\frac{\sqrt{\Lambda^2 + R^2} + \Lambda}{\sqrt{\Lambda^2 + R^2} - \Lambda} \right] \quad (51)$$

We can then define the electric field as

$$\vec{E}(R) = \lim_{\Lambda \rightarrow \infty} [-\vec{\nabla} V_\Lambda(R)] \quad (52)$$

It is easy to check that this prescription leads to the right result:

$$\vec{E}(R) = \lim_{\Lambda \rightarrow \infty} \hat{R} \frac{2\lambda}{R} \frac{L}{\sqrt{L^2 + R^2}} \rightarrow \frac{2\lambda}{R} \hat{R} \quad (53)$$

Notice that in this process we had to introduce a new variable Λ with the dimension of a length. This allows us to solve the puzzle first pointed out at the beginning. At the end, however, the dependence of the physical observable (i.e. the electric field) on this extra parameter disappears. Notice also that the object:

$$\delta V = \lim_{\Lambda \rightarrow \infty} [V_\Lambda(r_2) - V_\Lambda(r_1)] = \lambda \log \left(\frac{r_1^2}{r_2^2} \right) \quad (54)$$

is well defined. This suggests a way of defining the potential which is meaningful even in the $\Lambda \rightarrow \infty$ limit. We can *renormalize* the potential, by subtracting $V(R)$ at some fixed value of $R = R_0$, and taking the $\Lambda \rightarrow \infty$ limit:

$$V(R) \rightarrow V(R) - V(R_0) = \lambda \log \left(\frac{R_0^2}{R^2} \right) \quad (55)$$

The non-physical infinities present in $V(R)$ and $V(R_0)$ cancel each other, leaving a finite result, with a non-trivial R -dependence. Once again, this is possible because a dimensionful parameter (in this case R_0) has been introduced.

This example suggests a strategy for dealing with divergencies:

- i) Identify an appropriate way to *regularize* infinite integrals
- ii) Absorb the divergent terms into a redefinition of fields or parameters, e.g, via *subtractions*. This step is usually called *renormalization*.
- iii) Make sure the procedure is *consistent*, by checking that the physical results do not depend on the regularization prescription.

In the rest of this lecture I will explain how this strategy is applied to the case of ultraviolet divergencies encountered in perturbation theory.

3.2 Dimensional regularization

The typical expressions we have to deal with have the form:

$$I(M^2) = \int \frac{d^4 \ell}{(2\pi)^4} \frac{1}{[\ell^2 + M^2]^2} \quad (56)$$

You can easily show that the integral encountered in the quark self-energy diagram can be rewritten as:

$$\frac{1}{\ell^2} \frac{1}{(\ell - p)^2} = \int_0^1 dx \frac{1}{(L^2 + M^2)^2}, \text{ with } L = \ell - xp, M^2 = x(1-x)p^2 \quad (57)$$

The most straightforward extension of the ideas presented above in the case of the infinite charged wire is to regularize the integral using a momentum cutoff, and to renormalize it with a subtraction (for example $I(M^2) - I(M_0^2)$). Experience has shown, however, that the best way to regularize $I(M^2)$ is to take the analytic continuation of the integral in the number of space-time dimensions. In fact

$$I_D(M^2) = \int \frac{d^D \ell}{(2\pi)^D} \frac{1}{(\ell^2 + M^2)^2} \quad (58)$$

is finite $\forall D < 4$. If we could assign a formal meaning to $I_D(M^2)$ for *continuous* values of D away from $D = 4$, we could then perform all our manipulations in $D \neq 4$, regulate the divergences, renormalize fields and couplings, and then go back to $D = 4$.

To proceed, one defines (for Euclidean metrics):

$$d^D \ell = d\Omega_{D-1} \ell^{D-1} d\ell \quad (59)$$

with $d\Omega_{D-1}$ the differential solid angle in D dimensions. Ω_{D-1} is the surface of a D -dimensional sphere. It can be obtained by using the following formal identity:

$$\int d^D \ell e^{-\vec{\ell}^2} \equiv \left[\int d\ell e^{-\ell^2} \right]^D = \pi^{D/2} \quad (60)$$

The integral can also be evaluated, using eq. (59), as

$$\begin{aligned} \int d^D \ell e^{-\vec{\ell}^2} &= \Omega_D \int_0^\infty \ell^{D-1} e^{-\ell^2} d\ell = \Omega_D \frac{1}{2} \int_0^\infty d\ell^2 (\ell^2)^{\frac{D-2}{2}} e^{-\ell^2} \\ &= \Omega_D \frac{1}{2} \int_0^\infty dx e^{-x} x^{\frac{D-2}{2}} \equiv \frac{\Omega_D}{2} \Gamma\left(\frac{D}{2}\right) \end{aligned} \quad (61)$$

Comparing eqs. (60) and (61) we get:

$$I_D(M^2) = \frac{1}{(4\pi)^{D/2}} \frac{1}{\Gamma(D/2)} \int_0^\infty dx x^{\frac{D-2}{2}} (x + M^2)^{-2} = \frac{1}{(4\pi)^{D/2}} \frac{\Gamma(2 - D/2)}{\Gamma(2)} (M^2)^{\frac{D}{2}-2} \quad (62)$$

Defining $D = 4 - 2\epsilon$ (with the understanding that ϵ will be taken to 0 at the end of the day), and using the small- ϵ expansion:

$$\Gamma(\epsilon) = \frac{1}{\epsilon} - \gamma_\epsilon + \mathcal{O}(\epsilon) \quad (63)$$

we finally obtain:

$$(4\pi)^2 I_D(M^2) \rightarrow \frac{1}{\epsilon} - \log 4\pi M^2 - \gamma_\epsilon \quad (64)$$

The divergent part of the integral is then regularized as a pole in $(D-4)$. The M -dependent part of the integral behaves logarithmically, as expected because the integral itself was dimensionless in $D = 4$. The $1/\epsilon$ pole can be removed by a subtraction:

$$I(M^2) = I(\mu^2) + (4\pi)^2 \log\left(\frac{\mu^2}{M^2}\right) \quad (65)$$

where the subtraction scale μ^2 is usually referred to as the ‘‘renormalization scale’’.

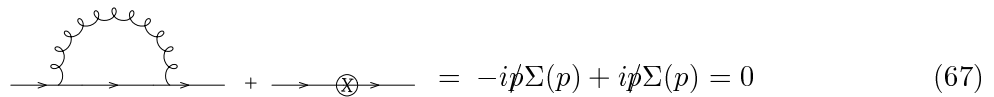
One can prove (and you will find this in the quoted textbooks) that other divergent integrals which appear in other loop diagrams can be regularized in a similar fashion, with the appearance of $1/\epsilon$ poles. Explicit calculations and more details on this technique can be found in the literature quoted at the end.

3.3 Renormalization

Let us come back now to our quark self-energy diagram, eq. (47). After regulating the divergence using dimensional regularization, we can eliminate it by adding a counterterm to the Lagrangian:

$$\mathcal{L} \rightarrow \mathcal{L} + \Sigma(p) \bar{\psi} i \not{\partial} \psi = [1 + \Sigma(p)] \bar{\psi} i \not{\partial} \psi + \dots \quad (66)$$

In this way, the corrections at $O(g^2)$ to the inverse propagator are finite:



$$\text{---} \text{---} \text{---} + \text{---} \text{---} \text{---} = -i\not{p}\Sigma(p) + i\not{p}\Sigma(p) = 0 \quad (67)$$

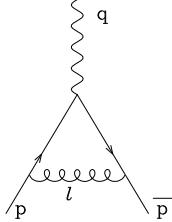
The inclusion of this counterterm can be interpreted as a renormalization of the quark wave function. To see this, it is sufficient to define:

$$\psi_R = \left(1 + \Sigma(p^2)\right)^{1/2} \psi \quad (68)$$

and verify that the kinetic part of the Lagrangian written in terms of ψ_R takes again the canonical form.

It may seem that this regularization/renormalization procedure can always be carried out, with all possible infinities being removed by ad hoc counterterms. This is not true. That these subtractions can be performed consistently for any possible type of divergence which develops in PT is a highly non-trivial fact. To convince you of this, consider the following example.

Let us study the QCD corrections to the interaction of quarks with a photon:



$$\begin{aligned}
&= (-ig)^2 C_F \int \frac{d^4 \ell}{(2\pi)^4} \left[\gamma^p \frac{i}{\not{p} + \not{\ell}} \overbrace{(-i e \gamma^\mu)}^{\Gamma^\mu} \frac{i}{\not{\bar{p}} + \not{\ell}} \gamma^p \right] \left(\frac{-i}{\ell^2} \right) \\
&= -ig^2 C_F \int \frac{d^4 \ell}{(2\pi)^4} (-2)(\not{\bar{p}} + \not{\ell}) \Gamma^\mu (\not{p} + \not{\ell}) \frac{1}{\ell^2 (p + \ell)^2 (\bar{p} + \ell)^2} \\
&\xrightarrow{\text{leading div.}} -ig^2 (-2) C_F \int \frac{d^4 \ell}{(2\pi)^4} \frac{\not{\ell} \Gamma^\mu \not{\ell}}{\ell^2 (p + \ell)^2 (\bar{p} + \ell)^2} \stackrel{\text{def}}{=} ie \gamma^\mu V(q^2)
\end{aligned}$$

It is easily recognized that $V(q^2)$ is divergent. The divergence can be removed by adding a counterterm to the bare Lagrangian:

$$\begin{aligned}
\mathcal{L}_{\text{int}} &= -e A_\mu \bar{\psi} \gamma^\mu \psi \rightarrow -e A_\mu \bar{\psi} \gamma^\mu \psi - e V(q^2) A_\mu \bar{\psi} \gamma^\mu \psi \\
&= -[1 + V(q^2)] e A_\mu \bar{\psi} \gamma^\mu \psi
\end{aligned} \tag{69}$$

If we take into account the counterterm that was introduced to renormalize the quark self-energy, the part of the quark Lagrangian describing the interaction with photons is now:

$$\mathcal{L}_{q,\gamma} = [1 + \Sigma(p^2)] \bar{\psi} i \not{\partial} \psi - [1 + V(q^2)] e A_\mu \bar{\psi} \gamma^\mu \psi \tag{70}$$

Defining a renormalized charge by:

$$e_R = e \frac{1 + V(p^2)}{1 + \Sigma(q^2)} \tag{71}$$

we are left with the renormalized Lagrangian:

$$\mathcal{L}_R = \bar{\psi}_R i \not{\partial} \psi_R + e_R A_\mu \bar{\psi}_R \gamma^\mu \psi_R \tag{72}$$

Can we blindly accept this result, regardless of the values of the counterterms $V(p^2)$ and $\Sigma(q^2)$? The answer to this question is NO! Charge conservation, in fact, requires $e_R = e$. The electric charge carried by a quark cannot be affected by the QCD corrections, and cannot be affected by the renormalization of QCD-induced divergencies. There are many ways to see that if $e_R \neq e$ the electric charge would not be conserved in strong interactions. The simplest way is to consider the process $e^+ \nu_e \rightarrow W^+ \rightarrow u \bar{d}$. The electric charge of the initial state is +1 in units of e . After including QCD corrections (which in the case of the interaction with a W are the same as those for the interaction of quarks with a photon), the charge of the final state is +1 in units of e_R . Unless $e_R = e$, the total electric charge would not be conserved in this process! The non-renormalization of the electric charge in presence of strong interactions is the fact that makes the charge of the proton to equal the sum of the charges of its constituent quarks, in spite of the complex QCD dynamics that holds the quarks together.

As a result, the renormalization procedure is consistent with charge conservation if and only if

$$\frac{V(q^2)}{\Sigma(p^2)} \stackrel{q^2 \rightarrow 0}{=} 1. \tag{73}$$

This identity should hold at all orders of perturbation theory. It represents a fundamental constraint on the consistency of the theory, and shows that the removal of infinities, by itself, is not a trivial trick which can be applied to arbitrary theories. Fortunately, the previous identity can be shown to hold. You can prove it explicitly at the one-loop order by explicitly evaluating the integrals defining $V(q)$ and $\Sigma(p)$.

To carry out the renormalization program for QCD at 1-loop order, several other diagrams in addition to the quark self-energy need to be evaluated. One needs the corrections to the gluon self-energy, to the coupling of a quark pair to a gluon, and to the 3-gluon coupling. Each of these corrections gives rise to infinities, which can be regulated in dimensional regularization. For the purposes of renormalization, it is useful to apply the concept of D dimensions not only to the evaluation of the infinite integrals, but to the full theory as well. In other words we should consider the Lagrangian as describing the interactions of fields in D -dimensions. Nothing changes in its form, but the canonical dimensions of fields and couplings will be shifted. This is because the action (defined as the integral over space-time of the lagrangian), is a dimensionless quantity. As a result, the canonical dimensions of the fields, and of the coupling constants, have to depend on D :

$$\begin{aligned} \left[\int d^D x \mathcal{L}(x) \right] &= 0 \Rightarrow [\mathcal{L}] = D = 4 - 2\epsilon \\ [\partial_\mu \phi \partial^\mu \phi] = D &\Rightarrow [\phi] = 1 - \epsilon \\ [\bar{\psi} \not{\partial} \psi] = D &\Rightarrow [\psi] = 3/2 - \epsilon \\ [\bar{\psi} \not{A} \psi g] = D &\Rightarrow [g] = \epsilon \end{aligned}$$

The gauge coupling constant acquires dimensions! This is a prelude to the non-trivial behaviour of the renormalized coupling constant as a function of the energy scale (“running”). But before we come to this, let us go back to the calculation of the counter-terms and the construction of the renormalized Lagrangian.

Replace the bare fields and couplings with renormalized ones¹:

$$\begin{aligned} \psi_{\text{bare}} &= Z_2^{1/2} \psi_R \\ A_{\text{bare}}^\mu &= Z_3^{1/2} A_R^\mu \\ g_{\text{bare}} &= Z_g \mu^\epsilon g_R \end{aligned}$$

We explicitly extracted the dimensions out of g_{bare} , introducing the dimensional parameter μ (renormalization scale). In this way the renormalized coupling g_R is dimensionless (as it should be once we go back to 4-dimensions).

The Lagrangian, written in terms of renormalized quantities, becomes:

$$\mathcal{L} = Z_2 \bar{\psi} i \not{\partial} \psi - \frac{1}{4} Z_3 F_{\mu\nu}^a F_a^{\mu\nu} + Z_g Z_2 Z_3^{1/2} \mu^\epsilon g \bar{\psi} \not{A} \psi + (\text{gauge fixing, ghosts, } \dots) \quad (74)$$

It is customary to define

$$Z_1 = Z_g Z_2 Z_3^{1/2} \quad (75)$$

If we set $Z_n = 1 + \delta_n$, we then obtain:

$$\begin{aligned} \mathcal{L} &= \bar{\psi} i \not{\partial} \psi - \frac{1}{4} F_{\mu\nu}^a F^{\mu\nu a} + \mu^\epsilon g \bar{\psi} \not{A} \psi + [\text{ghosts, GM}] \\ &+ \delta_2 \bar{\psi} i \not{\partial} \psi - \frac{1}{4} \delta_3 F_{\mu\nu}^a F^{\mu\nu a} + \delta_1 \mu^\epsilon g \bar{\psi} \not{A} \psi \end{aligned} \quad (76)$$

¹For the sake of simplicity, here and in the following we shall assume the quarks as massless. The inclusion of the mass terms does not add any interesting new feature in what follows.

The counter-terms δ_i are fixed by requiring the 1-loop Green functions to be finite. The explicit evaluation, which you can find carried out in detail, for example, in refs. [7,3], gives:

$$\text{quark self-energy} \Rightarrow \delta_2 = -C_F \left(\frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \right) \quad (77)$$

$$\text{gluon self-energy} \Rightarrow \delta_3 = \left(\frac{5}{3}C_A - \frac{4}{3}n_f T_F \right) \frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \quad (78)$$

$$q\bar{q}g \text{ vertex corrections} \Rightarrow \delta_1 = -(C_A + C_F) \frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \quad (79)$$

As usual we introduced the notation $\alpha_s = g^2/4\pi$. The strong-coupling renormalization constant Z_g can be obtained using these results and eq. (75):

$$Z_g = \frac{Z_1}{Z_2 Z_3^{1/2}} = 1 + \delta_1 - \delta_2 - \frac{1}{2}\delta_3 = 1 + \frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \left[-\frac{11}{6}C_A + \frac{2}{3}n_f T_F \right] \stackrel{\text{def}}{=} 1 - \frac{1}{\epsilon} \left(\frac{b_0}{2} \right) \alpha_s \quad (80)$$

Notice the cancellation of the terms proportional to C_F , between the quark self-energy (Z_2) and the abelian part of the vertex correction (Z_1). This is the same as in the case of the QCD non-renormalization of the electric coupling, discussed at the beginning of the lecture. The non-abelian part of the vertex correction contributes viceversa to the QCD coupling renormalization. This is a consequence of gauge invariance. The separation of the non-abelian contributions to the self-energy and to the vertex is not gauge-invariant, only their sum is. Notice also that the consistency of the renormalization procedure requires that the renormalized strong coupling g defining the strength of the interaction of quarks and gluons should be the same as that defining the interaction of gluons among themselves. If this didn't happen, the gauge invariance of the $q\bar{q} \rightarrow gg$ process so painfully achieved in the first lecture by fixing the coefficient of the 3-gluon coupling would not hold anymore at 1-loop! Once again, this additional constraint can be shown to hold through an explicit calculation.

3.4 Running of α_s

The running of α_s is a consequence of the renormalization-scale independence of the renormalization process. The bare coupling g_{bare} knows nothing about our choice of μ . The parameter μ is an artifact of the regularization prescription, introduced to define the dimensionful coupling in D dimensions, and should not enter in measurable quantities. As a result:

$$\frac{dg_{bare}}{d\mu} = 0 \quad (81)$$

Using the definition of g : $g_{bare} = \mu^\epsilon Z_g g$, we then get

$$\epsilon \mu^{2\epsilon} Z_g^2 \alpha_s + \mu^{2\epsilon} \alpha_s 2Z_g \frac{dZ_g}{dt} + \mu^{2\epsilon} Z_g^2 \frac{d\alpha_s}{dt} = 0 \quad (82)$$

where

$$\frac{d}{dt} = \mu^2 \frac{d}{d\mu^2} = \frac{d}{d \log \mu^2} \quad (83)$$

Z_g depends upon μ only via the presence of α_s . If we define

$$\beta(\alpha_s) = \frac{d\alpha_s}{dt} \quad (84)$$

we then get:

$$\beta(\alpha_s) + 2 \frac{\alpha_s}{Z_g} \frac{dZ_g}{d\alpha_s} \beta(\alpha_s) = -\epsilon \alpha_s \quad (85)$$

Using eq. (80) and expanding in powers of α_s , we get:

$$\beta(\alpha_s) = \frac{-\epsilon\alpha_s}{1 + 2\frac{\alpha_s}{Z_g} \frac{dZ_g}{d\alpha_s}} = \frac{-\epsilon\alpha_s}{1 - \frac{b_0\alpha_s}{\epsilon}} = -b_0\alpha_s^2 + O(\alpha_s^2, \epsilon) \quad (86)$$

and finally:

$$\beta(\alpha_s) = -b_0\alpha_s^2 \quad \text{with} \quad b_0 = \frac{1}{2\pi} \left(\frac{11}{6} C_A - \frac{2}{3} n_f T_F \right) \stackrel{N=3}{=} \frac{1}{12\pi} (33 - 2n_f) \quad (87)$$

We can now solve eq. (84), assuming $b_0 > 0$ (which is true provided the number of quark flavours is less than 16) and get the famous *running* of α_s :

$$\alpha_s(\mu^2) = \frac{1}{b_0 \log(\mu^2/\Lambda^2)} \quad (88)$$

The parameter Λ describes the boundary condition of the first order differential equation defining the running of α_s , and corresponds to the scale at which the coupling becomes infinity.

3.5 Renormalization group invariance

The fact that the coupling constant α_s depends on the unphysical renormalization scale μ should not be a source of worry. This is because the coupling constant itself is not an observable. What we observe are decay rates, spectra, or cross sections. These are given by the product of the coupling constant times some matrix element, which in general will acquire a non-trivial renormalization-scale dependence through the renormalization procedure. We therefore just need to check that the scale dependence of the coupling constant and of the matrix elements cancel each other, leaving results which do not depend on μ .

Consider now a physical observable, for example the ratio $R = \sigma(e^+e^- \rightarrow \text{hadrons})/\sigma(e^+e^- \rightarrow \mu^+\mu^-)$. R can be calculated in perturbation theory within QCD, giving rise to an expansion in the renormalized coupling $\alpha_s(\mu)$:

$$R[\alpha_s, s/\mu^2] = 1 + \alpha_s f_1(t) + \alpha_s^2 f_2(t) + \dots = \sum_{n=0}^{\infty} \alpha_s^n f_n(t) \quad (89)$$

where $t = s/\mu^2$ (and we omitted a trivial overall factor $3 \sum_f Q_f^2$). R depends on μ explicitly via the functions $f_n(t)$ and implicitly through α_s . Since R is an observable, it should be independent of μ , and the functions $f_n(t)$ cannot be totally arbitrary. In particular, one should have:

$$\mu^2 \frac{dR}{d\mu^2} = 0 = \left[\mu^2 \frac{\partial}{\partial \mu^2} + \beta(\alpha_s) \frac{\partial}{\partial \alpha_s} \right] R[\alpha_s, s/\mu^2] = 0 \quad (90)$$

Before we give the general, formal solution to this differential equation, it is instructive to work out directly its form within perturbation theory.

$$\mu^2 \frac{dR}{d\mu^2} = 0 = \beta(\alpha_s) f_1(t) + \alpha_s \mu^2 \frac{df_1}{d\mu^2} + 2\alpha_s \beta(\alpha_s) f_2(t) + \alpha_s^2 \mu^2 \frac{df_2}{d\mu^2} + \dots \quad (91)$$

At order α_s (remember that β is of order α_s^2) we get

$$\frac{df_1}{d\mu^2} = 0 \Rightarrow f_1 = \text{constant} \equiv a_1 \quad (92)$$

This is by itself a non-trivial result! It says that the evaluation of R at one-loop is finite, all UV infinities must cancel without charge renormalization. If they didn't cancel, f_1 would depend

explicitly on μ . As we saw at the beginning, this is a consequence of the non-renormalization of the electric charge.

At order α_s^2 we have:

$$\beta(\alpha_s)f_1(t) + \alpha_s^2 \frac{df_2}{d \log \mu^2} = 0 \Rightarrow f_2 = b_0 a_1 \log \frac{\mu^2}{s} + a_2 \text{ (integration constant)} \quad (93)$$

So up to order α_s^2 we have:

$$R = 1 + \underbrace{a_1 \alpha_s}_{\text{one-loop}} + \underbrace{a_1 b_0 \alpha_s^2 \log \mu^2/s + a_2 \alpha_s^2}_{\text{two-loops}} + \dots \quad (94)$$

Notice that the requirement of renormalization group invariance allows us to know the coefficient of the logarithmic term at 2-loops without having to carry out the explicit 2-loop calculation! It is also important to notice that in the limit of high energy, $s \rightarrow \infty$, the logarithmic term of the two-loop contribution becomes very large, and this piece becomes numerically of order α_s as soon as $\log s/\mu^2 \gtrsim 1/b_0 \alpha_s$. You can easily check that renormalization scale invariance requires the presence of such logs at all orders of PT. In particular:

$$f_{(n)}(t) = a_1 \left[b_0 \log \frac{\mu^2}{s} \right]^n + \dots \quad (95)$$

We can collect all these logs as follows:

$$R = 1 + a_1 \alpha_s \left[1 + \alpha_s b_0 \log \frac{\mu^2}{s} + (\alpha_s b_0 \log \frac{\mu^2}{s})^2 + \dots \right] + a_2 \alpha_s^2 + \dots \quad (96)$$

$$= 1 + a_1 \frac{\alpha_s(\mu)}{1 + \alpha_s(\mu) b_0 \log \frac{s}{\mu^2}} + a_2 \alpha_s^2 + \dots \equiv 1 + a_1 \alpha_s(s) + a_2 \alpha_s^2 + \dots \quad (97)$$

In fact:

$$\frac{\alpha_s(\mu)}{1 + \alpha_s(\mu) b_0 \log \frac{s}{\mu^2}} = \frac{1}{b_0 \log \frac{\mu^2}{\Lambda^2} + b_0 \log \frac{s}{\mu^2}} = \frac{1}{b_0 \log \frac{s}{\Lambda^2}} \equiv \alpha_s(s) \quad (98)$$

RG invariance constrains the form of higher-order corrections. All of the higher-order logarithmic terms are determined in terms of lower-order finite coefficients. They can be resummed by simply setting the scale of α_s to s . You can check by yourself that this will work also for the higher-order terms, such as those proportional to a_2 . So the final result has the form:

$$R = 1 + a_1 \alpha_s(s) + a_2 \alpha_s^2(s) + a_3 \alpha_s^3(s) + \dots \quad (99)$$

Of course a_1, a_2, \dots have to be determined by an explicit calculation. However, the truncation of the series at order n has now an accuracy which is truly of order α_s^{n+1} , contrary to before when higher-order terms were as large as lower-order ones. The explicit calculation has been carried out up to the a_3 coefficient. In particular,

$$a_1 = \frac{3}{4} \frac{C_F}{\pi} \equiv \frac{1}{\pi} \quad (100)$$

The formal proof of the previous equation can be obtained by showing that the general form of the equation

$$\left[\mu^2 \frac{\partial}{\partial \mu^2} + \beta(\alpha_s) \right] R(\alpha_s, \frac{s}{\mu^2}) = 0 \quad (101)$$

is given by

$$\left\{ \begin{array}{l} R(\alpha_s(s), 1), \text{ with} \\ \frac{d\alpha_s}{d \log \frac{s}{\mu^2}} = \beta(\alpha_s) \end{array} \right. \quad (102)$$

4 QCD in e^+e^- collisions

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

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

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

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

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

4.1 Soft Gluon Emission

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

$$(103)$$

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

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

$$A_{soft} = g\lambda_{ij}^a \left(\frac{p\cdot\epsilon}{p\cdot k} - \frac{\bar{p}\cdot\epsilon}{\bar{p}\cdot k} \right) A_{Born} \quad (104)$$

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

$$(105)$$

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

$$(106)$$

Example: Consider the “decay” of a virtual gluon into a quark pair. One more diagram should be added to those considered in the case of the electroweak decay. The fact that the quark pair

is not in a colour-singlet state anymore makes things a bit more interesting:

$$(107)$$

$$\begin{aligned} &\stackrel{k \rightarrow 0}{=} \left[i g f^{abc} \lambda_{ij}^c \left(\frac{Q\epsilon}{Qk} \right) + g (\lambda^b \lambda^a)_{ij} \left(\frac{p\epsilon}{pk} \right) - g (\lambda^a \lambda^b)_{ij} \left(\frac{\bar{p}\epsilon}{pk} \right) \right] A_{Born} \\ &= g (\lambda^a \lambda^b)_{ij} \left[\frac{Q\epsilon}{Qk} - \frac{\bar{p}\epsilon}{pk} \right] + g (\lambda^b \lambda^a)_{ij} \left[\frac{p\epsilon}{pk} - \frac{Q\epsilon}{Qk} \right] \end{aligned} \quad (108)$$

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

$$(109)$$

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

$$\sum_{a,b,i,j} |(\lambda^a \lambda^b)_{ij}|^2 = \sum_{a,b} \text{tr} (\lambda^a \lambda^b \lambda^b \lambda^a) = \frac{N^2 - 1}{2} C_F = \mathcal{O}(N^3) \quad (110)$$

$$\sum_{a,b,i,j} (\lambda^a \lambda^b)_{ij} [(\lambda^b \lambda^a)_{ij}]^* = \sum_{a,b} \text{tr}(\lambda^a \lambda^b \lambda^a \lambda^b) = \frac{N^2 - 1}{2} \underbrace{\left(C_F - \frac{C_A}{2} \right)}_{-\frac{1}{2N}} = \mathcal{O}(N) \quad (111)$$

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

4.2 Angular ordering for soft-gluon emission

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

Let us start first by proving the property of colour ordering. Consider the $q\bar{q}$ pair produced by the decay of a rapidly moving virtual photon. The amplitude for the emission of a soft gluon was given in eq. (104). Squaring, summing over colours and including the gluon phase-space we get the following result:

$$\begin{aligned}
d\sigma_g &= \sum |A_{soft}|^2 \frac{d^3k}{(2\pi)^3 2k^0} \sum |A_0|^2 \frac{-2p^\mu \bar{p}^\nu}{(pk)(\bar{p}k)} g^2 \sum \epsilon_\mu \epsilon_\nu^* \frac{d^3k}{(2\pi)^3 2k^0} \\
&= d\sigma_0 \frac{2(p\bar{p})}{(pk)(\bar{p}k)} g^2 C_f \left(\frac{d\phi}{2\pi} \right) \frac{k^0 dk^0}{8\pi^2} d \cos \theta \\
&= d\sigma_0 \frac{\alpha_s C_F}{\pi} \frac{dk^0}{k^0} \frac{d\phi}{2\pi} \frac{1 - \cos \theta_{ij}}{(1 - \cos \theta_{ik})(1 - \cos \theta_{jk})} d \cos \theta
\end{aligned} \tag{112}$$

where $\theta_{\alpha\beta} = \theta_\alpha - \theta_\beta$, and i, j, k refer to the q, \bar{q} and gluon directions, respectively. We can write the following identity:

$$\frac{1 - \cos \theta_{ij}}{(1 - \cos \theta_{ik})(1 - \cos \theta_{jk})} = \frac{1}{2} \left[\frac{\cos \theta_{jk} - \cos \theta_{ij}}{(1 - \cos \theta_{ik})(1 - \cos \theta_{jk})} + \frac{1}{1 - \cos \theta_{ik}} \right] + \frac{1}{2} [i \leftrightarrow j] \equiv W_{(i)} + W_{(j)} \tag{113}$$

We would like to interpret the two functions $W_{(i)}$ and $W_{(j)}$ as radiation probabilities from the quark and antiquark lines. Each of them is in fact only singular in the limit of gluon emission parallel to the respective quark:

$$W_{(i)} \rightarrow \text{finite if } k \parallel j \text{ (} \cos \theta_{jk} \rightarrow 1 \text{)} \tag{114}$$

$$W_{(j)} \rightarrow \text{finite if } k \parallel i \text{ (} \cos \theta_{ik} \rightarrow 1 \text{)} \tag{115}$$

The interpretation as probabilities is however limited by the fact that neither $W_{(i)}$ nor $W_{(j)}$ are positive definite. However, you can easily prove that

$$\int \frac{d\phi}{2\pi} W_{(i)} = \begin{cases} \frac{1}{1 - \cos \theta_{ik}} & \text{if } \theta_{ik} < \theta_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{116}$$

where the integral is the azimuthal average around the q direction. A similar result holds for $W_{(j)}$:

$$\int \frac{d\phi}{2\pi} W_{(j)} = \begin{cases} \frac{1}{1 - \cos \theta_{jk}} & \text{if } \theta_{jk} < \theta_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{117}$$

As a result, the emission of soft gluons outside the two cones obtained by rotating the antiquark direction around the quark's, and viceversa, averages to 0. Inside the two cones, one can consider the radiation from the emitters as being uncorrelated. In other words, the two colour lines defined by the quark and antiquark currents act as independent emitters, and the quantum coherence (i.e. the effects of interference between the two graphs contributing to the gluon-emission amplitude) is accounted for by constraining the emission to take place within those fixed cones.

If one repeats now the exercise for emission of one additional gluon, one will find the same angular constraint, but this time applied to the colour lines defined by the previously established *antenna*. As shown in the previous subsection, the $q\bar{q}g$ state can be decomposed at the leading order in $1/N$ into two independent emitters, one given by the colour line flowing from the gluon to the quark, the other given by the colour line flowing from the antiquark to the gluon. So the emission of the additional gluon will be constrained to take place either within the cone formed by the quark and the gluon, or within the cone formed by the gluon and the antiquark. Either

way, the emission angle will be smaller than the angle of the first gluon emission. This leads to the concept of angular ordering, with successive emission of soft gluons taking place within cones which get smaller and smaller.

The fact that colour always flows directly from the emitting parton to the emitted one, the collimation of the jet, and the softening of the radiation emitted at later stages, ensure that partons forming a colour-singlet cluster are close in phase-space. As a result, hadronization (the non-perturbative process that will bind together colour-singlet parton pairs) takes place locally inside the jet and is not a collective process: only pairs of nearby partons are involved. The inclusive properties of jets (e.g. the particle multiplicity, jet mass, jet broadening, etc.) are independent of the hadronization model, up to corrections of order $(\Lambda/\sqrt{s})^n$ (for some integer power n , which depends on the observable), with $\Lambda \lesssim 1$ GeV.

4.3 Jet rates

We now present explicit calculations of interesting observables. For simplicity, we will work with the soft-gluon approximation for the matrix elements and the phase-space. As a result, the correction to the differential $e^+e^- \rightarrow q\bar{q}$ cross-section from one-gluon emission becomes:

$$d\sigma_g = \sigma_0 \frac{2\alpha_s}{\pi} C_F \frac{dk_0}{k_0} \frac{d\cos\theta}{1 - \cos^2\theta} \quad , \quad \text{with } \sigma_0 = \text{Born amplitude.} \quad (118)$$

In this equation we used the fact that in the soft- g limit the q and \bar{q} are back-to-back, and

$$q \cdot \bar{q} = 2q_0\bar{q}_0 \quad , \quad q \cdot k = q_0k_0(1 - \cos\theta), \quad \bar{q}k = \bar{q}_0k_0(1 + \cos\theta) \quad (119)$$

Notice the presence in $d\sigma_g$ of soft and collinear singularities. They will have to cancel in the total cross-section which, as we saw in the previous lecture, is finite. They do indeed cancel against the contribution to the total cross-section coming from the virtual correction diagram, where a gluon is exchanged between the two quarks. In the total cross-section (and for other sufficiently inclusive observables) the final states produced by the virtual diagrams and by the real emission diagrams in the soft or collinear limit are the same, and both contribute. In order for the total cross-section to be finite, the virtual contribution will need to take the following form:

$$\frac{d^2\sigma_v}{dk_0 d\cos\theta} = -\sigma_0 \frac{2\alpha_s}{\pi} C_F \int_0^{\sqrt{s/2}} \frac{dk'_0}{k'_0} \int_{-1}^1 \frac{d\cos\theta'}{1 - \cos^2\theta'} \times \frac{1}{2} \delta(k_0) [\delta(1 - \cos\theta) + \delta(1 + \cos\theta)] \quad (120)$$

plus finite corrections. In this way:

$$\int_0^{\sqrt{s/2}} dk_0 \int_{-1}^1 d\cos\theta \left[\frac{d^2\sigma_g}{dk_0 d\cos\theta} + \frac{d^2\sigma_v}{dk_0 d\cos\theta} \right] = \text{finite} \quad (121)$$

With the form of the virtual corrections available (at least in this simplified soft-gluon-dominated approximation) we can proceed and calculate other quantities.

Jets are usually defined as clusters of particles close-by in phase-space. A typical jet definition distributes particles in sets of invariant mass smaller than a given parameter M , requiring that one particle only belongs to one jet, and that no other particles (or jets) can be added to a given jet without its mass exceeding M . In the case of a three-particle final state, such as the one we are studying, we get three-jet events if $(q+k)^2$, $(\bar{q}+k)^2$ and $(q+\bar{q})^2$ are all larger than M^2 . We will have two-jet events when at least one of these quantities gets smaller than M^2 . For example emission of a gluon near the direction of the quark, with

$2qk = 2q^0k^0(1 - \cos \theta) < M^2$, defines a two-jet event, one jet being given by the \bar{q} , the other by the system $q + k$.

One usually introduces the parameter $y = M^2/s$, and studies the jet multiplicity as a function of y . Let us calculate the two- and three-jet rates at order α_s . The phase-space domain for two-jet events is given by two regions. The first one is defined by $2qk = 2q_0k_0(1 - \cos \theta) < ys$. This region consists of two parts:

$$(I)_a : \left\{ \begin{array}{l} k_0 < y\sqrt{s} \\ 0 < \cos \theta < 1 \end{array} \right. \oplus (I)_b : \left\{ \begin{array}{l} k_0 > y\sqrt{s} \\ 1 - \frac{y\sqrt{s}}{k_0} < \cos \theta < 1 \end{array} \right. \quad (122)$$

(I)_a corresponds to soft gluons at all angles smaller than $\pi/2$ (i.e. in the quark emisphere), while (I)_b corresponds to hard gluons emitted at small angles from the quark.

The second region, (II), is analogous to (I), but the angles are now referred to the direction of the antiquark. The integrals of $d\sigma$ over (I) and (II) are of course the same. The $\mathcal{O}(\alpha_s)$ contribution to the two-jet rate is therefore given by:

$$\begin{aligned} \frac{\sigma_{2\text{-jet}}^{(\alpha_s)}}{\sigma_0} &= \frac{1}{\sigma_0} \left[2 \int_{(I)_a} d\sigma_g + 2 \int_{(I)_b} d\sigma_g + \int_{\text{virtual}} d\sigma_v \right] \\ &= \frac{4\alpha_s C_F}{\pi} \left[\int_0^{y\sqrt{s}} \frac{dk_0}{k_0} \int_0^1 \frac{d\cos \theta}{1 - \cos^2 \theta} + \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0} \int_{1 - (\frac{y\sqrt{s}}{k_0})}^1 \frac{d\cos \theta}{1 - \cos^2 \theta} \right. \\ &\quad \left. - \int_0^{y\sqrt{s}} \frac{dk_0}{k_0} \int_0^1 \frac{d\cos \theta}{1 - \cos^2 \theta} \right] \\ &= \frac{4\alpha_s C_F}{\pi} \left\{ - \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0} \int_0^1 \frac{d\cos \theta}{1 - \cos^2 \theta} + \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0} \int_{1 - (\frac{y\sqrt{s}}{k_0})}^1 \frac{d\cos \theta}{1 - \cos^2 \theta} \right\} \\ &= \frac{4\alpha_s C_F}{\pi} \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0} \int_0^{1 - (\frac{y\sqrt{s}}{k_0})} \left(\frac{d\cos \theta}{1 - \cos^2 \theta} \right) \\ &= \frac{2\alpha_s C_F}{\pi} \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0} \left[(-) \log \frac{k_0}{y\sqrt{s}} + (\text{finite for } y \rightarrow 0) \right] = -\frac{\alpha_s C_F}{\pi} \log^2 2y \quad (123) \end{aligned}$$

Including the Born contribution, which always gives rise to two and only two jets, we finally have:

$$\begin{aligned} \sigma_{2\text{-jet}} &= \sigma_0 \left[1 - \frac{\alpha_s C_F}{\pi} \log^2 y + \dots \right] \\ \sigma_{3\text{-jet}} &= \sigma_0 \frac{\alpha_s C_F}{\pi} \log^2 y + \dots \end{aligned}$$

If $y \rightarrow 0$, $\sigma_{3\text{-jet}}$ becomes larger than $\sigma_{2\text{-jet}}$. If y is sufficiently small, we can even get $\sigma_{2\text{-jet}} < 0!$ This is a sign that higher-order corrections become important. In the soft-gluon limit, assuming that the emission of a second gluon will also factorize², we can repeat the calculation at higher orders and obtain:

$$\begin{aligned} \sigma_{2\text{-jet}} &\simeq \sigma_0 \left[1 - \frac{\alpha_s C_F}{\pi} \log^2 y + \frac{1}{2!} \left(\frac{\alpha_s C_F}{\pi} \log^2 y \right)^2 + \dots \right] = \sigma_0 e^{-\frac{\alpha_s C_F}{\pi} \log^2 y} \\ \sigma_{3\text{-jet}} &\sim \sigma_0 \frac{\alpha_s C_F}{\pi} \log^2 y e^{-\frac{\alpha_s C_F}{\pi} \log^2 y} \\ &\vdots \\ \sigma_{(n+2)\text{-jet}} &\sim \sigma_0 \frac{1}{n!} \left(\frac{\alpha_s C_F}{\pi} \log^2 y \right)^n e^{-\frac{\alpha_s C_F}{\pi} \log^2 y} \quad (124) \end{aligned}$$

²This is not true (see later on), but let us just accept it to see how things develop.

It is immediate to recognize in this series a Poisson distribution, leading to an average number of jets given by:

$$\langle n_{jet} \rangle \simeq 2 + \frac{\alpha_s C_F}{\pi} \log^2 y \quad (125)$$

The *smaller* the resolution parameter y , the *smaller* the mass of the jets, the *larger* the importance of higher order corrections. If we take the parameter M down to the scale of few hundred MeV ($M \sim \Lambda_{QCD}$), each particle gets identified with an independent jet. We can therefore estimate the s -dependence of the average multiplicity of particles produced:

$$\langle n_{part} \rangle \sim \frac{C_F \alpha_s}{\pi} \log^2 \frac{s}{\Lambda^2} = \frac{C_F}{\pi} \frac{1}{b_0 \log \frac{s}{\Lambda^2}} \log^2 \frac{s}{\Lambda^2} \simeq \frac{C_F}{\pi b_0} \log \frac{s}{\Lambda^2} \quad (126)$$

The final state particle multiplicity grows with $\log(s)$.

In practice, things are a bit more complicated than this. Once the first gluon is emitted, additional gluons can be emitted from it as well. Therefore the final-state multiplicity will be dominated by the emission of gluons from gluons. The analysis becomes more complicated (see e.g. refs. [6,8] for the details), and the final result is:

$$\langle n_{part}(s) \rangle \sim \exp \sqrt{\frac{2C_A}{\pi b} \log\left(\frac{s}{\Lambda^2}\right)} \quad (127)$$

for the particle multiplicity, and

$$\langle n_{jet}(y) \rangle = 2 + 2 \frac{C_F}{C_A} \left(\cosh \sqrt{\frac{\alpha_s C_A}{2\pi} \log^2 \frac{1}{y}} - 1 \right) \sim \frac{C_F}{C_A} \exp \sqrt{\frac{\alpha_s C_A}{2\pi} \log^2 \frac{1}{y}} \quad (128)$$

for the average jet multiplicity.

Other interesting quantities that can be calculated using the simple formulas we developed so far are the average jet mass and the thrust. To define the jet mass we just divide the final state into two emispheres, separated by the plane orthogonal to the thrust axis. We now call jets the two sets of particles on either side of the plane. The $\langle m^2 \rangle$ of the jet is then given by

$$\langle m_{jet}^2 \rangle = \frac{1}{2\sigma_0} \left\{ \int_{(I)} (q+k)^2 d\sigma_g + \int_{(II)} (\bar{q}+k)^2 d\sigma_g \right\} \quad (129)$$

The virtual correction does not enter here, since the pure $q\bar{q}$ final state has jet masses equal to 0. The result of this simple computation leads to

$$\langle m_{jet}^2 \rangle = \frac{\alpha_s C_F}{\pi} s \quad (130)$$

Another interesting variable often used in experimental studies is the thrust T , defined by:

$$T = \max_{\hat{T}} \sum_i |\vec{p}_i \cdot \hat{T}| / \sum_i |\vec{p}_i|$$

where \hat{T} is the thrust axis, defined so as to maximize T . For three-body final states, \hat{T} is the direction of the highest-energy parton, and T is proportional to twice its energy:

$$T = 2 \frac{\bar{q}_0}{\sqrt{s}} \equiv 1 - \frac{(q+k)^2}{s} = 1 - \frac{m_{jet}^2}{s} \quad (131)$$

As a result:

$$\langle 1 - T \rangle = \frac{\alpha_s C_F}{\pi} \quad (132)$$

At LEP, $\langle 1 - T \rangle \simeq \frac{0.120}{\pi} \times \frac{4}{3} \simeq 0.05$. The terms neglected in the soft-gluon approximation we used throughout can be calculated, and give some small correction to the above results. Corrections will likewise come from higher-order effects. State-of-the-art calculations exist which evaluate all these “shape variables” (and more!) up to $\mathcal{O}(\alpha_s^2)$ accuracy, including a full next-to-leading-log accurate resummation of higher-order logarithms (such as the $\log 1/y$ terms we encountered in the discussion of jet rates, or terms of the form $\log^n(1 - T)$ which appear at higher orders in the evaluation of the thrust distributions). These calculations allow a reliable estimate of several different observables directly proportional to α_s , and provide the theoretical input for the extraction of α_s from the LEP QCD data [8].

Notice that non-perturbative corrections proportional to $\frac{\Lambda}{\sqrt{s}}$, with $\Lambda \sim 1$ GeV, can have a significant impact on the extraction of α_s . For example, a $\frac{\Lambda}{\sqrt{s}}$ correction to $\langle 1 - T \rangle$ would be a 20 % effect:

$$\frac{\Lambda}{\sqrt{s}} \sim 0.01, \quad \langle 1 - T \rangle_{PT} \simeq 0.05$$

Indeed one measures $\langle 1 - T \rangle_{LEP} = 0.068 \pm 0.003$, vs the full PT QCD prediction of 0.055 (using $\alpha_s = 0.120$).

5 QCD and the proton structure at large Q^2

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

The idea that the parton language [1] and the use of perturbative QCD can be used to describe the structure of the proton at short distances was developed in the late 60’s and early 70’s (for a nice review, see ref. [17]). While I will not provide you with a rigorous proof of the legitimacy of this approach, I will try to justify it qualitatively to make it sound at least plausible. I will then proceed to extract some results based on the application of perturbative QCD to lepton-hadron interactions.

5.1 The parton model

We all know that quarks are deeply bound inside the proton. It is important to realise, however, that the binding forces responsible for the quark confinement are due to the exchange of rather soft gluons. If a quark were to exchange a hard virtual gluon with another quark, in fact, the recoil would tend to break the proton apart. It is easy to verify that the exchange of gluons with virtuality larger than Q is then proportional to some large power of m_p/Q , m_p being the proton mass. Since the gluon coupling constant gets smaller at large Q , exchange of hard gluons is significantly suppressed³. As a result, the typical time scale for quarks inside the proton

³The fact that the coupling decreases at large Q plays a fundamental role in this argument. Were this not true, the parton picture could not be used!

to interact among themselves is of the order of $1/m_p$, or longer. If we probe the proton with an off-shell photon, the interaction should take place during the limited lifetime of the virtual photon, given by the inverse of its virtuality as a result of the Heisenberg principle. Once the photon gets “inside” the proton and meets a quark, the struck quark has no time to negotiate a coherent response with the other quarks, because the time scale for it to “talk” to its pals is too long compared with the duration of the interaction with the photon itself. As a result, the struck quark has no option but to interact with the photon as if it were a free particle.

The one thing that the above picture does not tell us, obviously, is in which precise state the quark was once it got struck by the photon. This depends on the internal wave function of the proton, which perturbative QCD cannot easily predict. We can however say that the wave function of the proton, and therefore the state of the “free” quark, are determined by the dynamics of the soft-gluon exchanges inside the proton itself. Since the time scale of this dynamics is long relative to the time scale of the photon-quark interaction, we can safely argue that the photon sees to good approximation a static snapshot of the proton’s inner guts. In other words, the state of the quark had been prepared long before the photon arrived. This also suggests that the state of the quark will not depend on the precise nature of the external probe, provided the time scale of the hard interaction is very short compared to the time it would take for the quark to readjust itself. As a result, if we could perform some measurement of the quark state using, say, a virtual-photon probe, we could then use this knowledge on the state of the quark to perform predictions for the interaction of the proton with any other probe (e.g. a virtual W or even a gluon from an opposite beam of hadrons).

In order to make the measurement of the proton structure as simple as possible, it is therefore wise to use a probe as simple as possible. A virtual photon emitted from a beam of high-energy electrons provides such a probe. The relative process is called deeply inelastic scattering (DIS), and was historically the first phenomenon which led people to introduce the concept of partons [2].

Assuming the parton picture outlined above, we can describe the cross-section for the interaction of the virtual photon with the proton as follows:

$$\sigma_0 = \int_0^1 dx \sum_i e_i^2 f_i(x) \hat{\sigma}_0(\gamma^* q_i \rightarrow q'_i, x) \quad (133)$$

where the 0 subscript anticipates that this description represents a leading order approximation. In the above equation, $f_i(x)$ represents the density of quarks of flavour i carrying a fraction x of the proton momentum. The hatted cross-section represents the interaction between the photon and a free (massless) quark:

$$\begin{aligned} \hat{\sigma}_0(\gamma^* q_i \rightarrow q'_i) &= \frac{1}{flux} \overline{\sum} |M_0(\gamma^* q \rightarrow q')|^2 \frac{d^3 p'}{(2\pi)^3 2p'_0} (2\pi)^4 \delta^4(p' - q - p) \\ &= \frac{1}{flux} \overline{\sum} |M_0|^2 2\pi \delta(p'^2) \end{aligned} \quad (134)$$

Using $p' = xP + q$, where P is the proton momentum, we get

$$(p')^2 = 2xP \cdot q + q^2 \equiv 2xP \cdot q - Q^2 \quad (135)$$

$$\hat{\sigma}_0(\gamma^* q \rightarrow q') = \frac{2\pi}{flux} \overline{\sum} |M_0|^2 \frac{1}{2P \cdot q} \delta(x - x_{bj}) \quad (136)$$

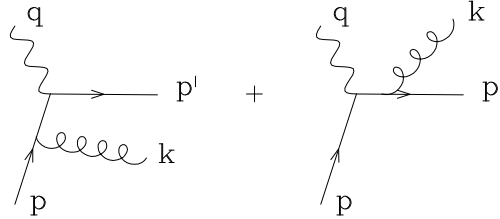
where $x_{bj} = \frac{Q^2}{2P \cdot q}$ is the so-called Bjorken- x variable. Finally:

$$\sigma_0 = \frac{2\pi}{flux} \frac{\overline{\sum} |M_0|^2}{Q^2} \sum_i x_{bj} f_i(x_{bj}) e_i^2 \equiv \frac{2\pi}{flux} \frac{\overline{\sum} |M_0|^2}{Q^2} F_2(x_{bj}) \quad (137)$$

The measurement of the inclusive ep cross-section as a function of Q^2 and $P \cdot q$ ($= m_p(E' - E)$ in the proton rest frame, with E' = energy of final-state lepton and E = energy of initial-state lepton) probes the quark momentum distribution inside the proton.

5.2 Parton evolution

Let us now study the QCD corrections to the LO parton-model description of DIS. This study will exhibit many important aspects of QCD (structure of collinear singularities, renormalization-group invariance) and will take us to an important element of the DIS phenomenology, namely scaling violations. We start from real-emission corrections to the Born level process:



$$(138)$$

The first diagram is proportional to $1/(p - k)^2 = 1/2(pk)$, which diverges when k is emitted parallel to p :

$$p \cdot k = p^0 k^0 (1 - \cos \theta) \xrightarrow{\cos \theta \rightarrow 1} 0 \quad (139)$$

The second diagram is also divergent, if k is emitted parallel to p' . This second divergence turns out to be harmless, since we are summing over all possible final states. Whether the final-state quark keeps all of its energy, or whether it decides to share it with a gluon emitted collinearly, an inclusive final-state measurement will not care. The collinear divergence can then be cancelled by a similar divergence appearing in the final-state quark self-energy corrections.

The first divergence is more serious, since from the point of view of the incoming photon (which only sees the quark, not the gluon) it *does* make a difference whether the momentum is all carried by the quark or is shared between the quark and the gluon. This means that no cancellation between collinear singularities in the real emission and virtual emission is possible. So let us go ahead, calculate explicitly the contribution of these diagrams, and learn how to deal with their singularities.

First of all note that while the second diagram is not singular in the region $k \cdot p \rightarrow 0$, its interference with the first one is. It is possible, however, to select a gauge for which the interference of the two diagrams is finite in this limit. You can show that the right choice is

$$\sum \epsilon_\mu \epsilon_\nu^*(k) = -g_{\mu\nu} + \frac{k^\mu p'^\nu + k^\nu p'^\mu}{k \cdot p'} \quad (140)$$

Notice that in this gauge not only $k \cdot \epsilon(k) = 0$, but also $p' \cdot \epsilon(k) = 0$. The key to getting to the end of a QCD calculation in a finite amount of time is choosing a proper gauge (which we just did) and the proper parametrization of the momenta involved. In our case, since we are interested in isolating the region where k becomes parallel with p , it is useful to set

$$k_\mu = (1 - z)p_\mu + \beta p'_\mu + (k_\perp)_\mu, \quad (141)$$

with $k_\perp \cdot p = k_\perp \cdot p' = 0$. β is obtained by imposing

$$k^2 = 0 = 2\beta(1 - z)p \cdot p' + k_\perp^2 \quad (142)$$

Defining $k_{\perp}^2 = -k_t^2$, we then get

$$\beta = \frac{k_t^2}{2(pp')(1-z)} \quad (143)$$

$$k_{\mu} = (1-x)p_{\mu} + \frac{k_t^2}{2(1-x)p \cdot p'} p'_{\mu} + (k_{\perp})_{\mu} \quad (144)$$

$(k_{\perp})_{\mu}$ is therefore the gluon momentum vector transverse to the incoming quark, in a frame where γ^* and q are aligned. k_t is the value of this transverse momentum. We also get

$$k \cdot p = \beta p \cdot p' = \frac{k_t^2}{2(1-z)} \quad \text{and} \quad k \cdot p' = (1-z)p \cdot p' \quad (145)$$

As a result $(p-k)^2 = -k_t^2/(1-z)$. The amplitude for the only diagram carrying the initial-state singularity is:

$$M_g = ig\lambda_{ij}^a \bar{u}(p')\Gamma \frac{\hat{p} - \hat{k}}{(p-k)^2} \hat{\epsilon}(k)u(p) \quad (146)$$

(where we introduced the notation $\hat{a} \equiv \not{a} \equiv a_{\mu}\gamma^{\mu}$). We indicated by Γ the interaction vertex with the external current q . It is important to keep Γ arbitrary, because we would like to get results which do not depend on the details of the interaction with the external probe. It is important that the singular part of the QCD correction, and therefore its renormalization, be process independent. Only in this way we can hope to achieve a true universality of the parton densities! So we will keep Γ generic, and make sure that our algebra does not depend on its form, at least in the $p \cdot k \rightarrow 0$ limit. Squaring the most singular part of the amplitude, and summing over colours and spins, we get:

$$\sum_{\substack{\text{g polariz.} \\ \text{and colours}}} |M_g|^2 = g^2 \overbrace{\sum_a^{N \times C_F} \text{tr}(\lambda^a \lambda^a)} \times \frac{1}{t^2} \times \sum_{\epsilon} \text{Tr}[\hat{p}'\Gamma(\hat{p} - \hat{k})\hat{\epsilon} p \hat{\epsilon}^* (\hat{p} - \hat{k})\Gamma^+] \quad (147)$$

with $t = (p-k)^2 = -k_t^2/(1-z)$. Let us look first at

$$\sum_{\epsilon} \hat{\epsilon} \hat{p} \hat{\epsilon}^* = \sum_{\epsilon} \epsilon_{\mu} \epsilon_{\nu}^* \gamma^{\mu} \hat{p} \gamma^{\nu} = -\gamma^{\mu} \hat{p} \gamma^{\mu} + \frac{1}{k \cdot p'} (\hat{p}' \hat{p} \hat{k} + \hat{k} \hat{p} \hat{p}') = \frac{2}{1-z} (\hat{k} + \beta \hat{p}') \quad (148)$$

(we used: $\hat{a}\hat{b}\hat{c} + \hat{c}\hat{b}\hat{a} = 2(a \cdot b)\hat{c} - 2(a \cdot c)\hat{b} + 2(b \cdot c)\hat{a}$ and some of the kinematical relations from the previous page). Then take

$$(\hat{p} - \hat{k})(\hat{k} + \beta \hat{p}')(\hat{p} - \hat{k}) = (\hat{p} - \hat{k})\hat{k}(\hat{p} - \hat{k}) + \beta(\hat{p} - \hat{k})\hat{p}'(\hat{p} - \hat{k}) \quad (149)$$

In the second term, proportional to β , we can approximate $\hat{k} = (1-z)\hat{p}$. This is because the other pieces $(\beta \hat{p}' + \hat{k}_{\perp})$ multiplied by β would cancel entirely the $\frac{1}{t^2}$ singularity, and would only contribute a non-singular term, which we are currently neglecting. So Eq. (149) becomes

$$\hat{p}\hat{k}\hat{p} + \beta z^2 \hat{p}\hat{p}'\hat{p} = 2(p \cdot k)\hat{p} + \beta z^2 2(p \cdot p')\hat{p} = 2(p \cdot k)(1+z^2)\hat{p} \quad (150)$$

and

$$\sum |M_g|^2 = 2g^2 C_F \frac{(1-z)}{k_t^2} \left(\frac{1+z^2}{1-z} \right) N \text{Tr}[\hat{p}'\Gamma\hat{p}\Gamma^+] \quad (151)$$

The last factor with the trace corresponds to the Born amplitude squared. So the one-gluon emission process factorizes in the collinear limit into the Born process times a factor which is independent of the beam's nature! If we add the gluon phase-space:

$$[dk] \equiv \frac{d^3k}{(2\pi)^3 2k^0} = \frac{dk_{\parallel}}{k^0} \frac{d\phi}{2\pi} \frac{1}{8\pi^2} \frac{dk_{\perp}^2}{2} = \frac{dz}{(1-z)} \frac{1}{16\pi^2} dk_{\perp}^2 \quad (152)$$

we get:

$$\overline{\sum} |M_g|^2 [dk] = \frac{dk_{\perp}^2}{k_{\perp}^2} dz \left(\frac{\alpha_s}{2\pi} \right) P_{qq}(z) \overline{\sum} |M_0|^2 \quad (153)$$

where

$$P_{qq}(z) = C_F \frac{1+z^2}{1-z} \quad (154)$$

is the so-called Altarelli-Parisi splitting function for the $q \rightarrow q$ transition (z is the momentum fraction of the original quark taken away by the quark after gluon emission). We are now ready to calculate the corrections to the parton-model cross-section:

$$\sigma_g = \int dx f(x) \frac{1}{flux} \int dz \frac{dk_{\perp}^2}{k_{\perp}^2} \left(\frac{\alpha_s}{2\pi} \right) P_{qq}(z) \overline{\sum} |M_0|^2 2\pi \delta(p'^2) \quad (155)$$

Using $(p')^2 = (p - k + q)^2 \sim (zp + q)^2 = (xzP + q)^2$ and

$$\delta(p'^2) = \frac{1}{2P \cdot q} \frac{1}{z} \delta(x - \frac{x_{bj}}{z}) = \frac{x_{bj}}{z} \delta(x - \frac{x_{bj}}{z}) \quad (156)$$

we finally obtain:

$$\sigma_g = \frac{2\pi}{flux} \left(\frac{\overline{\sum} |M_0|^2}{Q^2} \right) \sum_i e_i^2 x_{bj} \frac{\alpha_s}{2\pi} \int \frac{dk_{\perp}^2}{k_{\perp}^2} \int \frac{dz}{z} P_{qq}(z) f_i \left(\frac{x_{bj}}{z} \right) \quad (157)$$

We then find that the inclusion of the $\mathcal{O}(\alpha_s)$ correction is equivalent to a contribution to the parton density:

$$f_i(x) \rightarrow f_i(x) + \frac{\alpha_s}{2\pi} \int \frac{dk_{\perp}^2}{k_{\perp}^2} \int_x^1 \frac{dz}{z} P_{qq}(z) f_i \left(\frac{x}{z} \right) \quad (158)$$

Notice the presence of the integral $\int dk_{\perp}^2/k_{\perp}^2$. The upper limit of integration is proportional to Q^2 . The lower limit is 0. Had we included a quark mass, the propagator would have behaved like $1/(k_{\perp}^2 + m^2)$. But the quark is bound inside the hadron, so we do not quite know what m should be. Let us then assume that we cutoff the integral at a k_{\perp} value equal to some scale μ_0 , and see what happens. The effective parton density becomes:

$$f(x, Q^2) = f(x) + \log \left(\frac{Q^2}{\mu_0^2} \right) \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} P_{qq}(z) f \left(\frac{x}{z} \right) \quad (159)$$

The dependence on the scale μ_0 , which is a non-perturbative scale, can be removed by defining $f(x, Q^2)$ in terms of the parton density f measured at a large, perturbative scale μ^2 :

$$f(x, \mu^2) = f(x) + \log \left(\frac{\mu^2}{\mu_0^2} \right) \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} P_{qq}(z) f \left(\frac{x}{z} \right) \quad (160)$$

We can then perform a subtraction, and write:

$$f(x, Q^2) = f(x, \mu^2) + \log \left(\frac{Q^2}{\mu^2} \right) \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} P_{qq}(z) f \left(\frac{x}{z} \right) \quad (161)$$

The scale μ plays here a similar role to the renormalization scale introduced in the second lecture. Its choice is arbitrary, and $f(x, Q^2)$ should not depend on it. Requiring this independence, we get the following ‘‘renormalization-group invariance’’ condition:

$$\frac{df(x, Q^2)}{d \ln \mu^2} = \mu^2 \frac{df(x, \mu^2)}{d \mu^2} - \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} P_{qq}(z) f \left(\frac{x}{z} \right) \equiv 0 \quad (162)$$

and then

$$\mu^2 \frac{df(x, \mu^2)}{d\mu^2} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} P_{qq}(z) f\left(\frac{x}{z}, \mu^2\right) \quad (163)$$

This equation is usually called the DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) equation. As in the case of the resummation of leading logarithms in $R_{e^+e^-}$ induced by the RG invariance constraints, the DGLAP equation – which is the result of RG-invariance – resums a full tower of leading logarithms of Q^2 .

Proof: Let us define $t = \log \frac{Q^2}{\mu^2}$. We can then expand $f(x, t)$ in powers of t :

$$f(x, t) = f(x, 0) + t \frac{df}{dt}(x, 0) + \frac{t^2}{2!} \frac{d^2 f}{dt^2}(x, 0) + \dots \quad (164)$$

The first derivative is given by the DGLAP equation itself. Higher derivatives can be obtained by differentiating it:

$$\begin{aligned} f''(x, t) &= \frac{\alpha_s}{2\pi} \int \frac{dz}{z} P_{qq}(z) \frac{df}{dt}\left(\frac{x}{z}, t\right) \\ &= \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} P_{qq}(z) \frac{\alpha_s}{2\pi} \int_{\frac{x}{z}}^1 \frac{dz'}{z'} P_{qq}(z') f\left(\frac{x}{zz'}, t\right) \\ &\vdots \\ f^{(h)}(x, t) &= \frac{\alpha_s}{2\pi} \int_x^1 \dots \frac{\alpha_s}{2\pi} \int_{x/zz'\dots z^{(n-1)}}^1 \frac{dz^{(n)}}{z^{(n)}} P_{qq}(z^{(n)}) f\left(\frac{x}{zz'\dots}, t\right) \end{aligned} \quad (165)$$

The n -th term in this expansion, proportional to $(\alpha_s t)^n$, corresponds to the emission of n gluons (it is just the n -fold iteration of what we did studying the one-gluon emission case).

With similar calculations one can include the effect of the other $\mathcal{O}(\alpha_s)$ correction, originating from the splitting into a $q\bar{q}$ pair of a gluon contained in the proton. With the addition of this term, the evolution equation for the density of the i th quark flavour becomes:

$$\frac{df_q(x, t)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \left[P_{qq}(z) f_i\left(\frac{x}{z}, t\right) + P_{qg}(z) f_g\left(\frac{x}{z}, t\right) \right], \quad \text{with } P_{qg} = \frac{1}{2} [z^2 + (1-z)^2] \quad (166)$$

In the case of interactions with a coloured probe (say a gluon) we meet the following corrections, which affect the evolution of the gluon density $f_g(x)$:

$$\frac{df_g(x, t)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \left[P_{gq}(z) \sum_{i=q, \bar{q}} f_i\left(\frac{x}{z}, t\right) + P_{gg}(z) f_g\left(\frac{x}{z}, t\right) \right] \quad (167)$$

with

$$P_{gq}(z) = P_{qg}(1-z) = C_F \frac{1+(1-z)^2}{z} \quad \text{and} \quad P_{gg}(z) = 2C_A \left[\frac{1-z}{z} + \frac{z}{1-z} + z(1-z) \right] \quad (168)$$

Defining the moments of an arbitrary function $g(x)$ as follows:

$$g_n = \int_0^1 \frac{dx}{x} x^n g(x)$$

it is easy to prove that the evolution equations turn into ordinary linear differential equations:

$$\frac{df_i^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [P_{qq}^{(n)} f_i^{(n)} + P_{qg}^{(n)} f_g^{(n)}] \quad (169)$$

$$\frac{df_g^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [P_{gg}^{(n)} f_g^{(n)} + P_{gq}^{(n)} f_i^{(n)}] \quad (170)$$

5.3 Properties of the evolution equations

We now study some general properties of these equations. It is convenient to introduce the concepts of *valence* ($V(x, t)$) and *singlet* ($\Sigma(x, t)$) densities:

$$V(x) = \sum_i f_i(x) - \sum_{\bar{i}} f_{\bar{i}}(x) \quad (171)$$

$$\Sigma(x) = \sum_i f_i(x) + \sum_{\bar{i}} f_{\bar{i}}(x) \quad (172)$$

where the index \bar{i} refers to the antiquark flavours. The evolution equations then become:

$$\frac{dV^{(n)}}{dt} = \frac{\alpha_s}{2\pi} P_{qq}^{(n)} V^{(n)} \quad (173)$$

$$\frac{d\Sigma^{(n)}}{dt} = \frac{\alpha_s}{2\pi} \left[P_{qq}^{(n)} \Sigma^{(n)} + 2n_f P_{qg}^{(n)} f_g^{(n)} \right] \quad (174)$$

$$\frac{df_g^{(n)}}{dt} = \frac{\alpha_s}{2\pi} \left[P_{gq}^{(n)} \Sigma^{(n)} + P_{gg}^{(n)} f_g^{(n)} \right] \quad (175)$$

Note that the equation for the valence density decouples from the evolution of the gluon and singlet densities, which are coupled among themselves. This is physically very reasonable, since in perturbation theory the contribution to the quark and the antiquark densities coming from the evolution of gluons (via their splitting into $q\bar{q}$ pairs) is the same, and will cancel out in the definition of the valence. The valence therefore only evolves because of gluon emission. On the contrary, gluons and $q\bar{q}$ pairs in the proton *sea* evolve into one another.

The first moment of $V(x)$, $V^{(1)} = \int_0^1 dx V(x)$, counts the number of valence quarks. We therefore expect it to be independent of Q^2 :

$$\frac{dV^{(1)}}{dt} \equiv 0 = \frac{\alpha_s}{2\pi} P_{qq}^{(1)} V^{(1)} = 0 \quad (176)$$

Since $V^{(1)}$ itself is different from 0, we obtain a constraint on the first moment of the splitting function: $P_{qq}^{(1)} = 0$. This constraint is satisfied by including the effect of the virtual corrections, which generate a contribution to $P_{qq}(z)$ proportional to $\delta(1-z)$. This correction is incorporated in $P_{qq}(z)$ via the redefinition:

$$P_{qq}(z) \rightarrow \left(\frac{1+z^2}{1-z} \right)_+ \equiv \frac{1+z^2}{1-z} - \delta(1-z) \int_0^1 dy \left(\frac{1+y^2}{1-y} \right) \quad (177)$$

where the $+$ sign turns $P_{qq}(z)$ into a distribution. In this way, $\int_0^1 dz P_{qq}(z) = 0$ and the valence sum-rule is obeyed at all Q^2 .

Another sum rule which does not depend on Q^2 is the momentum sum rule, which imposes the constraint that all of the momentum of the proton is carried by its constituents (valence plus sea plus gluons):

$$\int_0^1 dx x \left[\sum_{i, \bar{i}} f_i(x) + f_g(x) \right] \equiv \Sigma^{(2)} + f_g^{(2)} = 1 \quad (178)$$

Once more this relation should hold for all Q^2 values, and you can prove by using the evolution equations that this implies:

$$P_{qq}^{(2)} + P_{gq}^{(2)} = 0 \quad (179)$$

$$P_{gg}^{(2)} + 2n_f P_{qg}^{(2)} = 0 \quad (180)$$

You can check using the definition of second moment, and the explicit expressions of the P_{qq} and P_{gq} splitting functions, that the first condition is automatically satisfied. The second condition is satisfied by including the virtual effects in the gluon propagator, which contribute a term proportional to $\delta(1-z)$. It is a simple exercise to verify that the final form of the $P_{gg}(z)$ splitting function, satisfying eq. (180), is:

$$P_{gg} \rightarrow 2C_A \left\{ \frac{x}{(1-x)_+} + \frac{1-x}{x} + x(1-x) \right\} + \delta(1-x) \left[\frac{11C_A - 2n_f}{6} \right] \quad (181)$$

5.4 Solution of the evolution equations

The evolution equations formulated in the previous section can be solved analytically in moment space. The boundary conditions are given by the moments of the parton densities at a given scale μ , where in principle they can be obtained from a direct measurement. The solution at different values of the scale Q can then be obtained by inverting numerically the expression for the moments back to x space. The resulting evolved densities can then be used to calculate cross sections for an arbitrary process involving hadrons, at an arbitrary scale Q . We shall limit ourselves here to studying some properties of the analytic solutions, and will present and comment some plots obtained from numerical studies available in the literature.

As an exercise, you can show that the solution of the evolution equation for the valence density is the following:

$$V^{(n)}(Q^2) = V^{(n)}(\mu^2) \left[\frac{\log Q^2/\Lambda^2}{\log \mu^2/\Lambda^2} \right]^{P_{qq}^{(n)}/2\pi b_0} = V^{(n)}(\mu^2) \left[\frac{\alpha_s(\mu^2)}{\alpha_s(Q^2)} \right]^{P_{qq}^{(n)}/2\pi b_0} \quad (182)$$

where the running of $\alpha_s(\mu^2)$ has to be taken into account to get the right result. Since all moments $P^{(n)}$ are negative, the evolution to larger values of Q makes the valence distribution softer and softer. This is physically reasonable, since the only thing that the valence quarks can do is to loose energy because of gluon emission.

The solutions for the gluon and singlet distributions f_g and Σ can be obtained by diagonalizing the 2×2 system in eqs. (174) and (175). We study the case of the second moments, which correspond to the momentum fractions carried by quarks and gluons separately. In the asymptotic limit $\Sigma^{(2)}$ goes to a constant, and $\frac{d\Sigma^{(2)}}{dt} = 0$. Then, using the momentum sum rule:

$$P_{qq}^{(2)} \Sigma^{(2)} + 2n_f P_{qg}^{(2)} f_g^{(2)} = 0 \quad (183)$$

$$\Sigma^{(2)} + f_g^{(2)} = 1 \quad (184)$$

The solution of this system is:

$$\Sigma^{(2)} = \frac{1}{1 + \frac{4C_F}{n_f}} \quad (= 15/31 \text{ for } n_f = 5) \quad (185)$$

$$f_g^{(2)} = \frac{4C_F}{4C_F + n_f} \quad (= 16/31 \text{ for } n_f = 5) \quad (186)$$

As a result, the fraction of momentum carried by gluons is asymptotically approximately 50% of the total proton momentum. It is interesting to note that, experimentally, this asymptotic value is actually reached already at rather low values of Q^2 . It was indeed observed already since the early days of the DIS experiments that only approximately 50% of the proton momentum was carried by charged constituents. This was one of the early evidences for the existence of gluons.

As I mentioned earlier, a complete solution for the evolved parton densities in x space can only be obtained from a numerical analysis. This work has been done in the past by several groups (see e.g. the discussions in ref. [8]), and is continuously being updated by including the most up-to-date experimental results used for the determination of the input densities at a fixed scale. Figure 2(a) describes the up-quark valence momentum density at different scales

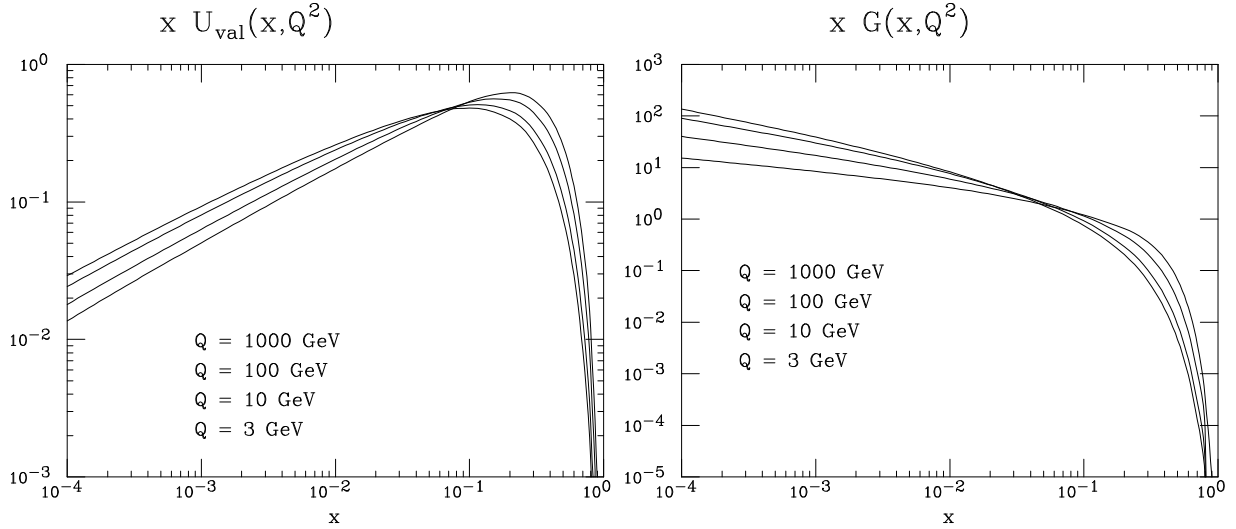


Fig. 2: Left: Valence up-quark momentum-density distribution, for different scales Q . Right: gluon momentum-density distribution.

Q . Note the anticipated softening at large scales, and the clear $\log Q^2$ evolution. The most likely momentum fraction carried by a valence up quark in the proton goes from $x \sim 20\%$ at $Q = 3$ GeV, to $x \lesssim 10\%$ at $Q = 1000$ GeV. Notice finally that the density vanishes at small x .

Figure 2b shows instead the gluon momentum density at different scales Q . This time the density grows at small- x , with an approximate $g(x) \sim 1/x^{1+\delta}$ behaviour, and $\delta > 0$ slowly increasing at large Q^2 . This low- x growth is due to the $1/x$ emission probability for the radiation of gluons, which was discussed in the previous lecture and which is represented by the $1/x$ factors in the $P_{gq}(x)$ and $P_{gg}(x)$ splitting functions.

Figure 3a shows the up-quark *sea* momentum density at different scales Q . Shape and evolution match those of the gluon density, a consequence of the fact that sea quarks come from the splitting of gluons. Since the gluon-splitting probability is proportional to α_s , the approximate ratio $sea/gluon \sim 0.1$ which can be obtained by comparing figs. 2b and 3a is perfectly justified.

Finally, the momentum densities for gluons, up-sea, charm and up-valence distributions are shown in fig.3b for $Q = 1000$ GeV. Note here that u_{sea} and charm are approximately the same at very large Q and small x , The proton momentum is mostly carried by valence quarks and by gluons. The contribution of sea quarks is negligible.

6 QCD in hadronic collisions

In hadronic collisions, all phenomena are QCD-related. The dynamics is more complex than in e^+e^- or DIS, since both beam and target have a non-trivial partonic structure. As a result, calculations (and experimental analyses) are more complicated. QCD phenomenology is however

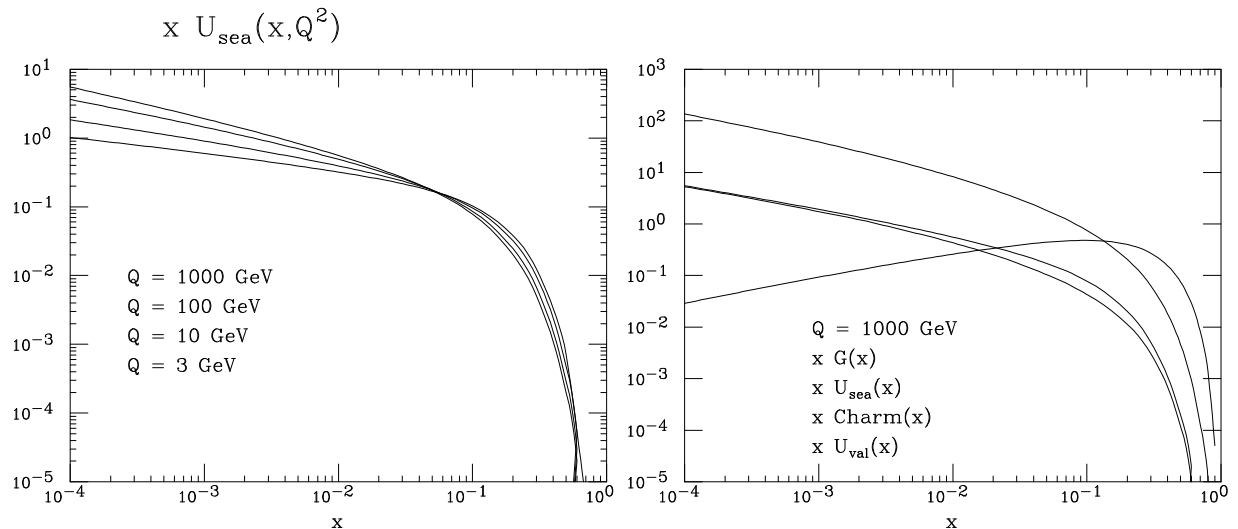


Fig. 3: Left: Sea up-quark momentum-density distribution, for different scales Q . Right: Momentum-density distribution for several parton species, at $Q = 1000$ GeV.

much richer, and the higher energies available in hadronic collisions allow to probe the structure of the proton and of its constituents at the smallest scales attainable in a laboratory.

Contrary to the case of e^+e^- and lepton-hadron collisions, where calculations are routinely available up to next-to-next-to-leading order (NNLO) accuracy, theoretical calculations for hadronic collisions are available at best with next-to-leading-order (NLO) accuracy. The only exception is the case of Drell-Yan production, where NNLO results are known for the total cross sections. So we generally have relatively small precision in the theoretical predictions, and theoretical uncertainties which are large when compared to LEP or HERA.

However, $p\bar{p}$ collider physics is primarily *discovery* physics, rather than precision physics (there are exceptions, such as the measurements of the W mass and of the properties of b -hadrons. But these are not QCD-related measurements). As such, knowledge of QCD is essential both for the estimate of the expected signals, and for the evaluation of the backgrounds. Tests of QCD in $p\bar{p}$ collisions confirm our understanding of perturbation theory, or, when they fail, point to areas where our approximations need to be improved. (see, e.g., the theory advances prompted by the measurements of ψ production at CDF!).

Finally, a reliable theoretical control over the details of production dynamics allows one to extract important information on the structure of the proton (parton densities) in regions of Q^2 and x otherwise inaccessible. Control of QCD at the current machines (the Tevatron at Fermilab) is therefore essential for the extrapolation of predictions to higher energies (say for applications at the future LHC, at CERN).

The key ingredients for the calculation of production rates and distributions in hadronic collisions are:

- the matrix elements for the hard, partonic process (e.g., $gg \rightarrow gg, gg \rightarrow b\bar{b}, q\bar{q}' \rightarrow W, \dots$),
- the hadronic parton densities, discussed in the previous lecture

Then the production rate for a given final state H is given by a factorization formula similar to the one used to describe DIS:

$$d\sigma(p\bar{p} \rightarrow H + X) = \int dx_1 dx_2 \sum_{i,j} f_i(x_1, Q) f_j(x_2, Q) d\hat{\sigma}(ij \rightarrow H) \quad (187)$$

where the parton densities f_i 's are evaluated at a scale Q typical of the hard process under consideration. For example $Q \simeq M_{DY}$ for production of a Drell-Yan pair, $Q \simeq E_T$ for high transverse-energy (E_T) jets, $Q^2 \simeq p_T^2 + m_Q^2$ for high- p_T heavy quarks, etc.

In this lecture we will briefly explore two of the QCD phenomena currently studied in hadronic collisions: Drell-Yan, and inclusive jet production. More details can be found in ref. [8,4].

6.1 Drell-Yan processes

While the Z boson has been recently studied with great precision by the LEP experiments, it was actually discovered, together with the W boson, by the CERN experiments UA1 and UA2 in $p\bar{p}$ collisions. W physics is now being studied in great detail at LEP2, but the best direct measurements of its mass by a single group still belong to $p\bar{p}$ experiments (CDF and D0 at the Tevatron). Even after the ultimate luminosity will have been accumulated at LEP2, with a great improvement in the determination of the parameters of the W boson, the monopoly of W studies will immediately return to hadron colliders, with the Tevatron data-taking resuming in the year 2000, and later on with the start of the LHC experiments.

Precision measurements of W production in hadronic collisions are important for several reasons:

- this is the only process in hadronic collisions which is known to NNLO accuracy
- the rapidity distribution of the charged leptons from W decays is sensitive to the ratio of the up and down quark densities, and can contribute to our understanding of the proton structure.
- deviations from the expected production rates of highly virtual W 's ($p\bar{p} \rightarrow W^* \rightarrow e\nu$) are a possible signal of the existence of new W bosons, and therefore of new gauge interactions.

The partonic cross-section for the production of a W boson from the annihilation of a $q\bar{q}$ pair can be easily calculated, giving the following result [8,4]:

$$\hat{\sigma}(q_i\bar{q}_j \rightarrow W) = \pi \frac{\sqrt{2}}{3} |V_{ij}|^2 G_F M_W^2 \delta(\hat{s} - M_W^2) = A_{ij} M_W^2 \delta(\hat{s} - M_W^2) \quad (188)$$

where \hat{s} is partonic center of mass energy squared, and V_{ij} is the element of the Cabibbo-Kobayashi-Maskawa matrix. The delta function comes from the $2 \rightarrow 1$ phase space, which forces the center-of-mass energy of the initial state to coincide with the W mass. It is useful to introduce the two variables

$$\tau = \frac{\hat{s}}{S_{had}} \equiv x_1 x_2 \quad (189)$$

$$y = \frac{1}{2} \log \left(\frac{E_W + p_W^z}{E_W - p_W^z} \right) \equiv \frac{1}{2} \log \left(\frac{x_1}{x_2} \right), \quad (190)$$

where S_{had} is the hadronic center of mass energy squared. The variable y is called *rapidity*. For slowly moving objects it reduces to the standard velocity, but, contrary to the velocity, it transforms additively even at high energies under Lorentz boosts along the direction of motion. Written in terms of τ and y , the integration measure over the initial-state parton momenta becomes: $dx_1 dx_2 = d\tau dy$. Using this expression and eq. (188) in eq. (187), we obtain the following result for the LO total W production cross section:

$$\sigma_{DY} = \sum_{i,j} \frac{\pi A_{ij}}{M_W^2} \tau \int_{\tau}^1 \frac{dx}{x} f_i(x) f_j \left(\frac{\tau}{x} \right) \equiv \sum_{i,j} \frac{\pi A_{ij}}{M_W^2} \tau \mathcal{L}_{ij}(\tau) \quad (191)$$

where the function $\mathcal{L}_{ij}(\tau)$ is usually called *partonic luminosity*. In the case of $u\bar{d}$ collisions, the overall factor in front of this expression has a value of approximately 6.5 nb. It is interesting to study the partonic luminosity as a function of the hadronic CoM energy. This can be done by taking a simple approximation for the parton densities. Following the indications of the figures presented in the previous lecture, we shall assume that $f_i(x) \sim 1/x^{1+\delta}$, with $\delta < 1$. Then

$$\mathcal{L}(\tau) = \int_{\tau}^1 \frac{dx}{x} \frac{1}{x^{1+\delta}} \left(\frac{x}{\tau}\right)^{1+\delta} = \frac{1}{\tau^{1+\delta}} \int_{\tau}^1 \frac{dx}{x} = \frac{1}{\tau^{1+\delta}} \log\left(\frac{1}{\tau}\right) \quad (192)$$

and

$$\sigma_W \sim \tau^{-\delta} \log\left(\frac{1}{\tau}\right) = \left(\frac{S_{had}}{M_W^2}\right)^{\delta} \log\left(\frac{S_{had}}{M_W^2}\right) \quad (193)$$

The DY cross-section grows therefore at least logarithmically with the hadronic CM energy. This is to be compared with the behaviour of the Z production cross section in e^+e^- collisions, which is steeply diminishing for values of s well above the production threshold. The reason for the different behaviour in hadronic collisions is that while the energy of the hadronic initial state grows, it will always be possible to find partons inside the hadrons with the appropriate energy to produce the W directly on-shell. The number of partons available for the production of a W is furthermore increasing with the increase in hadronic energy, since the larger the hadron energy, the smaller will be the value of hadron momentum fraction x necessary to produce the W . The increasing number of partons available at smaller and smaller values of x causes then the growth of the total W production cross section.

A comparison between the best available prediction for the production rates of W and Z bosons in hadronic collisions, and the experimental data, is shown in fig. 4. The experimental uncertainties will soon be dominated by the limited knowledge of the machine luminosity, and will exceed the accuracy of the NNLO predictions. This suggests that in the future the total rate of produced W bosons could be used as an accurate luminometer.

It is also interesting to note that an accurate measurement of the relative W and Z production rates (which is not affected by the knowledge of the total integrated luminosity, that will cancel in their ratio) provides a tool to measure the total W width. This can be seen from the following equation:

$$\Gamma_W = \frac{N^{obs}(Z \rightarrow e^+e^-)}{N^{obs}(W \rightarrow e^{\pm}\nu)} \left(\frac{\sigma_{W^{\pm}}}{\sigma_Z}\right) \left(\frac{\Gamma_{e\nu}^W}{\Gamma_{e^+e^-}^Z}\right) \Gamma_Z$$

\uparrow
measure

$\nwarrow \nearrow$
calculable

\uparrow
LEP/SLC

As of today, this technique provides the best measurement of Γ_W : $\Gamma_W = 2.06 \pm 0.06$ GeV, which is a factor of 5 more accurate than the current best direct measurements from LEP2.

6.2 W Rapidity Asymmetry

The measurement of the charge asymmetry in the rapidity distribution of W bosons produced in $p\bar{p}$ collisions can provide an important measurement of the ratio of the u-quark and d-quark momentum distributions. Using the formulas provided above, you can in fact easily check as an exercise that:

$$\frac{d\sigma_{W^+}}{dy} \propto f_u^p(x_1) f_{\bar{d}}^{\bar{p}}(x_2) + f_d^p(x_1) f_{\bar{u}}^{\bar{p}}(x_2) \quad (194)$$

$$\frac{d\sigma_{W^-}}{dy} \propto f_{\bar{u}}^p(x_1) f_d^{\bar{p}}(x_2) + f_u^p(x_1) f_{\bar{d}}^{\bar{p}}(x_2) \quad (195)$$

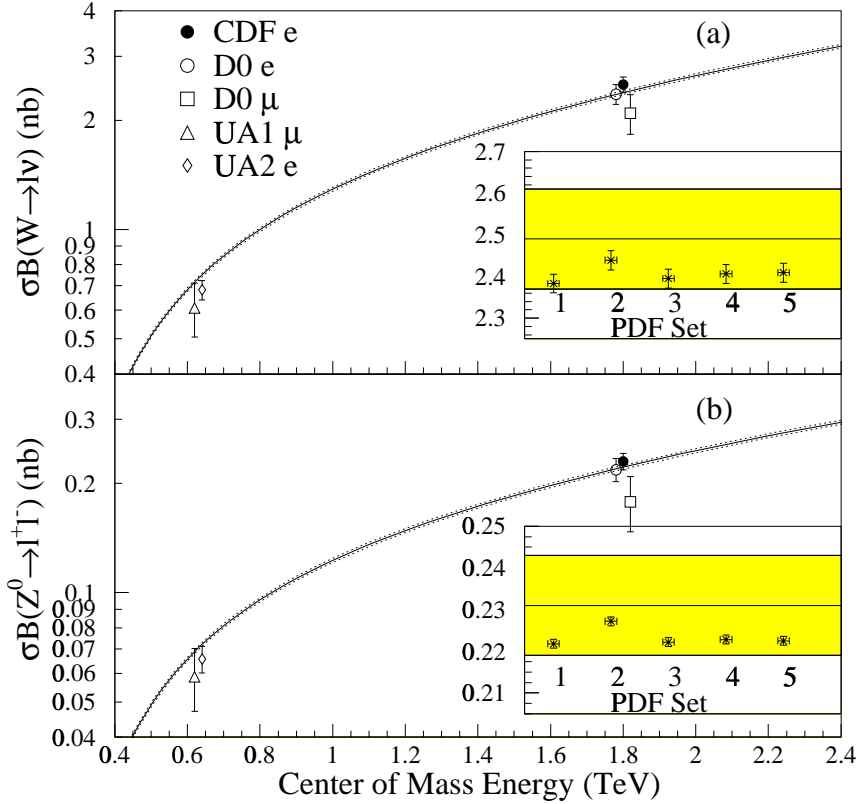


Fig. 4: Comparison of measured (a) $\sigma \cdot B(W \rightarrow e\nu)$ and (b) $\sigma \cdot B(Z^0 \rightarrow e^+e^-)$ to 2-loop theoretical predictions using MRSA parton distribution functions. The UA1 and UA2 measurements and D0 measurements are offset horizontally by ± 0.02 TeV for clarity. In the inset, the shaded area shows the 1σ region of the CDF measurement; the stars show the predictions using various parton distribution function sets (1) MRSA, (2) MRSD0', (3) MRSD-', (4) MRSH and (5) CTEQ2M. The theoretical points include a common uncertainty in the predictions from choice of renormalization scale ($M_W/2$ to $2M_W$).

We can then construct the following charge asymmetry (assuming the dominance of the quark densities over the antiquark ones, which is valid in the kinematical region of interest for W production at the Tevatron):

$$A(y) = \frac{\frac{d\sigma_{W^+}}{dy} - \frac{d\sigma_{W^-}}{dy}}{\frac{d\sigma_{W^+}}{dy} + \frac{d\sigma_{W^-}}{dy}} = \frac{f_u^p(x_1) f_d^p(x_2) - f_d^p(x_1) f_u^p(x_2)}{f_u^p(x_1) f_d^p(x_2) + f_d^p(x_1) f_u^p(x_2)} \quad (196)$$

Setting $f_d(x) = f_u(x) R(x)$ we then get:

$$A(y) = \frac{R(x_2) - R(x_1)}{R(x_2) + R(x_1)}. \quad (197)$$

which measures the $R(x)$ ratio since $x_{1,2}$ are known in principle from the kinematics: $x_{1,2} = \sqrt{\tau} \exp(\pm y)$ ⁴. The current CDF data provide the most accurate measurement to date of this quantity (see ref. [8]).

⁴In practice one cannot determine $x_{1,2}$ with arbitrary precision on an event-by-event basis, since the longitudinal momentum of the neutrino cannot be easily measured. The actual measurement is therefore done by studying the charge asymmetry in the rapidity distribution of the charged lepton.

6.3 Jet Production

Jet production is the hard process with the largest rate in hadronic collisions. For example, the cross section for producing at the Tevatron ($\sqrt{S_{had}} = 1.8$ TeV) jets of transverse energy $E_T^{jet} \lesssim 50$ GeV is of the order of a μb . This means 50 events/sec at the luminosities available at the Tevatron. The data collected at the Tevatron so far extend all the way up to the E_T values of the order of 450 GeV. These events are generated by collisions among partons which carry over 50% of the available $p\bar{p}$ energy, and allow to probe the shortest distances ever reached. The leading mechanisms for jet production are shown in fig. 5.

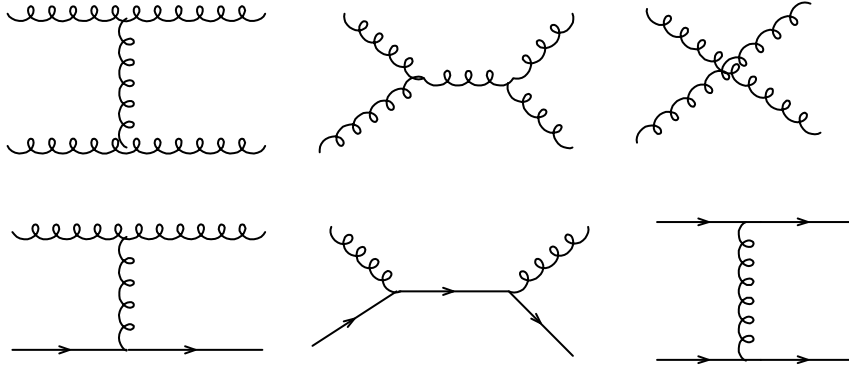


Fig. 5: Representative diagrams for the production of jet pairs in hadronic collisions.

The 2-jet inclusive cross section can be obtained from the formula

$$d\sigma = \sum_{ijkl} dx_1 dx_2 f_i^{(H_1)}(x_1, \mu) f_j^{(H_2)}(x_2, \mu) \frac{d\hat{\sigma}_{ij \rightarrow k+l}}{d\Phi_2} d\Phi_2 \quad (198)$$

that has to be expressed in terms of the rapidity and transverse momentum of the quarks (or jets), in order to make contact with physical reality. The two-particle phase space is given by

$$d\Phi_2 = \frac{d^3k}{2k^0(2\pi)^3} 2\pi \delta((p_1 + p_2 - k)^2), \quad (199)$$

and, in the CM of the colliding partons, we get

$$d\Phi_2 = \frac{1}{2(2\pi)^2} d^2k_T dy 2\delta(\hat{s} - 4(k^0)^2), \quad (200)$$

where k_T is the transverse momentum of the final-state partons. Here y is the rapidity of the produced parton in the parton CM frame. It is given by

$$y = \frac{y_1 - y_2}{2} \quad (201)$$

where y_1 and y_2 are the rapidities of the produced partons in the laboratory frame (in fact, in any frame). One also introduces

$$y_0 = \frac{y_1 + y_2}{2} = \frac{1}{2} \log \frac{x_1}{x_2}, \quad \tau = \frac{\hat{s}}{S_{had}} = x_1 x_2. \quad (202)$$

We have

$$dx_1 dx_2 = dy_0 d\tau. \quad (203)$$

We obtain

$$d\sigma = \sum_{ijkl} dy_0 \frac{1}{S_{had}} f_i^{(H_1)}(x_1, \mu) f_j^{(H_2)}(x_2, \mu) \frac{d\hat{\sigma}_{ij \rightarrow k+l}}{d\Phi_2} \frac{1}{2(2\pi)^2} 2 dy d^2 k_T \quad (204)$$

which can also be written as

$$\frac{d\sigma}{dy_1 dy_2 d^2 k_T} = \frac{1}{S_{had} 2(2\pi)^2} \sum_{ijkl} f_i^{(H_1)}(x_1, \mu) f_j^{(H_2)}(x_2, \mu) \frac{d\hat{\sigma}_{ij \rightarrow k+l}}{d\Phi_2}. \quad (205)$$

The variables x_1, x_2 can be obtained from y_1, y_2 and k_T from the equations

$$y_0 = \frac{y_1 + y_2}{2} \quad (206)$$

$$y = \frac{y_1 - y_2}{2} \quad (207)$$

$$x_T = \frac{2k_T}{\sqrt{S_{had}}} \quad (208)$$

$$x_1 = x_T e^{y_0} \cosh y \quad (209)$$

$$x_2 = x_T e^{-y_0} \cosh y. \quad (210)$$

For the partonic variables, we need \hat{s} and the scattering angle in the parton CM frame θ , since

$$t = -\frac{\hat{s}}{2} (1 - \cos \theta), \quad u = -\frac{\hat{s}}{2} (1 + \cos \theta). \quad (211)$$

Neglecting the parton masses, you can show that the rapidity can also be written as:

$$y = -\log \tan \frac{\theta}{2} \equiv \eta, \quad (212)$$

with η being usually referred to as pseudorapidity.

The leading-order Born cross sections for parton parton scattering are reported in table 1. It is interesting to note that a good approximation to the exact results can be easily obtained

Process	$\frac{d\hat{\sigma}}{d\Phi_2}$
$qq' \rightarrow qq'$	$\frac{1}{2\hat{s}} \frac{4}{9} \frac{\hat{s}^2 + \hat{u}^2}{\hat{t}^2}$
$qq \rightarrow qq$	$\frac{1}{2} \frac{1}{2\hat{s}} \left[\frac{4}{9} \left(\frac{\hat{s}^2 + \hat{u}^2}{\hat{t}^2} + \frac{\hat{s}^2 + \hat{t}^2}{\hat{u}^2} \right) - \frac{8}{27} \frac{\hat{s}^2}{\hat{u}\hat{t}} \right]$
$q\bar{q} \rightarrow q'\bar{q}'$	$\frac{1}{2\hat{s}} \frac{4}{9} \frac{\hat{t}^2 + \hat{u}^2}{\hat{s}^2}$
$q\bar{q} \rightarrow q\bar{q}$	$\frac{1}{2\hat{s}} \left[\frac{4}{9} \left(\frac{\hat{s}^2 + \hat{u}^2}{\hat{t}^2} + \frac{\hat{t}^2 + \hat{u}^2}{\hat{s}^2} \right) - \frac{8}{27} \frac{\hat{u}^2}{\hat{s}\hat{t}} \right]$
$q\bar{q} \rightarrow gg$	$\frac{1}{2} \frac{1}{2\hat{s}} \left[\frac{32}{27} \frac{\hat{t}^2 + \hat{u}^2}{\hat{t}\hat{u}} - \frac{8}{3} \frac{\hat{t}^2 + \hat{u}^2}{\hat{s}^2} \right]$
$gg \rightarrow q\bar{q}$	$\frac{1}{2\hat{s}} \left[\frac{1}{6} \frac{\hat{t}^2 + \hat{u}^2}{\hat{t}\hat{u}} - \frac{3}{8} \frac{\hat{t}^2 + \hat{u}^2}{\hat{s}^2} \right]$
$qq \rightarrow qq$	$\frac{1}{2\hat{s}} \left[-\frac{4}{9} \frac{\hat{s}^2 + \hat{u}^2}{\hat{s}\hat{u}} + \frac{\hat{u}^2 + \hat{s}^2}{\hat{t}^2} \right]$
$gg \rightarrow gg$	$\frac{1}{2} \frac{1}{2\hat{s}} \frac{9}{2} \left(3 - \frac{\hat{t}\hat{u}}{\hat{s}^2} - \frac{\hat{s}\hat{u}}{\hat{t}^2} - \frac{\hat{s}\hat{t}}{\hat{u}^2} \right)$

Table 1: Cross sections for light parton scattering. The notation is $p_1 p_2 \rightarrow kl$, $\hat{s} = (p_1 + p_2)^2$, $\hat{t} = (p_1 - k)^2$, $\hat{u} = (p_1 - l)^2$.

by using the soft-gluon techniques introduced in the third lecture. Based on the fact that even at 90° $\min(|t|, |u|)$ does not exceed $s/2$, and that therefore everything else being equal a propagator in the t or u channel contributes to the square of an amplitude 4 times more than a propagator in the s channel, it is reasonable to assume that the amplitudes are dominated by the diagrams with a gluon exchanged in the t (or u) channel. It is easy to calculate the amplitudes in this limit using the soft-gluon approximation. For example, the amplitude for the exchange of a soft gluon among a qq' pair is given by:

$$(\lambda_{ij}^a) (\lambda_{kl}^a) 2p_\mu \frac{1}{t} 2p'_\mu = \lambda_{ij}^a \lambda_{kl}^a \frac{4p \cdot p'}{t} = \frac{2s}{t} \lambda_{ij}^a \lambda_{kl}^a \quad (213)$$

The p_μ and p'_μ factors represent the coupling of the exchanged gluon to the q and q' quark lines, respectively (see eq. (105)). Squaring, and summing and averaging over spins and colours, gives

$$\overline{\sum_{\text{colours, spin}} |M_{qq'}|^2} = \frac{1}{N^2} \left(\frac{N^2 - 1}{4} \right) \frac{4s^2}{t^2} = \frac{8}{9} \frac{s^2}{t^2} \quad (214)$$

Since for this process the diagram with a t -channel gluon exchange is symmetric for $s \leftrightarrow u$ exchange, and since $u \rightarrow -s$ in the $t \rightarrow 0$ limit, the above result can be rewritten in an explicitly (s, u) symmetric way as

$$\frac{4}{9} \frac{s^2 + u^2}{t^2} \quad (215)$$

which indeed exactly agrees with the result of the exact calculation, as given in table 1. The corrections which appear from s or u gluon exchange when the quark flavours are the same or when we study a $q\bar{q}$ process are small, as can be seen by comparing the above result to the expressions in the table.

As another example we consider the case of $qg \rightarrow qg$ scattering. The amplitude will be exactly the same as in the $qq' \rightarrow qq'$ case, up to the different colour factors. A simple calculation then gives:

$$\overline{\sum_{\text{colours, spin}} |M_{qg}|^2} = \frac{9}{4} \overline{\sum} |M_{qq'}|^2 = \frac{s^2 + u^2}{t^2} \quad (216)$$

The exact result is

$$\frac{u^2 + s^2}{t^2} - \frac{4}{9} \frac{u^2 + s^2}{us} \quad (217)$$

which even at 90° , the point where the t -channel exchange approximation is worse, only differs from this latter by no more than 25%.

As a final example we consider the case of $gg \rightarrow gg$ scattering, which in our approximation gives:

$$\overline{\sum} |M_{gg}|^2 = \frac{9}{2} \frac{s^2}{t^2} \quad (218)$$

By $u \leftrightarrow t$ symmetry we should expect the simple improvement:

$$\overline{\sum} |M_{gg}|^2 \sim \frac{9}{2} \left(\frac{s^2}{t^2} + \frac{s^2}{u^2} \right). \quad (219)$$

This only differs by 20% from the exact result at 90° .

Notice that at small t the following relation holds:

$$\hat{\sigma}_{gg} : \hat{\sigma}_{qg} : \hat{\sigma}_{q\bar{q}} = \left(\frac{9}{4} \right) : 1 : \left(\frac{4}{9} \right) \quad (220)$$

The $9/4$ factors are simply the ratios of the colour factors for the coupling to gluons of a gluon (C_A) and of a quark (T_F), after including the respective colour-average factors ($1/(N^2 - 1)$ for the gluon, and $1/N$ for the quark). Using eq. (220), we can then write:

$$d\sigma_{hadr} = \int dx_1 dx_2 \sum_{i,j} f_i(x_1) f_j(x_2) d\hat{\sigma}_{ij} = \int dx_1 dx_2 F(x_1) F(x_2) d\hat{\sigma}_{gg}(gg \rightarrow \text{jets}) \quad (221)$$

where the object:

$$F(x) = f_g(x) + \frac{4}{9} \sum_f [q_f(x) + \bar{q}_f(x)] \quad (222)$$

is usually called the *effective structure function*. This result indicates that the measurement of the inclusive jet cross section does not allow in principle to disentangle the independent contribution of the various partonic components of the proton, unless of course one is considering a kinematical region where the production is dominated by a single process. The relative contributions of the different channels, as predicted using the global fits of parton densities available in the literature, are shown in fig. 6

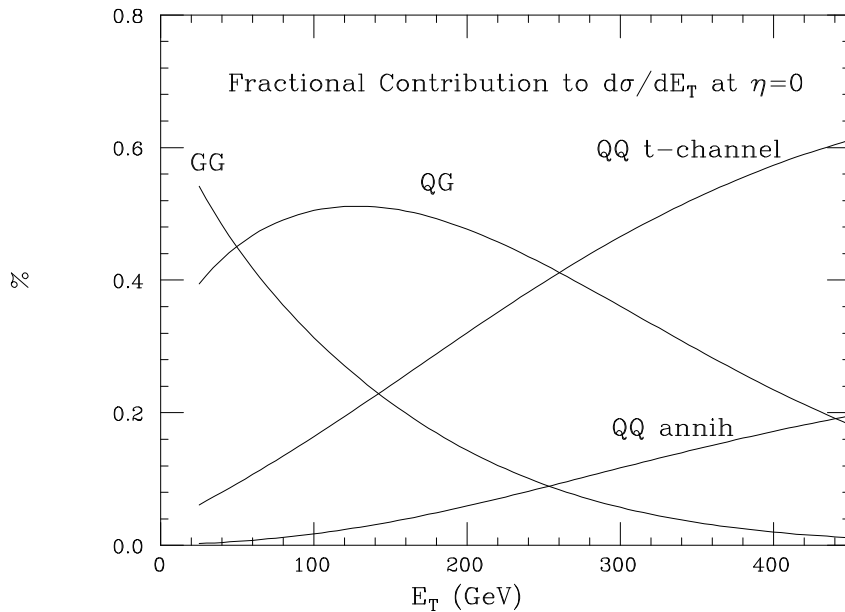


Fig. 6: Relative contribution to the inclusive jet- E_T rates from the different production channels.

Predictions for jet production at colliders are available today at the next-to-leading order in QCD. A comparison between these calculations and the available data is given in figs. 7 and 8. At the Tevatron, jets up to 450 GeV transverse momentum have been observed. That is $x \gtrsim 0.5$ and $Q^2 \simeq 160,000 \text{ GeV}^2$. This is a domain of x and Q^2 not accessible to HERA. The current agreement between theory and data is at the level of 30 % over 8 orders of magnitude of cross-section, from $E_T \sim 20$ to $E_T \sim 450$ GeV. The small deviation observed by CDF at high E_T is under active investigation both experimentally and theoretically. It is still premature to say whether it can be a signal of new phenomena, or whether it is the result of our incomplete knowledge of the gluon density at large x . Either way, future higher-statistics measurements at the Tevatron will provide some important input on these fundamental questions. The resulting knowledge will enable theorists to reliably predict production rates for all interesting processes that will take place at the LHC.

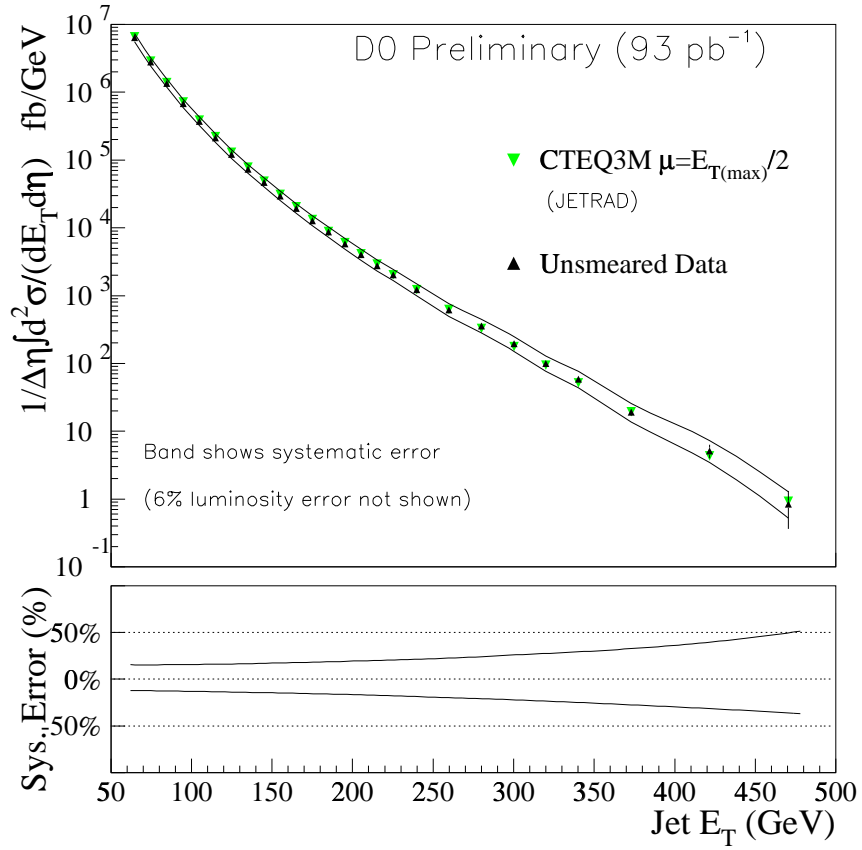


Fig. 7: Inclusive E_T spectra for central jets at the Tevatron

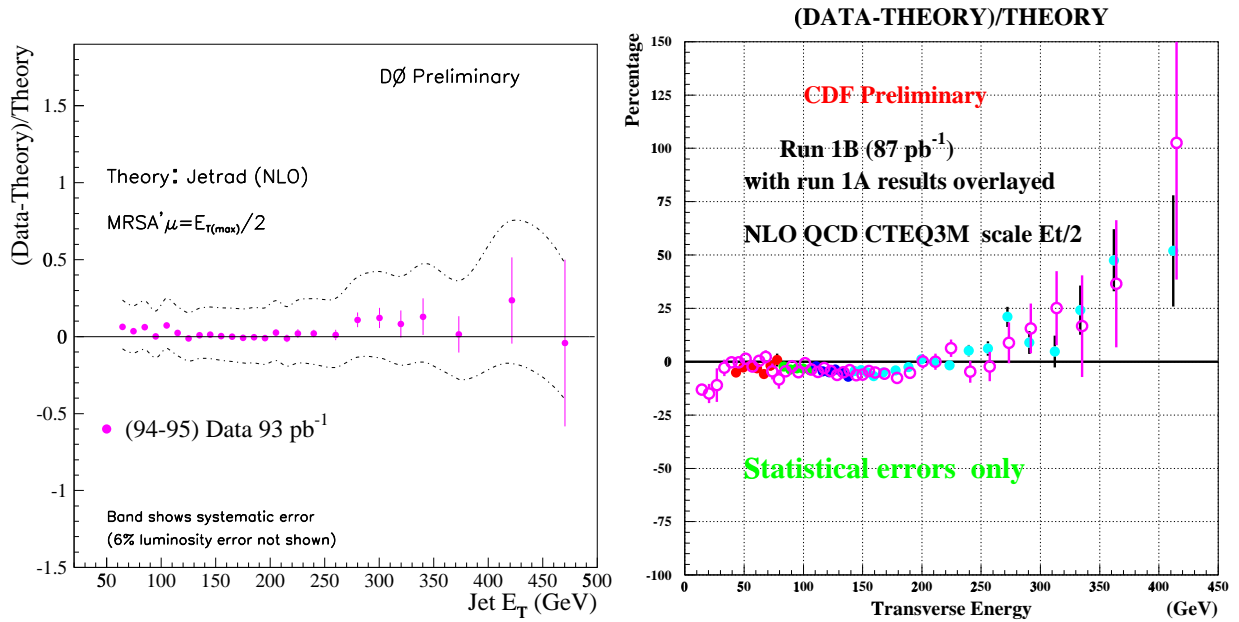


Fig. 8: Comparison of inclusive jets cross sections with QCD calculations at the Tevatron.

Acknowledgements

It is a pleasure to thank the organizers of this School, for the successful efforts made to bring top-quality students together, and to provide a great environment for physics discussions and for a pleasant time as well.

REFERENCES

Standard Textbooks:

1. R.P. Feynman, Photon-Hadron Interactions, W.A. Benjamin, NY (1972).
2. B.L. Ioffe, V.A. Khoze and L.N. Lipatov, Hard Processes, North Holland (1984).
3. T. Muta, Foundations of QCD, World Scientific (1987).
4. V. Barger and R.J.N. Phillips, Collider Physics, Addison Wesley (1987).
5. R. Field, Applications of Perturbative QCD, Addison Wesley (1989).
6. Yu.L. Dokshitzer, V.A. Khoze, A.H. Mueller and S.I. Troyan, Basics of Perturbative QCD, Editions Frontieres (1991).
7. M.E. Peskin and D.V. Schroeder: An Introduction to Quantum Field Theory, Addison-Wesley (1995).
8. R.K. Ellis, W.J. Stirling and B.R. Webber: QCD and Collider Physics, Cambridge University Press (1996).

Pedagogical Reviews

9. P. Nason, lectures delivered at the 1997 CERN School.
10. Yu.L. Dokshitzer, lectures delivered at the 1995 CERN School.
11. M. Neubert, lectures delivered at the 1994 CERN School.

Review Articles

12. G. Altarelli, *Phys. Rep.* **81** (1982) 1.
13. A.H. Mueller, *Phys. Rev.* **73** (1981) 237.
14. Yu.L. Dokshitzer, D.I. Dyakonov and S.I. Trojan, *Phys. Rep.* **58** (1980) 270.
15. A. Bassetto, M. Ciafaloni and G. Marchesini, *Phys. Rep.* **100** (1983) 201.
16. M.L. Mangano and S.J. Parke, *Phys. Rep.* **200** (1991) 301.

Historical Reviews

17. D.J. Gross, [hep-th/9809060](#).
18. G. 't Hooft, [hep-th/9808154](#).

FLAVOR PHYSICS AND CP VIOLATION

Y. Nir

Department of Particle Physics

Weizmann Institute of Science, Rehovot 76100, Israel

ftnir@clever.weizmann.ac.il

Abstract

This is a written version of a series of three lectures aimed at graduate students in the field of experimental high energy physics. The emphasis is on physics that is relevant to B -factories.

The main topics covered are:

- (i) The flavor sector of the Standard Model;
- (ii) Determination of the mixing parameters;
- (iii) CP violation in meson decays (a model independent description);
- (iv) CP violation in the Standard Model;
- (v) CP violation as a probe of new physics.

1 FLAVOR PHYSICS

1.1 What is Flavor Physics and Why is it Interesting?

The Standard Model fermions appear in three generations. *Flavor physics* describes interactions that distinguish between the fermion generations.

The fermions experience two types of interactions: gauge interactions, where two fermions couple to a gauge boson, and Yukawa interactions, where two fermions couple to a scalar. Within the Standard Model [1-3], there are twelve gauge bosons, related to the gauge symmetry

$$G_{\text{SM}} = SU(3)_C \times SU(2)_L \times U(1)_Y, \quad (1)$$

and a single Higgs scalar, related to the spontaneous symmetry breaking

$$G_{\text{SM}} \rightarrow SU(3)_C \times U(1)_{\text{EM}}. \quad (2)$$

In the *interaction basis*, gauge interactions are diagonal (and universal, namely described by a single gauge coupling for each factor in G_{SM} : g_s , g and g'). By definition, the interaction eigenstates have no gauge couplings between fermions of different generations. The Yukawa interactions are, however, quite complicated in the interaction basis. In particular, there are Yukawa couplings that involve fermions of different generations and, consequently, the interaction eigenstates do not have well-defined masses. *Flavor Physics* refers to the part of the Standard Model that depends on the Yukawa couplings.

In the *mass basis*, Yukawa interactions are diagonal (though not universal). The mass eigenstates have, by definition, well-defined masses. The gauge interactions related to spontaneously broken symmetries can, however, be quite complicated in the mass basis. In particular, the $SU(2)_L$ gauge couplings are not diagonal, that is they *mix* quarks of different generations. *Flavor Physics* here refers to fermion masses and mixings.

Why is flavor physics interesting?

- (i) Flavor physics has not been well tested yet. For the gauge interactions, experiments (particularly LEP and SLD) have provided us with tests at or even below an accuracy level of one percent, where radiative corrections become essential. In contrast, several

flavor parameters are only known to an accuracy level of $\mathcal{O}(30\%)$. Many rare decay processes that are sensitive to the flavor parameters have not been measured yet. In the near future, various experiments (particularly CLEO, BaBar and Belle) will substantially improve the determination of the flavor parameters and will measure various rare B decays, thus providing much more stringent tests of this sector of the Standard Model.

- (ii) Most of the Standard Model flavor parameters are small and hierarchical. The Standard Model does not provide any explanation of these features. This is *the flavor puzzle*. It may be a hint for physics beyond the Standard Model, where the smallness and hierarchy of the flavor parameters find a natural explanation. For example, horizontal symmetries (that is, symmetries under which different generations transform differently) that are broken by a small parameter give selection rules for the Yukawa couplings.
- (iii) Flavor changing neutral current processes (FCNC) depend on the flavor parameters. For vanishing Yukawa couplings, FCNC would be absent to all orders in the gauge couplings. Consequently, within the Standard Model FCNC are suppressed by small mixing angles and, in some cases, small quark masses. Furthermore, within the Standard Model FCNC vanish at tree level. Consequently, they are further suppressed by powers of the weak coupling. Many extensions of the Standard Model allow significant new contributions to these processes that modify the Standard Model predictions. Therefore, the flavor sector is a very sensitive probe of New Physics.
- (iv) CP violation is closely related to flavor physics. It is one of the least tested aspects of the Standard Model: Even though the Kobayashi-Maskawa phase [4] can account for the CP violation that has been measured in K decays [5], the Standard Model picture of CP violation could still be completely wrong. Almost any extension of the Standard Model provides new sources of CP violation. The observed baryon asymmetry of the universe *requires* new sources of CP violation [6]. (The motivation to study CP violation is described in more detail in Section 3.)

1.2 What are the Flavor Parameters?

The Standard Model fermions appear in three generations. Each generation is made of five different representations of the Standard Model gauge group G_{SM} of eq. (1):

$$Q_{Li}^I(3, 2)_{+1/6}, \quad u_{Ri}^I(3, 1)_{+2/3}, \quad d_{Ri}^I(3, 1)_{-1/3}, \quad L_{Li}^I(1, 2)_{-1/2}, \quad \ell_{Ri}^I(1, 1)_{-1}. \quad (3)$$

Our notations mean that, for example, the left-handed quarks, Q_L^I , are in a triplet (3) of the $SU(3)_C$ group, a doublet (2) of $SU(2)_L$ and carry hypercharge $Y = Q_{\text{EM}} - T_3 = +1/6$. The index I denotes *interaction eigenstates*. The index $i = 1, 2, 3$ is the *flavor* (or generation) index. (The above representations describe quarks and leptons and include, therefore, left-handed and right-handed fields. An alternative way to write down the various representations, which is particularly useful for the supersymmetric extension of the Standard Model, is to describe left-handed fields only. Now the fields include also antiquarks and antileptons:

$$Q_i^I(3, 2)_{+1/6}, \quad \bar{u}_i^I(\bar{3}, 1)_{-2/3}, \quad \bar{d}_i^I(\bar{3}, 1)_{+1/3}, \quad L_i^I(1, 2)_{-1/2}, \quad \bar{\ell}_i^I(1, 1)_{+1}. \quad (4)$$

The Standard Model gauge interactions do not distinguish between the different generations. Another way to state this is to say that the gauge interactions are flavor-blind. The strength of the gauge interactions depends on the gauge quantum numbers given in (3) and not on the flavor index i . Most important for our purposes, the interaction of the $SU(2)_L$ gauge bosons (W_μ^a , $a = 1, 2, 3$) with quarks is given by

$$-\mathcal{L}_W = \frac{g}{2} \overline{Q_{Li}^I} \gamma^\mu \tau^a Q_{Li}^I W_\mu^a. \quad (5)$$

The 4×4 matrix γ^μ operates in Lorentz space (it describes the combination of two spin-1/2 quark fields and one spin-1 gauge boson field into a Lorentz scalar) and the 2×2 matrix τ^a operates in the $SU(2)_L$ space (it describes the combination of the two quark doublets and the W^a -triplet into an $SU(2)_L$ singlet). The coupling $\overline{Q_{Li}^I} Q_{Li}^I$ can be equivalently written as $\overline{Q_{Li}^I} \mathbf{1}_{ij} Q_{Lj}^I$ where the 3×3 unit matrix $\mathbf{1}$ operates in flavor space and makes the universality of the gauge interactions manifest.

The Yukawa interactions have a complicated form in this basis:

$$-\mathcal{L}_Y = Y_{ij}^d \overline{Q_{Li}^I} \phi d_{Rj}^I + Y_{ij}^u \overline{Q_{Li}^I} \tilde{\phi} u_{Rj}^I + Y_{ij}^\ell \overline{L_{Li}^I} \phi \ell_{Rj}^I, \quad (6)$$

where $\phi(1, 2)_{+1/2}$ is the Standard Model Higgs doublet, and $\tilde{\phi} = i\sigma_2 \phi^*$. The Yukawa matrices Y^d , Y^u and Y^ℓ are general (and, in particular, complex) 3×3 matrices. Note that, in the absence of right-handed neutrinos, $N_i(1, 1)_0$, one cannot write (renormalizable) Yukawa interactions for the neutrinos.

To transform to the mass basis, one has to take into account spontaneous symmetry breaking (2). Within the Standard Model this breaking is the result of a vacuum expectation value assumed by the neutral component of the Higgs doublet, $\langle \phi^0 \rangle = \frac{v}{\sqrt{2}}$ with the electroweak breaking scale of order $v \approx 246 \text{ GeV}$. Upon the replacement $\mathcal{R}e(\phi^0) \rightarrow (v + H^0)/\sqrt{2}$, the Yukawa interactions (6) give rise to mass terms:

$$-\mathcal{L}_M = (M_d)_{ij} \overline{d_{Li}^I} d_{Rj}^I + (M_u)_{ij} \overline{u_{Li}^I} u_{Rj}^I + (M_\ell)_{ij} \overline{\ell_{Li}^I} \ell_{Rj}^I, \quad (7)$$

where

$$M_f = \frac{v}{\sqrt{2}} Y^f, \quad (8)$$

and we decomposed the $SU(2)_L$ doublets into their components:

$$Q_{Li}^I = \begin{pmatrix} u_{Li}^I \\ d_{Li}^I \end{pmatrix}, \quad L_{Li}^I = \begin{pmatrix} \nu_{Li}^I \\ \ell_{Li}^I \end{pmatrix}. \quad (9)$$

Since neutrinos have no Yukawa interactions, they are massless.

The mass basis corresponds, by definition, to diagonal mass matrices. We can always find unitary matrices V_{fL} and V_{fR} such that

$$V_{fL} M_f V_{fR}^\dagger = M_f^{\text{diag}}, \quad (10)$$

with M_f^{diag} diagonal and real. The mass eigenstates are then identified as

$$\begin{aligned} d_{Li} &= (V_{dL})_{ij} d_{Lj}^I, & d_{Ri} &= (V_{dR})_{ij} d_{Rj}^I, \\ u_{Li} &= (V_{uL})_{ij} u_{Lj}^I, & u_{Ri} &= (V_{uR})_{ij} u_{Rj}^I, \\ \ell_{Li} &= (V_{\ell L})_{ij} \ell_{Lj}^I, & \ell_{Ri} &= (V_{\ell R})_{ij} \ell_{Rj}^I, \\ \nu_{Li} &= (V_{\nu L})_{ij} \nu_{Lj}^I. \end{aligned} \quad (11)$$

Note that, since the neutrinos are massless, $V_{\nu L}$ is arbitrary.

The charged current interactions (that is the interactions of the charged $SU(2)_L$ gauge bosons $W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2)$), which in the interaction basis are described by (5), have a complicated form in the mass basis:

$$-\mathcal{L}_{W^\pm} = \frac{g}{\sqrt{2}} \overline{u_{Li}} \gamma^\mu (V_{uL} V_{dL}^\dagger)_{ij} d_{Lj} W_\mu^+ + \text{h.c.} \quad (12)$$

The 3×3 unitary matrix,

$$V_{\text{CKM}} = V_{uL}V_{dL}^\dagger, \quad (13)$$

is the CKM *mixing matrix* for quarks [7,4]. It generally depends on nine parameters: three real angles and six phases.

The form of the matrix is not unique. Usually, the following two conventions are employed:

- (i) There is freedom in defining V_{CKM} in that we can permute between the various generations. This freedom is fixed by ordering the up quarks and the down quarks by their masses, i.e. $m_{u_1} < m_{u_2} < m_{u_3}$ and $m_{d_1} < m_{d_2} < m_{d_3}$. (Usually, we call $(u_1, u_2, u_3) \rightarrow (u, c, t)$ and $(d_1, d_2, d_3) \rightarrow (d, s, b)$.) It is an interesting fact that with this convention V_{CKM} is close to a unit matrix. (See, for example, ref. [8] for a discussion of this point in the framework of horizontal symmetries.)
- (ii) There is further freedom in the phase structure of V_{CKM} . Let us define P_f ($f = u, d, \ell$) to be diagonal unitary (phase) matrices. Then, if instead of using V_{fL} and V_{fR} for the rotation (11) to the mass basis we use \tilde{V}_{fL} and \tilde{V}_{fR} , defined by $\tilde{V}_{fL} = P_f V_{fL}$ and $\tilde{V}_{fR} = P_f V_{fR}$, we still maintain a legitimate mass basis since M_f^{diag} remains unchanged by such transformations. However, V_{CKM} does change:

$$V_{\text{CKM}} \rightarrow P_u V_{\text{CKM}} P_d^*. \quad (14)$$

This freedom is fixed by demanding that V_{CKM} will have the minimal number of phases. In the three generation case V_{CKM} has a single phase. (There are five phase differences between the elements of P_u and P_d and, therefore, five of the six phases in the CKM matrix can be removed.) This is the Kobayashi-Maskawa phase δ_{KM} which is the single source of *CP violation* in the Standard Model [4].

As a result of the fact that V_{CKM} is not diagonal, the W^\pm gauge bosons can couple to quark (mass eigenstates) of different generations. Within the Standard Model, this is the only source of *flavor changing* interactions. In principle, there could be additional sources of flavor mixing in the lepton sector and in Z^0 interactions. We now explain why, within the Standard Model, this does not happen.

Mixing in the lepton sector: An analysis similar to the above applies also to the left-handed leptons. The mixing matrix is $(V_{\nu L}V_{\ell L}^\dagger)$. However, we can use the arbitrariness of $V_{\nu L}$ (related to the masslessness of neutrinos) to choose $V_{\nu L} = V_{\ell L}$, and the mixing matrix becomes a unit matrix. We conclude that the masslessness of neutrinos (if true) implies that there is no mixing in the lepton sector. If neutrinos have masses then the leptonic charged current interactions will exhibit mixing and CP violation.

Mixing in neutral current interactions: Defining $\tan \theta_W \equiv g'/g$, the Standard Model gives

$$Z^\mu = \cos \theta_W W_3^\mu - \sin \theta_W B^\mu. \quad (15)$$

(B is the gauge boson related to $U(1)_Y$.) Therefore, to study the interactions of the Z boson, we need to know the W_3 -interactions (given in (5)) and the B interactions:

$$-\mathcal{L}_B = -g' \left[\frac{1}{6} \overline{Q_{Li}^I} \gamma^\mu \mathbf{1}_{ij} Q_{Lj}^I + \frac{2}{3} \overline{u_{Ri}^I} \gamma^\mu \mathbf{1}_{ij} u_{Rj}^I - \frac{1}{3} \overline{d_{Ri}^I} \gamma^\mu \mathbf{1}_{ij} d_{Rj}^I \right] B_\mu. \quad (16)$$

Let us examine, for example, the Z -interactions with d_L in the mass basis:

$$\begin{aligned} -\mathcal{L}_Z &= \frac{g}{\cos \theta_W} \left(-\frac{1}{2} + \frac{1}{3} \sin^2 \theta_W \right) \overline{d_{Li}} \gamma^\mu (V_{dL}^\dagger V_{dL})_{ij} d_{Lj} Z_\mu \\ &= \frac{g}{\cos \theta_W} \left(-\frac{1}{2} + \frac{1}{3} \sin^2 \theta_W \right) \overline{d_{Li}} \gamma^\mu d_{Li} Z_\mu. \end{aligned} \quad (17)$$

We learn that the neutral current interactions remain universal in the mass basis and there are no additional flavor parameters in their description. This situation goes beyond the Standard Model to all models where all left-handed quarks are in $SU(2)_L$ doublets and all right-handed ones in singlets. The Z -boson does have flavor changing couplings in models where this is not the case.

How many flavor parameters are there in the Standard Model? In the interaction basis, the flavor parameters come from the three Yukawa matrices. Since each of these is a 3×3 complex matrix, there are 27 real and 27 imaginary parameters in these matrices. Not all of them are, however, physical. If we switch off the Yukawa matrices, there is a global symmetry added to the Standard Model,

$$G_{\text{global}}(Y^f = 0) = U(3)_Q \times U(3)_{\bar{d}} \times U(3)_{\bar{u}} \times U(3)_L \times U(3)_{\bar{\ell}}. \quad (18)$$

A unitary rotation of the three generations for each of the five representations in (3) would leave the Standard Model Lagrangian invariant. This means that the physics described by a given set of Yukawa matrices (Y^d, Y^u, Y^ℓ) , and the physics described by another set,

$$\tilde{Y}^d = V_Q^\dagger Y^d V_{\bar{d}}, \quad \tilde{Y}^u = V_Q^\dagger Y^u V_{\bar{u}}, \quad \tilde{Y}^\ell = V_L^\dagger Y^\ell V_{\bar{\ell}}, \quad (19)$$

where V are all unitary matrices, is the same. One can use this freedom to remove, at most, 15 real and 30 imaginary parameters (the number of parameters in five 3×3 unitary matrices). However, the fact that the Standard Model with the Yukawa matrices switched on has still a global symmetry of

$$G_{\text{global}} = U(1)_B \times U(1)_e \times U(1)_\mu \times U(1)_\tau \quad (20)$$

means that only 26 imaginary parameters can be removed. We conclude that there are 13 flavor parameters: 12 real ones and a single phase.

Examining the mass basis one can easily identify the flavor parameters. In the quark sector, we have six quark masses, three mixing angles (the number of real parameters in V_{CKM}) and the single phase δ_{KM} mentioned above. In the lepton sector, we have the three charged lepton masses.

2 THE MIXING PARAMETERS

While the fermion masses are determined from kinematics of various processes so that the values are model independent, the mixing parameters can only be determined from weak interaction processes and could be affected by new physics. There is an intensive experimental effort to measure the elements of the CKM matrix,

$$V_{\text{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}. \quad (21)$$

There are three ways to determine the CKM parameters:

- (i) Direct measurements: Standard Model tree level processes;
- (ii) Unitarity: relations among the CKM elements following from $V_{\text{CKM}}^\dagger V_{\text{CKM}} = \mathbf{1}$;
- (iii) Indirect measurements: Standard Model loop processes.

Direct measurements are expected to hold almost model independently. The reason is that viable extensions of the Standard Model have built-in mechanisms to suppress flavor changing processes in order that the strong constraints from FCNC are satisfied. These mechanisms make the contributions to Standard Model tree level processes highly suppressed. In most extensions of the Standard Model, the new physics takes place at a scale Λ_{NP} that is much higher than

the EW breaking scale and consequently the contributions to decay amplitudes are suppressed by $\mathcal{O}(m_Z^2/\Lambda_{\text{NP}}^2) \ll 1$. In the next subsection we briefly describe the determination of CKM elements by direct measurements.

Unitarity holds if the only quarks (namely, color triplets with electric charges $+2/3$ or $-1/3$) are those of the three generations of the Standard Model. In many extensions of the Standard Model, *e.g.* the minimal supersymmetric extension of the Standard Model, this is indeed the situation and unitarity is a valid way of determining CKM elements.

If there are additional quarks, which could be either sequential (fourth generation) or non-sequential (*e.g.* vector-like down quarks $D(3,1)_{-1/3} + \bar{D}(\bar{3},1)_{+1/3}$), and if these extra quarks mix with the observed quarks, then CKM unitarity is violated.

Indirect measurements are very sensitive to new physics. Take, for example, the $B - \bar{B}$ mixing amplitude. Within the Standard Model, the leading contribution comes from an EW box diagram and is therefore of $\mathcal{O}(g^4)$ and depends on small mixing angles, $|V_{tb}V_{td}|^2$. These suppression factors do not necessarily persist in extensions of the Standard Model. For example, in supersymmetric models there could be contributions of $\mathcal{O}(g_s^4)$ (gluino-mediated) and the mixing angles could be comparable to (or even larger than) the Standard Model ones. The validity of indirect measurements is then model dependent.

One can make however a generic statement about the relation between violation of CKM unitarity and the validity of indirect measurements [9]. Let us consider again the measurement of $|V_{tb}V_{td}|$ from Δm_B . In models with vector-like down quarks, there is a *tree level* (Z -mediated) contribution to this amplitude. In four generation models, the heavy mass of the t' quark gives an enhancement factor. In either case, whenever there is a non-negligible violation of the CKM unitarity, there will be a much more significant modification of the Standard Model predictions through large contributions to FCNC processes.

The most efficient way to investigate the mixing parameters is then the following:

- (1) Measure as many parameters as possible by direct measurements. At present we have $|V_{ud}|$, $|V_{us}|$, $|V_{ub}|$, $|V_{cd}|$, $|V_{cs}|$, $|V_{cb}|$ and $|V_{tb}|$.
- (2) Test whether the directly measured elements are consistent with unitarity. If there is consistency, determine the ‘missing’ parameters (or improve the determination of those measured with large errors) by using unitarity. At present we do so for $|V_{td}|$, $|V_{ts}|$, $|V_{tb}|$ and $|V_{cs}|$. If there is inconsistency, then most likely the quark sector extends beyond the three generations of the Standard Model.
- (3) Test the predictions for FCNC processes. If there is consistency, one can further improve the determination of poorly known CKM parameters. This is the case at present for $|V_{tb}V_{td}|$ (from Δm_B and Δm_{B_s}) and for δ_{KM} (from ε_K). If there is inconsistency, then New Physics has been discovered.

2.1 Direct Measurements

Seven of the nine absolute values of the CKM entries are measured directly, namely by tree level processes. (All numbers below are taken from [10].) Nuclear beta decays give

$$|V_{ud}| = 0.9740 \pm 0.0010. \quad (22)$$

Semileptonic kaon and hyperon decays give

$$|V_{us}| = 0.2196 \pm 0.0023. \quad (23)$$

Neutrino and antineutrino production of charm off valence d quarks give

$$|V_{cd}| = 0.224 \pm 0.016. \quad (24)$$

Semileptonic D decays give

$$|V_{cs}| = 1.04 \pm 0.16. \quad (25)$$

Semileptonic exclusive and inclusive B decays give

$$|V_{cb}| = 0.0395 \pm 0.0017. \quad (26)$$

The endpoint spectrum in semileptonic B decays gives

$$|V_{ub}/V_{cb}| = 0.08 \pm 0.02. \quad (27)$$

The decay $t \rightarrow b\ell^+\nu_\ell$ gives

$$|V_{tb}|^2/(|V_{tb}|^2 + |V_{ts}|^2 + |V_{td}|^2) = 0.99 \pm 0.29. \quad (28)$$

2.2 Unitarity of the CKM Matrix

The requirement of CKM unitarity is simply stated as $V_{\text{CKM}}^\dagger V_{\text{CKM}} = \mathbf{1}$. This leads to various relations among the matrix elements. The orthogonality between any two columns will be very useful in our discussion:

$$V_{ud}V_{us}^* + V_{cd}V_{cs}^* + V_{td}V_{ts}^* = 0, \quad (29)$$

$$V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* = 0, \quad (30)$$

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0. \quad (31)$$

Another class of unitarity constraints is given by $\sum_{i=1}^3 |V_{ij}|^2 = \sum_{j=1}^3 |V_{ij}|^2 = 1$. A particularly useful relation is

$$|V_{ub}|^2 + |V_{cb}|^2 + |V_{tb}|^2 = 1. \quad (32)$$

Using unitarity constraints, one can narrow down some of the ranges determined from direct measurements (most noticeably, that of $|V_{cs}|$) and put constraints on the top mixings $|V_{ti}|$. For example, the relation (32) and the very small measured values of $|V_{ub}|$ and $|V_{cb}|$ imply that, to an excellent approximation,

$$|V_{tb}| = 1. \quad (33)$$

The relation (30) and the small measured value of $|V_{us}V_{ub}|$ imply that, to a good approximation,

$$|V_{ts}| \approx |V_{cb}|. \quad (34)$$

The relation (31), together with $|V_{ub}/V_{cb}| \leq 0.10$ and $|V_{cd}/V_{ud}| = 0.22$, gives

$$|V_{td}V_{tb}| \approx 0.0085 \pm 0.0045. \quad (35)$$

The full information on the absolute values of the CKM elements from both direct measurements and three generation unitarity is summarized by [10]:

$$|V| = \begin{pmatrix} 0.9745 - 0.9760 & 0.217 - 0.224 & 0.0018 - 0.0045 \\ 0.217 - 0.224 & 0.9737 - 0.9753 & 0.036 - 0.046 \\ 0.004 - 0.013 & 0.035 - 0.042 & 0.9991 - 0.9994 \end{pmatrix}. \quad (36)$$

The unitarity of the CKM matrix is manifest using an explicit parameterization. There are various useful ways to parameterize it, but the standard choice [10] is the following [11]:

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix}, \quad (37)$$

where $c_{ij} \equiv \cos \theta_{ij}$ and $s_{ij} \equiv \sin \theta_{ij}$. A test of the CKM picture is then whether there is a range for the *four* parameters s_{12} , s_{23} , s_{13} and δ that is consistent with the *seven* direct measurements described in the previous subsection. Indeed, the following ranges are consistent with (36):

$$s_{12} = 0.2196 \pm 0.0023, \quad s_{23} = 0.0395 \pm 0.0017, \quad s_{13}/s_{23} = 0.08 \pm 0.02. \quad (38)$$

(The phase δ is not constrained at present by direct measurements.) Another useful parametrization is in terms of the four Wolfenstein parameters (λ, A, ρ, η) with $\lambda = |V_{us}| = 0.22$ playing the role of an expansion parameter and η representing the CP violating phase [12]:

$$V = \begin{pmatrix} 1 - \frac{\lambda^2}{2} & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1 - \frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3(1 - \rho - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4). \quad (39)$$

Because of the smallness of λ and the fact that for each element the expansion parameter is actually λ^2 , it is sufficient to keep only the first few terms in this expansion. The ranges in (36) can be translated into the following ranges of the Wolfenstein parameters:

$$\lambda = 0.2196 \pm 0.0023, \quad A = 0.819 \pm 0.035, \quad (\rho^2 + \eta^2)^{1/2} = 0.36 \pm 0.09. \quad (40)$$

The relation between the parameters of (37) and (39) is given by

$$s_{12} \equiv \lambda, \quad s_{23} \equiv A\lambda^2, \quad s_{13}e^{-i\delta} \equiv A\lambda^3(\rho - i\eta). \quad (41)$$

This specifies the higher order terms in (39).

2.3 Neutral Meson Mixing

The presently useful indirect measurements (Δm_B , Δm_{B_s} and ε_K) are all related to neutral meson mixing. Before presenting the implications of these measurements for the CKM parameters, we briefly discuss then the physics and formalism of neutral meson mixing. We refer specifically to the neutral B meson system, but most of our discussion applies equally well to the neutral K , B_s and D systems.

Our phase convention for the CP transformation law of the neutral B mesons is defined by

$$\text{CP}|B^0\rangle = \omega_B|\bar{B}^0\rangle, \quad \text{CP}|\bar{B}^0\rangle = \omega_B^*|B^0\rangle, \quad (|\omega_B| = 1). \quad (42)$$

Physical observables do not depend on the phase factor ω_B . An arbitrary linear combination of the neutral B -meson flavor eigenstates,

$$a|B^0\rangle + b|\bar{B}^0\rangle, \quad (43)$$

is governed by a time-dependent Schrödinger equation,

$$i\frac{d}{dt} \begin{pmatrix} a \\ b \end{pmatrix} = H \begin{pmatrix} a \\ b \end{pmatrix} \equiv \left(M - \frac{i}{2}\Gamma \right) \begin{pmatrix} a \\ b \end{pmatrix}, \quad (44)$$

for which M and Γ are 2×2 Hermitian matrices.

The off-diagonal terms in these matrices, M_{12} and Γ_{12} , are particularly important in the discussion of mixing and CP violation. M_{12} is the dispersive part of the transition amplitude from B^0 to \bar{B}^0 . In the Standard Model it arises only at order g^4 . In the language of quark diagrams, the leading contribution is from box diagrams. At sufficiently high loop momentum, $k \gg \Lambda_{\text{QCD}}$, these diagrams are a very good approximation to the Standard Model contribution to M_{12} . This, or any other contribution from heavy intermediate states from new physics, is

the *short distance* contribution. For small loop momenta, $k \lesssim 1$ GeV, we do not expect quark hadron duality to hold. The box diagram is a poor approximation to the contribution from light intermediate states, namely to *long distance* contributions. Fortunately, in the B and B_s systems, the long distance contributions are expected to be negligible. (This is not the case for K and D mesons. Consequently, it is difficult to extract useful information from the measurement of Δm_K and from the bound on Δm_D .) Γ_{12} is the absorptive part of the transition amplitude. Since the cut of a diagram always involves on-shell particles and thus long distance physics, the cut of the quark box diagram is a poor approximation to Γ_{12} . However, it does correctly give the suppression from small electroweak parameters such as the weak coupling. In other words, though the hadronic uncertainties are large and could change the result by order 50%, the cut in the box diagram is expected to give a reasonable order of magnitude estimate of Γ_{12} . (For $\Gamma_{12}(B_s)$ it has been shown that local quark-hadron duality holds exactly in the simultaneous limit of small velocity and large number of colors. We thus expect an uncertainty of $\mathcal{O}(1/N_C) \sim 30\%$ [13-14]. For $\Gamma_{12}(B_d)$ the small velocity limit is not as good an approximation but an uncertainty of order 50% still seems a reasonable estimate.) New physics is not expected to affect Γ_{12} significantly because it usually takes place at a high energy scale and is relevant to the short distance part only.

The light B_L and heavy B_H mass eigenstates are given by

$$|B_{L,H}\rangle = p|B^0\rangle \pm q|\bar{B}^0\rangle. \quad (45)$$

The complex coefficients q and p obey the normalization condition $|q|^2 + |p|^2 = 1$. Note that $\arg(q/p^*)$ is just an overall common phase for $|B_L\rangle$ and $|B_H\rangle$ and has no physical significance. The mass difference and the width difference between the physical states are given by

$$\Delta m \equiv M_H - M_L, \quad \Delta \Gamma \equiv \Gamma_H - \Gamma_L. \quad (46)$$

Solving the eigenvalue equation gives

$$(\Delta m)^2 - \frac{1}{4}(\Delta \Gamma)^2 = (4|M_{12}|^2 - |\Gamma_{12}|^2), \quad (47)$$

$$\Delta m \Delta \Gamma = 4\mathcal{R}e(M_{12}\Gamma_{12}^*),$$

$$\frac{q}{p} = -\frac{2M_{12}^* - i\Gamma_{12}^*}{\Delta m - \frac{i}{2}\Delta \Gamma} = -\frac{\Delta m - \frac{i}{2}\Delta \Gamma}{2M_{12} - i\Gamma_{12}}. \quad (48)$$

In the B system, $|\Gamma_{12}| \ll |M_{12}|$ (see discussion below), and then, to leading order in $|\Gamma_{12}/M_{12}|$, (47) and (48) can be written as

$$\Delta m_B = 2|M_{12}|, \quad \Delta \Gamma_B = 2\mathcal{R}e(M_{12}\Gamma_{12}^*)/|M_{12}|, \quad (49)$$

$$\frac{q}{p} = -\frac{M_{12}^*}{|M_{12}|}. \quad (50)$$

2.4 Indirect Measurements

The most useful CP conserving indirect measurement is that of Δm_B [15]:

$$\Delta m_B = 0.471 \pm 0.016 \text{ ps}^{-1}. \quad (51)$$

The Standard Model accounts for $\Delta m_B = 2|M_{12}|$ by box diagrams with intermediate top quarks [16]:

$$\Delta m_B = \frac{G_F^2}{6\pi^2} \eta_B m_B m_W^2 (B_B f_B^2) S_0(x_t) |V_{tb} V_{td}^*|^2, \quad (52)$$

where G_F is the Fermi constant, η_B is a QCD correction factor calculated in NLO [17], $S_0(x_t)$ is a kinematic function calculated from the box graphs [16], and $x_t = \bar{m}_t^2/m_W^2$. We use [18] $\bar{m}_t(m_t) = 167 \pm 6 \text{ GeV}$, giving $S_0(x_t) \approx 2.36$. $B - \bar{B}$ mixing is dominated by short distance physics (an intermediate top), so that the main source of theoretical uncertainty lies in the matrix element of the four quark operator between the meson states. The value of the matrix element is parameterized by $B_B f_B^2$ and is estimated by *e.g.* lattice calculations [10], $B_B f_B^2 = (1.4 \pm 0.1)(175 \pm 25 \text{ MeV})^2$. The constraint on the CKM parameters from (52) can be written as

$$|V_{tb}^* V_{td}| = 0.0086 \left[\frac{\Delta m_B}{0.471 \text{ ps}^{-1}} \right]^{1/2} \left[\frac{0.2 \text{ GeV}}{\sqrt{B_B f_B}} \right] \left[\frac{2.4}{S_0(x_t)} \right]^{1/2} \left[\frac{0.55}{\eta_B} \right]^{1/2}, \quad (53)$$

This constraint gives at present

$$|V_{tb}^* V_{td}| = 0.0084 \pm 0.0018, \quad (54)$$

which is consistent with, and actually significantly improves the unitarity constraint (35).

Another useful indirect measurement is that of Δm_{B_s} . The expression for Δm_{B_s} is very similar to (51), except for the CKM dependence and an $SU(3)$ breaking factor (that is an approximate global symmetry of the strong interactions that holds in the limit $m_u = m_d = m_s = 0$):

$$\frac{\Delta m_{B_d}}{\Delta m_{B_s}} = \frac{m_{B_d}}{m_{B_s}} \frac{B_{B_d} f_{B_d}^2}{B_{B_s} f_{B_s}^2} |V_{td}/V_{ts}|^2. \quad (55)$$

The uncertainty in the ratio between the matrix elements is smaller than the uncertainty in each of them separately [19]:

$$\frac{B_{B_s} f_{B_s}^2}{B_{B_d} f_{B_d}^2} = 1.30 \pm 0.18. \quad (56)$$

At present, there is only a lower bound [15], $\Delta m_{B_s} \geq 12.4 \text{ ps}^{-1}$, leading to

$$|V_{td}/V_{ts}| \leq 0.24 \implies |V_{td}| \leq 0.0096, \quad (57)$$

which further improves the upper bound of (54).

The imaginary part of the $K - \bar{K}$ mixing amplitude corresponds to the CP violating observable ε_K discussed in the next chapter:

$$\varepsilon_K = \frac{\exp(i\pi/4)}{\sqrt{2}} \frac{\text{Im } M_{12}}{\Delta m_K}. \quad (58)$$

The off-diagonal mass matrix element M_{12} is obtained from the $\Delta S = 2$ effective Hamiltonian with contributions from both the c -quark and the t -quark in the EW loop, yielding

$$M_{12} = \frac{G_F^2}{12\pi^2} f_K^2 B_K m_K m_W^2 \times \left[(V_{cd}^* V_{cs})^2 \eta_1 S_0(x_c) + (V_{td}^* V_{ts})^2 \eta_2 S_0(x_t) + 2(V_{cd}^* V_{cs})(V_{td}^* V_{ts}) \eta_3 S_0(x_c, x_t) \right], \quad (59)$$

where f_K is the kaon decay constant and η_i are QCD factors calculated in NLO [17,20]. $\text{Im } M_{12}$ is dominated by short distance physics (intermediate top quark), so that the main source of theoretical uncertainty lies in the matrix element [19], $B_K = 0.6 - 1$. The resulting constraint on the CKM parameters can be written (with the convention that $(V_{ud}^* V_{us})$ is real) as

$$\varepsilon_K = \exp(i\pi/4) C_{\varepsilon_K} B_K \text{Im}(V_{td}^* V_{ts}) \times \{ \mathcal{R}e(V_{cd}^* V_{cs}) [\eta_1 S_0(x_c) - \eta_3 S_0(x_c, x_t)] - \mathcal{R}e(V_{td}^* V_{ts}) \eta_2 S_0(x_t) \}, \quad (60)$$

where all well-known quantities have been combined in the numerical constant,

$$C_{\varepsilon_K} = \frac{G_F^2}{6\sqrt{2}\pi^2} \frac{f_K^2 m_K m_W^2}{\Delta m_K} = 3.78 \times 10^4. \quad (61)$$

In the future, we may get useful information about the CKM parameters from the two rare kaon decays, $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ [21] and $K_L \rightarrow \pi^0 \nu \bar{\nu}$ [22], which are theoretically very clean. Both modes are dominated by short distance Z -penguins and box diagrams. The branching ratio for $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ can be expressed in terms of ρ and η [18]:

$$BR(K^+ \rightarrow \pi^+ \nu \bar{\nu}) = 8.33 \times 10^{-6} |V_{cb}|^4 [X(x_t)]^2 [\eta^2 + (\rho_0 - \rho)^2], \quad (62)$$

where

$$\rho_0 = 1 + \frac{P_0(X)}{X(x_t)} \frac{\lambda^4}{|V_{cb}|^2}, \quad (63)$$

and $X(x_t)$ and $P_0(X)$ represent the electroweak loop contributions in NLO for the top quark and for the charm quark, respectively. The main theoretical uncertainty is related to the strong dependence of the charm contribution on the renormalization scale and the QCD scale, $P_0(X) = 0.40 \pm 0.06$. First evidence for $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ was presented recently [23]. The large experimental error does not yet give a useful CKM constraint and is consistent with the Standard Model prediction.

The $K_L \rightarrow \pi^0 \nu \bar{\nu}$ decay is CP violating and will be discussed later in detail. The branching ratio can be expressed in terms of η [18]:

$$BR(K_L \rightarrow \pi^0 \nu \bar{\nu}) = 3.29 \times 10^{-5} |V_{cb}|^4 [X(x_t)]^2 \eta^2. \quad (64)$$

The present experimental bound, $BR(K_L \rightarrow \pi^0 \nu \bar{\nu}) \leq 1.6 \times 10^{-6}$ [24] lies about five orders of magnitude above the Standard Model prediction [25] and about two orders of magnitude above the bound that can be deduced using model independent isospin relations [26] from the experimental upper bound on the charged mode.

2.5 The Unitarity Triangle

Each of the three relations (29)-(31) requires the sum of three complex quantities to vanish and so can be geometrically represented in the complex plane as a triangle. These are “the unitarity triangles”, though the term “unitarity triangle” is usually reserved for the relation (31) only. It is a surprising feature of the CKM matrix that all unitarity triangles are equal in area. For any choice of $i, j, k, l = 1, 2, 3$, one can define a quantity J according to [27]

$$\mathcal{I}m[V_{ij}V_{kl}V_{il}^*V_{kj}^*] = J \sum_{m,n=1}^3 \epsilon_{ikm}\epsilon_{jln}. \quad (65)$$

Then, the area of each unitarity triangle equals $|J|/2$ while the sign of J gives the direction of the complex vectors around the triangles. As will be discussed below, CP is violated in the Standard Model only if $J \neq 0$. The area of the triangles is then related to the size of the Standard Model CP violation.

The rescaled unitarity triangle is derived from (31) by (a) choosing a phase convention such that $(V_{cd}V_{cb}^*)$ is real, and (b) dividing the lengths of all sides by $|V_{cd}V_{cb}^*|$. Step (a) aligns one side of the triangle with the real axis, and step (b) makes the length of this side 1. The form of the triangle is unchanged. Two vertices of the rescaled unitarity triangle are thus fixed at (0,0) and (1,0). The coordinates of the remaining vertex correspond to the Wolfenstein parameters (ρ, η) (see (39)).

Depicting the rescaled unitarity triangle in the (ρ, η) plane, the lengths of the two complex sides are

$$R_u \equiv \sqrt{\rho^2 + \eta^2} = \frac{1}{\lambda} \left| \frac{V_{ub}}{V_{cb}} \right|, \quad R_t \equiv \sqrt{(1 - \rho)^2 + \eta^2} = \frac{1}{\lambda} \left| \frac{V_{td}}{V_{cb}} \right|. \quad (66)$$

The three angles of the unitarity triangle are denoted by α, β and γ [28]:

$$\alpha \equiv \arg \left[-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*} \right], \quad \beta \equiv \arg \left[-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*} \right], \quad \gamma \equiv \arg \left[-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*} \right]. \quad (67)$$

They are physical quantities and, we will soon see, can be independently measured by CP asymmetries in B decays.

The only large uncertainties in the present determination of the CKM elements are in $|V_{ub}|$ and $|V_{td}|$. However, the two are related through (31). Thus, the unitarity triangle is a very convenient tool for presenting constraints on these poorly determined parameters. In particular, Δm_{B_d} and Δm_{B_s} constrain R_t , the semileptonic $b \rightarrow u$ rates constrain R_u , and the ε_K constraint can be written as

$$\eta \{ (1 - \rho)\eta_2 S_0(x_t) |V_{cb}|^2 + \eta_3 S_0(x_c, x_t) - \eta_1 S_0(x_c) \} |V_{cb}|^2 B_K = 1.24 \times 10^{-6}. \quad (68)$$

Examining the Wolfenstein parametrization, we learn that $|J| = \mathcal{O}(\lambda^6) \times \sin \delta$. More precisely, the ranges specified above for the mixing angles give the following 90% CL range:

$$|J| = (2.7 \pm 0.7) \times 10^{-5} \sin \delta. \quad (69)$$

The measurement of ε_K is consistent with this range provided that

$$\sin \delta = \mathcal{O}(1). \quad (70)$$

(The phase δ is defined in eq. (37) and equals γ of eq. (67).) When all available information, including the ε_K constraint, is taken into account, we find the following allowed ranges for the CKM parameters [29-31]:

$$-0.15 \leq \rho \leq +0.35, \quad +0.20 \leq \eta \leq +0.45, \quad (71)$$

$$0.4 \leq \sin 2\beta \leq 0.8, \quad -0.9 \leq \sin 2\alpha \leq 1.0, \quad 0.23 \leq \sin^2 \gamma \leq 1.0. \quad (72)$$

3 CP VIOLATION IN MESON DECAYS: A MODEL INDEPENDENT DISCUSSION

3.1 Introduction

CP violation arises naturally in the three generation Standard Model. The CP violation that has been measured in neutral K -meson decays (ε_K) is accommodated in the Standard Model in simple way [4]. Yet, CP violation is one of the least tested aspects of the Standard Model. The value of the ε_K parameter [5] as well as bounds on other CP violating parameters (most noticeably, the electric dipole moments of the neutron, d_N , and of the electron, d_e) can be accounted for in models where CP violation has features that are very different from the Standard Model ones.

It is unlikely that the Standard Model provides the complete description of CP violation in nature. First, it is quite clear that there exists New Physics beyond the Standard Model. Almost any extension of the Standard Model has additional sources of CP violating effects. In addition there is a great puzzle in cosmology that relates to CP violation, and that is the baryon asymmetry of the universe [6]. Theories that explain the observed asymmetry must include

new sources of CP violation [32]: the Standard Model cannot generate a large enough matter-antimatter imbalance to produce the baryon number to entropy ratio observed in the universe today [33-35].

In the near future, significant new information on CP violation will be provided by various experiments. The main source of information will be measurements of CP violation in various B decays, particularly neutral B decays into final CP eigenstates [36-38]. Another piece of valuable information might come from a measurement of the $K_L \rightarrow \pi^0 \nu \bar{\nu}$ decay [39,40,22,26]. For the first time, the pattern of CP violation that is predicted by the Standard Model will be tested. Basic questions such as whether CP is an approximate symmetry in nature will be answered.

It could be that the scale where new CP violating sources appear is too high above the Standard Model scale (*e.g.* the GUT scale) to give any observable deviations from the Standard Model predictions. In such a case, the outcome of the experiments will be a (frustratingly) successful test of the Standard Model and a significant improvement in our knowledge of the CKM matrix.

A much more interesting situation will arise if the new sources of CP violation appear at a scale that is not too high above the electroweak scale. Then they might be discovered in the forthcoming experiments. Once enough independent observations of CP violating effects are made, we will find that there is no single choice of CKM parameters that is consistent with all measurements. There may even be enough information in the pattern of the inconsistencies to tell us something about the nature of the new physics contributions [9,41-43].

The aim of this and the next two chapters is to explain the theoretical tools with which we will analyze new information about CP violation. In this chapter, we give a brief, model-independent discussion of CP violating observables. In the next chapter, we discuss CP violation in the Standard Model. In the last chapter, we describe CP violation beyond the Standard Model and, in particular, in Supersymmetric models. The latter enables us to elucidate the uniqueness of the Standard Model description of CP violation and how little it has been tested so far. It further demonstrates how the information from CP violation can help us probe in detail models of New Physics.

3.2 Notations and Formalism

To understand the experimental and theoretical aspects of CP violation in meson decays, we first introduce some formalism. We continue the discussion of eqs. (42)-(47). Again, we specifically discuss the neutral B meson system, but large parts of our analysis apply equally well to the other meson systems.

To discuss CP violation in mixing (see below), it is useful to write (50) to first order in $|\Gamma_{12}/M_{12}|$:

$$\frac{q}{p} = -\frac{M_{12}^*}{|M_{12}|} \left[1 - \frac{1}{2} \mathcal{I}m \left(\frac{\Gamma_{12}}{M_{12}} \right) \right]. \quad (73)$$

To discuss CP violation in decay (see below), we need to consider decay amplitudes. The CP transformation law for a final state f is

$$\text{CP}|f\rangle = \omega_f |\bar{f}\rangle, \quad \text{CP}|\bar{f}\rangle = \omega_f^* |f\rangle, \quad (|\omega_f| = 1). \quad (74)$$

For a final CP eigenstate $f = f_{\text{CP}}$, the phase factor ω_f is replaced by $\eta_{f_{\text{CP}}} = \pm 1$, the CP eigenvalue of the final state. We define the decay amplitudes A_f and \bar{A}_f according to

$$A_f = \langle f | \mathcal{H}_d | B^0 \rangle, \quad \bar{A}_f = \langle f | \mathcal{H}_d | \bar{B}^0 \rangle, \quad (75)$$

where \mathcal{H}_d is the decay Hamiltonian.

To discuss CP violation in the interference of decays with and without mixing (see below), we introduce a complex quantity λ_f defined by

$$\lambda_f = \frac{q}{p} \frac{\bar{A}_f}{A_f}. \quad (76)$$

The effective Hamiltonian that is relevant to M_{12} is of the form

$$H_{\text{eff}}^{\Delta b=2} \propto e^{+2i\phi_B} [\bar{d}\gamma^\mu(1-\gamma_5)b]^2 + e^{-2i\phi_B} [\bar{b}\gamma^\mu(1-\gamma_5)d]^2, \quad (77)$$

where $2\phi_B$ is a CP violating (weak) phase. (We use the Standard Model $V-A$ amplitude, but the results can be generalized to any Dirac structure.) For the B system, where $|\Gamma_{12}| \ll |M_{12}|$, this leads to

$$q/p = \omega_B e^{-2i\phi_B}. \quad (78)$$

(We implicitly assumed that the vacuum insertion approximation gives the correct sign for M_{12} . In general, there is a $\text{sign}(B_B)$ factor on the right hand side of (78) [44].) The decay Hamiltonian is of the form

$$H_d \propto e^{+i\phi_f} [\bar{q}\gamma^\mu(1-\gamma_5)d] [\bar{b}\gamma_\mu(1-\gamma_5)q] + e^{-i\phi_f} [\bar{q}\gamma^\mu(1-\gamma_5)b] [\bar{d}\gamma_\mu(1-\gamma_5)q], \quad (79)$$

where ϕ_f is the appropriate weak phase. (Again, for simplicity we use a $V-A$ structure, but the results hold for any Dirac structure.) Then

$$\bar{A}_f/A_f = \omega_f \omega_B^* e^{-2i\phi_f}. \quad (80)$$

Eqs. (78) and (80) together imply that for a final CP eigenstate,

$$\lambda_{f\text{CP}} = \eta_{f\text{CP}} e^{-2i(\phi_B + \phi_f)}. \quad (81)$$

3.3 The Three Types of CP Violation in Meson Decays

There are three different types of CP violation in meson decays:

- (i) CP violation in mixing, which occurs when the two neutral mass eigenstate admixtures cannot be chosen to be CP-eigenstates;
- (ii) CP violation in decay, which occurs in both charged and neutral decays, when the amplitude for a decay and its CP-conjugate process have different magnitudes;
- (iii) CP violation in the interference of decays with and without mixing, which occurs in decays into final states that are common to B^0 and \bar{B}^0 . (It often occurs in combination with the other two types but there are cases when, to an excellent approximation, it is the only effect.)

(i) CP violation in mixing:

$$|q/p| \neq 1. \quad (82)$$

This results from the mass eigenstates being different from the CP eigenstates, and requires a relative phase between M_{12} and Γ_{12} . For the neutral B system, this effect could be observed through the asymmetries in semileptonic decays:

$$a_{\text{SL}} = \frac{\Gamma(\bar{B}_{\text{phys}}^0(t) \rightarrow \ell^+ \nu X) - \Gamma(B_{\text{phys}}^0(t) \rightarrow \ell^- \nu X)}{\Gamma(\bar{B}_{\text{phys}}^0(t) \rightarrow \ell^+ \nu X) + \Gamma(B_{\text{phys}}^0(t) \rightarrow \ell^- \nu X)}. \quad (83)$$

In terms of q and p ,

$$a_{\text{SL}} = \frac{1 - |q/p|^4}{1 + |q/p|^4}. \quad (84)$$

CP violation in mixing has been observed in the neutral K system ($\mathcal{R}e \varepsilon_K \neq 0$).

In the neutral B system, the effect is expected to be small, $\lesssim \mathcal{O}(10^{-2})$. The reason is that, model independently, the effect cannot be larger than $\mathcal{O}(\Delta\Gamma_B/\Delta m_B)$. The difference in width is produced by decay channels common to B^0 and \bar{B}^0 . The branching ratios for such channels are at or below the level of 10^{-3} . Since various channels contribute with differing signs, one expects that their sum does not exceed the individual level. Hence $\Delta\Gamma_B/\Gamma_B = \mathcal{O}(10^{-2})$ is a rather safe and model independent assumption. On the other hand, it is experimentally known that $\Delta m_B/\Gamma_B \approx 0.7$.

To calculate the deviation of $|q/p|$ from a pure phase (see (73)),

$$1 - \left| \frac{q}{p} \right| = \frac{1}{2} \mathcal{I}m \frac{\Gamma_{12}}{M_{12}}, \quad (85)$$

one needs to calculate M_{12} and Γ_{12} . This involves large hadronic uncertainties, in particular in the hadronization models for Γ_{12} .

(ii) CP violation in decay:

$$|\bar{A}_{\bar{f}}/A_f| \neq 1. \quad (86)$$

This appears as a result of interference among various terms in the decay amplitude, and will not occur unless at least two terms have different weak phases and different strong phases. CP asymmetries in charged B decays,

$$a_f = \frac{\Gamma(B^+ \rightarrow f^+) - \Gamma(B^- \rightarrow f^-)}{\Gamma(B^+ \rightarrow f^+) + \Gamma(B^- \rightarrow f^-)}, \quad (87)$$

are purely an effect of CP violation in decay. In terms of the decay amplitudes,

$$a_{f\pm} = \frac{1 - |\bar{A}_{f-}/A_{f+}|^2}{1 + |\bar{A}_{f-}/A_{f+}|^2}. \quad (88)$$

There is as yet no unambiguous experimental evidence for CP violation in decays. A measurement of $\mathcal{R}e \varepsilon'_K \neq 0$ [45,46] would constitute such evidence. It is also possible that the first unambiguous evidence for such CP violation will come from B decays, *e.g.* for $f^+ = \pi^+ K^0$.

There are two types of phases that may appear in A_f and $\bar{A}_{\bar{f}}$. Complex parameters in any Lagrangian term that contributes to the amplitude will appear in complex conjugate form in the CP-conjugate amplitude. Thus their phases appear in A_f and $\bar{A}_{\bar{f}}$ with opposite signs. In the Standard Model these phases occur only in the CKM matrix which is part of the electroweak sector of the theory, hence these are often called “weak phases”. The weak phase of any single term is convention dependent. However the difference between the weak phases in two different terms in A_f is convention independent because the phase rotations of the initial and final states are the same for every term. A second type of phase can appear in scattering or decay amplitudes even when the Lagrangian is real. Such phases do not violate CP, since they appear in A_f and $\bar{A}_{\bar{f}}$ with the same sign. Their origin is the possible contribution from intermediate on-shell states in the decay process, that is an absorptive part of an amplitude that has contributions from coupled channels. Usually the dominant rescattering is due to strong interactions and hence the designation “strong phases” for the phase shifts so induced. Again only the relative strong phases of different terms in a scattering amplitude have physical content, an overall phase rotation of the entire amplitude has no physical consequences.

Thus it is useful to write each contribution to A in three parts: its magnitude A_i ; its weak phase term $e^{i\phi_i}$; and its strong phase term $e^{i\delta_i}$. Then, if several amplitudes contribute to $B \rightarrow f$, we have

$$\left| \frac{\bar{A}_{\bar{f}}}{A_f} \right| = \left| \frac{\sum_i A_i e^{i(\delta_i - \phi_i)}}{\sum_i A_i e^{i(\delta_i + \phi_i)}} \right|. \quad (89)$$

The magnitude and strong phase of any amplitude involve long distance strong interaction physics, and our ability to calculate these from first principles is limited. Thus quantities that depend only on the weak phases are much cleaner than those that require knowledge of the relative magnitudes or strong phases of various amplitude contributions, such as CP violation in decay. There is however a large literature and considerable theoretical effort that goes into the calculation of amplitudes and strong phases. In many cases we can only relate experiment to Standard Model parameters through such calculations. The techniques that are used are expected to be more accurate for B decays than for K decays, because of the larger B mass, but theoretical uncertainty remains significant. The calculations generally contain two parts. First the operator product expansion and QCD perturbation theory are used to write any underlying quark process as a sum of local quark operators with well-determined coefficients. Then the matrix elements of the operators between the initial and final hadron states must be calculated. This is where theory is weakest and the results are most model dependent. Ideally lattice calculations should be able to provide accurate determinations for the matrix elements, and in certain cases this is already true, but much remains to be done.

(iii) CP violation in the interference between decays with and without mixing:

$$|\lambda_{f_{\text{CP}}}| = 1, \quad \mathcal{I}m \lambda_{f_{\text{CP}}} \neq 0. \quad (90)$$

Any $\lambda_{f_{\text{CP}}} \neq \pm 1$ is a manifestation of CP violation. The special case (90) isolates the effects of interest since both CP violation in decay (86) and in mixing (82) lead to $|\lambda_{f_{\text{CP}}}| \neq 1$. For the neutral B system, this effect can be observed by comparing decays into final CP eigenstates of a time-evolving neutral B state that begins at time zero as B^0 to those of the state that begins as \bar{B}^0 :

$$a_{f_{\text{CP}}} = \frac{\Gamma(\bar{B}_{\text{phys}}^0(t) \rightarrow f_{\text{CP}}) - \Gamma(B_{\text{phys}}^0(t) \rightarrow f_{\text{CP}})}{\Gamma(\bar{B}_{\text{phys}}^0(t) \rightarrow f_{\text{CP}}) + \Gamma(B_{\text{phys}}^0(t) \rightarrow f_{\text{CP}})}. \quad (91)$$

This time dependent asymmetry is given (for $|\lambda_{f_{\text{CP}}}| = 1$) by

$$a_{f_{\text{CP}}} = -\mathcal{I}m \lambda_{f_{\text{CP}}} \sin(\Delta m_B t). \quad (92)$$

CP violation in the interference of decays with and without mixing has been observed for the neutral K system ($\mathcal{I}m \varepsilon_K \neq 0$). It is expected to be an effect of $\mathcal{O}(1)$ in various B decays. For such cases, the contribution from CP violation in mixing is clearly negligible. For decays that are dominated by a single CP violating phase (for example, $B \rightarrow \psi K_S$ and $K_L \rightarrow \pi^0 \nu \bar{\nu}$), so that the contribution from CP violation in decay is also negligible, $a_{f_{\text{CP}}}$ is cleanly interpreted in terms of purely electroweak parameters. Explicitly, $\mathcal{I}m \lambda_{f_{\text{CP}}}$ gives the difference between the phase of the $B - \bar{B}$ mixing amplitude ($2\phi_B$) and twice the phase of the relevant decay amplitude ($2\phi_f$) (see eq. (81)):

$$\mathcal{I}m \lambda_{f_{\text{CP}}} = -\eta_{f_{\text{CP}}} \sin[2(\phi_B + \phi_f)]. \quad (93)$$

3.4 The ε_K Parameter

Historically, a different language from the one used by us has been employed to describe CP violation in $K \rightarrow \pi\pi$ and $K \rightarrow \pi\ell\nu$ decays. In this section we ‘translate’ the language of ε_K and ε'_K to our notations. Doing so will make it easy to understand which type of CP violation is related to each quantity.

The two CP violating quantities measured in neutral K decays are

$$\eta_{00} = \frac{\langle \pi^0 \pi^0 | \mathcal{H} | K_L \rangle}{\langle \pi^0 \pi^0 | \mathcal{H} | K_S \rangle}, \quad \eta_{+-} = \frac{\langle \pi^+ \pi^- | \mathcal{H} | K_L \rangle}{\langle \pi^+ \pi^- | \mathcal{H} | K_S \rangle}. \quad (94)$$

Define

$$\begin{aligned}
A_{00} &= \langle \pi^0 \pi^0 | \mathcal{H} | K^0 \rangle, & \bar{A}_{00} &= \langle \pi^0 \pi^0 | \mathcal{H} | \bar{K}^0 \rangle, \\
A_{+-} &= \langle \pi^+ \pi^- | \mathcal{H} | K^0 \rangle, & \bar{A}_{+-} &= \langle \pi^+ \pi^- | \mathcal{H} | \bar{K}^0 \rangle,
\end{aligned} \tag{95}$$

$$\lambda_{00} = \left(\frac{q}{p} \right)_K \frac{\bar{A}_{00}}{A_{00}}, \quad \lambda_{+-} = \left(\frac{q}{p} \right)_K \frac{\bar{A}_{+-}}{A_{+-}}. \tag{96}$$

Then

$$\eta_{00} = \frac{1 - \lambda_{00}}{1 + \lambda_{00}}, \quad \eta_{+-} = \frac{1 - \lambda_{+-}}{1 + \lambda_{+-}}. \tag{97}$$

The η_{00} and η_{+-} parameters get contributions from CP violation in mixing ($|(q/p)|_K \neq 1$) and from the interference of decays with and without mixing ($\mathcal{I}m \lambda_{ij} \neq 0$) at $\mathcal{O}(10^{-3})$ and from CP violation in decay ($|\bar{A}_{ij}/A_{ij}| \neq 1$) at $\mathcal{O}(10^{-6})$.

There are two isospin channels in $K \rightarrow \pi\pi$ leading to final $(2\pi)_{I=0}$ and $(2\pi)_{I=2}$ states:

$$\begin{aligned}
\langle \pi^0 \pi^0 | &= \sqrt{\frac{1}{3}} \langle (\pi\pi)_{I=0} | - \sqrt{\frac{2}{3}} \langle (\pi\pi)_{I=2} |, \\
\langle \pi^+ \pi^- | &= \sqrt{\frac{2}{3}} \langle (\pi\pi)_{I=0} | + \sqrt{\frac{1}{3}} \langle (\pi\pi)_{I=2} |.
\end{aligned} \tag{98}$$

The fact that there are two strong phases allows for CP violation in decay. The possible effects are, however, small (on top of the smallness of the relevant CP violating phases) because the final $I = 0$ state is dominant (this is the $\Delta I = 1/2$ rule). Defining

$$A_I = \langle (\pi\pi)_I | \mathcal{H} | K^0 \rangle, \quad \bar{A}_I = \langle (\pi\pi)_I | \mathcal{H} | \bar{K}^0 \rangle, \tag{99}$$

we have, experimentally,

$$|A_2/A_0| \approx 1/20. \tag{100}$$

Instead of η_{00} and η_{+-} we may define two combinations, ε_K and ε'_K , in such a way that the possible effects of CP violation in decay (mixing) are isolated into ε'_K (ε_K).

The experimental definition of the ε_K parameter is

$$\varepsilon_K \equiv \frac{1}{3}(\eta_{00} + 2\eta_{+-}). \tag{101}$$

To zeroth order in A_2/A_0 , we have $\eta_{00} = \eta_{+-} = \varepsilon_K$. However, the specific combination (101) is chosen in such a way that the following relation holds to *first* order in A_2/A_0 :

$$\varepsilon_K = \frac{1 - \lambda_0}{1 + \lambda_0}. \tag{102}$$

Since, by definition, only one strong channel contributes to λ_0 , there is indeed no CP violation in decay in (102). It is simple to show that $\mathcal{R}e \varepsilon_K \neq 0$ is a manifestation of CP violation in mixing while $\mathcal{I}m \varepsilon_K \neq 0$ is a manifestation of CP violation in the interference between decays with and without mixing. Since experimentally $\arg \varepsilon_K \approx \pi/4$, the two contributions are comparable.

The experimental definition of the ε'_K parameter is

$$\varepsilon'_K \equiv \frac{1}{3}(\eta_{+-} - \eta_{00}). \tag{103}$$

The theoretical expression is

$$\varepsilon'_K \approx \frac{1}{6}(\lambda_{00} - \lambda_{+-}). \tag{104}$$

Obviously, any type of CP violation which is independent of the final state does not contribute to ε'_K . Consequently, there is no contribution from CP violation in mixing to (104). It is simple to show that $\mathcal{R}e \varepsilon'_K \neq 0$ is a manifestation of CP violation in decay while $\mathcal{I}m \varepsilon'_K \neq 0$ is a manifestation of CP violation in the interference between decays with and without mixing.

3.5 CP violation in $K \rightarrow \pi\nu\bar{\nu}$

CP violation in the rare $K \rightarrow \pi\nu\bar{\nu}$ decays is very interesting. It is very different from the CP violation that has been observed in $K \rightarrow \pi\pi$ decays which is small and involves theoretical uncertainties. Similar to the CP asymmetry in $B \rightarrow \psi K_S$, it is predicted to be large and can be cleanly interpreted. Furthermore, observation of the $K_L \rightarrow \pi^0\nu\bar{\nu}$ decay at the rate predicted by the Standard Model will provide evidence that CP violation cannot be attributed to mixing ($\Delta F = 2$) processes only, as in superweak models.

Define the decay amplitudes

$$A_{\pi^0\nu\bar{\nu}} = \langle \pi^0\nu\bar{\nu} | \mathcal{H} | K^0 \rangle, \quad \bar{A}_{\pi^0\nu\bar{\nu}} = \langle \pi^0\nu\bar{\nu} | \mathcal{H} | \bar{K}^0 \rangle, \quad (105)$$

and the related $\lambda_{\pi\nu\bar{\nu}}$ quantity:

$$\lambda_{\pi\nu\bar{\nu}} = \left(\frac{q}{p} \right)_K \frac{\bar{A}_{\pi^0\nu\bar{\nu}}}{A_{\pi^0\nu\bar{\nu}}}. \quad (106)$$

The decay amplitudes of $K_{L,S}$ into a final $\pi^0\nu\bar{\nu}$ state are then

$$\langle \pi^0\nu\bar{\nu} | \mathcal{H} | \bar{K}_{L,S} \rangle = pA_{\pi^0\nu\bar{\nu}} \mp q\bar{A}_{\pi^0\nu\bar{\nu}}, \quad (107)$$

and the ratio between the corresponding decay rates is

$$\frac{\Gamma(K_L \rightarrow \pi^0\nu\bar{\nu})}{\Gamma(K_S \rightarrow \pi^0\nu\bar{\nu})} = \frac{1 + |\lambda_{\pi\nu\bar{\nu}}|^2 - 2\mathcal{R}e\lambda_{\pi\nu\bar{\nu}}}{1 + |\lambda_{\pi\nu\bar{\nu}}|^2 + 2\mathcal{R}e\lambda_{\pi\nu\bar{\nu}}}. \quad (108)$$

We learn that the $K_L \rightarrow \pi^0\nu\bar{\nu}$ decay rate vanishes in the CP limit ($\lambda_{\pi\nu\bar{\nu}} = 1$), as expected on general grounds [39].

Since the effects of CP violation in decay and in mixing are expected to be negligibly small, $\lambda_{\pi\nu\bar{\nu}}$ is, to an excellent approximation, a pure phase. Defining θ_K to be the relative phase between the $K - \bar{K}$ mixing amplitude and the $s \rightarrow d\nu\bar{\nu}$ decay amplitude, namely $\lambda_{\pi\nu\bar{\nu}} = e^{2i\theta_K}$, we get from (108):

$$\frac{\Gamma(K_L \rightarrow \pi^0\nu\bar{\nu})}{\Gamma(K_S \rightarrow \pi^0\nu\bar{\nu})} = \frac{1 - \cos 2\theta_K}{1 + \cos 2\theta_K} = \tan^2 \theta_K. \quad (109)$$

Using the isospin relation $A(K^0 \rightarrow \pi^0\nu\bar{\nu})/A(K^+ \rightarrow \pi^+\nu\bar{\nu}) = 1/\sqrt{2}$, we get

$$a_{\pi\nu\bar{\nu}} \equiv \frac{\Gamma(K_L \rightarrow \pi^0\nu\bar{\nu})}{\Gamma(K^+ \rightarrow \pi^+\nu\bar{\nu})} = \frac{1 - \cos 2\theta_K}{2} = \sin^2 \theta_K. \quad (110)$$

Note that $a_{\pi\nu\bar{\nu}} \leq 1$, and consequently a measurement of $\Gamma(K^+ \rightarrow \pi^+\nu\bar{\nu})$ can be used to set a model independent upper limit on $\Gamma(K_L \rightarrow \pi^0\nu\bar{\nu})$ [26].

3.6 CP Violation in $D \rightarrow K\pi$ Decays

Within the Standard Model, $D - \bar{D}$ mixing is expected to be well below the experimental bound. Furthermore, effects related to CP violation in $D - \bar{D}$ mixing are expected to be negligibly small since this mixing is described to a good approximation by physics of the first two generations. An experimental observation of $D - \bar{D}$ mixing close to the present bound (and, obviously, of related CP violation) will then be evidence for New Physics. We now give the formalism of the neutral D system for the case that the mixing is close to the present bounds.

We define the neutral D mass eigenstates:

$$|D_{1,2}\rangle = p|D^0\rangle \pm q|\bar{D}^0\rangle. \quad (111)$$

We define the following four decay amplitudes:

$$\begin{aligned}
A_{K^+\pi^-} &= \langle K^+\pi^- | \mathcal{H} | D^0 \rangle, & \bar{A}_{K^+\pi^-} &= \langle K^+\pi^- | \mathcal{H} | \bar{D}^0 \rangle, \\
A_{K^-\pi^+} &= \langle K^-\pi^+ | \mathcal{H} | D^0 \rangle, & \bar{A}_{K^-\pi^+} &= \langle K^-\pi^+ | \mathcal{H} | \bar{D}^0 \rangle.
\end{aligned}
\tag{112}$$

We introduce the following two quantities:

$$\lambda_{K^+\pi^-} = \left(\frac{q}{p}\right)_D \frac{\bar{A}_{K^+\pi^-}}{A_{K^+\pi^-}}, \quad \lambda_{K^-\pi^+} = \left(\frac{q}{p}\right)_D \frac{\bar{A}_{K^-\pi^+}}{A_{K^-\pi^+}}.
\tag{113}$$

The following approximations are all experimentally confirmed:

$$\Delta m_D \ll \Gamma_D; \quad \Delta \Gamma_D \ll \Gamma_D; \quad |\lambda_{K^+\pi^-}^{-1}| \ll 1; \quad |\lambda_{K^-\pi^+}| \ll 1.
\tag{114}$$

We are interested in the case that Δm_D is found to be close to the present bound. (As mentioned above, this could only happen with new physics beyond the Standard Model.) Then, we can make a second approximation:

$$\Delta \Gamma_D \ll \Delta m_D.
\tag{115}$$

We further make the reasonable assumptions that CP violation in decay is negligible:

$$\left| \frac{A_{K^+\pi^-}}{\bar{A}_{K^-\pi^+}} \right| = \left| \frac{\bar{A}_{K^+\pi^-}}{A_{K^-\pi^+}} \right| = 1,
\tag{116}$$

and that CP violation in mixing is negligible:

$$\left| \left(\frac{q}{p}\right)_D \right| = 1.
\tag{117}$$

With (116) and (117), we find

$$|\lambda_{K^+\pi^-}^{-1}| = |\lambda_{K^-\pi^+}| \equiv |\lambda_{K\pi}|.
\tag{118}$$

For the observation of mixing, we are interested in the state $|D^0(t)\rangle$ that starts out as a pure $|D^0\rangle$ at $t = 0$ and in the state $|\bar{D}^0(t)\rangle$ that starts out as a pure $|\bar{D}^0\rangle$. The result of the above discussion is the following form for the (time dependent) ratio between the doubly Cabibbo suppressed and Cabibbo allowed rates:

$$\begin{aligned}
\frac{\Gamma[D^0(t) \rightarrow K^+\pi^-]}{\Gamma[D^0(t) \rightarrow K^-\pi^+]} &= |\lambda_{K\pi}|^2 + \frac{(\Delta m_D)^2}{4} t^2 + \mathcal{I}m(\lambda_{K^+\pi^-}^{-1}) \Delta m_D t, \\
\frac{\Gamma[\bar{D}^0(t) \rightarrow K^-\pi^+]}{\Gamma[\bar{D}^0(t) \rightarrow K^+\pi^-]} &= |\lambda_{K\pi}|^2 + \frac{(\Delta m_D)^2}{4} t^2 + \mathcal{I}m(\lambda_{K^-\pi^+}) \Delta m_D t.
\end{aligned}
\tag{119}$$

(These are approximate expressions that hold for time $t \lesssim \frac{1}{\Gamma_D}$.)

The linear term is potentially CP violating. There are four possibilities concerning this term [47,48]:

- (i) $\mathcal{I}m(\lambda_{K^+\pi^-}^{-1}) = \mathcal{I}m(\lambda_{K^-\pi^+}) = 0$: both strong and weak phases play no role in these processes.
- (ii) $\mathcal{I}m(\lambda_{K^+\pi^-}^{-1}) = \mathcal{I}m(\lambda_{K^-\pi^+}) \neq 0$: weak phases play no role in these processes. There is a different strong phase shift in $D^0 \rightarrow K^+\pi^-$ and $D^0 \rightarrow K^-\pi^+$.
- (iii) $\mathcal{I}m(\lambda_{K^+\pi^-}^{-1}) = -\mathcal{I}m(\lambda_{K^-\pi^+}) \neq 0$: strong phases play no role in these processes. CP violating phases affect the mixing amplitude.
- (iv) $|\mathcal{I}m(\lambda_{K^+\pi^-}^{-1})| \neq |\mathcal{I}m(\lambda_{K^-\pi^+})|$: both strong and weak phases play a role in these processes.

4 CP VIOLATION IN THE STANDARD MODEL

4.1 Introduction

The irremovable phase in the CKM matrix allows CP violation. Recalling the CP transformation laws,

$$\bar{\psi}_i \psi_j \rightarrow \bar{\psi}_j \psi_i, \quad \bar{\psi}_i \gamma^\mu W_\mu (1 - \gamma_5) \psi_j \rightarrow \bar{\psi}_j \gamma^\mu W_\mu (1 - \gamma_5) \psi_i, \quad (120)$$

we learn that mass terms and gauge interactions can be CP invariant if the masses and couplings are real. In particular, consider the coupling of W^\pm to quarks. It has the form

$$g V_{ij} \bar{u}_i \gamma_\mu W^{+\mu} (1 - \gamma_5) d_j + g V_{ij}^* \bar{d}_j \gamma_\mu W^{-\mu} (1 - \gamma_5) u_i. \quad (121)$$

The CP operation interchanges the two terms except that V_{ij} and V_{ij}^* are not interchanged. Thus, CP is a good symmetry only if there is a mass basis where all couplings and masses are real.

CP is not necessarily violated in the three generation SM. If two quarks of the same charge had equal masses, one mixing angle and the phase could be removed from V_{CKM} . Thus CP violation requires

$$(m_t^2 - m_c^2)(m_c^2 - m_u^2)(m_t^2 - m_u^2)(m_b^2 - m_s^2)(m_s^2 - m_d^2)(m_b^2 - m_d^2) \neq 0. \quad (122)$$

If the value of any of the three mixing angles were 0 or $\pi/2$, then again the phase could be removed. Finally, CP would not be violated if the value of the single phase were 0 or π . These last eight conditions are elegantly incorporated into one, parameterization independent condition, that is (see (65) for the definition of J):

$$J \neq 0. \quad (123)$$

(In the parameterization (37) $J = c_{12} c_{23} c_{13}^2 s_{12} s_{23} s_{13} \sin \delta$. This shows explicitly that $J \neq 0$ is equivalent to $\theta_{ij} \neq 0, \pi/2$ and $\delta \neq 0, \pi$.) The fourteen conditions incorporated in (122) and (123) can all be written as a single condition on the mass matrices in the interaction basis [27]:

$$\mathcal{I}m\{\det[M_d M_d^\dagger, M_u M_u^\dagger]\} \neq 0 \Leftrightarrow \text{CP violation}. \quad (124)$$

The quantity J is of much interest in the study of CP violation from the CKM matrix. The maximum value that J might assume is $1/(6\sqrt{3}) \approx 0.1$, but in reality it is $\sim 3 \times 10^{-5}$.

Since the Standard Model contains only a single independent CP-violating phase, all possible CP-violating effects in this theory are very closely related. Consequently, the pattern of CP-violations in B decays is strongly constrained. The goal of B factories is to test whether this pattern occurs in Nature.

4.2 CP Violation in Mixing

In the B_d system we expect that $\Gamma_{12} \ll M_{12}$ model independently. Using the SM box diagrams to estimate the two quantities [49], one gets

$$\frac{\Gamma_{12}}{M_{12}} = -\frac{3\pi}{2} \frac{1}{f_2(m_t^2/m_W^2)} \frac{m_b^2}{m_t^2} \left(1 + \frac{8}{3} \frac{m_c^2}{m_b^2} \frac{V_{cb} V_{cd}^*}{V_{tb} V_{td}^*} \right). \quad (125)$$

This confirms our order of magnitude estimate, $|\Gamma_{12}/M_{12}| \lesssim 10^{-2}$. CP violation in mixing is proportional to $\mathcal{I}m(\Gamma_{12}/M_{12})$ which is even further suppressed:

$$1 - \left| \frac{q}{p} \right| = \frac{1}{2} \mathcal{I}m \frac{\Gamma_{12}}{M_{12}} = \frac{4\pi}{f_2(m_t^2/m_W^2)} \frac{m_c^2}{m_t^2} \frac{J}{|V_{tb} V_{td}^*|^2} \sim 10^{-3}. \quad (126)$$

Note that the suppression comes from the (m_c^2/m_t^2) factor. The last term is the ratio of the area of the unitarity triangle to the length of one of its sides squared, so it is $\mathcal{O}(1)$. In contrast, for the B_s system, where (126) holds except that V_{td} is replaced by V_{ts} , there is an additional suppression from $J/|V_{tb}V_{ts}^*|^2 \sim 10^{-2}$ (see the corresponding unitarity triangle).

The above estimate of CP violation in mixing suffers from large uncertainties (of order 30% [13] or even higher [50]) related to the use of a quark diagram to describe Γ_{12} .

4.3 CP Violation in Hadronic Decays of Neutral B

In the previous subsection we estimated the effect of CP violation in mixing to be of $\mathcal{O}(10^{-3})$ within the Standard Model, and $\leq \mathcal{O}(|\Gamma_{12}/M_{12}|) \sim 10^{-2}$ model independently [51]. In semileptonic decays, CP violation in mixing is the leading effect and therefore it can be measured through a_{SL} . In purely hadronic B decays, however, CP violation in decay and in the interference of decays with and without mixing is $\gtrsim \mathcal{O}(10^{-2})$. We can therefore safely neglect CP violation in mixing in the following discussion and use

$$\frac{q}{p} = -\frac{M_{12}^*}{|M_{12}|} = \frac{V_{tb}^*V_{td}}{V_{tb}V_{td}^*}\omega_B. \quad (127)$$

A crucial question is then whether CP violation in decay is comparable to the CP violation in the interference of decays with and without mixing or negligible. In the first case, we can use the corresponding charged B decays to observe effects of CP violation in decay. In the latter case, CP asymmetries in neutral B decays are subject to clean theoretical interpretation: we will either have precise measurements of CKM parameters or be provided with unambiguous evidence for new physics. The question of the relative size of CP violation in decay can only be answered on a channel by channel basis, which is what we do in this section.

Most channels have contributions from both tree- and three types of penguin-diagrams, the latter classified according to the identity of the quark in the loop, as diagrams with different intermediate quarks may have both different strong phases and different weak phases [52]. On the other hand, the subdivision of tree processes into spectator, exchange and annihilation diagrams is unimportant in this respect since they all carry the same weak phase.

While quark diagrams can be easily classified in this way, the description of B decays is not so neatly divided into tree and penguin contributions once long distance physics effects are taken into account. Rescattering processes can change the quark content of the final state and confuse the identification of a contribution. There is no physical distinction between rescattered tree diagrams and long-distance contributions to the cuts of a penguin diagram. While these issues complicate estimates of various rates, they can always be avoided in describing the weak phase structure of B -decay amplitudes. The decay amplitudes for $b \rightarrow q\bar{q}q'$ can always be written as a sum of three terms with definite CKM coefficients:

$$A(q\bar{q}q') = V_{tb}V_{tq'}^*P_{q'}^t + V_{cb}V_{cq'}^*(T_{c\bar{c}q'}\delta_{qc} + P_{q'}^c) + V_{ub}V_{uq'}^*(T_{u\bar{u}q'}\delta_{qu} + P_{q'}^u). \quad (128)$$

Here P and T denote contributions from tree and penguin diagrams, excluding the CKM factors. As they stand, the P terms are not well defined because of the divergences of the penguin diagrams. Only differences of penguin diagrams are finite and well defined. However already we see that diagrams that can be mixed by rescattering effects always appear with the same CKM coefficients and hence that a separation of these terms is not needed when discussing weak phase structure. Now it is useful to use eqs. (30) and (31) to eliminate one of the three terms, by writing its CKM coefficient as minus the sum of the other two.

In the case of $q\bar{q}s$ decays it is convenient to remove the $V_{tb}V_{ts}^*$ term. Then

$$\begin{aligned}
A(c\bar{c}s) &= V_{cb}V_{cs}^*(T_{c\bar{c}s} + P_s^c - P_s^t) + V_{ub}V_{us}^*(P_s^u - P_s^t), \\
A(u\bar{u}s) &= V_{cb}V_{cs}^*(P_s^c - P_s^t) + V_{ub}V_{us}^*(T_{u\bar{u}s} + P_s^u - P_s^t), \\
A(s\bar{s}s) &= V_{cb}V_{cs}^*(P_s^c - P_s^t) + V_{ub}V_{us}^*(P_s^u - P_s^t).
\end{aligned} \tag{129}$$

In these expressions only differences of penguin contributions occur, which makes the cancellation of the ultraviolet divergences of these diagrams explicit. Furthermore, the second term has a CKM coefficient that is much smaller, by $\mathcal{O}(\lambda^2)$, than the first. Hence this grouping is useful in classifying the expected CP violation in decay. (Note that terms $b \rightarrow d\bar{d}s$, which have only penguin contributions, mix strongly with the $u\bar{u}s$ terms and hence cannot be separated from them. Thus P terms in $A(u\bar{u}s)$ include contributions from both $d\bar{d}s$ and $u\bar{u}s$ diagrams.)

In the case of $q\bar{q}d$ decays the three CKM coefficients are of similar magnitude. The convention is then to retain the $V_{tb}V_{td}^*$ term because, in the Standard Model, the phase difference between this weak phase and half the mixing weak phase is zero. Thus only one unknown weak phase enters the calculation of the interference between decays with and without mixing. We can choose to eliminate which of the other terms does not have a tree contribution. In the cases $q = s$ or d , since neither has a tree contribution either term can be removed. Thus we write

$$\begin{aligned}
A(c\bar{c}d) &= V_{tb}V_{td}^*(P_d^t - P_d^u) + V_{cb}V_{cd}^*(T_{c\bar{c}d} + P_d^c - P_d^u), \\
A(u\bar{u}d) &= V_{tb}V_{td}^*(P_d^t - P_d^c) + V_{ub}V_{ud}^*(T_{u\bar{u}d} + P_d^u - P_d^c), \\
A(s\bar{s}d) &= V_{tb}V_{td}^*(P_d^t - P_d^u) + V_{cb}V_{cd}^*(P_d^c - P_d^u).
\end{aligned} \tag{130}$$

Again only differences of penguin amplitudes occur. Furthermore the difference of penguin terms that occurs in the second term would vanish if the charm and up quark masses were equal, and thus is GIM suppressed [53]. However, even in modes with no tree contribution, ($s\bar{s}d$), the interference of the two terms can still give significant CP violation in the interference of decays with and without mixing.

The penguin processes all involve the emission of a neutral boson, either a gluon (strong penguins) or a photon or Z boson (electroweak penguins). Excluding the CKM coefficients, the ratio of the contribution from the difference between a top and light quark strong penguin diagram to the contribution from a tree diagram is of order

$$r_{PT} = \frac{P^t - P^{light}}{T_{q\bar{q}q'}} \approx \frac{\alpha_s}{12\pi} \ln \frac{m_t^2}{m_b^2}. \tag{131}$$

This is a factor of $\mathcal{O}(0.03)$. However this estimate does not include the effect of hadronic matrix elements, which are the probability factor to produce a particular final state particle content from a particular quark content. Since this probability differs for different kinematics, color flow and spin structures, it can be different for tree and penguin contributions and may partially compensate the coupling constant suppression of the penguin term. Recent CLEO results on $BR(B \rightarrow K\pi)$ and $BR(B \rightarrow \pi\pi)$ [54] suggest that the matrix element of penguin operators is indeed enhanced compared to that of tree operators. The enhancement could be by a factor of a few, leading to

$$r_{PT} \sim \lambda^2 - \lambda. \tag{132}$$

(Note that r_{PT} does not depend on the CKM parameters. We use powers of the Wolfenstein parameter λ to quantify our estimate for r_{PT} is order to simplify the comparison between the size of CP violation in decay and CP violation in the interference between decays with and without mixing.) Electroweak penguin difference terms are even more suppressed since they have an α_{EM} or α_W instead of the α_s factor in (131), but certain Z -contributions are enhanced by the large top quark mass and so can be non-negligible.

We thus classify B decays into four classes. Classes (i) and (ii) are expected to have relatively small CP violation in decay and hence are particularly interesting for extracting CKM parameters from interference of decays with and without mixing. In the remaining two classes, CP violation in decay could be significant and the neutral decay asymmetries cannot be cleanly interpreted in terms of CKM phases.

- (i) Decays dominated by a single term: $b \rightarrow c\bar{c}s$ and $b \rightarrow s\bar{s}s$. The Standard Model cleanly predicts very small CP violation in decay: $\mathcal{O}(\lambda^4 - \lambda^3)$ for $b \rightarrow c\bar{c}s$ and $\mathcal{O}(\lambda^2)$ for $b \rightarrow s\bar{s}s$. Any observation of large CP asymmetries in charged B decays for these channels would be a clue to physics beyond the Standard Model. The corresponding neutral modes have cleanly predicted relationships between CKM parameters and the measured asymmetry from interference between decays with and without mixing. The modes $B \rightarrow \psi K$ and $B \rightarrow \phi K$ are examples of this class.
- (ii) Decays with a small second term: $b \rightarrow c\bar{c}d$ and $b \rightarrow u\bar{u}d$. The expectation that penguin-only contributions are suppressed compared to tree contributions suggests that these modes will have small effects of CP violation in decay, of $\mathcal{O}(\lambda^2 - \lambda)$, and an approximate prediction for the relationship between measured asymmetries in neutral decays and CKM phases can be made. Examples here are $B \rightarrow DD$ and $B \rightarrow \pi\pi$.
- (iii) Decays with a suppressed tree contribution: $b \rightarrow u\bar{u}s$. The tree amplitude is suppressed by small mixing angles, $V_{ub}V_{us}$. The no-tree term may be comparable or even dominate and give large interference effects. An example is $B \rightarrow \rho K$.
- (iv) Decays with no tree contribution: $b \rightarrow s\bar{s}d$. Here the interference comes from penguin contributions with different charge 2/3 quarks in the loop and gives CP violation in decay that could be as large as 10% [55,56]. An example is $B \rightarrow KK$.

Note that if the penguin enhancement is significant, then some of the decay modes listed in class (ii) might actually fit better in class (iii). For example, it is possible that $b \rightarrow u\bar{u}d$ decays have comparable contributions from tree and penguin amplitudes. On the other hand, this would also mean that some modes listed in class (iii) could be dominated by a single penguin term. For such cases an approximate relationship between measured asymmetries in neutral decays and CKM phases can be made.

4.4 CP violation in the interference between B decays with and without mixing

Let us first discuss an example of class (i), $B \rightarrow \psi K_S$. A new ingredient in the analysis is the effect of $K - \bar{K}$ mixing. For decays with a single K_S in the final state, $K - \bar{K}$ mixing is essential because $B^0 \rightarrow K^0$ and $\bar{B}^0 \rightarrow \bar{K}^0$, and interference is possible only due to $K - \bar{K}$ mixing. This adds a factor of

$$\left(\frac{p}{q}\right)_K = \frac{V_{cs}V_{cd}^*}{V_{cs}^*V_{cd}}\omega_K^* \quad (133)$$

into (\bar{A}/A) . The quark subprocess in $\bar{B}^0 \rightarrow \psi\bar{K}^0$ is $b \rightarrow c\bar{c}s$ which is dominated by the W -mediated tree diagram:

$$\frac{\bar{A}_{\psi K_S}}{A_{\psi K_S}} = \eta_{\psi K_S} \left(\frac{V_{cb}V_{cs}^*}{V_{cb}^*V_{cs}}\right) \left(\frac{V_{cs}V_{cd}^*}{V_{cs}^*V_{cd}}\right)\omega_B^*. \quad (134)$$

The CP-eigenvalue of the state is $\eta_{\psi K_S} = -1$. Combining (127) and (134), we find

$$\lambda(B \rightarrow \psi K_S) = - \left(\frac{V_{tb}^*V_{td}}{V_{tb}V_{td}^*}\right) \left(\frac{V_{cb}V_{cs}^*}{V_{cb}^*V_{cs}}\right) \left(\frac{V_{cd}^*V_{cb}}{V_{cd}V_{cb}^*}\right) \implies \mathcal{I}m\lambda_{\psi K_S} = \sin(2\beta). \quad (135)$$

The second term in (129) is of order $\lambda^2 r_{PT}$ for this decay and thus eq. (135) is clean of hadronic uncertainties to $\mathcal{O}(10^{-3})$. Consequently, this measurement can give the theoretically

cleanest determination of a CKM parameter, even cleaner than the determination of $|V_{us}|$ from $K \rightarrow \pi \ell \nu$. (If $\text{BR}(K_L \rightarrow \pi \nu \bar{\nu})$ is measured, it will give a comparably clean determination of η .)

A second example of a theoretically clean mode in class (i) is $B \rightarrow \phi K_S$. The quark subprocess involves FCNC and cannot proceed via a tree level SM diagram. The leading contribution comes from penguin diagrams. The two terms in eq. (129) are now both differences of penguins but the second term is CKM suppressed and thus of $\mathcal{O}(\lambda^2)$ compared to the first. Thus CP violation in the decay is at most a few percent, and can be neglected in the analysis of asymmetries in this channel. The analysis is similar to the ψK_S case, and the asymmetry is proportional to $\sin(2\beta)$:

The same quark subprocesses give theoretically clean CP asymmetries also in B_s decays. These asymmetries are, however, very small since the relative phase between the mixing amplitude and the decay amplitudes (β_s defined below) is very small.

The best known example of class (ii) is $B \rightarrow \pi\pi$. The quark subprocess is $b \rightarrow u\bar{u}d$ which is dominated by the W -mediated tree diagram. Neglecting for the moment the second, pure penguin, term in eq. (130) we find

$$\frac{\bar{A}_{\pi\pi}}{A_{\pi\pi}} = \eta_{\pi\pi} \frac{V_{ub}V_{ud}^*}{V_{ub}^*V_{ud}} \omega_B^* \quad (136)$$

The CP eigenvalue for two pions is $+1$. Combining (127) and (136), we get

$$\lambda(B \rightarrow \pi^+\pi^-) = \left(\frac{V_{tb}^*V_{td}}{V_{tb}V_{td}^*} \right) \left(\frac{V_{ud}^*V_{ub}}{V_{ud}V_{ub}^*} \right) \implies \mathcal{I}m\lambda_{\pi\pi} = \sin(2\alpha). \quad (137)$$

The pure penguin term in eq. (130) has a weak phase, $\arg(V_{td}^*V_{tb})$, different from the term with the tree contribution, so it modifies both $\mathcal{I}m\lambda$ and (if there are non-trivial strong phases) $|\lambda|$. The recent CLEO results mentioned above suggest that the penguin contribution to $B \rightarrow \pi\pi$ channel is significant, probably 10% or more. This then introduces CP violation in decay, unless the strong phases cancel (or are zero, as suggested by factorization arguments). The resulting hadronic uncertainty can be eliminated using isospin analysis [57]. This requires a measurement of the rates for the isospin-related channels $B^+ \rightarrow \pi^+\pi^0$ and $B^0 \rightarrow \pi^0\pi^0$ as well as the corresponding CP-conjugate processes. The rate for $\pi^0\pi^0$ is expected to be small and the measurement is difficult, but even an upper bound on this rate can be used to limit the magnitude of hadronic uncertainties [58].

Related but slightly more complicated channels with the same underlying quark structure are $B \rightarrow \rho^0\pi^0$ and $B \rightarrow a_1^0\pi^0$. Again an analysis involving the isospin-related channels can be used to help eliminate hadronic uncertainties from CP violations in the decays [59,60]. Channels such as $\rho\rho$ and $a_1\rho$ could in principle also be studied, using angular analysis to determine the mixture of CP-even and CP-odd contributions.

The analysis of $B \rightarrow D^+D^-$ proceeds along very similar lines. The quark subprocess here is $b \rightarrow c\bar{c}d$, and so the tree contribution gives

$$\lambda(B \rightarrow D^+D^-) = \eta_{D^+D^-} \left(\frac{V_{tb}^*V_{td}}{V_{tb}V_{td}^*} \right) \left(\frac{V_{cd}^*V_{cb}}{V_{cd}V_{cb}^*} \right) \implies \mathcal{I}m\lambda_{DD} = -\sin(2\beta). \quad (138)$$

since $\eta_{D^+D^-} = +1$. Again, there are hadronic uncertainties due to the pure penguin term in (130), but they are estimated to be small.

In all cases the above discussions have neglected the distinction between strong penguins and electroweak penguins. The CKM phase structure of both types of penguins is the same. The only place where this distinction becomes important is when an isospin argument is used

to remove hadronic uncertainties due to penguin contributions. These arguments are based on the fact that gluons have isospin zero, and hence strong penguin processes have definite ΔI . Photons and Z -bosons on the other hand contribute to more than one ΔI transition and hence cannot be separated from tree terms by isospin analysis. In most cases electroweak penguins are small, typically no more than ten percent of the corresponding strong penguins and so their effects can safely be neglected. However in cases (iii) and (iv), where tree contributions are small or absent, their effects may need to be considered. (A full review of the role of electroweak penguins in B decays has been given in ref. [61].)

4.5 Unitarity Triangles

One can obtain an intuitive understanding of the Standard Model CP violation in the interference between decays with and without mixing by examining the unitarity triangles. It is instructive to draw the three triangles, (29), (30) and (31), knowing the experimental values (within errors) for the various $|V_{ij}|$. In the first triangle (29), one side is of $\mathcal{O}(\lambda^5)$ and therefore much shorter than the other, $\mathcal{O}(\lambda)$, sides. In the second triangle (30), one side is of $\mathcal{O}(\lambda^4)$ and therefore shorter than the other, $\mathcal{O}(\lambda^2)$, sides. In the third triangle (31), all sides have lengths of $\mathcal{O}(\lambda^3)$. The first two triangles then almost collapse to a line while the third one is open.

Let us examine the CP asymmetries in the leading decays into final CP eigenstates. For the B mesons, the size of these asymmetries (*e.g.* $\mathcal{I}m\lambda_{\psi K_S}$) depends on β because it gives the difference between half the phase of the $B - \bar{B}$ mixing amplitude and the phase of the decay amplitudes. The form of the third unitarity triangle, (31), implies that $\beta = \mathcal{O}(1)$, which explains why these asymmetries are expected to be large.

It is useful to define the analog phases for the B_s meson, β_s , and the K meson, β_K :

$$\beta_s \equiv \arg \left[-\frac{V_{ts}V_{tb}^*}{V_{cs}V_{cb}^*} \right], \quad \beta_K \equiv \arg \left[-\frac{V_{cs}V_{cd}^*}{V_{us}V_{ud}^*} \right]. \quad (139)$$

The angles β_s and β_K can be seen to be the small angles of the second and first unitarity triangles, (30) and (29), respectively. This gives an intuitive understanding of why CP violation is small in the leading K decays (that is ε_K measured in $K \rightarrow \pi\pi$ decays) and is expected to be small in the leading B_s decays (*e.g.* $B_s \rightarrow \psi\phi$). Decays related to the short sides of these triangles are rare but could exhibit significant CP violation. Actually, the large angles in the (29) triangle are approximately β and $\pi - \beta$, which explains why CP violation in $K \rightarrow \pi\nu\bar{\nu}$ is related to β and expected to be large. The large angles in the (30) triangle are approximately γ and $\pi - \gamma$. This explains why the CP asymmetry in $B_s \rightarrow \rho K_S$ is related to γ and expected to be large. (Note, however, that this mode gets comparable contributions from penguin and tree diagrams and does not give a clean CKM measurement [56].)

5 CP VIOLATION BEYOND THE STANDARD MODEL

The Standard Model picture of CP violation is rather unique and highly predictive. In particular, we would like to point out the following features:

- (i) CP is broken explicitly.
- (ii) All CP violation arises from a single phase, that is δ_{KM} .
- (iii) The measured value of ε_K requires that δ_{KM} is of order one. (In other words, CP is not an approximate symmetry of the Standard Model.)
- (iv) The values of all other CP violating observables can be predicted. In particular, CP violation in $B \rightarrow \psi K_S$ (and similarly various other CP asymmetries in B decays), and in $K \rightarrow \pi\nu\bar{\nu}$ are expected to be of order one.

The commonly repeated statement that CP violation is one of the least tested aspects of the Standard Model is well demonstrated by the fact that none of the above features necessarily holds in the presence of New Physics. In particular, there are viable models of new physics (*e.g.* certain supersymmetric models) with the following features:

- (i) CP is broken spontaneously.
- (ii) There are many CP violating phases (even in the low energy effective theory).
- (iii) CP is an approximate symmetry, with all CP violating phases small (usually $10^{-3} \lesssim \phi_{\text{CP}} \lesssim 10^{-2}$).
- (iv) Values of CP violating observables can be predicted and could be very different from the Standard Model predictions (except, of course, ε_K). In particular, $\mathcal{I}m\lambda_{\psi K_S}$ and $a_{\pi\nu\bar{\nu}}$ could both be $\ll 1$.

To understand how the Standard Model predictions could be modified by New Physics, we will focus on CP violation in the interference between decays with and without mixing. As explained above, it is this type of CP violation which, due to its theoretical cleanliness, may give unambiguous evidence for New Physics most easily.

5.1 CP Violation as a Probe of Flavor Beyond the Standard Model

Let us consider five specific CP violating observables.

- (i) $\mathcal{I}m\lambda_{\psi K_S}$, the CP asymmetry in $B \rightarrow \psi K_S$. This measurement will cleanly determine the relative phase between the $B - \bar{B}$ mixing amplitude and the $b \rightarrow c\bar{c}s$ decay amplitude ($\sin 2\beta$ in the Standard Model). The $b \rightarrow c\bar{c}s$ decay has Standard Model tree contributions and therefore is very unlikely to be significantly affected by new physics. On the other hand, the mixing amplitude can be easily modified by new physics. We parametrize such a modification by a phase θ_d :

$$\mathcal{I}m\lambda_{\psi K_S} = \sin[2(\beta + \theta_d)]. \quad (140)$$

- (ii) $\mathcal{I}m\lambda_{\phi K_S}$, the CP asymmetry in $B \rightarrow \phi K_S$. This measurement will cleanly determine the relative phase between the $B - \bar{B}$ mixing amplitude and the $b \rightarrow s\bar{s}s$ decay amplitude. The $b \rightarrow s\bar{s}s$ decay has only Standard Model penguin contributions and therefore is sensitive to new physics. We parametrize the modification of the decay amplitude by a phase θ_A [62]:

$$\mathcal{I}m\lambda_{\phi K_S} = \sin[2(\beta + \theta_d + \theta_A)]. \quad (141)$$

- (iii) $a_{\pi\nu\bar{\nu}}$, the CP violating ratio of $K \rightarrow \pi\nu\bar{\nu}$ decays. This measurement will cleanly determine the relative phase between the $K - \bar{K}$ mixing amplitude and the $s \rightarrow d\nu\bar{\nu}$ decay amplitude. The experimentally measured small value of ε_K requires that the phase of the $K - \bar{K}$ mixing amplitude is not modified from the Standard Model prediction. On the other hand, the decay, which in the Standard Model is a loop process with small mixing angles, can be easily modified by new physics.
- (iv) $\mathcal{I}m(\lambda_{K-\pi^+})$, the CP violating quantity in $D \rightarrow K^-\pi^+$ decay. The ratio

$$a_{D \rightarrow K\pi} = \frac{\mathcal{I}m(\lambda_{K-\pi^+})}{|\lambda_{K-\pi^+}|} \quad (142)$$

depends on the relative phase between the $D - \bar{D}$ mixing amplitude and the $c \rightarrow d\bar{s}u$ decay amplitude. Within the Standard Model, this decay channel is tree level. It is unlikely that it is affected by new physics. On the other hand, the mixing amplitude can be easily modified by new physics.

- (v) d_N , the electric dipole moment of the neutron. We did not discuss this quantity so far because, unlike CP violation in meson decays, flavor changing couplings are not necessary for d_N . In other words, the CP violation that induces d_N is *flavor diagonal*. It does in general get contributions from flavor changing physics, but it could be induced by sectors that are flavor blind. Within the Standard Model (and ignoring the strong CP angle θ_{QCD}), the contribution from δ_{KM} arises at the three loop level and is at least six orders of magnitude below the experimental bound [10] d_N^{exp} ,

$$d_N^{\text{exp}} = 1.1 \times 10^{-25} \text{ e cm.} \quad (143)$$

The various CP violating observables discussed above are sensitive then to new physics in the mixing amplitudes for the $B - \bar{B}$ and $D - \bar{D}$ systems, in the decay amplitudes for $b \rightarrow s\bar{s}s$ and $s \rightarrow d\nu\bar{\nu}$ channels and to flavor diagonal CP violation. If information about all these processes becomes available and deviations from the Standard Model predictions are found, we can ask rather detailed questions about the nature of the new physics that is responsible to these deviations:

- (i) Is the new physics related to the down sector? the up sector? both?
- (ii) Is the new physics related to $\Delta B = 1$ processes? $\Delta B = 2$? both?
- (iii) Is the new physics related to the third generation? to all generations?
- (iv) Are the new sources of CP violation flavor changing? flavor diagonal? both?

It is no wonder then that with such rich information, flavor and CP violation provide an excellent probe of new physics.

5.2 Supersymmetry

A generic supersymmetric extension of the Standard Model contains a host of new flavor and CP violating parameters. (For reviews on supersymmetry see refs. [63-66]. The following section is based on [67].) The requirement of consistency with experimental data provides strong constraints on many of these parameters. For this reason, the physics of flavor and CP violation has had a profound impact on supersymmetric model building. A discussion of CP violation in this context can hardly avoid addressing the flavor problem itself. Indeed, many of the supersymmetric models that we analyze below were originally aimed at solving flavor problems.

As concerns CP violation, one can distinguish two classes of experimental constraints. First, bounds on nuclear and atomic electric dipole moments determine what is usually called the *supersymmetric CP problem*. Second, the physics of neutral mesons and, most importantly, the small experimental value of ε_K pose the *supersymmetric ε_K problem*. The latter is closely related to the flavor structure of supersymmetry.

The contribution to the CP violating ε_K parameter in the neutral K system is dominated by diagrams involving Q and \bar{d} squarks in the same loop [68-72]. The corresponding effective four-fermi operator involves fermions of both chiralities, so that its matrix elements are enhanced by $\mathcal{O}(m_K/m_s)^2$ compared to the chirality conserving operators. For $m_{\tilde{q}} \simeq m_Q \simeq m_D = \tilde{m}$ (our results depend only weakly on this assumption) and focusing on the contribution from the first two squark families, one gets (we use the results in ref. [72])

$$\frac{(\Delta m_K \varepsilon_K)^{\text{SUSY}}}{\Delta m_K \varepsilon_K} \sim 10^7 \left(\frac{300 \text{ GeV}}{\tilde{m}} \right)^2 \left| \frac{(\delta m_D^2)_{12}}{\tilde{m}^2} \right|^2 |K_{12}^d|^2 \sin \phi, \quad (144)$$

where \tilde{m} is the typical scale of squark and gluino masses, $(\delta m_D^2)_{12}$ is the mass-squared difference between the first two down squark generations, K_{12}^d is the mixing angle in the gluino-quark-squark coupling and ϕ is the relevant CP violating phase in the mixing. In a generic supersymmetric framework, we expect $\tilde{m} = \mathcal{O}(m_Z)$, $\delta m_D^2/\tilde{m}^2 = \mathcal{O}(1)$, $K_{ij}^d = \mathcal{O}(1)$ and $\sin \phi = \mathcal{O}(1)$.

Table 1: CP violating observables in various classes of Supersymmetric flavor models.

Model	d_N/d_N^{exp}	θ_d	θ_A	$a_{D \rightarrow K\pi}$	$a_{K \rightarrow \pi\nu\bar{\nu}}$
Standard Model	$\lesssim 10^{-6}$	0	0	0	$\mathcal{O}(1)$
Exact Universality	$\gtrsim 10^{-6}$	0	0	0	=SM
Approximate CP	$\sim 10^{-1}$	$-\beta$	0	$\mathcal{O}(10^{-3})$	$\mathcal{O}(10^{-5})$
Alignment	$\gtrsim 10^{-3}$	$\mathcal{O}(0.2)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	\approx SM
Approx. Universality	$\gtrsim 10^{-2}$	$\mathcal{O}(0.2)$	$\mathcal{O}(1)$	0	\approx SM
Heavy Squarks	$\sim 10^{-1}$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(10^{-2})$	\approx SM

Then the constraint (144) is generically violated by about seven orders of magnitude. Eq. (144) also shows what are the possible solutions to the supersymmetric flavor and CP problems:

- (i) *Universality*: At some high scale, the soft supersymmetry breaking terms are universal. In other words, the different squark generations are degenerate [73-74]. There are two very different ways to achieve such a situation. First, the mechanism that communicates supersymmetry breaking to the observable sector could be flavor blind. This is the case with gauge mediated supersymmetry breaking [75-78], but it is also possible (though not generic) that similar boundary conditions occur when supersymmetry breaking is communicated to the observable sector up at the Planck scale [79-85]. RGE effects will introduce some splitting at low energy which, for the first two squark generations, is typically of $\mathcal{O}(m_c^2/m_W^2)$. Second, the Yukawa hierarchy could be a result of a non-Abelian flavor symmetry with the first two generations forming a doublet [86-94]. In this framework, the first two squark generations are approximately degenerate with splitting which could be as high as $\mathcal{O}(\lambda^2)$. The third generation could be widely split from the first two.
- (ii) *Alignment* [95-97]: The mixing angles in the gluino-quark-squark couplings are small. This is usually achieved in models where the Yukawa hierarchy is explained by Abelian flavor symmetries. In the symmetry limit, both the quark mass matrices and the squark mass-squared matrices are diagonal, so that mixing is suppressed by small breaking parameters. Typically, the alignment is required to be very precise between the first two down generations, while all other supersymmetric mixing angles are similar to the corresponding CKM angles.
- (iii) *Heavy Squarks* [86,89,98-100]: If the masses of the first and second generation squarks m_i are larger than the other soft masses, $m_i^2 \sim 100 \tilde{m}^2$, then the Supersymmetric CP problem is solved and the ε_K problem is relaxed (but not eliminated). This does not necessarily lead to naturalness problems, since these two generations are almost decoupled from the Higgs sector.
- (iv) *Approximate CP* [101-103].: Both supersymmetric CP problems are solved if CP is an approximate symmetry, broken by a small parameter of order 10^{-3} . Of course, some mechanism to solve the supersymmetric flavor problems has to be invoked.

Measurements of CP violation will provide us with an excellent probe of the flavor and CP structure of supersymmetry. This is clearly demonstrated in Table 1.

5.3 Final Comments

The unique features of CP violation are well demonstrated by examining the CP asymmetry in $B \rightarrow \psi K_S$, $\mathcal{I}m\lambda_{\psi K_S}$, and CP violation in $K \rightarrow \pi\nu\bar{\nu}$, $\mathcal{I}m\lambda_{\pi\nu\bar{\nu}}$. Model independently, $\mathcal{I}m\lambda_{\psi K_S}$ measures the relative phase between the $B - \bar{B}$ mixing amplitude and the $b \rightarrow c\bar{c}d$ decay

amplitude (more precisely, the $b \rightarrow c\bar{c}s$ decay amplitude times the $K - \bar{K}$ mixing amplitude), while $\mathcal{I}m\lambda_{\pi\nu\bar{\nu}}$ measures the relative phase between the $K - \bar{K}$ mixing amplitude and the $s \rightarrow d\nu\bar{\nu}$ decay amplitude. We would like to emphasize the following three points:

- (i) *The two measurements are theoretically clean to better than $\mathcal{O}(10^{-2})$.* Thus they can provide the most accurate determination of CKM parameters.
- (ii) *As concerns CP violation, the Standard Model is a uniquely predictive model.* In particular, it predicts that the seemingly unrelated $\mathcal{I}m\lambda_{\psi K_S}$ and $\mathcal{I}m\lambda_{\pi\nu\bar{\nu}}$ measure the same parameter, that is the angle β of the unitarity triangle.
- (iii) *In the presence of New Physics, there is in general no reason for a relation between $\mathcal{I}m\lambda_{\psi K_S}$ and $\mathcal{I}m\lambda_{\pi\nu\bar{\nu}}$.* Therefore, a measurement of both will provide a sensitive probe of New Physics.

Acknowledgements

My understanding of flavor physics and CP violation has benefitted from numerous discussions with Helen Quinn. Large parts of this review are based on previous reviews written in collaboration with her. Parts of this review are based on contributions to the BaBar Physics Book that were written in collaboration with Gerald Eigen, Ben Grinstein, Stephane Plaszczynski and Marie-Helene Schune. Other parts of this review are based on a review written in collaboration with Yuval Grossman and Riccardo Rattazzi. I thank Sven Bergmann, Galit Eyal, Eilam Gross and Yael Shadmi for their help in preparing these lectures. Y.N. is supported in part by the United States – Israel Binational Science Foundation (BSF), by the Israel Science Foundation and by the Minerva Foundation (Munich).

References

- [1] S.L. Glashow, Nucl. Phys. 22 (1961) 579.
- [2] S. Weinberg, Phys. Rev. Lett. 19 (1967) 1264.
- [3] A. Salam, in *Proc. 8th Nobel Symp.* (Stockholm), ed. N. Swartholm (Almqvist and Wiksells, Stockholm 1968).
- [4] M. Kobayashi and T. Maskawa, Prog. Theo. Phys. 49 (1973) 652.
- [5] J.H. Christenson, J.W. Cronin, V.L. Fitch and R. Turlay, Phys. Rev. Lett. 13 (1964) 138.
- [6] A.D. Sakharov, ZhETF Pis. Red. 5 (1967) 32; JETP Lett. 5 (1967) 24.
- [7] N. Cabibbo, Phys. Rev. Lett. 10 (1963) 531.
- [8] Y. Grossman and Y. Nir, Nucl. Phys. B448 (1995) 30.
- [9] Y. Nir and D. Silverman, Nucl. Phys. B345 (1990) 301.
- [10] C. Caso *et al.*, Particle Data Group, Eur. Phys. Soc. J. C3 (1998) 1.
- [11] L.L. Chau and W.Y. Keung, Phys. Rev. Lett. 53 (1984) 1802.
- [12] L. Wolfenstein, Phys. Rev. Lett. 51 (1983) 1945.
- [13] R. Aleksan *et al.*, Phys. Lett. B316 (1993) 567.
- [14] M. Beneke *et al.*, hep-ph/9808385.
- [15] J. Alexander, talk given at 29th Int. Conf. on HEP in Vancouver, July 22-29, 1998.
- [16] T. Inami and C.S. Lim, Prog. Theo. Phys. 65 (1981) 297; (E) 65 (1982) 772.
- [17] A.J. Buras, M. Jamin and P.H. Weisz, Nucl. Phys. B347 (1990) 491.
- [18] A.J. Buras and R. Fleischer, hep-ph/9704376, in *Heavy Flavours II*, eds. A.J. Buras and M. Lindner (World Scientific, 1998).
- [19] C.T. Sachrajda, in the *proceedings of the 18th International Symposium on Lepton Photon Interactions*, (Hamburg, 1997), hep-ph/9711386.
- [20] S. Herrlich and U. Niereste, Phys. Rev. D52 (1995) 6505; Nucl. Phys. B476 (1996) 27.
- [21] G. Buchalla and A.J. Buras, Nucl. Phys. B412 (1996) 106.
- [22] G. Buchalla and A.J. Buras, Nucl. Phys. B398 (1993) 285; Nucl. Phys. B400 (1993) 225.
- [23] S. Adler *et al.*, E787 Collaboration, Phys. Rev. Lett. 79 (1997) 2204.
- [24] J. Adams *et al.*, KTeV Collaboration, hep-ex/9806007.
- [25] A.J. Buras, hep-ph/9610461.
- [26] Y. Grossman and Y. Nir, Phys. Lett. B398 (1997) 163.
- [27] C. Jarlskog, Phys. Rev. Lett. 55 (1985) 1039.
- [28] C.O. Dib, I. Dunietz, F.J. Gilman and Y. Nir, Phys. Rev. D41 (1990) 1522.
- [29] Y. Nir, in the *proceedings of the 18th International Symposium on Lepton Photon Interactions*, (Hamburg, 1997), hep-ph/9709301.
- [30] The BaBar Physics Book, SLAC-R-504, eds. P.F. Harrison and H.R. Quinn (1998).
- [31] Y. Grossman, Y. Nir, S. Plaszczynski and M.-H. Schune, Nucl. Phys. B511 (1998) 69, hep-ph/9709288.
- [32] For a review see, A.G. Cohen, D.B. Kaplan and A.E. Nelson, Ann. Rev. Nucl. Part. Sci. 43 (1993) 27.
- [33] G.R. Farrar and M.E. Shaposhnikov, Phys. Rev. D50 (1994) 774.
- [34] M.B. Gavela *et al.*, Nucl. Phys. B430 (1994) 382.

- [35] P. Huet and E. Sather, Phys. Rev. D51 (1995) 379.
- [36] A.B. Carter and A.I. Sanda, Phys. Rev. Lett. 45 (1980) 952; Phys. Rev. D23 (1981) 1567.
- [37] I.I. Bigi and A.I. Sanda, Nucl. Phys. B193 (1981) 85; Nucl. Phys. B281 (1987) 41.
- [38] I. Dunietz and J. Rosner, Phys. Rev. D34 (1986) 1404.
- [39] L.S. Littenberg, Phys. Rev. D39 (1989) 3322.
- [40] G. Buchalla and A.J. Buras, Phys. Lett. B333 (1994) 476.
- [41] C.O. Dib, D. London and Y. Nir, Int. J. Mod. Phys. A6 (1991) 1253.
- [42] Y. Nir and H.R. Quinn, Ann. Rev. Nucl. Part. Sci. 42 (1992) 211.
- [43] M. Gronau and D. London, Phys. Rev. D55 (1997) 2845, hep-ph/9608430.
- [44] Y. Grossman, B. Kayser and Y. Nir, Phys. Lett. B415 (1997) 90, hep-ph/9708398.
- [45] G.D. Barr *et al.*, NA31 collaboration, Phys. Lett. B317 (1993) 233.
- [46] L.K. Gibbons *et al.*, E731 collaboration, Phys. Rev. Lett. 70 (1993) 1203.
- [47] G. Blaylock, A. Seiden and Y. Nir, Phys. Lett. B355 (1995) 555, hep-ph/9504306.
- [48] L. Wolfenstein, Phys. Rev. Lett. 75 (1995) 2460, hep-ph/9505285.
- [49] I.I. Bigi, V.A. Khoze, N.G. Uraltsev and A.I. Sanda, in *CP Violation*, ed. C. Jarlskog (World Scientific, Singapore, 1992).
- [50] L. Wolfenstein, Phys. Rev. D57 (1998) 5453.
- [51] For a recent discussion, see L. Randall and S. Su, hep-ph/9807377.
- [52] M. Bander, S. Silverman and A. Soni, Phys. Rev. Lett. 43 (1979) 242.
- [53] S.L. Glashow, J. Iliopoulos and L. Maiani, Phys. Rev. D2 (1970) 1285.
- [54] R. Godang *et al.*, CLEO Collaboration, Phys. Rev. Lett. 80 (1998) 3456, hep-ex/9711010.
- [55] R. Fleischer, Phys. Lett. B341 (1995) 205.
- [56] A.J. Buras and R. Fleischer, Phys. Lett. B341 (1995) 379.
- [57] M. Gronau and D. London, Phys. Rev. Lett. 65 (1990) 3381.
- [58] Y. Grossman and H. Quinn, Phys. Rev. D58 (1998) 017504, hep-ph/9712306.
- [59] H.J. Lipkin, Y. Nir, H.R. Quinn and A. Snyder, Phys. Rev. D44 (1991) 1454.
- [60] H.R. Quinn and A. Snyder, Phys. Rev. D48 (1993) 2139.
- [61] R. Fleischer, Int. J. Mod. Phys. A12 (1997) 2459, hep-ph/9612446.
- [62] Y. Grossman and M.P. Worah, Phys. Lett. B395 (1997) 241.
- [63] H.P. Nilles, Phys. Rep. 110 (1984) 1.
- [64] H.E. Haber and G.L. Kane, Phys. Rep. 117 (1985) 75.
- [65] R. Barbieri, Riv. Nuovo Cim. 11 (1988) 1.
- [66] H.E. Haber, SCIPP 92/33, Lectures given at TASI 92.
- [67] Y. Grossman, Y. Nir and R. Rattazzi, hep-ph/9701231, in *Heavy Flavours II*, eds. A.J. Buras and M. Lindner (World Scientific, 1998).
- [68] J. Donoghue, H. Nilles and D. Wyler, Phys. Lett. B128 (1983) 55.
- [69] F. Gabbiani and A. Masiero, Nucl. Phys. B322 (1989) 235.
- [70] J.S. Hagelin, S. Kelley and T. Tanaka, Nucl. Phys. B415 (1994) 293.
- [71] E. Gabrielli, A. Masiero and L. Silvestrini, Phys. Lett. B374 (1996) 80.
- [72] F. Gabbiani, E. Gabrielli, A. Masiero and L. Silvestrini, Nucl. Phys. B477 (1996) 321.

- [73] S. Dimopoulos and H. Georgi, Nucl. Phys. B193 (1981) 150.
- [74] N. Sakai, Z. Phys. C11 (1981) 153.
- [75] M. Dine and A. Nelson, Phys. Rev. D48 (1993) 1277.
- [76] M. Dine and A. Nelson and Y. Shirman, Phys. Rev. D51 (1994) 1362.
- [77] M. Dine, A. Nelson, Y. Nir and Y. Shirman, Phys. Rev. D53 (1996) 2658.
- [78] M. Dine, Y. Nir and Y. Shirman, Phys. Rev. D55 (1997) 1501.
- [79] A.H. Chamseddine, R. Arnowitt and P. Nath, Phys. Rev. Lett. 49 (1982) 970; Nucl. Phys. B227 (1983) 1219.
- [80] R. Barbieri, S. Ferrara and C.A. Savoy, Phys. Lett. B119 (1982) 343.
- [81] L. Hall, J. Lykken and S. Weinberg, Phys. Rev. D27 (1983) 235.
- [82] J. Ellis, C. Kounnas and D.V. Nanopoulos, Nucl. Phys. B247 (1984) 373.
- [83] M. Lanzaogorta and G.G. Ross, Phys. Lett. B364 (1995) 163.
- [84] V. Kaplunovsky and J. Louis, Phys. Lett. B306 (1993) 269.
- [85] R. Barbieri, J. Louis and M. Moretti, Phys. Lett. B312 (1993) 451; (E) *ibid.* B316 (1993) 632.
- [86] M. Dine, A. Kagan and R.G. Leigh, Phys. Rev. D48 (1993) 4269.
- [87] P. Pouliot and N. Seiberg, Phys. Lett. B318 (1993) 169.
- [88] L.J. Hall and H. Murayama, Phys. Rev. Lett. 75 (1995) 3985.
- [89] A. Pomarol and D. Tommasini, Nucl. Phys. B466 (1996) 3, hep-ph/9507462.
- [90] R. Barbieri, G. Dvali and L.J. Hall, Phys. Lett. B377 (1996) 76, hep-ph/9512388.
- [91] C. Carone, L.J. Hall and H. Murayama, Phys. Rev. D54 (1996) 2328, hep-ph/9602364.
- [92] Z.G. Berezhiani, hep-ph/9609342.
- [93] R. Barbieri, L.J. Hall, S. Raby and A. Romanino, Nucl. Phys. B493 (1997) 3, hep-ph/9610449.
- [94] G. Eyal, hep-ph/9807308.
- [95] Y. Nir and N. Seiberg, Phys. Lett. B309 (1993) 337, hep-ph/9304307.
- [96] M. Leurer, Y. Nir and N. Seiberg, Nucl. Phys. B420 (1994) 468, hep-ph/9410320.
- [97] Y. Nir and R. Rattazzi, Phys. Lett. B382 (1996) 363, hep-ph/9603233.
- [98] M. Dine, A. Kagan and S. Samuel, Phys. Lett. B243 (1990) 250.
- [99] A.G. Cohen, D.B. Kaplan and A.E. Nelson, Phys. Lett. B388 (1996) 588, hep-ph/9607394.
- [100] A.G. Cohen, D.B. Kaplan, F. Lepeintre and A.E. Nelson, Phys. Rev. Lett. 78 (1997) 2300, hep-ph/9610252.
- [101] K.S. Babu and S.M. Barr, Phys. Rev. Lett. 72 (1994) 2831.
- [102] S.A. Abel and J.M. Frere, Phys. Rev. D55 (1997) 1623.
- [103] G. Eyal and Y. Nir, Nucl. Phys. B528 (1998) 21, hep-ph/9801411.

John ELLIS

Theoretical Physics Division, CERN

CH - 1211 Geneva 23

Abstract

In the first lecture, the Standard Model is reviewed, with the aim of seeing how its successes constrain possible extensions, the significance of the apparently low Higgs mass indicated by precision electroweak experiments is discussed, and *defects of the Standard Model* are examined. The second lecture includes a general discussion of the electroweak vacuum and an *introduction to supersymmetry*, motivated by the gauge hierarchy problem. In the third lecture, the *phenomenology of supersymmetric models* is discussed in more detail, with emphasis on the information provided by LEP data. The fourth lecture introduces *Grand Unified Theories*, with emphases on general principles and on neutrino masses and mixing. Finally, the last lecture contains short discussions of some *further topics*, including supersymmetry breaking, gauge-mediated messenger models, supergravity, strings and M phenomenology.

1 GETTING MOTIVATED

There have been many reviews of different subjects in particle physics ‘for pedestrians’. At this school many of us had the fun experience of walking in the Scottish hills, which is more strenuous than a stroll across the Old Course at St Andrews, though less dangerous than mountain climbing in the Alps. The spirit of these lectures is similar: an invigorating introduction to modern phenomenological trends, but not too close to the theoretical precipices.

1.1 Recap of the Standard Model

The Standard Model continues to survive all experimental tests at accelerators. However, despite its tremendous successes, no-one finds the Standard Model [1] satisfactory, and many present and future experiments are being aimed at some of the Big Issues raised by the Standard Model: is there a Higgs boson? is there supersymmetry? why are there only six quarks and six leptons? what is the origin of flavour mixing and CP violation? can the different interactions be unified? does the proton decay? are there neutrino masses? For the first time, clear evidence for new physics beyond the Standard Model may be emerging from non-accelerator neutrino physics [2]. Nevertheless the Standard Model remains the rock on which our quest for new physics must be built, so let us start by reviewing its basic features and examine whether its successes offer any hint of the direction in which to search for new physics.

We first review the electroweak gauge bosons and the Higgs mechanism of spontaneous symmetry breaking by which we believe they acquire masses [3]. The gauge bosons are described by the action

$$\mathcal{L} = -\frac{1}{4} G_{\mu\nu}^i G^{i\mu\nu} - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \quad (1)$$

where $G_{\mu\nu}^i \equiv \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + ig\epsilon_{ijk} W_\mu^j W_\nu^k$ is the field strength for the $SU(2)$ gauge boson W_μ^i , and $F_{\mu\nu} \equiv \partial_\mu W_\nu^i - \partial_\nu W_\mu^i$ is the field strength for the $U(1)$ gauge boson B_μ . The action (1) contains bilinear terms that yield the boson propagators, and also trilinear and quartic

gauge-boson interactions. The gauge bosons couple to quarks and leptons via

$$\mathcal{L}_F = - \sum_f i [\bar{f}_L \gamma^\mu D_\mu f_L + \bar{f}_R \gamma^\mu D_\mu f_R] \quad (2)$$

where the D_μ are covariant derivatives:

$$D_\mu \equiv \partial_\mu - i g \sigma_i W_\mu^i - i g' Y B_\mu \quad (3)$$

The $SU(2)$ piece appears only for the left-handed fermions f_L , which are isospin doublets, while the right-handed fermions f_R are isospin singlets, and hence couple only to the $U(1)$ gauge boson B_μ , via hypercharges Y .

The origin of all the masses in the Standard Model is an isodoublet scalar Higgs field, whose kinetic term in the action is

$$\mathcal{L}_\phi = -|D_\mu \phi|^2 \quad (4)$$

and which has the magic potential:

$$\mathcal{L}_V = -V(\phi) : V(\phi) = -\mu^2 \phi^\dagger \phi + \frac{\lambda}{2} (\phi^\dagger \phi)^2 \quad (5)$$

Because of the negative sign for the quadratic term in (5), the symmetric solution $\langle 0|\phi|0 \rangle = 0$ is unstable, and if $\lambda > 0$ the favoured solution has a non-zero vacuum expectation value which we may write in the form:

$$\langle 0|\phi|0 \rangle = \langle 0|\phi^\dagger|0 \rangle = v \left(\frac{0}{\frac{1}{\sqrt{2}}} \right) : v^2 = \frac{\mu^2}{2\lambda} \quad (6)$$

corresponding to spontaneous breakdown of the electroweak gauge symmetry.

Expanding around the vacuum: $\phi = \langle 0|\phi|0 \rangle + \hat{\phi}$, the kinetic term (4) for the Higgs field yields mass terms for the gauge bosons:

$$\mathcal{L}_\phi \ni -\frac{g^2 v^2}{2} W_\mu^+ W^{\mu-} - g'^2 \frac{v^2}{2} B_\mu B^\mu + g g' v^2 B_\mu W^{\mu 3} - g^2 \frac{v^2}{2} W_\mu^3 W^{\mu 3} \quad (7)$$

There are also bilinear derivative couplings of the gauge bosons to the massless Goldstone bosons η , e.g., in the charged-boson sector we have

$$-\partial_\mu \eta^+ \partial_\mu \eta^- + \left(\frac{igv}{2} \partial_\mu \eta^+ W^{\mu-} + h.c. \right) \quad (8)$$

Combining these with the first term in (7), we see a quadratic mass term for the combination

$$W_\mu^+ - 2i \frac{\partial_\mu \eta^+}{gv} \quad (9)$$

of charged bosons. This clearly gives a mass to the W^\pm bosons:

$$m_{W^\pm} = \frac{gv}{2} \quad (10)$$

whilst the neutral gauge bosons (W_μ^3, B_μ) have a 2×2 mass-squared matrix:

$$\begin{pmatrix} \frac{g^2}{2} & -\frac{gg'}{2} \\ -\frac{gg'}{2} & \frac{g'^2}{2} \end{pmatrix} v^2 \quad (11)$$

This is easily diagonalized to yield the mass eigenstates:

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2 + g'^2}} : m_Z = \frac{1}{2}\sqrt{g^2 + g'^2}v ; A_\mu = \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2 + g'^2}} : m_A = 0 \quad (12)$$

that we identify with the Z and γ , respectively. It is useful to introduce the electroweak mixing angle θ_W defined by

$$\sin \theta_W = \frac{g'}{\sqrt{g^2 + g'^2}} \quad (13)$$

in terms of the $SU(2)$ gauge coupling g and g' . Many other quantities can be expressed in terms of $\sin \theta_W$ (13): for example, $m_W^2/m_Z^2 = \cos^2 \theta_W$. The charged-current interactions are of the current-current form:

$$\frac{1}{4}\mathcal{L}_{cc} = \frac{G_F}{\sqrt{2}} J_\mu^+ J^{-\mu} : \frac{G_F}{\sqrt{2}} \equiv \frac{g^2}{8m_W^2} \quad (14)$$

as are the neutral-current interactions:

$$\frac{1}{4}\mathcal{L}_{NC} = \frac{G_F^{NC}}{\sqrt{2}} J_\mu^0 J^{\mu 0} : J_\mu^0 \equiv J_\mu^3 - \sin^2 \theta_W J_\mu^{em} , G_F^{NC} \equiv \frac{g^2 + g'^2}{8m_Z^2} \quad (15)$$

The ratio of neutral- and charged-current interaction strengths is often expressed as

$$\rho = \frac{G_F^{NC}}{G_F} = \frac{m_W^2}{m_Z^2 \cos^2 \theta_W} \quad (16)$$

which takes the value unity in the Standard Model with only Higgs doublets [4], as assumed here. However, this and the other tree-level relations given above are modified by quantum corrections (loop effects), as we discuss later.

Figures 1 and 2 compile the most important precision electroweak measurements [5]. It is striking that m_Z (Fig. 1) is now known more accurately than the muon decay constant. Precision measurements of Z decays also restrict possible extensions of the Standard Model. For example, the number of effective equivalent light-neutrino species is measured very accurately:

$$N_\nu = 2.994 \pm 0.011 \quad (17)$$

I had always hoped that N_ν might turn out to be non-integer: $N_\nu = \pi$ would have been good, and $N_\nu = e$ would have been even better, but this was not to be! The constraint (17) is also important for possible physics beyond the Standard Model, such as supersymmetry as we discuss later. The measurement (17) implies, by extension, that there can only be three charged leptons and hence, in order to keep triangle anomalies cancelled, no more quarks [6]. Hence a fourth conventional matter generation is not a possible extension of the Standard Model.

There are by now many precision measurements of $\sin^2 \theta_W$ (Fig. 2): this is a free parameter in the Standard Model, whose value [7] is a suggestive hint for grand unification [8] and supersymmetry [9], as we discuss later. Notice also in Fig. 2 that consistency of the data seems to prefer a relatively low value for the Higgs mass, which is another possible suggestion of supersymmetry, as we also discuss later.

The previous field-theoretical discussion of the Higgs mechanism can be rephrased in more physical language. It is well known that a massless vector boson such as the photon γ or gluon g has just two polarization states: $\lambda = \pm 1$. However, a massive vector boson such as the ρ has three polarization states: $\lambda = 0, \pm 1$. This third polarization state is provided by a spin-0 field as seen in (9). In order to make $m_{W^\pm, Z^0} \neq 0$, this should have non-zero electroweak isospin $I \neq 0$, and the simplest possibility is a complex isodoublet (ϕ^+, ϕ^0) , as assumed above.

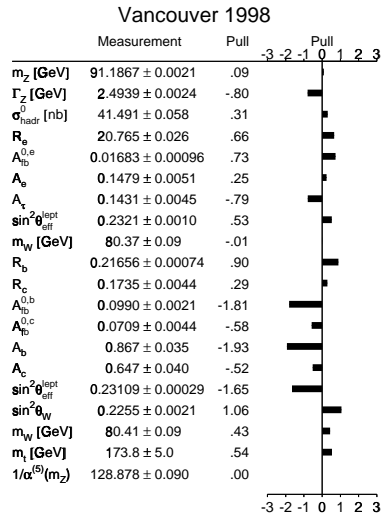


Fig. 1: Precision electroweak measurements and the pulls they exert in a global fit [5].

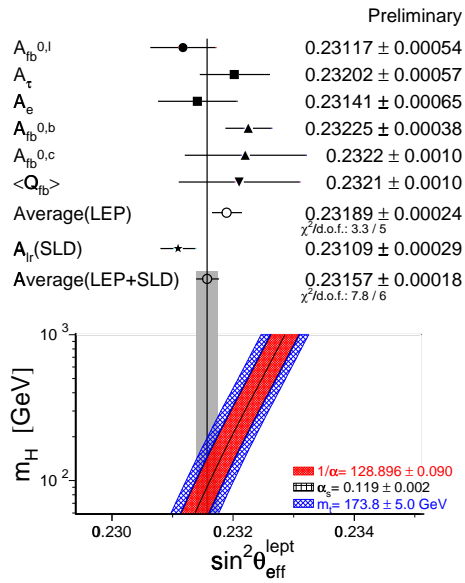


Fig. 2: Precision determinations of $\sin^2 \theta_W$ [5].

This has four degrees of freedom, three of which are eaten by the W^\pm and Z^0 as their third polarization states, leaving us with one physical Higgs boson H . Once the vacuum expectation value $|\langle 0|\phi|0\rangle| = v/\sqrt{2}$: $v = \mu/\sqrt{2\lambda}$ is fixed, the mass of the remaining physical Higgs boson is given by

$$m_H^2 = 2\mu^2 = 4\lambda v^2 \quad (18)$$

which is a free parameter in the Standard Model.

The necessity for such a physical Higgs boson may be further demonstrated by considering the scattering amplitude for $\bar{f}f \rightarrow W^+W^-$ [10]. By unitarity this contributes to elastic $\bar{f}f \rightarrow \bar{f}f$ scattering at the one-loop level. This contribution would be divergent and unrenormalizable in the absence of a direct-channel spin-0 contribution to cancel mass-dependent contributions from the established ν, γ and Z^0 exchanges. If these spin-0 contributions to $\bar{f}f \rightarrow W^+W^-$ and analogously $W^+W^- \rightarrow W^+W^-$ are due to a single Higgs boson, as in the Standard Model, its couplings to fermions and gauge bosons are completely determined:

$$g_{H\bar{f}f} = \frac{g}{2} \frac{m_f}{m_W}, \quad g_{HW^+W^-} = gm_W, \quad g_{HZ^0Z^0} = gm_Z \quad (19)$$

Thus the Higgs production and decay rates are completely fixed as functions of the unknown mass M_H (18) [11]. This unitarity argument actually requires that $m_H \leq 1$ TeV in order to accomplish its one-loop cancellation mission [12, 13, 14].

The search for the Higgs boson is one of the main objectives of the LEP 2 experimental programme. The dominant production mechanism is $e^+e^- \rightarrow Z^0 + H$ [11, 15], which has the tree-level cross section [15, 13]

$$\sigma_{ZM} = \frac{G_F^2 m_Z^4}{96\pi s} (1 + (1 - 4\sin^2\theta_W)^2) \lambda^{1/2} \frac{\lambda + 12m_Z^2/s}{(1 - m_Z^2/s)^2} \quad (20)$$

the prefactor comes from the known HZ^0Z^0 vertex (19), and the phase-space factor

$$\lambda \equiv \left(1 - \left(\frac{m_H^2}{s}\right) - \left(\frac{m_Z^2}{s}\right)\right)^2 - \frac{4m_H^2 m_Z^2}{s^2} \quad (21)$$

which gives us sensitivity to $m_H \lesssim E_{cm} - M_Z$. With the current LEP 2 running at 189 GeV, each individual LEP experiment has established a lower limit $m_H \gtrsim 96$ GeV, and the four experiments could probably be combined to yield $m_H \gtrsim 98$ GeV [16]. The next two years of LEP 2 running at energies ~ 200 GeV should enable the Higgs to be discovered if $m_H \lesssim 110$ GeV, if as much luminosity is accumulated as in 1998. As we see shortly, this covers the range of m_H where the precision electroweak data [5] indicate the highest probability density. Hence, the integrated probability that LEP 2 discovers the Higgs boson is not negligible, though we must brace ourselves for the likelihood that it is too heavy to be discovered at LEP.

1.2 Interpretation of the Precision Electroweak Data

The precision of the electroweak data shown in Figure 1 is so high – of order 0.1 % in some cases – that quantum corrections are crucial for their interpretation [17]. At the one-loop level, these include vacuum-polarization, vector and box diagrams. The dominant contributions from two- and higher-loop diagrams must also be taken into account. These loop diagrams must be renormalized, and this is achieved by fixing three quantities at their physical values: $m_Z = 91.1867 \pm 0.0021$ GeV, $\alpha_{em}^{-1} = 137.03599959(38)13$, $G_\mu = 1.166389(22) \times 10^{-5}$ GeV⁻². In the case of experiments at the Z^0 peak, one needs to calculate the renormalization of α_{em} over scales $m_e \lesssim Q \lesssim m_Z$ due to vacuum polarization diagrams. The principal uncertainty in

this renormalization is due to hadronic diagrams in the range of Q where perturbative QCD calculations are not directly applicable. The renormalized value used by the LEP electroweak working group is

$$\alpha_{em}^{-1}(m_Z) = 128.878 \pm 0.090 \quad (22)$$

However, this may be refined to $\alpha_{em}^{-1}(m_Z) = 128.933 \pm 0.021$ by more complete use of constraints from perturbative QCD and data on τ decays [18]. Beyond the tree level, the parameter $\sin^2 \theta_W$ may be defined in several different ways. One option is the “on-shell” definition $\sin^2 \theta_W \equiv 1 = m_W^2/m_Z^2$ [19]. The LEP experiments often use another physical definition more closely related to their experimental observables, as in Fig. 2, but theorists often favour the $\overline{\text{MS}}$ definition [19], which is more convenient for comparison with QCD and GUT calculations.

Consistency between the different measurements shown in Fig. 1 – e.g., the $\sin^2 \theta_W$ measurements shown displayed in Fig. 2 – imposes constraints on the masses of heavy virtual particles that appear in loop diagrams, such as the top quark and the Higgs boson [20, 21]. As examples of this, consider their contributions to m_W and m_Z in the “on-shell” renormalization scheme:

$$m_W^2 \sin^2 \theta_W = m_Z^2 \cos^2 \theta_W \sin^2 \theta_W = \frac{\pi\alpha}{\sqrt{2}G_\mu} (1 + \Delta r) \quad (23)$$

In the absence of the top quark, the gauge symmetry of the Standard Model would be lost, since the b quark would occupy an incomplete doublet of weak isospin, destroying the renormalizability of the theory at the one-loop level. This is reflected in the contributions of the one-loop vacuum-polarization diagrams [20]:

$$\Delta r \ni \frac{3G_\mu}{8\pi^2\sqrt{2}} m_t^2 + \dots \quad (24)$$

in the limit $m_t \gg m_b$. Likewise, the Standard Model would be non-renormalizable in the absence of a physical Higgs boson, so Δr must also blow up as $m_H \rightarrow \infty$. As pointed out by Veltman [21], a screening theorem restricts this to a logarithmic dependence at the one-loop level

$$\Delta r \ni \frac{\sqrt{2}G_\mu}{16\pi^2} m_W^2 \left\{ \frac{4}{3} \ln \frac{m_H^2}{m_W^2} + \dots \right\} \quad (25)$$

for $m_H \gg m_W$, though there is a quadratic dependence at the two-loop level.

Comparing (24) and (25), we see that the dependence on m_t is much greater than that on m_H . A measurement of Δr gives in principle an estimate of m_t , though with some uncertainty if m_H is allowed to vary between 10 GeV and 1 TeV. Before the start-up of LEP, we gave the upper bound $m_t \lesssim 170$ GeV [22, 23, 26]. By combining several different types of precision electroweak measurement, it is in principle possible to estimate independently both m_t and m_H . The present world data set implies [5]

$$m_t = 161_{-8}^{+9} \text{ GeV} \quad (26)$$

which is compatible with both the pre-LEP estimate and the direct measurements by CDF and $D\phi$ [24]:

$$m_t = 173.8 \pm 5.0 \text{ GeV} \quad (27)$$

Combining this with the precision electroweak data enables a more precise estimate of m_H to be made.

A key rôle in this estimate is being played by direct measurements of m_W . Until now, the most precise of these has been that from $\bar{p}p$ colliders, dominated by the Fermilab Tevatron collider [24]:

$$m_W = 80.41 \pm 0.09 \text{ GeV} \quad (28)$$

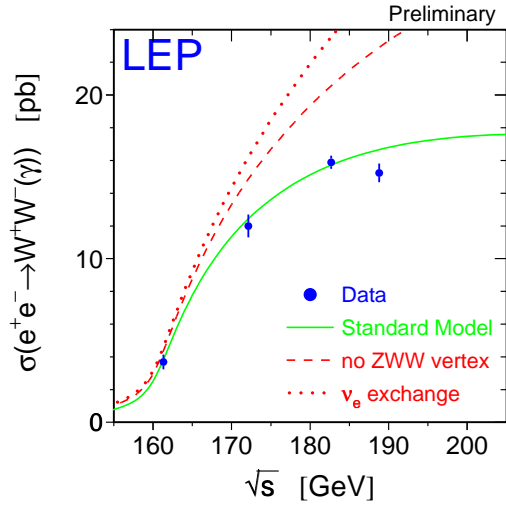


Fig. 3: Measurements of $\sigma(e^+e^- \rightarrow W^+W^-)$ [5].

with an honourable mention for the indirect determinations from deep-inelastic ν scattering:

$$m_W = 80.25 \pm 0.11 \text{ GeV} \quad (29)$$

with some slight dependence on m_t and m_H . These values can be compared with the indirect prediction based on other precision electroweak data [5], within the framework of the Standard Model:

$$m_W = 80.329 \pm 0.029 \text{ GeV} . \quad (30)$$

Reducing the error in the direct measurement (28) would constrain further the estimate of m_H within the Standard Model, and could constrain significantly its possible extensions, such as supersymmetry, with the error in (30) providing a relevant target for the experimental precision. This is also demonstrated by the implications for the error in the estimate of m_H corresponding to a given error in m_W :

$$\begin{aligned} \Delta m_W = & \quad 25 \quad 50 \quad \text{MeV} \\ m_H = 100 & \quad +86 \quad +140 \quad \text{GeV} \\ & \quad -54 \quad -72 \\ m_H = 300 & \quad +196 \quad +323 \quad \text{GeV} \\ & \quad -126 \quad -168 \end{aligned} \quad (31)$$

It is a major goal of the LEP 2 experimental programme to achieve such precision [25].

The measured cross-section for $e^+e^- \rightarrow W^+W^-$ is shown in Fig. 3 [5]. We see that ν exchange alone does not fit the data: one also needs to include both the γW^+W^- and $Z^0 W^+W^-$ vertices present in the Standard Model¹. The first LEP 2 measurement of m_W was obtained by measuring the cross section at $E_{cm} = 161$ GeV, close to the threshold, but this has now been surpassed in accuracy by the direct reconstruction of W^\pm decays at higher E_{cm} . The current LEP 2 average is [5]

$$m_W = 80.37 \pm 0.90 \text{ GeV} \quad (32)$$

which now matches the $\bar{p}p$ measurement error (28).

¹It is surely too soon to cry “new physics” on the basis of the cross-section measurement at $E_{cm} = 189$ GeV, particularly since the more recent data shown at the LEPC [16] indicate a lesser discrepancy!

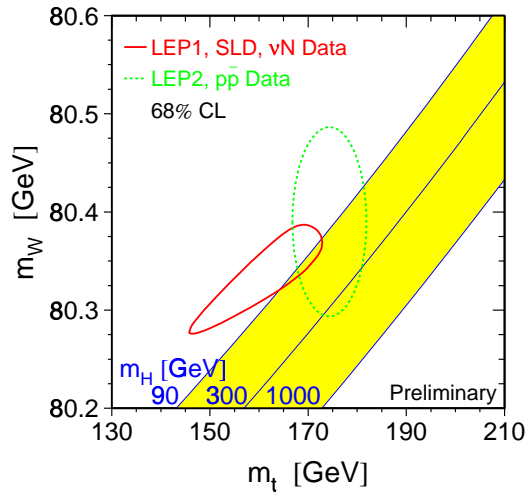


Fig. 4: Contours of m_W and m_t from direct measurements (dotted ellipse) compared with predictions based on the precision electroweak data analyzed within the Standard Model (solid curve) [5]. Note the sensitivity to m_H .

We see in Fig. 4 that the m_W measurements favour qualitatively $m_H < 300$ GeV, though not at a high level of significance [5]. Stronger evidence for a light Higgs boson [26] is provided by the lower energy LEP 1, SLD and νN data, as also seen in Fig. 4. Combining all the precision electroweak data, one finds

$$m_H = 76^{+85}_{-47} \pm 10 \text{ GeV} \quad (33)$$

as seen in Fig. 5, corresponding to $m_H < 260$ GeV at the 95% confidence level, if one uses a conservative error in $\alpha_{em}(m_Z)$ and makes due allowance for unknown higher-loop uncertainties in the analysis [5].

The range (33) may be compared with upper and lower bounds derived within the Standard Model. The tree-level unitarity limit $m_H \lesssim 1$ TeV [12, 13, 14] may be strengthened by including loop effects via renormalization-group calculations [27]. We see in Fig. 6 the upper bound on m_H that is obtained by requiring the Standard Model couplings to remain finite at all energies up to some cutoff Λ : $m_H \lesssim 200$ GeV if $\Lambda \simeq m_P$ and $m_H \lesssim 700$ GeV if $\Lambda \simeq m_H$, corresponding to upper limits from lattice calculations [28]. Also shown in Fig. 6 are lower limits on m_H obtained by requiring that the effective Higgs potential remain positive for $|\phi| \lesssim \Lambda$: $m_H \gtrsim 140$ GeV if $\Lambda \simeq m_P$ [27].

It is depressing to note that the range (33) of m_H estimated on the basis of the precision electroweak data is compatible with the Standard Model remaining valid all the way up to the Planck scale: $\Lambda \simeq m_P$ ². Moreover, the range (33) also imposes strong constraints on possible extensions of the Standard Model. For example, Fig. 7 shows that it effectively excludes a fourth generation [27]. Note that this renormalization-group argument is independent of the neutrino-counting argument (17) given earlier. In particular, this argument still holds if $m_{\nu_4} > m_{Z/2}$: in fact, it even becomes slightly stronger!

²Nevertheless, the range (33) is even more compatible with supersymmetry, which is one possible example physics of new physics at $\Lambda \lesssim 1$ TeV.

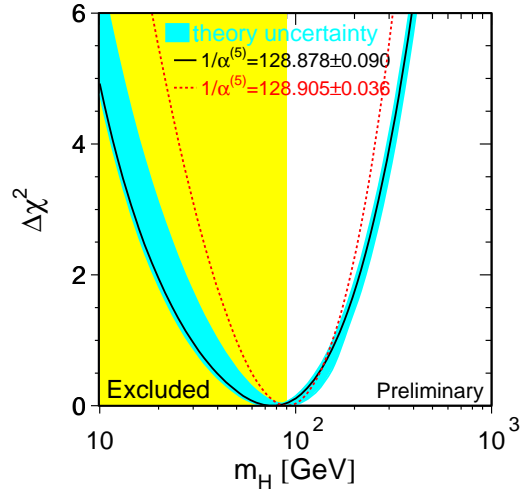


Fig. 5: The χ^2 function for a global fit to the precision electroweak data prefers $m_H \sim 100$ GeV [5].

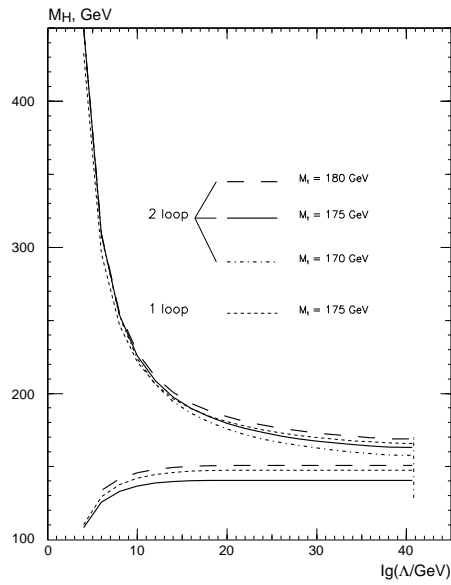


Fig. 6: The range of m_H compatible with the Standard Model remaining valid up to a high scale Λ [27].

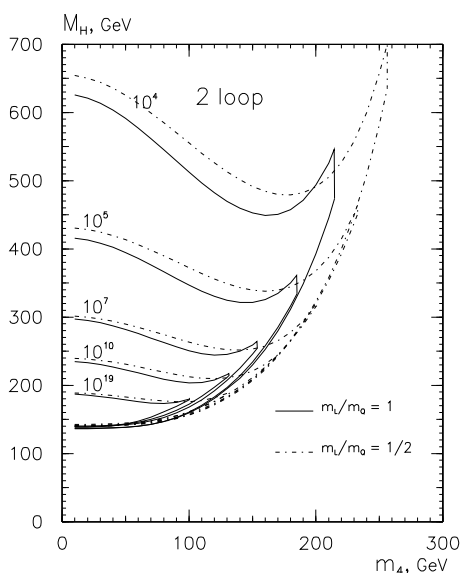


Fig. 7: The regions of fourth-generation mass m_4 and Higgs mass m_H compatible with the Standard Model remaining valid up to a high scale Λ [27].

1.3 Defects of the Standard Model

It has been said repeatedly that there is no confirmed experimental evidence from accelerators against the Standard Model, and several possible extensions have been ruled out. Nevertheless, no thinking physicist could imagine that the Standard Model is the end of physics. Even if one accepts the rather bizarre set of group representations and hypercharges that it requires, the Standard Model contains at least 19 parameters: 3 gauge couplings $g_{1,2,3}$ and 1 CP-violating non-perturbative vacuum angle θ_3 , 6 quark and 3 charged-lepton masses with 3 charged weak mixing angles and 1 CP-violating phase δ , and 2 parameters: (μ, λ) or (m_H, m_W) to characterize the Higgs sector. Moreover, many more parameters are required if one wishes to accommodate non-accelerator observations. For example, neutrino masses and mixing introduce at least 7 parameters: 3 masses, 3 mixing angles and 1 CP-violating phase, cosmological inflation introduces at least 1 new mass scale of order 10^{16} GeV, the cosmological baryon asymmetry is not explicable within the Standard Model, so one or more additional parameters are needed, and the cosmological constant may be non-zero. The ultimate “Theory of Everything” should explain all these as well as the parameters of the Standard Model.

It is convenient to organize the questions raised by the Standard Model into three broad categories. One is the **Problem of Mass**: do particle masses really originate from a Higgs boson, and, if so, why are these masses not much closer to the Planck mass $m_P \simeq 10^{19}$ GeV? This is the main subject of the next two lectures. Another is the **Problem of Unification**: can all the particle interactions be unified in a simple gauge group, and, if so, does it predict observable new phenomena such as baryon decay and/or neutrino masses, and does it predict relations between parameters of the Standard Model such as gauge couplings or fermion masses? This is the main subject of the fourth lecture. Then there is the **Problem of Flavour**: what is the origin of the six flavours each of quarks and leptons, and what explains their weak charged-current mixing and CP violation? This is the main subject of Yossi Nir’s lectures [29]. Finally, the quest for the **Theory of Everything** seems most promising in the context of string theory, particularly in its most recent incarnation of M theory, as discussed in the fifth lecture, and by Michael Green [30]. In addition to all the above problems, this should also reconcile quantum mechanics with general relativity, explain the origin of space-time and the number of dimensions,

2 INTRODUCTION TO SUPERSYMMETRY

2.1 The Electroweak Vacuum

We have discussed in Lecture 1 the fact that generating particle masses requires breaking the electroweak gauge symmetry spontaneously:

$$m_{W,Z} \neq 0 \Leftrightarrow \langle 0|X_{I,I_3}|0 \rangle \neq 0 \quad (34)$$

for some spin-0 quantity X with non-zero isospin I and third component I_3 . The fact that experimentally $\rho \equiv m_W^2/m_Z^2 \cos^2 \theta_W \simeq 1$ is consistent with the Standard Model expectation that X has mainly $I = \frac{1}{2}$ [4]. This is also what is required to give non-zero fermion masses: $m_f \bar{f}_L f_R + h.c.$, since the $f_{L,R}$ have $I = \frac{1}{2}, 0$. The question then remains: what is the nature of X ? In particular, is it elementary or composite?

The former is the option chosen in the Standard Model: $X = H : \langle 0|H^0|0 \rangle \neq 0$. However, as discussed in more detail later, quantum corrections to the squared mass of an elementary Higgs boson diverge quadratically:

$$\delta m_H^2 = 0 \left(\frac{\alpha}{\pi} \right) \Lambda^2 \quad (35)$$

where Λ is some cutoff, corresponding physically to the scale up to which the Standard Model remains valid. We discuss later the possibility that Λ can be identified with the energy threshold for symmetry. This should occur at $\lambda \lesssim 1$ TeV, in order that the quantum corrections (35) have the same magnitude as the physical Higgs boson mass.

The alternative option is that X is composite, namely a fermion-antifermion condensate $\langle 0|\bar{F}F|0 \rangle \neq 0$. This idea is motivated by the existence of a quark-antiquark condensate $\langle 0|\bar{q}q|0 \rangle \neq 0$ in QCD, and the rôle of Cooper pairs $\langle 0|e^-e^-|0 \rangle \neq 0$ in conventional superconductivity. Two major possibilities for the condensate have been considered: a top-antitop condensate $\langle 0|\bar{t}t|0 \rangle \neq 0$ held together by a large Yukawa coupling $\lambda_{H\bar{t}t}$ [31], and technicolour [32], in which new interactions that become strong at an energy scale $\Lambda \sim 1$ TeV bind new strongly-interacting technifermions: $\langle 0|\bar{T}T|0 \rangle \neq 0$. The $\bar{t}t$ condensate idea is currently disfavoured, since simple implementations require $m_t > 200$ GeV in contradiction with experiment, so we concentrate here on the technicolour alternative.

The technicolour idea [32] was initially modelled on the known dynamics of QCD:

$$\langle 0|\bar{q}_L q_R + h.c.|0 \rangle \neq 0 \rightarrow \langle 0|\bar{F}_L F_R + h.c.|0 \rangle \neq 0 \quad (36)$$

which breaks isospin with $I = \frac{1}{2}$, if the electroweak multiplet assignments of the F_L and F_R are the same as those of the q_L and q_R . The scale of this breaking will be appropriate if $\Lambda_{QCD} \rightarrow \Lambda_{TC} \sim 1$ TeV. Just as QCD contains (if $m_q = 0$) massless pions with the axial-current matrix element $\langle 0|A_\mu|\pi \rangle = ip_\mu f_\pi$, one expects a similar coupling

$$\langle 0|J_\mu|\pi_{TC} \rangle = ip_\mu F_\pi \quad (37)$$

for the technipion π_{TC} , which does the same Goldstone-eating job as (8), (9) if $F_\pi = v \simeq 250$ GeV. If there are two massless flavours of technifermions, one expects 3 massless technipions to be eaten by the W^\pm and Z^0 , and the physical Higgs boson is replaced by an effective massive composite scalar, analogous to the scalar mesons of QCD and weighing $0(1)$ TeV. However, a single technidoublet is not enough when one imposes the necessary cancellation of anomalies and tries to give masses to conventional fermions [33]. For these reasons, the Standard Technicolour Model used to include a full technigeneration: $[(N, E), (U, D)_{1,2,3}]_{1,\dots,N_{TC}}$, where the indices

denote colour and technicolour indices. For generality, one can study models as functions of the numbers of techniflavours and technicolours: (N_{TF}, N_{TC}) .

Their effects via one-loop quantum corrections can be parametrized in terms of their contributions to electroweak observables via three combinations of vacuum polarizations [34, 35]. One example is:

$$T \equiv \frac{\epsilon_1}{\alpha} \equiv \frac{\Delta\rho}{\alpha} : \Delta\rho = \frac{\pi_{ZZ}(0)}{m_Z^2} - \frac{\pi_{WW}(0)}{m_W^2} - 2 \tan \theta_W \frac{\pi_{\gamma Z}(0)}{m_Z^2} \quad (38)$$

which describes deviations from the tree-level relation $\rho \equiv m_W^2/m_Z^2 \cos^2 \theta = 1$ and measures isospin-breaking effects. This is related to Δr (23) and receives contributions from Standard-Model particles:

$$T \ni \frac{3}{16\pi} \frac{1}{\sin^2 \theta_W \cos^2 \theta_W} \left(\frac{m_t^2}{m_W^2} \right) - \frac{3}{16\pi \cos^2 \theta_W} \ln \left(\frac{m_H^2}{m_Z^2} \right) + \dots \quad (39)$$

The other relevant combinations of vacuum polarizations are [34, 35]

$$S \equiv \frac{4 \sin^2 \theta_W}{\alpha} \quad \epsilon_3 \ni \frac{1}{12\pi} \ln \left(\frac{m_H^2}{m_Z^2} \right) + \dots \quad (40)$$

and [34, 35]

$$U \equiv -\frac{4 \sin^2 \theta_W}{\alpha} \quad \epsilon_2 \quad (41)$$

The precision electroweak data may be used to constrain (S, T, U) (or $\epsilon_{1,2,3}$), and thereby possible extensions of the Standard Model with the same $SU(2) \times U(1)$ gauge group and additional matter particles, such as technicolour. Note, however, that this approach is not adequate for precision analyses of theories with important vertex diagrams such as the Standard Model or its minimal supersymmetric extension, to be discussed later. These have important vertex and box diagrams, as well as the vacuum-polarization diagrams taken care of by $S, T, U(\epsilon_{1,2,3})$. Some of these be treated by introducing further parameters such as ϵ_b for the $Z\bar{b}b$ vertex. Even so, two-loop and other higher-order effects are not treated exactly in this approach.

Figure 8 compiles the constraints on $\sin^2 \theta_W$ and the overall Z weak coupling strength coming from the precision electroweak data (top panel) and the resulting constraints on S and T (bottom panel) [36]. We see that the lower-energy data are compatible with the high-energy data at around the one- σ level, and that the high-energy data impose strong constraints on S, T, U (equivalent to $\epsilon_{1,2,3}$). Figure 9 shows the available constraints in the (ϵ_1, ϵ_2) plane [37]. We see that the one-loop corrections are certainly needed, since the data lie many σ away from the (improved) Born approximation. We also see that the data are quite consistent with the Standard Model. A compilation of determinations of the ϵ_i are shown in Fig. 10 [37], where we see a discrepancy only in ϵ_b , but even this is only slightly more than one σ . Finally, Fig. 11 compares the data constraint in the (S, T) plane not only with the Standard Model but also with various technicolour models [36]. The models chosen all have one technidoublet, and hence $N_{TF} = 2$, and varying values of $N_{TC} = 2, 3, 4$. We see that even the least disfavoured model is at least four σ away from the data, and models with larger N_{TC} (shown) and N_{TF} (not shown) deviate even further from experiment.

This large discrepancy has almost been the death of technicolour models, but various suggestions have been made that one could respect the experimental constraints if the technicolour dynamics is somewhat different from that of QCD. Specifically, it has been suggested that the technicolour coupling may not run as rapidly as the strong coupling [38]. Unfortunately, calculations in this framework of “walking technicolour” cannot be made as precisely as in the

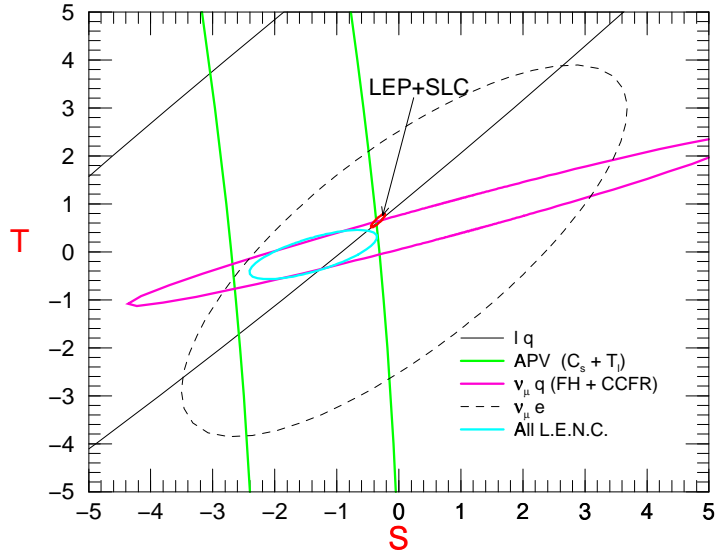


Fig. 8: Constraints on S and T inferred from precision electroweak measurements at low and high energies [36].

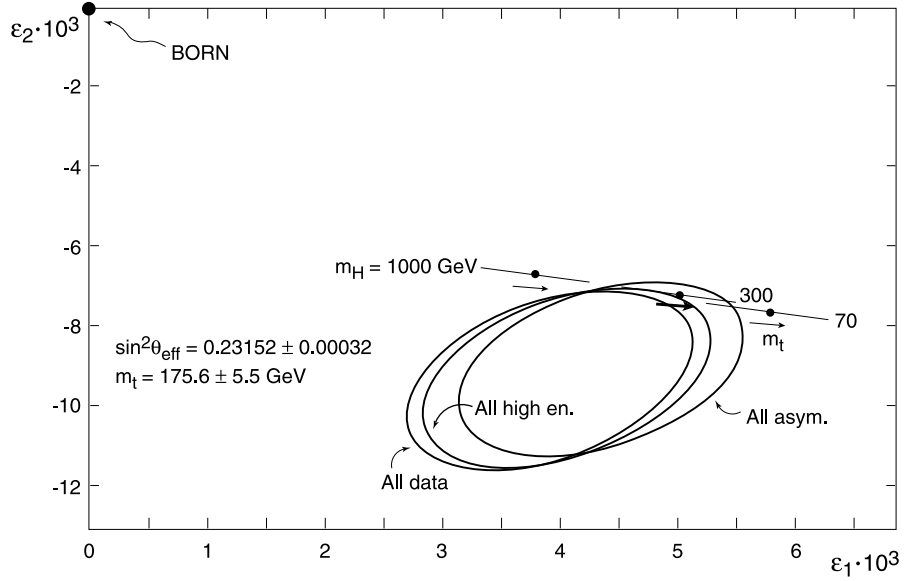


Fig. 9: Constraints on (ϵ_1, ϵ_2) from precision electroweak measurements, which agree with the Standard Model predictions [37]. Note that the data lie many standard deviations away from the values expected (“Born”) if quantum corrections are neglected.

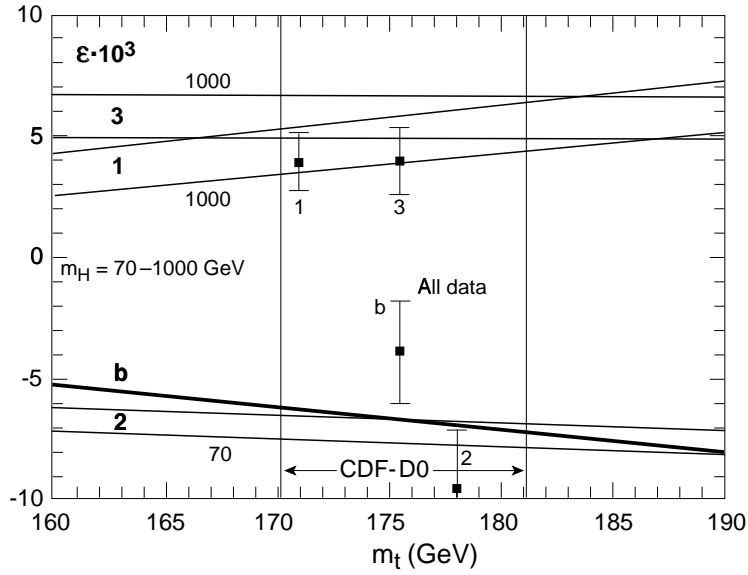


Fig. 10: Values of the quantum corrections ϵ_i inferred from the data [37], compared with the Standard Model predictions.

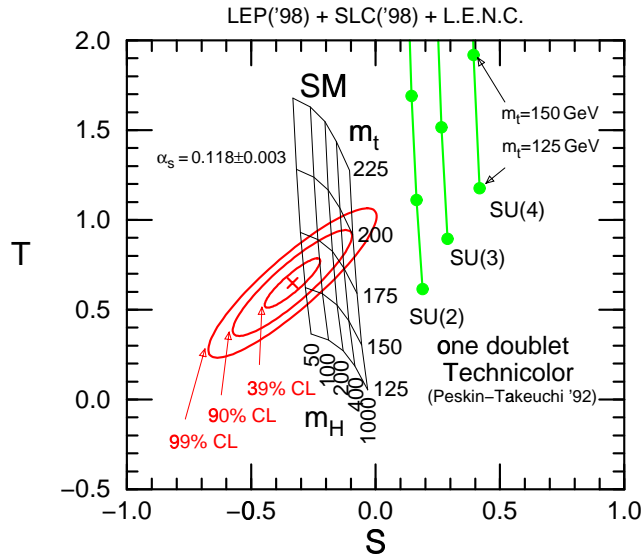


Fig. 11: Values of S and T extracted from the data (ellipses) [36], confronted with the predictions of the Standard Model and of one-doublet technicolour models.

conventional technicolour models discussed above, rendering it difficult to test or disprove. For the moment, no calculable technicolour model is consistent with the precision electroweak data, so we turn to supersymmetry.

2.2 Introduction to supersymmetry

Back in the 1960's there were many (forgettable) attempts to combine internal symmetries such as flavour isospin or $SU(3)$ with relativistic external symmetries such as Lorentz invariance. However, in 1967 Coleman and Mandula [39] proved that this could not be done using only bosonic charges. The way to avoid this no-go theorem was found in 1971, when Gol'fand and Likhtman [40] showed that one could extend the Poincaré algebra using fermionic charges. In the same year, Neveu, Schwarz and Ramond [41] invented supersymmetry in two dimensions when they discovered how to incorporate fermions in string models. Supersymmetric field theories in four dimensions were discovered in 1973, by Volkov and Akulov [42] in a non-linear realization, and by Wess and Zumino [43] in the linear realization now used in most model-building. Soon afterwards, Wess, Zumino, Iliopoulos and Ferrara [44] realized that supersymmetric models were free of many of the divergences found in other four-dimensional field theories. Then, in 1976, Freedman, van Nieuwenhuizen and Ferrara [45] and independently Deser and Zumino [46] showed how supersymmetry can be realized locally (by analogy with gauge theories) in the context of supergravity.

Many of the ideas for using supersymmetry were motivated by the desire to unify known bosons and fermions: for example, unifying mesons and baryons motivated the early string work [41] and that of Wess and Zumino [43]. It was initially suggested that neutrino could be a Goldstone fermion in a non-linear realization of supersymmetry [42], but it was soon pointed out that experimental data on ν interaction cross sections conflicted with theorems on the low-energy behaviour in such theories [47]. The fact that supersymmetric theories had fewer (in some cases, no) divergences offered to some people who never liked infinite renormalizations hope that one could construct a finite theory. Others were attracted by the idea that supersymmetry might relate the odd-person-out Higgs boson to fermionic matter and perhaps gauge bosons. At a more fundamental level, the fact that local supersymmetry involves gravity suggested to many the idea of unifying all the particles and their interactions in some supergravity theory. However, this motivation did not provide a clear clue as to the mass scale of supersymmetry breaking, so there was no obvious reason why the sparticle masses should not be as heavy as $m_P \simeq 10^{19}$ GeV.

Such a reason was eventually provided by the mass hierarchy problem [48]: why is $m_W \ll m_P$? The latter is the only candidate we have for a fundamental mass scale in physics, where gravity is expected to become as strong as other particle interactions, e.g., graviton exchange at LEP 10^{19} would be comparable to γ and Z^0 exchange. The hierarchy problem can be rephrased as: “why is $G_F \gg G_N$?”, since $G_F \sim 1/m_W^2$ and $G_N = 1/m_P^2$. Alternatively, for the benefit of atomic, molecular and condensed-matter physicists, not to mention chemists and biologists, one can ask: why is the Coulomb potential in an atom so much larger than the Newton potential? The former is $e^2/r : e^2 = 0(1)$, whereas the latter is $G_N m_p m_e / r$, so the Newton potential is negligible just because conventional particle masses $m_{p,e}$ are much lighter than m_P .

You might think that one could just set $m_W \ll m_P$ by hand, and ignore the problem. However, there is a threat from radiative corrections [48]. Each of the one-loop diagrams in Fig. 12 is individually quadratically divergent, implying

$$\delta m_{H,W}^2 = \mathcal{O} \left(\frac{g^2}{16\pi^2} \right) \int^\Lambda d^4 k \frac{1}{k^2} = \mathcal{O} \left(\frac{\alpha}{\pi} \right) \Lambda^2 \quad (42)$$

where the cutoff Λ in the integral represents the scale up to which the Standard Model remains

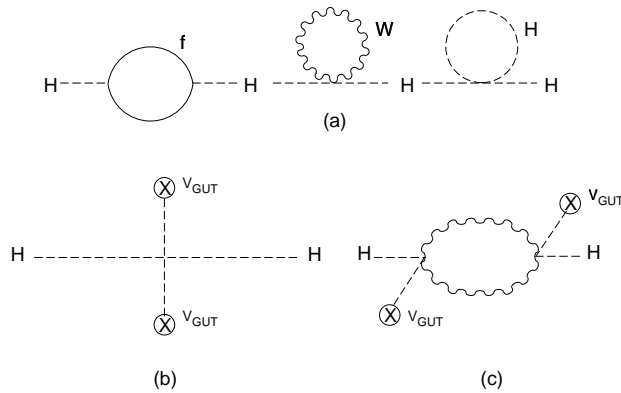


Fig. 12: (a) One-loop quantum corrections to m_H^2 in the Standard Model. (b) Tree-level and (c) one-loop corrections to m_H^2 in a GUT.

valid, and beyond which new physics sets in. If we think $\Lambda \simeq m_P$ or the grand unification scale, the quantum correction (42) is much larger than the physical value of $m_{H,W} \sim 100$ GeV. This is not a problem for renormalization theory: there could be a large bare contribution with the opposite sign, and one could fine-tune its value to many significant figures so that the physical value $m_{H,W}^2 \simeq \delta m_{H,W}^2$ (42). However, this seems unnatural, and would have to be repeated order by order in perturbation theory. In contrast, the one-loop corrections to a fundamental fermion mass m_f are proportional to m_f itself, and only logarithmically divergent:

$$\delta m_f = \mathcal{O}\left(\frac{g^2}{16\pi^2}\right) m_f \int^\Lambda d^4k \frac{1}{k^4} = \mathcal{O}\left(\frac{\alpha}{\pi}\right) m_f \ln \frac{\Lambda}{m_f} \quad (43)$$

This correction is no larger numerically than the physical value, for any $\Lambda \lesssim m_P$. This is because there is a chiral symmetry reflected in the m_f factor in (43) that keeps the quantum corrections naturally (logarithmically) small. The hope is to find a corresponding symmetry principle to make small boson masses natural: $\delta m_{H,W}^2 \lesssim m_{H,W}^2$.

This is achieved by supersymmetry [49], exploiting the fact that the boson and fermion loop diagrams in Fig. 12a have opposite signs. If there are equal numbers of fermions and bosons, and if they have equal couplings as in a supersymmetric theory, the quadratic divergences (42) cancel:

$$\delta m_{H,W}^2 = -\left(\frac{g_F^2}{16\pi^2}\right) (\Lambda^2 + M_F^2) + \left(\frac{g_B^2}{16\pi^2}\right) (\Lambda^2 + M_B^2) = \mathcal{O}\left(\frac{\alpha}{4\pi}\right) |m_B^2 - m_F^2| \quad (44)$$

This is no larger than the physical value: $\delta m_{H,W}^2 \lesssim m_{H,W}^2$, and hence naturally small³, if

$$|m_B^2 - m_F^2| \lesssim 1 \text{ TeV}^2 \quad (45)$$

This naturalness argument [48] is the only available theoretical motivation for thinking that supersymmetry may manifest itself at an accessible energy scale.

However, this argument is qualitative, and a matter of taste. It does not tell us whether sparticles should appear at 900 GeV, 1 TeV or 2 TeV, and some theorists reject it altogether. They say that, in a renormalizable theory such as the Standard Model, one need not worry about the fine-tuning of a bare parameter, since it is not physical. However, I take naturalness seriously as a physical argument: it is telling us that a large hierarchy is intrinsically unstable,

³There is a logarithmic multiplicative factor in the right-hand side of (44) that is reflected in the discussion below of renormalization-group corrections to supersymmetric particle masses.

and supersymmetry is the most plausible way of stabilizing it. Moreover, many logarithmic divergences are absent in supersymmetry, which stabilizes the possible GUT Higgs corrections to m_H shown in Fig. 12b arising from the loops shown in Fig. 12c, which is also important for stabilizing the hierarchy $m_W \ll m_{GUT}$, as we see later.

2.3 What is supersymmetry?

After all this introduction and motivation, just what is supersymmetry [49]? It is a symmetry that links bosons and fermions via spin- $\frac{1}{2}$ charges Q_α (where α is a spinorial index). It seems to be the last possible symmetry of the particle scattering matrix [50]. As such, many would argue that it must inevitably play a rôle in physics, and it has in fact already appeared at a phenomenological level in condensed-matter, atomic and nuclear physics. All previously-known symmetries are generated by bosonic charges, which are, apart from the momentum operator P_μ associated with Lorentz invariance, scalar charges Q that relate different particles of the same spin J : $Q|J\rangle = |J'\rangle$, $Q \in U(1), SU(2), SU(3), \dots$. Indeed, as already mentioned, Coleman and Mandula [39] showed that it was impossible to mix such internal symmetries with Lorentz invariance using bosonic charges. The essence of their proof is easy to grasp.

Consider $2 \rightarrow 2$ scattering: $1 + 2 \rightarrow 3 + 4$, and consider the possibility that there is a conserved tensor charge $\Sigma_{\mu\nu}$ corresponding to some higher bosonic symmetry (there can be no other charge with one vector index, besides P_μ , and higher tensor charges can be discussed analogously to $\Sigma_{\mu\nu}$). Its diagonal matrix elements are required by Lorentz invariance to have the following tensor decomposition:

$$\langle a|\Sigma_{\mu\nu}|a\rangle = \alpha p_\mu^a p_\nu^a + \beta g_{\mu\nu} \quad (46)$$

where p_μ^a is the four-momentum of the particle a and α, β are unknown reduced matrix elements. For $\Sigma_{\mu\nu}$ to be conserved in the scattering process, as long as $\alpha \neq 0$ ⁴ one must require

$$p_\mu^1 p_\nu^1 + p_\mu^2 p_\nu^2 = p_\mu^3 p_\nu^3 + p_\mu^4 p_\nu^4 \quad (47)$$

as well as $p_\mu^1 + p_\mu^2 = p_\mu^3 + p_\mu^4$. It is easy to convince oneself that the only possible simultaneous solutions to these linear and quadratic conservation conditions correspond to purely forward scattering. This conflicts [39] with the basic principles of quantum field theory as well as experiment.

This argument is fine as far as it goes, but it does not apply to any spinorial charge Q_α , since the diagonal matrix elements vanish: $\langle a|Q_\alpha|a\rangle = 0$.

Let us explore now what is the possible algebra of an algebra of such spinorial charges $Q_\alpha^i : i = 1, 2, \dots, N$ [50]. If they are to be symmetry generators, they must commute with the Hamiltonian:

$$[Q_\alpha^i, H] = 0 \quad (48)$$

Hence, their anticommutator (which is bosonic) must also commute with H :

$$[\{Q_\alpha^i, Q_\beta^j\}, H] = 0 \quad (49)$$

By the Coleman-Mandula theorem [39], this anticommutator must be a combination of the conserved Lorentz vector charge P_μ and some scalar charge Z^{ij} . The only possible form is in fact

$$\{Q_\alpha^i, Q_\beta^j\} = 2\delta^{ij}(\gamma^\mu C)_{\alpha\beta} P_\mu + Z^{ij} \quad (50)$$

⁴The case $\alpha = 0$ corresponds to a scalar charge $\Sigma_{\mu\nu} = \hat{\Sigma}g_{\mu\nu}$.

where we use four-component spinors, C is the charge-conjugation matrix and Z^{ij} is antisymmetric in the supersymmetry indices $\{i, j\}$. Thus, this so-called ‘‘central charge’’ vanishes for the $N = 1$ case of phenomenological relevance.

The basic building blocks of $N = 1$ supersymmetric theories are supermultiplets containing the following helicity states [49]:

$$\text{chiral} : \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, \quad \text{gauge} : \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}, \quad \text{graviton} : \begin{pmatrix} 2 \\ \frac{3}{2} \end{pmatrix}, \quad (51)$$

which are used to describe matter and Higgses, gauge fields and gravity, respectively. You may wonder why one does not use theories with extended supersymmetry: $N \geq 2$. The building blocks for $N = 2$ are:

$$\text{matter} : \begin{pmatrix} \frac{1}{2} \\ 0 \\ -\frac{1}{2} \end{pmatrix}, \quad \text{gauge} : \begin{pmatrix} 1 \\ \frac{1}{2} \\ 0 \end{pmatrix}, \quad \text{gravity} : \begin{pmatrix} 2 \\ \frac{3}{2} \\ 1 \end{pmatrix}, \quad (52)$$

and it is apparent that left- and right-handed particles (helicities $\mp 1/2$) must be in identical representations of the gauge group. This is immediate for the matter supermultiplet in (52), and must also be the case for fermions in the gauge supermultiplet, since the helicity ∓ 1 must be in identical adjoint representations. Hence an $N = 2$ theory cannot accommodate parity violation, and is not suitable for phenomenology ⁵.

The simplest $N = 1$ supersymmetric field theory contains a free fermion and a free boson [49, 52]:

$$\mathcal{L} = \partial_\mu \phi^* \partial^\mu \phi + i \psi^\dagger \bar{\sigma} \cdot \partial \psi \quad (53)$$

where we work with two-component spinors and denote $\sigma_\mu = (1, \underline{\sigma})$, $\bar{\sigma}_\mu = (1, -\underline{\sigma})$, where the $\underline{\sigma}$ are Pauli matrices. The simple supersymmetry transformation laws are

$$\delta_\xi \phi = \sqrt{2} \xi^T C \psi, \quad \delta_\xi \psi = \sqrt{2} i \sigma \cdot \partial \phi C \xi^* \quad (54)$$

where ξ is an infinitesimal spinor parameter and C is the conjugation matrix: $C = -i\sigma^2 = C^*$, $C^{-1} = C^T = -C$. It is easy to check that under (54) the Lagrangian (53) changes by a total derivative $\partial_\mu(\dots)$, and hence the action $A = \int d^4x \mathcal{L}(x)$ is invariant. We can also see in (54) a reflection of the supersymmetry algebra (50): after two supersymmetry transformations, the fields (ϕ, ψ) are transformed by derivatives $(\partial\phi, \partial\psi)$, corresponding to the action of the momentum operator $P_\mu = i\partial_\mu$.

The example (54) can easily be extended to include interactions [49, 52]:

$$\mathcal{L} = \partial_\mu \phi^* \partial^\mu \phi + i \psi^\dagger \bar{\sigma} \cdot \partial \psi + F^\dagger F + \left(F \frac{\partial W}{\partial \phi} - \frac{1}{2} \psi^T C \psi \frac{\partial^2 W}{\partial \phi^2} + \text{herm.conj.} \right) \quad (55)$$

with supersymmetry transformations:

$$\delta_\xi \phi = \sqrt{2} \xi^T C \psi, \quad \delta_\xi \psi = \sqrt{2} i \sigma \cdot \partial \phi C \xi^* + \xi F, \quad \delta_\xi F = -\sqrt{2} i \xi^\dagger \bar{\sigma} \cdot \partial \psi \quad (56)$$

The field F is called an auxiliary field: notice that it has no kinetic term, and so may be eliminated by using an equation of motion:

$$F^\dagger = -\frac{\partial W}{\partial \phi} \quad (57)$$

⁵Moreover, there are severe lower limits, in the context of unified theories, on the possible renormalization scale down to which $N = 2$ supersymmetry may remain valid [51].

Thus all the matter interactions are characterized by the analytic function $W(\phi)$, which is called the superpotential. Renormalizability of the field theory requires the superpotential to be a cubic function: for $W = \lambda\phi_1\phi_2\phi_3$, one obtains from (55) the following particle interactions:

$$\lambda \left[(\psi_1^T C \psi_2) \phi_3 + (\psi_2^T C \psi_3) \phi_1 + (\psi_3^T C \psi_1) \phi_2 \right] + |\lambda\phi_1\phi_2|^2 + |\lambda\phi_2\phi_3|^2 + |\lambda\phi_3\phi_1|^2 \quad (58)$$

where the last terms provide a quartic potential for the scalar fields ϕ_i and are called in the jargon “ F terms”.

We shall not discuss here in detail the construction of the interactions of a chiral supermultiplet with a gauge supermultiplet [49], limiting ourselves to quoting the results. In addition to the gauge interactions of the chiral fermions and their bosonic partners, there are gaugino interactions

$$\sqrt{2}g \left[(\psi_i^T C (T^a)_j^i V_a) \phi^{j*} + \text{herm.conj.} \right] \quad (59)$$

where $(T^a)_j^i$ is the gauge representation matrix for the chiral fields. There is also another quartic potential term for the scalars:

$$V = \frac{g^2}{2} \sum_a |\phi^{i*} (T^a)_i^j \phi_j|^2 \quad (60)$$

which are called in the jargon “ D terms”. Finally, we note for completeness that the conventional gauge-boson kinetic term and the gauge interactions of fermions in the adjoint representation of the gauge group, such as the gauginos \tilde{V}_a , are automatically supersymmetric.

2.4 Minimal Supersymmetric Extension of the Standard Model

Let us now return to phenomenology. If one is to construct a minimal supersymmetric model, the first natural question is: can one construct it out of the Standard Model particles alone? It is easy to see that this is impossible, because the known bosons and fermions have different conserved quantum numbers [53]. For example, gluons are in an octet ($\underline{8}$) representation of colour, whereas quarks are in triplet ($\underline{3}$) representation of colour. Similarly, there are no known weak-isotriplet fermions, as would be needed to partner the electroweak gauge bosons. The known leptons are isodoublets like the Higgs boson, but they carry lepton number, unlike the Higgs. For these reasons, new particles must be postulated [53] as supersymmetric partners of known particles, as seen in the Table.

The minimal supersymmetric extension of the Standard Model (MSSM) [54] has the same gauge interactions as the Standard Model. In addition, there are couplings of the form (58) derived from the following superpotential:

$$W = \lambda_d Q D^C H + \lambda_\ell L E^C H + \lambda_u Q U^C \bar{H} + \mu \bar{H} H \quad (61)$$

Here, $Q[L]$ denote isodoublets of supermultiplets containing $(u, d)_L[(\nu, \ell)_L]$, $D^C[U^C, E^C]$ are singlets containing the left-handed conjugates $d_L^C[u_L^C, e_L^C]$ of the right-handed $d_R[u_R, e_R]$, and the superpotential couplings $\lambda_d[\lambda_{u,\ell}]$ correspond to the Yukawa couplings of the Standard Model that give masses to the $d[u, \ell^-]$, respectively:

$$m_d = \lambda_d \langle H \rangle, \quad m_u = \lambda_u \langle \bar{H} \rangle, \quad m_\ell = \lambda_\ell \langle H \rangle. \quad (62)$$

Each of these should be understood as a 3×3 matrix in generation space, which is to be diagonalized as in the Standard Model.

In addition to the Standard-Model-like superpotential interactions shown in (61), the following superpotential couplings [55] are also permitted by the gauge symmetries of the Standard Model:

$$W \ni \lambda L L E^C + \lambda' Q D^C L + \lambda'' U^C D^C D^C \quad (63)$$

Particle	Spin	Spartner	Spin
quark: q	$\frac{1}{2}$	squak: \tilde{q}	0
lepton: ℓ	$\frac{1}{2}$	slepton: $\tilde{\ell}$	0
photon: γ	1	photino: $\tilde{\gamma}$	$\frac{1}{2}$
W	1	wino: \tilde{W}	$\frac{1}{2}$
Z	1	zino: \tilde{Z}	$\frac{1}{2}$
Higgs: H	0	higgsino: \tilde{H}	$\frac{1}{2}$

Table 1: Particles in the Standard Model and their supersymmetric partners.

Each of these violate conservation of either lepton number L or baryon number B . The possible presence of such interactions attracted some interest in 1997 [56, 57] with the discovery of unexpectedly many events at HERA at large x and Q^2 [58], but interest has now subsided. Their potential significance is only discussed intermittently in these lectures.

Note that (61) requires two Higgs doublets H, \bar{H} with opposite hypercharges in order to give masses to all the matter fermions. In the Standard Model, one doublet ϕ and its complex conjugate ϕ^\dagger would have sufficed. This does not work in the MSSM, because the superpotential W must be an analytic function of the fields. Moreover, Higgs supermultiplets include Higgsino fermions that generate triangle anomalies which must cancel among themselves, requiring at least two Higgs doublets. These couple via the μ term in (61). Note also that the ratio of Higgs vacuum expectation values

$$\tan \beta \equiv \frac{\langle \bar{H} \rangle}{\langle H \rangle} \quad (64)$$

is undetermined and should be treated as a free parameter. Finally, we comment that the superpotential and gauge couplings determine the MSSM's quartic scalar couplings, providing important constraints on the Higgs masses, as we see later.

Before discussing in more detail the phenomenology of the MSSM, it is appropriate to mention two important but indirect experimental indications that favour supersymmetry. One is the relatively light mass of the Higgs boson inferred from the analysis of precision electroweak data [5], as seen in Fig. 5. As discussed in more detail in the next Lecture, the lightest MSSM Higgs boson must weigh $\lesssim 150$ GeV [59], in good agreement with the range favoured by the data. The other indication in favour of supersymmetry is the measured value of $\sin^2 \theta_W$ [5], as shown in Fig. 2. As discussed in more detail in Lecture 4, Grand Unified Theories predict $\sin^2 \theta_W$ as a function of $\alpha_s(m_Z)$. For the measured value of $\alpha_s(m_Z)$, GUTs without supersymmetry predict $\sin^2 \theta_W \sim 0.21$ to 0.22 [7], whereas GUTs with supersymmetry at the TeV scale predict $\sin^2 \theta_W \sim 0.23$ [9], in much better agreement with the data [60].

These two experimental arguments buttress the theoretical argument given earlier, which was based on the hierarchy problem. Put together, these provide ample motivation for studying the phenomenology of the MSSM in more detail, as we do in the next Lecture.

3.1 Soft Supersymmetry Breaking

The first issue that must be addressed in the phenomenology of supersymmetry is the sad fact that no sparticles have ever been detected. This means that sparticles do not weigh the same as their supersymmetric partners: $m_{\tilde{e}} \neq m_e$, $m_{\tilde{\gamma}} \neq m_\gamma$, etc., and hence that supersymmetry must be broken. We return in Lecture 5 to review some theoretical ideas about the origin of supersymmetry breaking, restricting ourselves here to a phenomenological parametrization [61]. Any such parametrization should retain the desirable features of supersymmetry, particularly the absence of power-law divergences. This “softness” requirement means that any supersymmetry-breaking interactions $\mathcal{L}_{\text{susyX}}$ should have quantum field dimension < 4 (recall that the quantum field dimension of a boson (derivative) (fermion) is $1(1)(\frac{3}{2})$), and hence a positive power of some numerical mass parameters, so that $\int d^4x \mathcal{L}_{\text{susyX}}$ is dimensionless. There are in fact further restrictions on soft supersymmetry-breaking parameters [62], and a general parametrization comprises scalar mass terms: $m_{0_i}^2 |\phi_i|^2$, gaugino masses: $\frac{1}{2} M_a \tilde{V}_a^T C \tilde{V}_a$, and trilinear or bilinear scalar interactions proportional to superpotential terms: $A_\lambda \lambda \phi^3$, $B_\mu \mu \phi^2$. Note some absences from this list, including masses for fermions in chiral supermultiplets $m_\psi \psi^T C \psi$ and non-analytic trilinear scalar couplings $\propto \phi^* \phi^2$.

We shall adopt for now the hypothesis (to be discussed in Lecture 5) that the soft supersymmetry-breaking masses $m_{0_i}^2, M_a, A_\lambda, B_\mu$ originate at some high GUT or gravity scale, perhaps from some supergravity or superstring mechanism. The physical values of the soft supersymmetry-breaking parameters are then subject to logarithmic renormalizations that may be calculated and resummed using the renormalization-group techniques familiar from QCD [63], which also figure in the GUT calculations of $\sin^2 \theta_W$ that are reviewed in Lecture 4. Renormalizations by gauge interactions have the general structure

$$m_{0_i}^2 \rightarrow m_{0_i}^2 + C_i^a M_a^2, \quad M_a \rightarrow \frac{\alpha_a}{\alpha_{GUT}} M_a \quad (65)$$

at the one-loop level, and higher-loop renormalizations are also well understood [64].

It is often assumed that the soft supersymmetry-breaking masses are universal at the GUT or supergravity scale:

$$m_{0_i}^2 \equiv m_0^2, \quad M_a \equiv m_{1/2}, \quad A_\lambda \equiv A, \quad B_\mu \equiv B \quad (66)$$

but this hypothesis is not very well motivated, since, in particular, general supergravity models give no theoretical hint why they should be universal. Some superstring models give hints of universality for the gaugino masses $m_{1/2}$, but universality for the scalar masses $m_{0_i}^2$ is more questionable. Since a high degree of universality is suggested (at least for the first two generations) by flavour-changing neutral-current (FCNC) constraints [65], this provides some impetus for models guaranteeing scalar-mass universality, such as the gauge-mediated or messenger models [66] discussed briefly in Lecture 5. If one assumes universality, the parameters $\mu, \tan \beta, m_0, m_{1/2}, A$ suffice to characterize MSSM phenomenology.

Figure 13 shows the results of some typical renormalization-group calculations assuming universal inputs [67]. We see that scalar masses are generally renormalized to larger values as the scale is reduced, but this is not necessarily the case if there are large Yukawa interactions such as those of the top quark, which may modify (65) in the case of Higgs masses. Such Yukawa effects involving the top quark must certainly be taken into account, and could also be important for the bottom quark and the τ lepton if $\tan \beta$ is large. The potential significance of these Yukawa interactions is that they tend to drive m_H^2 to smaller values at smaller renormalization scales μ [68]

$$\mu \frac{d}{\ln \mu} m_h^2 = \frac{1}{(4\pi)^2} \left(3\lambda_t^2 (m_h^2 + m_q^2 + m_t^2) + \dots \right) \quad (67)$$

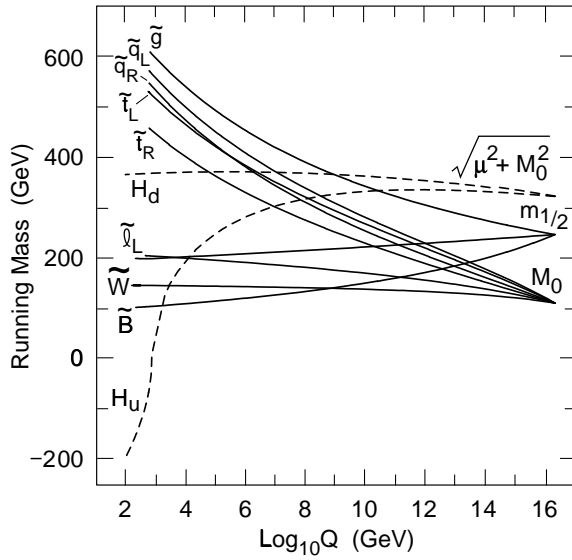


Fig. 13: Renormalization-group evolution of soft supersymmetry-breaking mass parameters [67].

where $m_{\tilde{q}}$ is a squark mass.

This makes it possible to generate electroweak symmetry breaking dynamically, even if $m_H^2 > 0$ at the input scale along with the other scalar mass-squared parameters [68], as seen in Fig. 13. The appropriate renormalization scale for discussing the effective Higgs potential of the MSSM is $Q \lesssim 1$ TeV, and the electroweak gauge symmetry will be broken if either or both of $m_{H_{1,2}}^2(Q) < 0$, as in the model potential (5). This is certainly possible for $m_t \sim 175$ GeV as observed.

3.2 Supersymmetric Higgs Bosons

As was discussed in Lecture 2, one expects two complex Higgs doublets $H_2 \equiv (H_2^+, H_2^0)$, $H_1 \equiv (H_1^+, H_1^0)$ in the MSSM, with a total of 8 real degrees of freedom. Of these, 3 are eaten via the Higgs mechanism to become the longitudinal polarization states of the W^\pm and Z^0 , leaving 5 physical Higgs bosons to be discovered by experiment. Three of these are neutral: the lighter CP-even neutral h , the heavier CP-even neutral H , the CP-odd neutral A , and charged bosons H^\pm . The quartic potential is completely determined by the D terms (59)

$$V_4 = \frac{g^2 + g'^2}{8} (|H_1^0|^2 - |H_2^0|^2)^2 \quad (68)$$

for the neutral components, whilst the quadratic terms may be parametrized at the tree level by

$$\frac{1}{2} = m_{H_1}^2 |H_1|^2 + m_{H_2}^2 |H_2|^2 + (m_3^2 H_1 H_2 + \text{herm.conj.}) \quad (69)$$

where $m_3^2 = B_\mu \mu$. One combination of the three parameters ($m_{H_1}^2, m_{H_2}^2, m_3^2$) is fixed by the Higgs vacuum expectation $v = \sqrt{v_1^2 + v_2^2} = 246$ GeV, and the other two combinations may be rephrased as $(m_A, \tan \beta)$. These characterize all Higgs masses and couplings in the MSSM at the tree level. Looking back at (18), we see that the gauge coupling strength of the quartic interactions (68) suggests a relatively low mass for at least the lightest MSSM Higgs boson h , and this is indeed the case, with $m_h \leq m_Z$ at the tree level:

$$m_h^2 = m_Z^2 \cos^2 2\beta \quad (70)$$

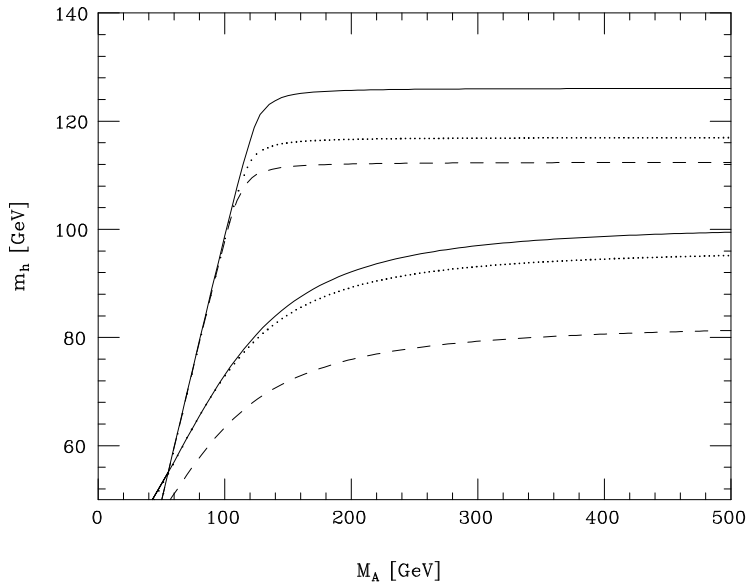


Fig. 14: The lightest Higgs boson mass in the MSSM, for different values of $\tan\beta$ and the CP-odd Higgs boson mass M_A [25].

This raised considerable hope that the lightest MSSM Higgs boson could be discovered at LEP, with its prospective reach to $m_H \sim 100$ GeV.

However, radiative corrections to the Higgs masses are calculable in a supersymmetric model (this was, in some sense, the whole point of introducing supersymmetry!), and they turn out to be non-negligible for $m_t \sim 175$ GeV [59]. Indeed, the leading one-loop corrections to m_h^2 depend quartically on m_t :

$$\Delta m_h^2 = \frac{3m_t^4}{4\pi^2 v^2} \ln \left(\frac{m_{\tilde{t}_1} m_{\tilde{t}_2}}{m_t^2} \right) + \frac{3m_t^4 \hat{A}_t^2}{8\pi^2 v^2} \left[2h(m_{\tilde{t}_1}^2, m_{\tilde{t}_2}^2) + \hat{A}_t^2 f(m_{\tilde{t}_1}^2, m_{\tilde{t}_2}^2) \right] + \dots \quad (71)$$

where $m_{\tilde{t}_{1,2}}$ are the physical masses of the two stop squarks $\tilde{t}_{1,2}$ to be discussed in more detail shortly, $\hat{A}_t \equiv A_t - \mu \cot\beta$, and

$$h(a, b) \equiv \frac{1}{a-b} \ln \left(\frac{a}{b} \right), \quad f(a, b) = \frac{1}{(a-b)^2} \left[2 - \frac{a+b}{a-b} \ln \left(\frac{a}{b} \right) \right] \quad (72)$$

Non-leading one-loop corrections to the MSSM Higgs masses are also known, as are corrections to coupling vertices, two-loop corrections and renormalization-group resummations [69]. For $m_{\tilde{t}_{1,2}} \lesssim 1$ TeV and a plausible range of A_t , one finds

$$m_h \lesssim 130 \text{ GeV} \quad (73)$$

as seen in Fig. 14. There we see the sensitivity of m_h to $(m_A, \tan\beta)$, and we also see how m_A, m_H and m_{H^\pm} approach each other for large m_A .

The radiative corrections (71), (72) have major implications for experiments and accelerators. They may push the MSSM Higgs sector beyond the reach of LEP 2 and into the lap of the LHC [70]. They motivate the optimization of LHC detectors for the Higgs mass range (73). They may motivate the orientation of future e^+e^- linear-collider construction so as to study such an MSSM Higgs boson in more detail than is possible at the LHC [71].

The decay modes of the MSSM Higgs bosons have been carefully studied, as seen in Fig. 15 [71]. Like the single Higgs boson of the Standard Model, the lightest MSSM Higgs boson h

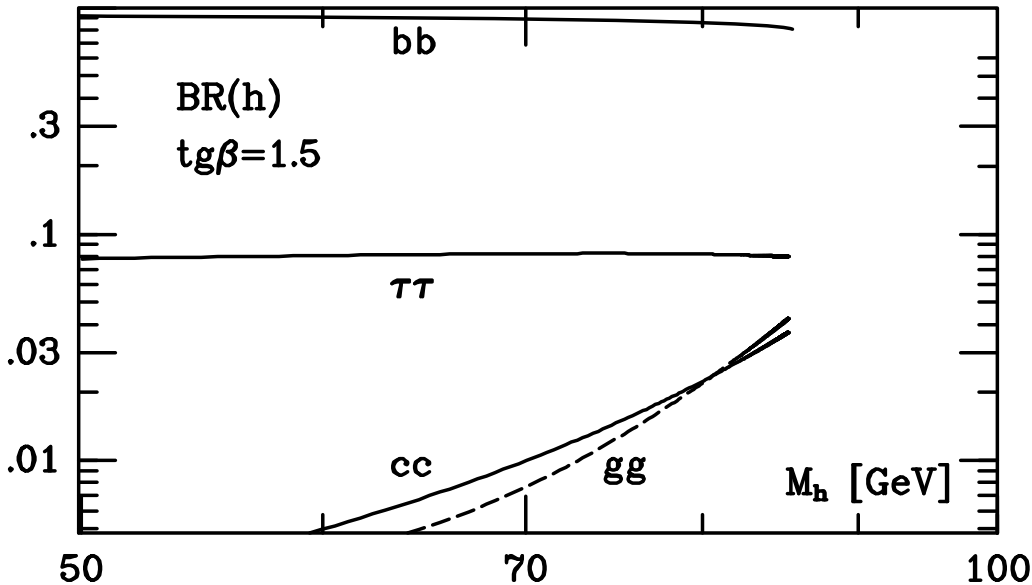


Fig. 15: The expected decay modes of the lightest MSSM Higgs boson h [71].

prefers to decay into the heaviest particles available, typically $h \rightarrow \bar{b}b$, and this has been the primary focus of searches at LEP 2. However, there are “blind spots” in MSSM parameter space where this decay mode is suppressed by cancellations, complicating the search at LEP 2. Ignoring this possible complication, Fig. 16 shows the regions of MSSM parameter that may be explored at LEP 2 for different centre-of-mass energies and luminosities. After LEP 2, the Fermilab Tevatron collider has a chance of observing $h \rightarrow \bar{b}b$ and possibly other decays [72], if it accumulates sufficient luminosity. The potential for LHC searches for MSSM Higgs bosons is shown in Fig. 17 for one choice of the MSSM parameters [70]. We see that the entire parameter space is covered at maximum luminosity, though with considerable reliance on the rare decay mode $h \rightarrow \gamma\gamma$.

3.3 Sparticle Masses and Mixing

We now progress to a more complete discussion of sparticle masses and mixing.

Sfermions : Each flavour of charged lepton or quark has both left- and right-handed components $f_{L,R}$, and these have separate spin-0 boson superpartners $\tilde{f}_{L,R}$. These have different isospins $I = \frac{1}{2}, 0$, but may mix as soon as the electroweak gauge symmetry is broken. Thus, for each flavour we should consider a 2×2 mixing matrix for the $\tilde{f}_{L,R}$, which takes the following general form [73]:

$$M_{\tilde{f}}^2 \equiv \begin{pmatrix} m_{\tilde{f}_{LL}}^2 & m_{\tilde{f}_{LR}}^2 \\ m_{\tilde{f}_{LR}}^2 & m_{\tilde{f}_{RR}}^2 \end{pmatrix} \quad (74)$$

The diagonal terms may be written in the form

$$m_{\tilde{f}_{LL,RR}}^2 = m_{\tilde{f}_{L,R}}^2 + m_{\tilde{f}_{L,R}}^{D^2} + m_f^2 \quad (75)$$

where m_f is the mass of the corresponding fermion, $\tilde{m}_{\tilde{f}_{L,R}}^2$ is the soft supersymmetry-breaking mass discussed in the previous section, and $m_{\tilde{f}_{L,R}}^{D^2}$ is a contribution due to the quartic D terms in the effective potential:

$$m_{\tilde{f}_{L,R}}^{D^2} = m_Z^2 \cos 2\beta (I_3 + \sin^2 \theta_W Q_{em}) \quad (76)$$

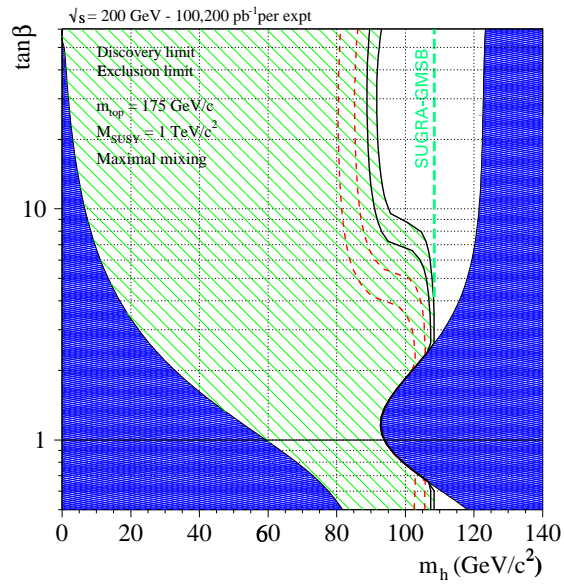


Fig. 16: Prospective discovery and exclusion limits for MSSM Higgs searches at LEP.

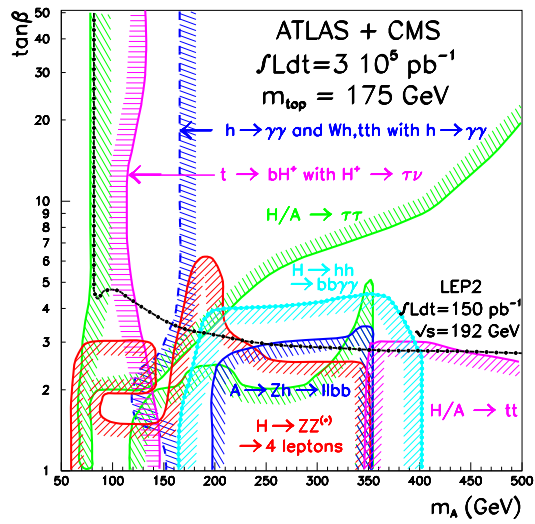


Fig. 17: Prospective coverage of the MSSM Higgs sector at the LHC [70].

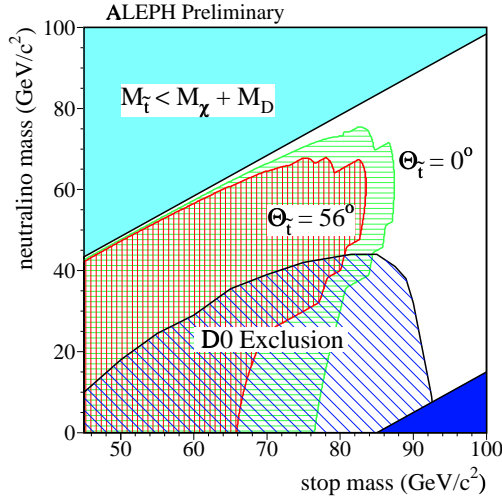


Fig. 18: Excluded domains in the search for light stops and neutralinos, as a function of the stop mixing angle $\theta_{\tilde{t}}$ [16].

where the term $\propto I_3$ is non-zero only for the \tilde{f}_L . Finally, the off-diagonal mixing term takes the general form

$$m_{\tilde{f}_{L,R}}^2 = m_f \left(A_f + \mu \frac{\tan \beta}{\cot \beta} \right) \quad \text{for } f = \begin{matrix} e, \mu, \tau, d, s, b \\ u, c, t \end{matrix} \quad (77)$$

It is clear that $\tilde{f}_{L,R}$ mixing is likely to be important for the \tilde{t} , and it may also be important for the $\tilde{b}_{L,R}$ and $\tilde{\tau}_{L,R}$ if $\tan \beta$ is large. We also see from (75) that the diagonal entries for the $\tilde{t}_{L,R}$ would be different from those of the $\tilde{u}_{L,R}$ and $\tilde{c}_{L,R}$, even if their soft supersymmetry-breaking masses were universal, because of the m_f^2 contribution. In fact, we also expect non-universal renormalization of $m_{\tilde{t}_{LL,RR}}^2$ (and also $m_{\tilde{b}_{LL,RR}}^2$ and $m_{\tilde{\tau}_{LL,RR}}^2$ if $\tan \beta$ is large), because of Yukawa effects analogous to those discussed in the previous section for the renormalization of the soft Higgs masses.

For these reasons, the $\tilde{t}_{L,R}$ are not usually assumed to be degenerate with the other squark flavours. Indeed, one of the \tilde{t} could well be the lightest squark, perhaps even lighter than the t quark itself [73]. The mass limits [16] combined in Fig. 18 assume degenerate $(\tilde{u}, \tilde{d}, \tilde{s}, \tilde{c}, \tilde{b})_{L,R}$, even though this degeneracy should also be broken by the flavour-universal D terms (76) and by renormalization effects that are different for $\tilde{f}_{L,R}$. The search for the stop mass eigenstates $\tilde{t}_{1,2}$ requires a separate analysis. Figure 19 shows the experimental lower limits on $m_{\tilde{t}_1}$ from ALEPH and $D\phi$ for different assumed values of the \tilde{t} mixing angle $\theta_{\tilde{t}}$ [16], and assuming that $\tilde{t} \rightarrow c\chi$ decay dominates, where χ is the lightest neutralino.

Charginos: These are the supersymmetric partners of the W^\pm and H^\pm , which mix through a 2×2 matrix

$$-\frac{1}{2} (\tilde{W}^-, \tilde{H}^-) M_C \begin{pmatrix} \tilde{W}^+ \\ \tilde{H}^+ \end{pmatrix} + \text{herm.conj.} \quad (78)$$

where

$$M_C \equiv \begin{pmatrix} M_2 & \sqrt{2}m_W \sin \beta \\ \sqrt{2}m_W \cos \beta & \mu \end{pmatrix} \quad (79)$$

Here M_2 is the unmixed $SU(2)$ gaugino mass and μ is the Higgs mixing parameter introduced in (61). Figure 20 displays (among other lines to be discussed later) the contour $m_{\chi^\pm} = 91$ GeV

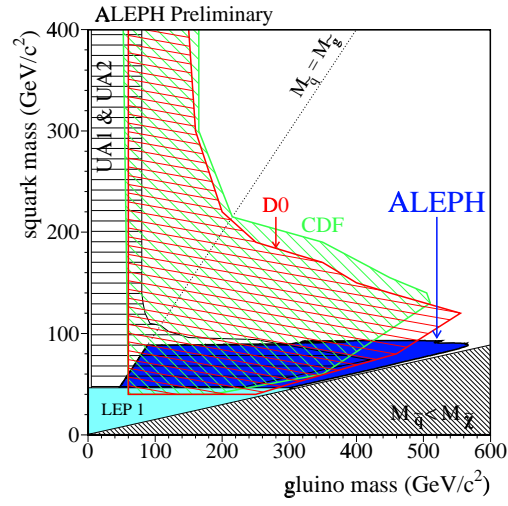
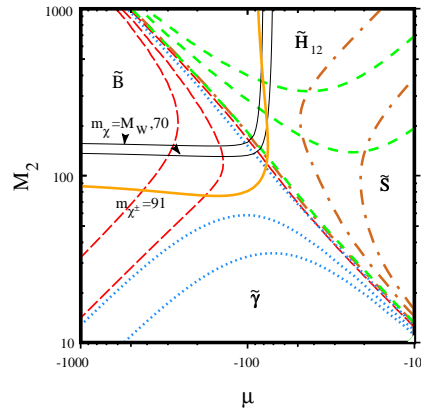
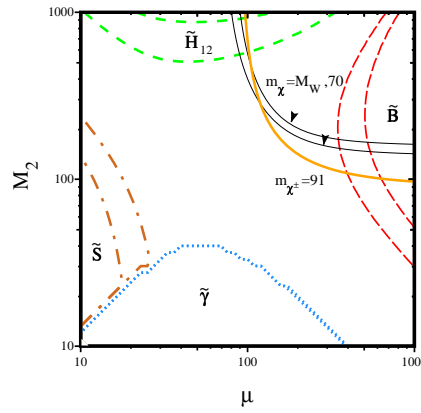


Fig. 19: Excluded regions in gluino and squark masses [16].



(a)



(b)

Fig. 20: The (μ, M_2) plane characterizing charginos and neutralinos, for (a) $\mu < 0$ and (b) $\mu > 0$, including contours of m_χ and m_{χ^\pm} , and of neutralino purity [74].

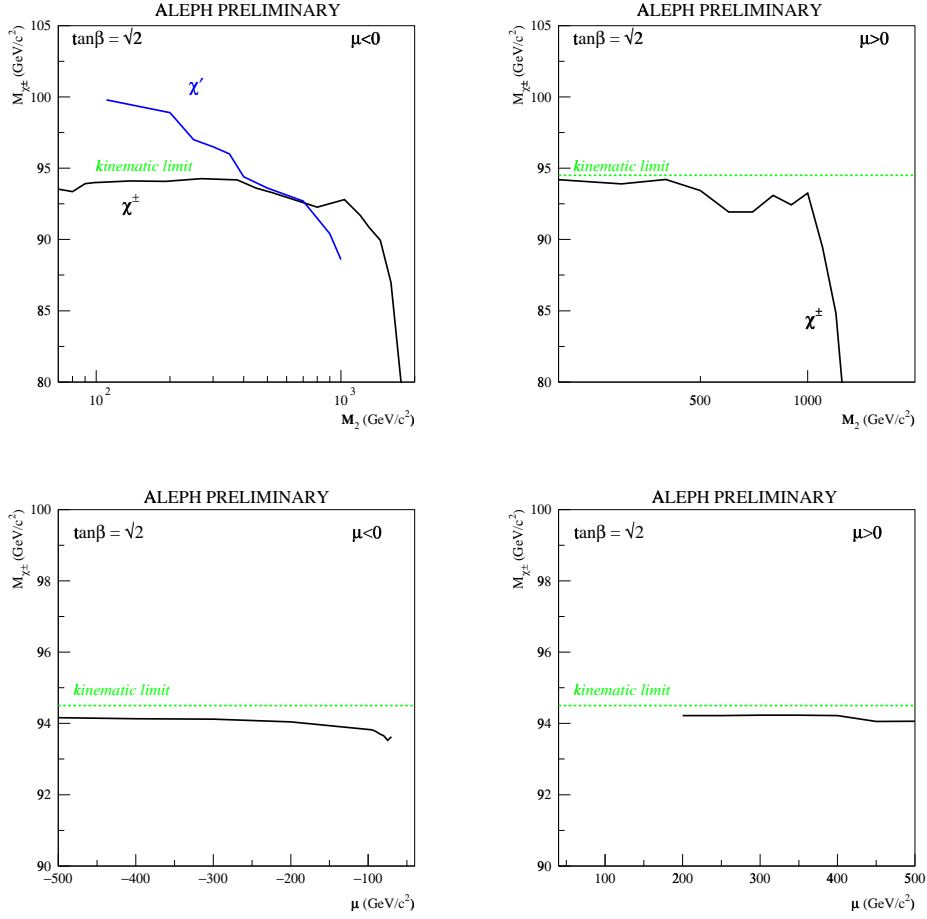


Fig. 21: Lower limits on the chargino mass inferred in the higgsino region (top panels) and the gaugino region (bottom panels) [16].

for the lighter of the two chargino mass eigenstates [74]. Some recent experimental lower limits on m_{χ^\pm} as functions of the other MSSM parameters are shown in Fig. 21 [16].

Neutralinos: These are characterized by a 4×4 mass mixing matrix [75], which takes the following form in the $(\tilde{W}^3, \tilde{B}, \tilde{H}_2^0, \tilde{H}_1^0)$ basis :

$$m_N = \begin{pmatrix} M_2 & 0 & \frac{-g_2 v_2}{\sqrt{2}} & \frac{g_2 v_1}{\sqrt{2}} \\ 0 & M_1 & \frac{g' v_2}{\sqrt{2}} & \frac{-g' v_1}{\sqrt{2}} \\ \frac{-g_2 v_2}{\sqrt{2}} & \frac{g' v_2}{\sqrt{2}} & 0 & \mu \\ \frac{g_2 v_1}{\sqrt{2}} & \frac{-g' v_1}{\sqrt{2}} & \mu & 0 \end{pmatrix} \quad (80)$$

Note that this has a structure similar to M_C (79), but with its entries replaced by 2×2 submatrices. As has already been mentioned, one conventionally assumes that the $SU(2)$ and $U(1)$ gaugino masses $M_{1,2}$ are universal at the GUT or supergravity scale, so that

$$M_1 \simeq M_2 \frac{\alpha_1}{\alpha_2} \quad (81)$$

so the relevant parameters of (80) are generally taken to be $M_2 = (\alpha_2/\alpha_{GUT})m_{1/2}$, μ and $\tan \beta$.

Figure 20 also displays contours of the mass of the lightest neutralino χ , as well as contours of its gaugino and Higgsino contents [74]. In the limit $M_2 \rightarrow 0$, χ would be approximately a photino and it would be approximately a Higgsino in the limit $\mu \rightarrow 0$. Unfortunately, these idealized limits are excluded by unsuccessful LEP and other searches for neutralinos and charginos, as we now discuss in more detail.

3.4 The Lightest Supersymmetric Particle

This is expected to be stable in the MSSM, and hence should be present in the Universe today as a cosmological relic from the Big Bang [76, 75]. Its stability arises because there is a multiplicatively-conserved quantum number called R parity, that takes the values $+1$ for all conventional particles and -1 for all sparticles [53]. The conservation of R parity can be related to that of baryon number B and lepton number L , since

$$R = (-1)^{3B+L+2S} \quad (82)$$

where S is the spin. Note that R parity could be violated either spontaneously if $\langle 0|\tilde{\nu}|0\rangle \neq 0$ or explicitly if one of the supplementary couplings (63) is present. There could also be a coupling HL , but this can be defined away by choosing a field basis such that \tilde{H} is defined as the superfield with a bilinear coupling to H . Note that R parity is not violated by the simplest models for neutrino masses, which have $\Delta L = 0, \pm 2$, nor by the simple GUTs discussed in the next Lecture, which violate combinations of B and L that leave R invariant. There are three important consequences of R conservation:

1. sparticles are always produced in pairs, e.g., $\bar{p}p \rightarrow \tilde{q}\tilde{g}X$, $e^+e^- \rightarrow \tilde{\mu} + \tilde{\mu}^-$,
2. heavier sparticles decay to lighter ones, e.g., $\tilde{q} \rightarrow q\tilde{g}$, $\tilde{\mu} \rightarrow \mu\tilde{\gamma}$, and
3. the lightest sparticles is stable,

because it has no legal decay mode.

This feature constrains strongly the possible nature of the lightest supersymmetric sparticle. If it had either electric charge or strong interactions, it would surely have dissipated its energy and condensed into galactic disks along with conventional matter. There it would surely

have bound electromagnetically or via the strong interactions to conventional nuclei, forming anomalous heavy isotopes that should have been detected. There are upper limits on the possible abundances of such bound relics, as compared to conventional nucleons [77]:

$$\frac{n(\text{relic})}{n(p)} \lesssim 10^{-15} \quad \text{to} \quad 10^{-29} \quad (83)$$

for $1 \text{ GeV} \lesssim m_{\text{relic}} \lesssim 1 \text{ TeV}$. These are far below the calculated abundances of such stable relics:

$$\frac{n(\text{relic})}{n(p)} \gtrsim 10^{-6} \quad (10^{-10}) \quad (84)$$

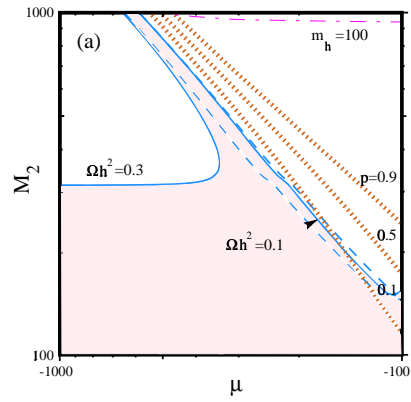
for relic particles with electromagnetic (strong) interactions. We may conclude [75] that any supersymmetric relic is probably electromagnetically neutral with only weak interactions, and could in particular not be a gluino. Whether the lightest hadron containing a gluino is charged or neutral, it would surely bind to some nuclei. Even if one pleads for some level of fractionation, it is difficult to see how such gluino nuclei could avoid the stringent bounds established for anomalous isotopes of many species [77].

Plausible scandidates of different spins are the sneutrinos $\tilde{\nu}$ of spin 0, the lightest neutralino χ of spin 1/2, and the gravitino \tilde{G} of spin 3/2. The sneutrinos have been ruled out by the combination of LEP experiments and direct searches for cosmological relics. Neutrino counting (17) requires $m_{\tilde{\nu}} \gtrsim 43 \text{ GeV}$ [78], in which case the direct relic searches in underground low-background experiments require $m_{\tilde{\nu}} \gtrsim 1 \text{ TeV}$ [79]. The gravitino cannot be ruled out, and its popularity has revived somewhat with the renaissance of gauge-mediated (messenger) models [66], as described in Lecture 5. For the rest of this Lecture, however, we concentrate on the neutralino possibility.

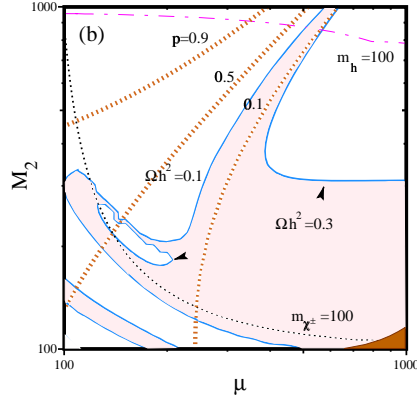
A very attractive feature of the neutralino candidature for the lightest supersymmetric particle is that it has a relic density of interest to astrophysicists and cosmologists: $\Omega_{\chi} h^2 = \mathcal{O}(0.1)$ over generic domains of the MSSM parameter space [75]. This feature is seen clearly in Fig. 22, where $0.1 < \Omega_{\chi} h^2 < 0.3$ is possible in a large area of the (μ, M_2) plane for suitable choices of the other MSSM parameters [74]. In this domain, the lightest neutralino χ could constitute the cold dark matter favoured by theories of cosmological structure formation [80].

We have already seen in Fig. 21 some of the experimental limits on chargino and neutralino production, that may be used to set interesting limits on m_{χ} . One example is shown in Fig. 23, where one particular choice of m_0 is assumed [16]. (This parameter is relevant because $\tilde{\nu}$ exchange contributes to $\sigma(e^+e^- \rightarrow \chi^+\chi^-)$, and the $\tilde{\nu}$ and \tilde{e} masses influence χ^{\pm} decay patterns [78].) It is interesting to note in Fig. 23 that LEP 1 data (e.g., neutrino counting in Z^0 decays (17)) did not by themselves provide an absolute lower limit on m_{χ} : this became possible only by combining them with higher-energy LEP data [78].

The lower limit on m_{χ} can be strengthened by combining the direct chargino/neutralino searches with other experimental and theoretical constraints [78, 74], as illustrated in Fig. 24. The dotted lines labelled LEP are the analogues of Fig. 23, but with m_0 allowed to float freely. The dotted lines marked H, C incorporate the experimental lower limit on m_h and the cosmological relic-density constraint $\Omega_{\chi} h^2 \leq 0.3$, respectively. The solid lines marked UHM further assume universal scalar masses for the Higgs multiplets. The lines marked *cosmo*, *DM* combine this assumption with the relic-density assumptions $\Omega_{\chi} h^2 < 0.3$, >0.1 , respectively. Figure 24 documents a lower limit $m_{\chi} > 40 \text{ GeV}$ [81], which can be strengthened using more recent LEP 2 data to about 45 GeV [74]. We expect that higher-energy runs of LEP will extend this sensitivity to $m_{\chi} \sim 50 \text{ GeV}$. We also see in Fig. 24 that this type of combined analysis of the MSSM parameter space imposes an absolute lower limit on $\tan \beta$. Data from LEP that have



(a)



(b)

Fig. 22: Contours of Ωh^2 , m_h , $m_{\chi\pm}$ and higgsino purity in the (μ, M_2) plane, for (a) $\mu < 0$ and (b) $\mu > 0$ [74].

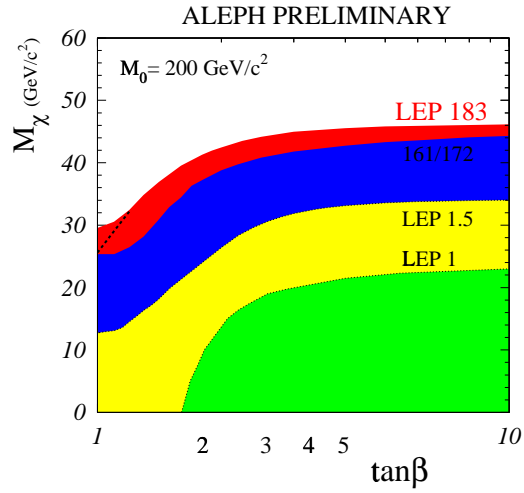


Fig. 23: Experimental lower limit on the lightest neutralino mass, as a function of $\tan\beta$, assuming $m_0 = 200$ GeV [16].

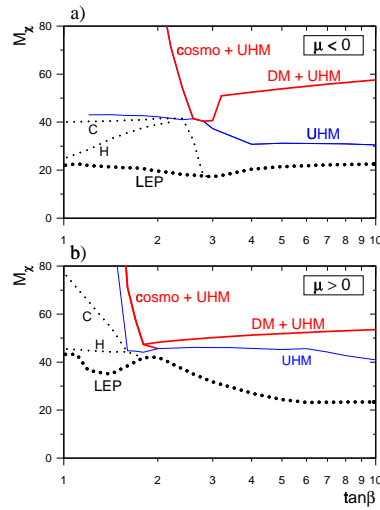


Fig. 24: Lower limits on m_χ from LEP (thick dots), Higgs searches (H), cosmological limits on Ωh^2 (c), which are strengthened by assuming universal scalar masses for Higgs bosons (UHM), also in combination with the constraints $\Omega h^2 < 0.3$ (cosmo + UHM) and $\Omega h^2 > 0.1$ (DM + UHM) [78].

been published so far indicate that $\tan \beta \gtrsim 1.8$ [74], and future LEP runs will be sensitive up to $\tan \beta \sim 3$, principally via Higgs searches.

We close with a few comments on the prospects for sparticle searches at the LHC. These should be able to extend the squark and gluino searches up to masses on 2.5 GeV, as seen in Fig. 25 [82]. By looking in several different channels: missing energy E_T^{miss} with 0, 1, 2, etc. leptons, it should be possible to explore several times over the domain of parameter space of interest to cosmologists where $\Omega_\chi h^2 \lesssim 0.4$, as also seen in Fig. 25⁶. Moreover, it should be possible to reconstruct several different sparticles via the cascade decays of squarks and gluinos, and even make detailed mass measurements that could test supergravity mass relations [84].

3.5 The “Anomaly” that Went Away...

For some time, measurements of $R_b = \Gamma(Z^0 \rightarrow \bar{b}b)/\Gamma(Z \rightarrow \text{all hadrons})$ seemed to be in significant disagreement with the Standard Model, generating considerable interest. It was suggested that the discrepancy might be explicable by one-loop supersymmetric radiative corrections, due either to Higgs exchange if m_A were small and $\tan \beta$ large, or to chargino and stop exchange if both m_{χ^\pm} and $m_{\tilde{t}}$ were small, as well as $\tan \beta$ [85]. The Higgs former scenario was early effectively excluded by early Higgs searches at LEP, but the χ^\pm/\tilde{t} scenario fitted well with theoretical prejudices and survived somewhat longer. It was particularly interesting, because it suggested that either a chargino or a stop might be light enough to be produced at LEP 2 or at the Fermilab Tevatron collider.

As time has progressed, the R_b anomaly has steadily decreased in significance, and is now barely a one- σ discrepancy, as seen in Fig. 26 [5]. In parallel, both LEP 2 and the Tevatron have explored considerable domains of MSSM parameter space, excluding significant domains of m_{χ^\pm} and $m_{\tilde{t}}$. Might there still be a significant supersymmetric contribution to R_b , comparable to the experimental error $\Delta R_b \sim 0.0010$? Even before the latest exclusion domains from LEP *et*

⁶This statement may require some re-examination in the light of co-annihilation effects on the relic χ density [83].

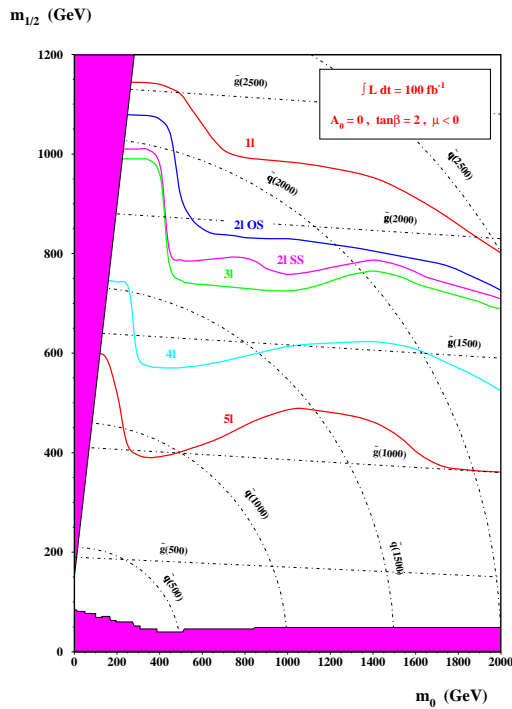


Fig. 25: Regions of the $(m_0, m_{1/2})$ plane that can be explored at the LHC [82].

al., we [86] found that, of about 500,000 possible choices of the basic MSSM model parameters, only 210 of those that respected the experimental constraints in early 1997 (including $b \rightarrow s\gamma$ and the cosmological relic density) could yield $\Delta R_b \geq 0.0010$. This already made the case for a significant supersymmetric contribution to R_b appear somewhat implausible. (Though everybody would have been happy if one of these “unusual” models was close to reality!) Another possible strike against these models was that they required a departure from the universality assumptions favoured in supergravity models, as seen in Fig. 27, where the “globular cluster” of “interesting” models with $\Delta R_b > 0.0010$ is outside the zone of parameter space accessible in such universal supergravity models, which can only yield [86]

$$\Delta R_b < 0.0003 \quad (85)$$

Moreover, the “interesting” models all had $m_{\tilde{t}_1} < 100$ GeV, and 90% of them have now been excluded by the more sensitive LEP 2 searches shown in Fig. 19. The conclusion must be that plausible parameter choices for the MSSM do not yield a significant contribution to R_b , and hence that it is legitimate to use the measurements in a global fit to the precision electroweak data in the Standard Model, as was assumed in Lecture 1.

4 GRAND UNIFICATION

4.1 Basic Strategy

The philosophy of grand unification [8, 87] is to seek a simple gauge group that includes the untidy $SU(3)$, $SU(2)$ and $U(1)$ gauge groups of QCD and the electroweak sector of the Standard Model. The hope is that this grand unification can be achieved while neglecting gravity, at least as a first approximation. If the grand unification scale turns out to be significantly less than the Planck mass, this is not obviously a false hope. We discuss later in this Lecture and the next

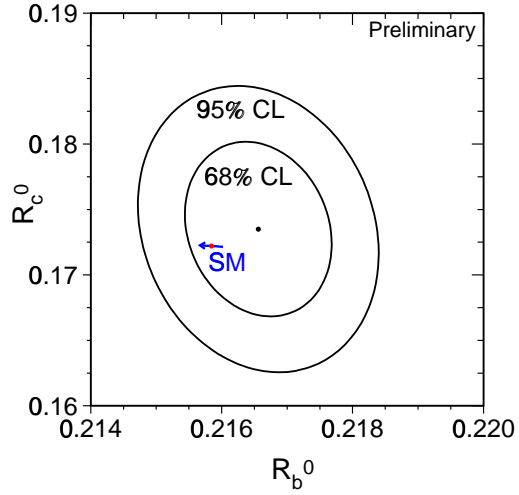


Fig. 26: Experimental constraints on $Z \rightarrow \bar{b}b$ and $\bar{c}c$ decays compared with the Standard Model prediction [5].

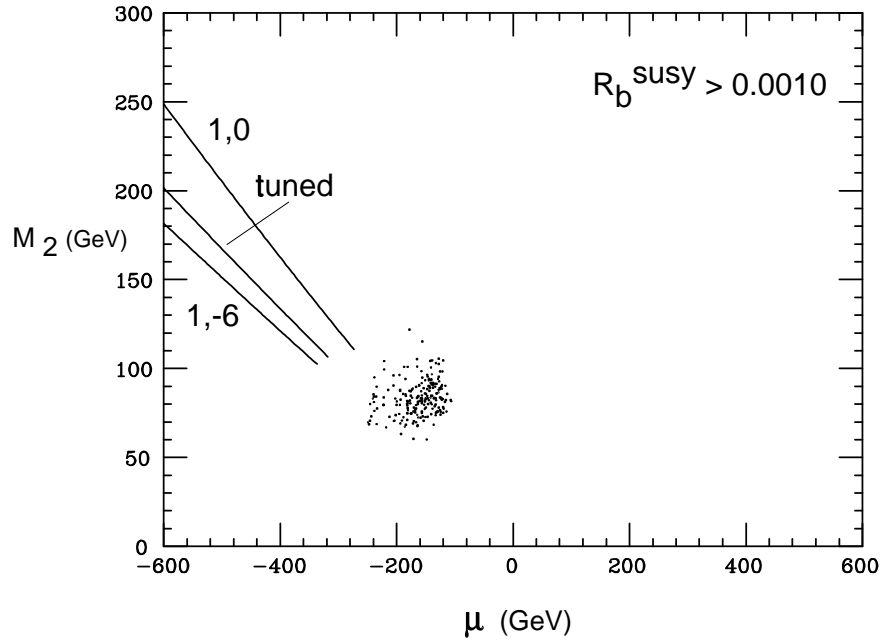


Fig. 27: Region of the (μ, M_2) plane where MSSM contributions to R_b larger than 0.0010 are possible which cannot be attained in models with universal scalar masses (solid lines) [86].

the extent to which this hope is indeed realistic: for the moment we just note that the grand unification scale is indeed expected to be exponentially large:

$$\frac{m_{GUT}}{m_W} = \exp\left(\mathcal{O}\left(\frac{1}{\alpha_{em}}\right)\right) \quad (86)$$

and typical estimates will be that $m_{GUT} = \mathcal{O}(10^{16} \text{ GeV})$. Such a calculation involves an extrapolation of known physics by many orders of magnitude further than, e.g., the extrapolation that Newton made from the apple to the Solar System. However, it is not excluded by our current knowledge of the Standard Model. For example, we see in Fig. 6 that the estimate (33) of the Higgs mass is consistent with the Standard Model remaining valid all the way to the Planck scale $m_P \simeq 10^{19} \text{ GeV}$, and even far beyond.

If the grand unification scale is indeed so large, most tests of it are likely to be indirect, and we meet some later, such as relations between Standard Model gauge couplings and between particle masses. Any new interactions, such as those that might cause protons to decay or give masses to neutrinos, are likely to be very strongly suppressed.

The first apparent obstacle to the philosophy of grand unification is the fact that the strong coupling $\alpha_3 = g_3^2/4\pi$ is indeed much stronger than the electroweak couplings at present-day energies: $\alpha_3 \gg \alpha_2, \alpha_1$. However, you have seen here in the lectures by Michelangelo Mangano [88] that the strong coupling is asymptotically free:

$$\alpha_3(Q) \simeq \frac{12\pi}{(33 - 2N_q) \ln(Q^2/\Lambda_3^2)} + \dots \quad (87)$$

where N_q is the number of quarks, $\Lambda_3 \simeq \text{few hundred MeV}$ is an intrinsic scale of the strong interactions, and the dots in (87) represent higher-loop corrections to the leading one-loop behaviour shown. The other Standard Model gauge couplings also exhibit logarithmic violations analogous to (87). For example, the effective value of $\alpha_{em}(m_Z) \sim 1/128$, with estimated ranges displayed in Fig. 5. The renormalization-group evolution for the $SU(2)$ gauge coupling is

$$\alpha_2(Q) \simeq \frac{12\pi}{(22 - 2N_q - N_{H/2}) \ln(Q^2/\Lambda_2^2)} + \dots \quad (88)$$

where we have assumed equal numbers of quarks and leptons, and N_H is the number of Higgs doublets. Taking the inverses of (87) and (88), and then taking their difference, we find

$$\frac{1}{\alpha_3(Q)} - \frac{1}{\alpha_2(Q)} = \left(\frac{11 + N_{H/2}}{12\pi}\right) \ln\left(\frac{Q^2}{m_X^2}\right) + \dots \quad (89)$$

We have absorbed the scales Λ_3 and Λ_2 into a single grand unification scale M_X where $\alpha_3 = \alpha_2$.

Evaluating (89) when $Q = \mathcal{O}(M_W)$, where $\alpha_3 \gg \alpha_2 = \mathcal{O}(\alpha_{em})$, we derive the characteristic feature (86) that the grand unification scale is exponentially large. As we see in more detail later, in most GUTs there are new interactions mediated by bosons weighing $\mathcal{O}(m_X)$ that cause protons to decay with a lifetime αm_X^4 . In order for the proton lifetime to exceed the experimental limit, we need $m_X \gtrsim 10^{14} \text{ GeV}$ and hence $\alpha_{em} \lesssim 1/120$ in (86) [89]. On the other hand, if the neglect of gravity is to be consistent, we need $m_X \lesssim 10^{19} \text{ GeV}$ and hence $\alpha_{em} \gtrsim 1/170$ in (86) [89]. The fact that the measured value of the fine-structure constant $\alpha_{em} \simeq 1/137.03599959(38)13$ lies in this allowed range may be another hint favouring the GUT philosophy.

Further empirical evidence for grand unification is provided by the previously-advertized prediction it makes for the neutral electroweak mixing angle [7]. Calculating the renormalization of the electroweak couplings, one finds:

$$\sin^2 \theta_W = \frac{\alpha_{em}(m_W)}{\alpha_2(m_W)} \simeq \frac{3}{8} \left[1 - \frac{\alpha_{em}}{4\pi} \frac{110}{9} \ln \frac{m_X^2}{m_W^2} \right] \quad (90)$$

which can be evaluated to yield $\sin^2 \theta_W \sim 0.210$ to 0.220 , if there are only Standard Model particles with masses $\lesssim m_X$ [7]. This is to be compared with the experimental value $\sin^2 \theta_W = 0.23155 \pm 0.00019$ shown in Fig. 2. Considering that $\sin^2 \theta_W$ could *a priori* have had any value between 0 and 1, this is an impressive qualitative success. The small discrepancy can be removed by adding some extra particles, such as the supersymmetric particles in the MSSM.

Another qualitative success is the prediction of the b quark mass [90, 91]. In many GUTs, such as the minimal $SU(5)$ model discussed shortly, the b quark and the τ lepton have equal Yukawa couplings when renormalized at the GUT scale. The renormalization group then tells us that

$$\frac{m_b}{m_\tau} \simeq \left[\ln \left(\frac{m_b^2}{m_X^2} \right) \right]^{\frac{12}{33-2N_q}} \quad (91)$$

Using $m_\tau = 1.78$ GeV, we predict that $m_b \simeq 5$ GeV, in agreement with experiment⁷. Happily, this prediction remains successful if the effects of supersymmetric particles are included in the renormalization-group calculations [92].

To examine the GUT predictions for $\sin^2 \theta_W$, etc. in more detail, one needs to study the renormalization-group equations beyond the leading one-loop order. Through two loops, one finds that

$$Q \frac{\partial \alpha_i(Q)}{\partial Q} = -\frac{1}{2\pi} \left(b_i + \frac{b_{ij}}{4\pi} \alpha_j(Q) \right) [\alpha_i(Q)]^2 \quad (92)$$

where the b_i receive the one-loop contributions

$$b_i = \begin{pmatrix} 0 \\ -\frac{22}{3} \\ -11 \end{pmatrix} + N_g \begin{pmatrix} \frac{4}{3} \\ \frac{4}{3} \\ \frac{4}{3} \end{pmatrix} + N_H \begin{pmatrix} \frac{1}{10} \\ \frac{1}{6} \\ 0 \end{pmatrix} \quad (93)$$

from gauge bosons, N_g matter generations and N_H Higgs doublets, respectively, and at two loops

$$b_{ij} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{136}{3} & 0 \\ 0 & 0 & -102 \end{pmatrix} + N_g \begin{pmatrix} \frac{19}{15} & \frac{3}{5} & \frac{44}{15} \\ \frac{1}{5} & \frac{49}{3} & 4 \\ \frac{4}{30} & \frac{3}{2} & \frac{76}{3} \end{pmatrix} + N_H \begin{pmatrix} \frac{9}{50} & \frac{9}{10} & 0 \\ \frac{3}{10} & \frac{13}{6} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (94)$$

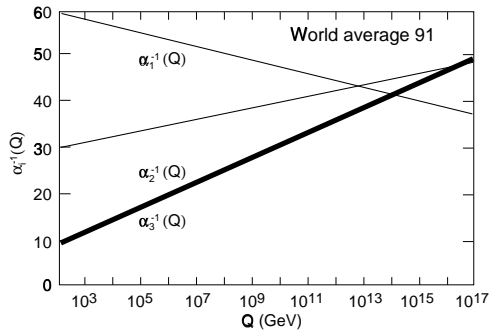
These coefficients are all independent of any specific GUT model, depending only on the light particles contributing to the renormalization. Including supersymmetric particles as in the MSSM, one finds [9]

$$b_i = \begin{pmatrix} 0 \\ -6 \\ -9 \end{pmatrix} + N_g \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + N_H \begin{pmatrix} \frac{3}{10} \\ \frac{1}{2} \\ 0 \end{pmatrix} \quad (95)$$

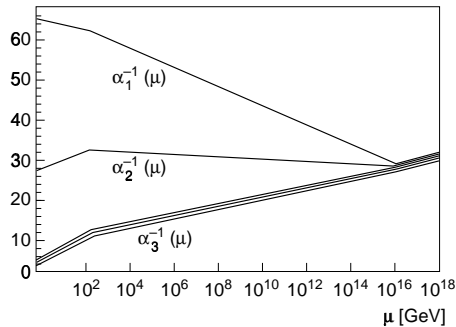
and

$$b_{ij} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -24 & 0 \\ 0 & 0 & -54 \end{pmatrix} + N_g \begin{pmatrix} \frac{38}{15} & \frac{6}{5} & \frac{88}{15} \\ \frac{2}{5} & 14 & 8 \\ \frac{11}{5} & 3 & \frac{68}{3} \end{pmatrix} + N_H \begin{pmatrix} \frac{9}{50} & \frac{9}{10} & 0 \\ \frac{3}{10} & \frac{7}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (96)$$

⁷This prediction was made [90] shortly before the b quark was discovered. When we received the proofs of this article, I gleefully wrote by hand in the margin our then prediction, which was already in the text, as 2 to 5. This was misread by the typesetter to become 2605: a spectacular disaster!



(a)



(b)

Fig. 28: The measurements of the gauge coupling strengths at LEP (a) do not evolve to a unified value if there is no supersymmetry but do (b) if supersymmetry is included [93].

again independent of any specific supersymmetric GUT.

One can use these two-loop equations to make detailed calculations of $\sin^2 \theta_W$ in different GUTs. These confirm that non-supersymmetric models are not consistent with the determinations of the gauge couplings from LEP and elsewhere [60]. Previously, we argued that these models predicted a wrong value for $\sin^2 \theta_W$, given the experimental value of α_3 . In Fig. 28a we see the converse, namely that extrapolating the experimental determinations of the α_i using the non-supersymmetric renormalization-group equations (93), (94) does not lead to a common value at any renormalization scale. In contrast, we see in Fig. 28b that extrapolation using the supersymmetric renormalization-group equations (95), (96) **does** lead to possible unification at $m_{GUT} \sim 10^{16}$ GeV [93].

Turning this success around, and assuming $\alpha_3 = \alpha_2 = \alpha_1$ at m_{GUT} with no threshold corrections at this scale, one may estimate that [94]

$$\begin{aligned} \sin^2 \theta_W(M_Z) \Big|_{\overline{\text{MS}}} &= 0.2029 + \frac{7\alpha_{em}}{15\alpha_3} + \frac{\alpha_{em}}{20\pi} \left[-3 \ln \left(\frac{m_t}{m_Z} \right) + \frac{28}{3} \ln \left(\frac{m_{\tilde{g}}}{m_Z} \right) \right. \\ &\quad \left. - \frac{32}{3} \ln \left(\frac{m_{\tilde{W}}}{m_Z} \right) - \ln \left(\frac{m_A}{m_Z} \right) - 4 \ln \left(\frac{\mu}{m_Z} \right) + \dots \right] \end{aligned} \quad (97)$$

Setting all the sparticle masses to 1 TeV reproduces approximately the value of $\sin^2 \theta_W$ observed experimentally. Can one invert this successful argument to estimate the supersymmetric particle mass scale? One can show [95] that the sparticle mass thresholds in (97) can be lumped into the parameter

$$T_{susy} \equiv |\mu| \left(\frac{m_{\tilde{W}}^2}{m_{\tilde{g}}^2} \right)^{14/19} \left(\frac{m_A^2}{\mu^2} \right)^{3/38} \left(\frac{m_{\tilde{W}}^2}{\mu^2} \right)^{2/19} \prod_{i=1}^3 \left(\frac{m_{\tilde{\ell}_{Li}}^3 m_{\tilde{q}_i}^7}{m_{\tilde{\ell}_{Ri}}^2 m_{\tilde{u}_i}^5 m_{\tilde{d}_i}^3} \right)^{1/19} \quad (98)$$

If one assumes sparticle mass universality at the GUT scale, then [95]

$$T_{susy} \simeq |\mu| \left(\frac{\alpha_2}{\alpha_3} \right)^{3/2} \simeq \frac{\mu}{7} \quad (99)$$

approximately. The measured value of $\sin^2 \theta_W$ is consistent with $T_{susy} \sim 100$ GeV to 1 TeV, roughly as expected from the hierarchy argument. However, the uncertainties are such that one cannot use this consistency to constrain T_{susy} very tightly [96]. In particular, even if one accepts the universality hypothesis, there could be important model-dependent threshold corrections around the GUT scale [94, 97]. We are at the limit of what one can say without studying specific models, so let us now do so.

4.2 GUT Models

Before embarking on their study, however, we first clarify some necessary technical points. As well as looking for a simple unifying group $G \supset SU(3) \times SU(2) \times U(1)$, we shall be looking for unifying representations R that contain both quarks and leptons. Since gauge interactions conserve helicity, any particles with the same helicity are fair game to appear in any GUT representation R , and it is convenient to work with states of just one helicity, say left-handed. The left-handed particle content of the Standard Model is as follows. In each generation, there is a quark doublet $(u, d)_L$ which transforms as $(3, 2)$ of $SU(3) \times SU(2)_L$. Instead of working with the right-handed singlets u_R, d_R that have $(3, 1)$ representations, it is convenient to work with their antiparticles, which are left-handed: the u_L^c and d_L^c transform as $(\bar{3}, 1)$ of $SU(3) \times SU(2)_L$. Similarly each generation contains a lepton doublet $(\nu, \ell^-)_L$ transforming as $(1, 2)$, and the right-handed charged lepton ℓ_R is replaced by its conjugate ℓ_L^c , which transforms as a $(1, 1)$ of $SU(3) \times SU(2)_L$. We should also keep track of the hypercharges $Y = Q - I_3$. One of the major puzzles of the Standard Model is why

$$\sum_{q, \ell} Q_i = 3Q_u + 3Q_d + Q_e = 0 \quad (100)$$

In the Standard Model, the hypercharge assignments are **a priori** independent of the $SU(3) \times SU(2)_L$ assignments, although constrained by the fact that quantum consistency requires the resulting triangle anomalies to cancel. In a simple GUT, the relation (100) is automatic: whenever Q is a generator of a simple gauge group, $\sum_R Q = 0$ for particles in any representation R (consider, e.g., the values of I_3 in any representation of $SU(2)$).

The basic rules of GUT model-building are that one must look for (a) a gauge group of rank 4 or more – to accommodate the Standard Model $SU(3) \times SU(2) \times U(1)$ gauge group – which (b) admits complex representations – to accommodate the known matter fermions. The rank of a gauge group is the number of generators that can be diagonalized simultaneously, i.e., the number of quantum numbers that it admits. For example, $SU(2)$ and $U(1)_{em}$ both have rank 1 corresponding to I_3 and Q_{em} , respectively, and $SU(3)$ has rank 2 corresponding to T_3 and Y . Complex representations are required to allow the violation of charge conjugation C , as required by the Standard Model, which has

$$(\nu, e)_L \in (1, 2) , \quad (u, d)_L \in (3, 2) , \quad e_L^c \in (1, 1) , \quad u_L^c , \quad d_L^c \in (\bar{3}, 1) \quad (101)$$

as discussed above.

The following is the mathematical catalogue [8] of rank-4 gauge groups which are either simple or the direct products of identical simple gauge groups:

$$Sp(8) , \quad SO(8) , \quad SO(9) , \quad F_4 , \quad SU(3) \times SU(3) , \quad SU(5) \quad (102)$$

Among these, only $SU(3) \times SU(3)$ and $SU(5)$ have complex representations. Moreover, if one tried to use $SU(3) \times SU(3)$, one would need to embed the electroweak gauge group in the second $SU(3)$ factor. This would be possible only if $\sum_q Q_q = 0 = \sum_\ell Q_\ell$, which is not the case for the known quarks and leptons. Therefore, attention has focussed on $SU(5)$ [8] as the only possible rank-4 GUT group.

The useful representations of $SU(5)$ are the complex vector $\underline{5}$ representation denoted by F_α , its conjugate $\bar{\underline{5}}$ denoted by \bar{F}^α , the complex two-index antisymmetric tensor $\underline{10}$ representation $T_{[\alpha\beta]}$, and the adjoint $\underline{24}$ representation A_β^α . The latter is used to accommodate the gauge bosons of $SU(5)$:

$$\begin{pmatrix} & & & \vdots & \bar{X} & \bar{Y} \\ & & & \vdots & \bar{X} & \bar{Y} \\ & g_{1,\dots,8} & & \vdots & \bar{X} & \bar{Y} \\ & & & \vdots & \bar{X} & \bar{Y} \\ \dots\dots\dots & & & & & \\ X & X & X & \vdots & & \\ & & & \vdots & W_{1,2,3} & \\ Y & Y & Y & \vdots & & \end{pmatrix} \quad (103)$$

where the $g_{1,\dots,8}$ are the gluons of $SU(3)$, the $W_{1,2,3}$ are the $SU(2)$ weak bosons, the $U(1)$ hypercharge boson is proportional to the traceless diagonal generator $(1, 1, 1, -3/2, -3/2)$, and the (X, Y) are $(3, 2)$ of new gauge bosons that we discuss in the next section.

The quarks and leptons of each generation are accommodated in $\bar{\underline{5}}$ and $\underline{10}$ representations of $SU(5)$:

$$\bar{F} = \begin{pmatrix} d_R^c \\ d_Y^c \\ d_B^c \\ \dots \\ -e^- \\ \nu_e \end{pmatrix}_L, \quad T = \begin{pmatrix} 0 & u_B^c & -u_Y^c & \vdots & -u_R & -d_R \\ -u_B^c & 0 & u_R^c & \vdots & -u_Y & -d_Y \\ u_Y^c & -u_R^c & 0 & \vdots & -u_B & -d_B \\ \dots\dots\dots & & & & & \\ u_R & u_Y & u_B & \vdots & 0 & -e^c \\ d_R & d_Y & d_B & \vdots & e^c & 0 \end{pmatrix}_L \quad (104)$$

The particle assignments are unique up to the effects of mixing between generations, which we do not discuss in detail here [98]. The uniqueness is because

$$\bar{\underline{5}} = (\bar{\underline{3}}, 1) + (1, 2), \quad \underline{10} = (3, 2) + (\bar{\underline{3}}, 1) + (1, 2) \quad (105)$$

in terms of $SU(3) \times SU(2)$ representations. Therefore, the $(\nu, e)_L$ doublet in (101) can only be assigned to the $\bar{\underline{5}}$, and since $\Sigma Q_{em} = 0$ in any GUT representation, the $(\bar{\underline{3}}, 1)$ in the $\bar{\underline{5}}$ must be assigned to the d^c in (101). The remaining $(u, d)_L \in (3, 2)$, $u^c \in (\bar{\underline{3}}, 1)$ and $e^c \in (1, 1)$ in (101) fit elegantly into the $\underline{10}$, as seen in (104) and (105)⁸.

The remaining steps in constructing an $SU(5)$ GUT are the choices of representations for Higgs bosons, first to break $SU(5) \rightarrow SU(3) \times SU(2) \times U(1)$ and subsequently to break the electroweak $SU(2) \times U(1)_Y \rightarrow U(1)_{em}$. The simplest choice for the first stage is an adjoint $\underline{24}$

⁸Different particle assignments are possible in the flipped $SU(5)$ model inspired and derived from string [99], because it contains an external $U(1)$ factor not included in the simple $SU(5)$ group.

of Higgs bosons Φ :

$$\langle 0|\Phi|0 \rangle = \begin{pmatrix} 1 & 0 & 0 & \vdots & 0 & 0 \\ 0 & 1 & 0 & \vdots & 0 & 0 \\ 0 & 0 & 1 & \vdots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \vdots & -\frac{3}{2} & 0 \\ 0 & 0 & 0 & \vdots & 0 & -\frac{3}{2} \end{pmatrix} \times \mathcal{O}(m_{GUT}) \quad (106)$$

It is easy to see that this v.e.v. preserves colour $SU(3)$ acting on the first three rows and columns, weak $SU(2)$ acting on the last two rows and columns, and the hypercharge $U(1)$ along the diagonal. The subsequent breaking of $SU(2) \times U(1)_Y \rightarrow U(1)_{em}$ is most economically accomplished by a $\underline{5}$ representation of Higgs bosons H :

$$\langle 0|\phi|0 \rangle = (0, 0, 0, 0, 1) \times 0(m_W) \quad (107)$$

It is clear that this has an $SU(4)$ symmetry which yields [90] the relation $m_b = m_\tau$ that leads, after renormalization (91), to a successful prediction for m_b in terms of m_τ . However, the same trick does not work for the first two generations, indicating a need for epicycles in this simplest GUT model [100].

Making the minimal $SU(5)$ GUT supersymmetric, as motivated by the naturalness of the gauge hierarchy, is not difficult [61]. One must replace the above GUT multiplets by supermultiplets: $\underline{5}\bar{F}$ and $\underline{10} T$ for the matter particles, $\underline{24} \Phi$ for the GUT Higgs fields that break $SU(5) \rightarrow SU(3) \times SU(2) \times U(1)$. The only complication is that one needs $\underline{5}$ and $\bar{\underline{5}}$ Higgs representations H and \bar{H} to break $SU(2) \times U(1)_Y \rightarrow U(1)_{em}$, just as two doublets were needed in the MSSM. The Higgs potential is specified by the appropriate choice of superpotential [61]:

$$W = \left(\mu + \frac{3\lambda}{2}M\right) + \lambda\bar{H}\Phi H + f(\Phi) \quad (108)$$

where $f(\Phi)$ is chosen so that $\partial f/\partial\Phi = 0$ when

$$\langle 0|\Phi|0 \rangle = M \begin{pmatrix} 1 & 0 & 0 & \vdots & 0 & 0 \\ 0 & 1 & 0 & \vdots & 0 & 0 \\ 0 & 0 & 1 & \vdots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \vdots & -\frac{3}{2} & 0 \\ 0 & 0 & 0 & \vdots & 0 & -\frac{3}{2} \end{pmatrix} \quad (109)$$

Inserting this into the second term of (108), one finds terms $\lambda M\bar{H}_3H_3$, $-3/2\lambda M\bar{H}_2H_2$ for the colour-triplet and weak-doublet components of \bar{H} and H , respectively. Combined with the bizarre coefficient of the first term, these lead to

$$W \ni \left(\mu + \frac{5\lambda}{2}M\right)\bar{H}_3H_3 + \mu\bar{H}_2H_2 \quad (110)$$

Thus we have heavy Higgs triplets (as needed for baryon stability, see the next section) and light Higgs doublets. This requires fine tuning the coefficient of the first term in W (108) to about 1 part in 10^{13} ! The advantage of supersymmetry is that its no-renormalization theorems [44] guarantee that this fine tuning is “natural”, in the sense that quantum corrections like those in Fig. 12c do not destroy it, unlike the situation without supersymmetry. On the other hand, supersymmetry alone does not explain the origin of the hierarchy.

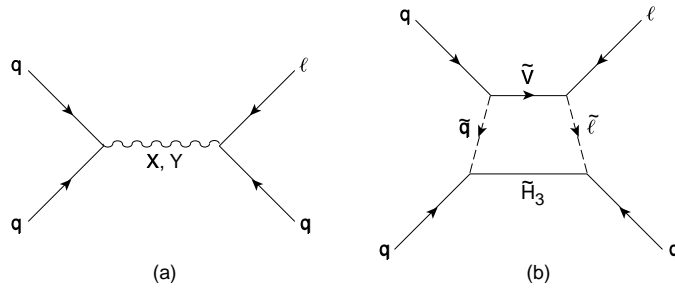


Fig. 29: Diagrams contributing to baryon decay (a) in minimal $SU(5)$ and (b) in minimal supersymmetric $SU(5)$.

4.3 Baryon Decay

Baryon instability is to be expected on general grounds, since there is no exact gauge symmetry to guarantee that baryon number B is conserved. Indeed, baryon decay is a generic prediction of GUTs, which we illustrate with the simplest $SU(5)$ model, that is anyway embedded in larger and more complicated GUTs. We see in (103) that there are two species of gauge bosons in $SU(5)$ that couple the colour $SU(3)$ indices (1,2,3) to the electroweak $SU(2)$ indices (4,5), called X and Y . As we can see from the matter representations (104), these may enable two quarks or a quark and lepton to annihilate, as seen in Fig. 29a. Combining these possibilities leads to interactions with $\Delta B = \Delta L = 1$. The forms of effective four-fermion interactions mediated by the exchanges of massive Z and Y bosons, respectively, are [91]:

$$\begin{aligned} & \left(\epsilon_{ijk} u_{Rk} \gamma^\mu u_{Lj} \right) \frac{g_X^2}{8m_X^2} \left(2e_R \gamma^\mu d_{Li} + e_L \gamma^\mu d_{Ri} \right) , \\ & \left(\epsilon_{ijk} u_{Rk} \gamma^\mu d_{Lj} \right) \frac{g_Y^2}{8m_X^2} \left(\nu_L \gamma^\mu d_{Ri} \right) . \end{aligned} \quad (111)$$

up to generation mixing factors.

Since the gauge couplings $g_X = g_Y = g_{3,2,1}$ in an $SU(5)$ GUT, and $m_X \simeq m_Y$, we expect that

$$G_X \equiv \frac{g_X^2}{8m_X^2} \simeq G_Y \equiv \frac{g_Y^2}{8m_Y^2} \quad (112)$$

It is clear from (111) that the baryon decay amplitude $A \propto G_X$, and hence the baryon $B \rightarrow \ell +$ meson decay rate

$$\Gamma_B = c G_X^2 m_p^5 \quad (113)$$

where the factor of m_p^5 comes from dimensional analysis, and c is a coefficient that depends on the GUT model and the non-perturbative properties of the baryon and meson.

The decay rate (113) corresponds to a proton lifetime

$$\tau_p = \frac{1}{c} \frac{m_X^4}{m_p^5} \quad (114)$$

It is clear from (114) that the proton lifetime is very sensitive to m_X , which must therefore be calculated very precisely. In minimal $SU(5)$, the best estimate was [101]

$$m_X \simeq (1 \text{ to } 2) \times 10^{15} \times \Lambda_{QCD} \quad (115)$$

where Λ_{QCD} is the characteristic QCD scale in the $\overline{\text{MS}}$ prescription with four active flavours. Making an analysis of the generation mixing factors [98], one finds that the preferred proton

(and bound neutron) decay modes in minimal $SU(5)$ are

$$\begin{aligned} p &\rightarrow e^+\pi^0, \quad e^+\omega, \quad \bar{\nu}\pi^+, \quad \mu^+K^0, \quad \dots \\ n &\rightarrow e^+\pi^-, \quad e^+\rho^-, \quad \bar{\nu}\pi^0, \quad \dots \end{aligned} \quad (116)$$

and the best numerical estimate of the lifetime is [101]

$$\tau(p \rightarrow e^+\pi^0) \simeq 2 \times 10^{31 \pm 1} \times \left(\frac{\Lambda_{QCD}}{400 \text{ MeV}} \right)^4 y \quad (117)$$

This is in *prima facie* conflict with the latest experimental lower limit

$$\tau(p \rightarrow e^+\pi^0) > 1.6 \times 10^{33} y \quad (118)$$

from super-Kamiokande [102]. However, this failure of minimal $SU(5)$ is not as conclusive as the failure of its prediction for $\sin^2 \theta_W$.

We saw earlier that supersymmetric GUTs, including $SU(5)$, fare better with $\sin^2 \theta_W$. They also predict a larger GUT scale [9]:

$$m_X \simeq 2 \times 10^{16} \text{ GeV} \quad (119)$$

so that $\tau(p \rightarrow e^+\pi^0)$ is considerably longer than the experimental lower limit. However, this is not the dominant proton decay mode in supersymmetric $SU(5)$ [103]. In this model, there are important $\Delta B = \Delta L = 1$ interactions mediated by the exchange of colour-triplet Higgsinos \tilde{H}_3 , dressed by gaugino exchange as seen in Fig. 29b [104]:

$$G_X \rightarrow \mathcal{O} \left(\frac{\lambda^2 g^2}{16\pi^2} \right) \frac{1}{m_{\tilde{H}_3} \tilde{m}} \quad (120)$$

where λ is a Yukawa coupling. Taking into account colour factors and the increase in λ for more massive particles, it was found [103] that decays into neutrinos and strange particles should dominate:

$$p \rightarrow \bar{\nu}K^+, \quad n \rightarrow \bar{\nu}K^0, \quad \dots \quad (121)$$

Because there is only one factor of a heavy mass $m_{\tilde{H}_3}$ in the denominator of (120), these decay modes are expected to dominate over $p \rightarrow e^+\pi^0$, etc., in minimal supersymmetric $SU(5)$. Calculating carefully the other factors in (120) [105], it seems that the modes (121) may be close to detectability in this model. The current experimental limit is $\tau(p \rightarrow \bar{\nu}K^+) > 10^{32} y$, and super-Kamiokande may soon be able to improve this significantly.

There are non-minimal supersymmetric GUT models such as flipped $SU(5)$ [99] in which the \tilde{H}_3 -exchange mechanism (120) is suppressed. In such models, $p \rightarrow e^+\pi^-$ may again be the preferred decay mode [106]. However, this is not necessarily the case, as colour-triplet Higgs boson exchange may be important, in which case $p \rightarrow \mu^+K^0$ could be dominant [107], or there may be non-intuitive generation mixing in the couplings of the X and Y bosons, offering the possibility $p \rightarrow \mu^+\pi^0$, etc. . Therefore, the continuing search for proton decay should be open-minded about the possible decay modes.

4.4 Neutrino Masses and Oscillations

The experimental upper limits on neutrino masses are far below the corresponding lepton masses [24]. From studies of the end-point of Tritium β decay, we have

$$m_{\nu_e} \lesssim 3.5 \text{ eV} \quad (122)$$

to be compared with $m_e = 0.511$ MeV. From studies of $\pi \rightarrow \mu\nu_\mu$ decays, we have

$$m_{\nu_\mu} < 160 \text{ keV} \quad (123)$$

to be compared with $m_\mu = 105$ MeV, and from studies [108] of $\tau \rightarrow$ pions + ν_τ we have

$$m_{\nu_\tau} < 18 \text{ MeV} \quad (124)$$

to be compared with $m_\tau = 1.78$ GeV. On the other hand, there is no good symmetry reason to expect the neutrino masses to vanish. We expect masses to vanish only if there is a corresponding exact gauge symmetry, cf., $m_\gamma = 0$ in QED with an unbroken $U(1)$ gauge symmetry.

Although there is no candidate gauge symmetry to ensure $m_\nu = 0$, this is a prediction of the Standard Model. We recall that the neutrino couplings to charged leptons take the form

$$J_\mu = \bar{e}\gamma_\mu(1 - \gamma_5)\nu_e + \bar{\mu}\gamma_\mu(1 - \gamma_5)\nu_\mu + \bar{\tau}\gamma_\mu(1 - \gamma_5)\nu_\tau \quad (125)$$

and that only left-handed neutrinos have ever been detected. In the cases of charged leptons and quarks, their masses arise in the Standard Model from couplings between left- and right-handed components via a Higgs field:

$$g_H \bar{f} f H_{\Delta I = \frac{1}{2}, \Delta L = 0} \bar{f}_R f_L + h.c. \rightarrow m_f = g_H \bar{f} f \langle 0 | H_{\Delta I = \frac{1}{2}, \Delta L = 0} | 0 \rangle \quad (126)$$

Such a left-right coupling is conventionally called a Dirac mass. The following questions arise for neutrinos: if there is no ν_R , can one have $m_\nu \neq 0$? and if there is a ν_R why are the neutrino masses so small?

The answer to the first question is positive, because it is possible to generate neutrino masses via the Majorana mechanism that involves the ν_L alone. This is possible because an (\bar{f}_R) field is in fact left-handed: $(\bar{f}_R) = (f^c)_L = f_L^T C$, where the superscript T denotes a transpose, and C is a 2×2 conjugation matrix. We can therefore imagine replacing

$$(\bar{f}_R) f_L \rightarrow f_L^T C f_L \quad (127)$$

which we denote by $f_L \cdot f_L$. In the cases of quarks and charged leptons, one cannot generate masses in this way, because $q_L \cdot q_L$ has $\Delta Q_{em}, \Delta(\text{colour}) \neq 0$ and $\ell_L \cdot \ell_L$ has $\Delta Q_{em} \neq 0$. However, the coupling $\nu_L \cdot \nu_L$ is not forbidden by such exact gauge symmetries from leading to a neutrino mass:

$$m^M \nu_L^T C \nu_L = m^M (\bar{\nu}^c)_L \nu_L = m^M \nu_L \cdot \nu_L \quad (128)$$

Such a combination has non-zero net lepton number $\Delta L = 2$ and weak isospin $\Delta I = 1$. There is no corresponding Higgs field in the Standard Model or in the minimal $SU(5)$ GUT, but there is no obvious reason to forbid one. If one were present, one could generate a Majorana neutrino mass via the renormalizable coupling

$$\tilde{g}_{H\nu\nu} H_{\Delta I = 1, \Delta L = L} \nu_L \cdot \nu_L \Rightarrow m^M = \tilde{g}_{H\nu\nu} \langle 0 | H_{\Delta I = 1, \Delta L = 2} | 0 \rangle \quad (129)$$

However, one could also generate a Majorana mass without such an additional Higgs field, via a non-renormalizable coupling to the conventional $\Delta I = \frac{1}{2}$ Standard Model Higgs field [109]:

$$\frac{1}{M} \left(H_{\Delta I = \frac{1}{2}} \nu_L \right) \cdot \left(H_{\Delta I = \frac{1}{2}} \nu_L \right) \Rightarrow m^M = \frac{1}{M} \langle 0 | H_{\Delta I = \frac{1}{2}} | 0 \rangle^2 \quad (130)$$

where M is some (presumably heavy: $M \gg m_W$) mass scale. The simplest possibility of generating a non-renormalizable interaction of the form (130) would be via the exchange of a heavy field N that is a singlet of $SU(3) \times SU(2) \times U(1)$ or $SU(5)$:

$$\frac{1}{M} \rightarrow \frac{\lambda^2}{M_N^2} \quad (131)$$

where one postulates a renormalizable coupling $\lambda H_{\Delta I=1/2} \nu_L \cdot N$. Such a heavy singlet field appears automatically in extensions of the $SU(5)$ GUT, such as $SO(10)$, but does not actually require the existence of any new GUT gauge bosons.

We now have all the elements we need for the see-saw mass matrix [110] favoured by GUT model-builders:

$$(\nu_L, N) \cdot \begin{pmatrix} m^M & m^D \\ m^D & M^M \end{pmatrix} \begin{pmatrix} \nu_L \\ N \end{pmatrix} \quad (132)$$

where the $\nu_L \cdot \nu_L$ Majorana mass m^M might arise from a $\Delta I = 1$ Higgs with coupling $\tilde{g}_{H\bar{\nu}\nu}$, (129), the $\nu_L \cdot N$ Dirac mass m^D could arise from a conventional Yukawa coupling λ (131) and should be of the same order as a conventional quark or lepton mass, and M^M could *a priori* be $\mathcal{O}(M_{GUT})$. Diagonalizing (132) and assuming that $m^M = 0$ or that $\langle 0|H_{\Delta I=1}|0 \rangle = \mathcal{O}(m_W^2/m_{GUT})$, as generically expected in GUTs, it is easy to diagonalize (132) and obtain the mass eigenstates

$$\begin{aligned} \nu_L + 0 \begin{pmatrix} m_W \\ m_X \end{pmatrix} N & : m = \mathcal{O} \left(\frac{m_W^2}{m_{GUT}} \right) \\ N + 0 \begin{pmatrix} m_W \\ m_X \end{pmatrix} \nu_L & : M = \mathcal{O}(M_{GUT}) \end{aligned} \quad (133)$$

So far, we have not touched on the generation structure of the neutrino masses. It is often suggested that m^M is negligible, M^M is (approximately) generation-independent, and $m^D \propto m_{2/3}$ (the u -quark mass matrix). If so, one sees that

$$m_{\nu_i} \sim \frac{m_{2/3_i}^2}{M_{GUT}} \quad (134)$$

and one might expect that

$$m_{\nu_e} \ll m_{\nu_\mu} \ll m_{\nu_\tau} \quad (135)$$

with mixing related to the Cabibbo-Kobayashi-Maskawa matrix.

As you know [111], evidence has recently been presented for atmospheric neutrino oscillations [2] between ν_μ and ν_τ with $\Delta m_A^2 \sim (10^{-2} \text{ to } 10^{-3}) \text{ eV}^2$ and a large mixing angle: $\sin^2 \theta_{\mu\tau} \gtrsim 0.8$. This is in addition to the previous evidence [112] for solar neutrino oscillations with $\Delta m_S^2 \simeq 10^{-5} \text{ eV}^2$ and $\sin^2 \theta \sim 10^{-3}$ or ~ 1 (Mikheev-Smirnov-Wolfenstein or MSW [113] oscillations) or $\Delta m_S^2 \sim 10^{-10} \text{ eV}^2$ and $\sin^2 \theta \sim 1$ (vacuum oscillations), as seen in Fig. 30.

Various theoretical groups [114] have restudied the previous see-saw prejudices in the light of the new data. In a hierarchical pattern of neutrino masses, one would expect

$$m_{\nu_3} \sim \sqrt{\Delta m_A^2} > m_{\nu_2} \sim \sqrt{\Delta m_S^2} > m_{\nu_1} \quad (136)$$

but is this compatible with the large mixing indicated (at least) for atmospheric neutrinos? Indeed it is [116], and theoretically it is difficult to see why any pair of neutrinos should be almost degenerate. On the other hand, there are perfectly natural 2×2 light-neutrino mass matrices that are compatible with large $\sin^2 2\theta_{\mu\tau}$ and the first mass hierarchy in (136) if $m_{\nu_2} \sim \sqrt{\Delta m_S^2} \sim 10^{-2.1/2} \text{ eV}$, particularly when it is observed [117] that renormalization-group effects below M_{GUT} may enhance $\sin^2 2\theta_{\mu\tau}$, as seen in Fig. 31 [116]. However, it is very difficult to understand the much larger hierarchy that would be needed for the vacuum solution to the solar neutrino problem with $m_{\nu_2} \sim \sqrt{\Delta m_S^2} \sim 10^{-5} \text{ eV}$. It is a more model-dependent question whether the large- or small-angle MSW solution is favoured. In one particular GUT model [116], we found the large-angle MSW solution more plausible, but the small-angle MSW solution could not be excluded. We still need more experimental information on neutrino masses and mixing, and this will surely be an active experimental field for years to come.

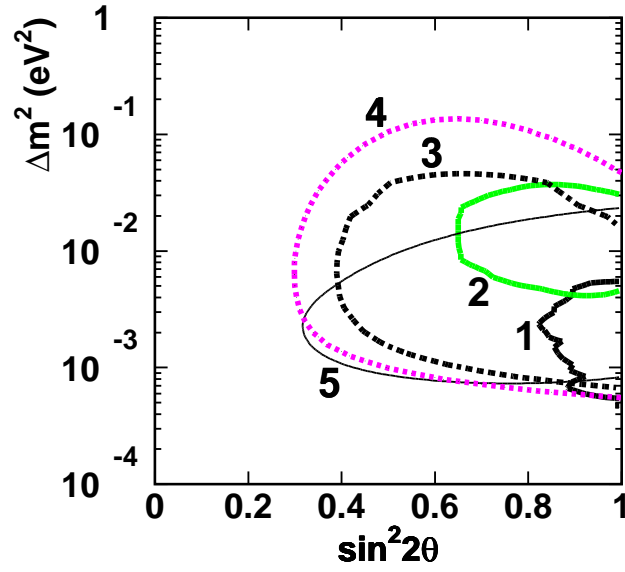


Fig. 30: Constraints on $\nu_\mu - \nu_\tau$ mixing from (1,2) contained events in super-Kamiokande and Kamiokande, (3,4) upward-going muons in super-Kamiokande and Kamiokande, and (5) stop/through upward muons in super-Kamiokande [115].

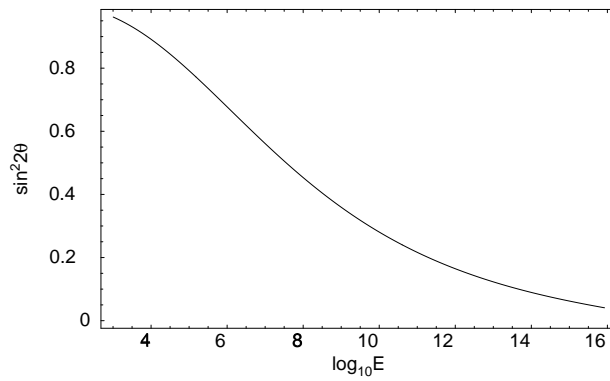


Fig. 31: Possible renormalization-group effect on the mixing angle in a two-generation neutrino model with supersymmetry at large $\tan\beta$ [116].

5 NONE OF THE ABOVE

Now is a good moment to review the progress made in addressing the defects of the Standard Model that were reviewed in Lecture 1, and to count how many parameters are still free. Grand unification reduces the three independent gauge couplings of the Standard Model to just one, so that is certainly progress. As far as the Standard Model Higgs sector is concerned, supersymmetry fixes the quartic Higgs coupling, which is progress, but the overall scale of electroweak gauge symmetry breaking is related to the undetermined scale of supersymmetry breaking. Indeed, in the absence of further theoretical input, the soft supersymmetry-breaking mass introduces $\mathcal{O}(100)$ new parameters unless (why?) one knows how to impose flavour universality.

Moreover, a complete treatment of neutrino masses involves additional parameters to describe the GUT Higgs potential [91, 61] and a see-saw mass matrix [110]. What is more, we still need at least one parameter to generate the cosmological baryon asymmetry and another to generate cosmological inflation. Finally, we should not forget gravity, whose parameters include the cosmological constant (if it vanishes, we need to understand why) as well as Newton's constant.

Thus, although progress has been made, there remain big questions in the supersymmetry-breaking sector and in quantum gravity, which are the main subjects addressed in this Lecture.

5.1 Supersymmetry Breaking

It is clear that supersymmetry must be broken: $m_{\tilde{e}} \neq m_e, m_{\tilde{\gamma}} \neq m_\gamma = 0$, etc. Could it be explicit or must it be spontaneous? Explicit supersymmetry breaking is not only ugly and unlike what occurs in gauge theories, it also induces inconsistencies at the quantum level when gravity is introduced. Therefore, attention has focused on spontaneous symmetry breaking. Since the supersymmetry charge Q is fermionic, this requires a non-zero matrix element for Q between the vacuum and some fermion χ called a Goldstone fermion or Goldstino:

$$\langle 0|Q|\chi \rangle = f_\chi^2 \neq 0 \quad (137)$$

The non-zero matrix element (137) has the immediate consequence, within global supersymmetry, that the vacuum energy is necessarily positive. This follows from the basic supersymmetry algebra:

$$\{Q, Q\} \propto \gamma_\mu P^\mu \quad (138)$$

Sandwiching (138) between vacuum states and inserting the intermediate state $|\chi \rangle \langle \chi|$, we find

$$|\langle 0|Q|\chi \rangle|^2 = f_\chi^4 > 0 \quad (139)$$

and hence the vacuum energy

$$\langle 0|P_0|0 \rangle \equiv E_0 > 0 \quad (140)$$

How may this spontaneous symmetry breaking be achieved? We recall the general form of the effective scalar potential in a globally supersymmetric theory:

$$V = \sum_i \left| \frac{\partial W}{\partial \phi^i} \right|^2 + \frac{1}{2} \sum_\alpha g_\alpha^2 |\phi^* T^\alpha \phi|^2 \quad (141)$$

where the former is called the F term and the latter the D term. In order to obtain (140), we need either the first term to be non-zero “ F -breaking” [118] - or the second term - “ D -breaking” [119]. The latter would require an extra $U(1)$ gauge group factor and many new matter fields, so we illustrate global supersymmetry breaking with the simplest F -breaking model [118].

Consider the superpotential

$$W = \alpha AB^2 + \beta C(B^2 - m^2) \quad (142)$$

where A, B, C denote gauge-singlet matter supermultiplets. It is easy to see that (142) yields

$$F_A^\dagger = \alpha B^2, \quad F_B^\dagger = 2B(\alpha A + \beta C), \quad F_C^\dagger = \beta(B^2 - m^2) \quad (143)$$

and hence an effective scalar potential

$$V = \sum_{i=A,B,C} |F_i|^2 = |2B(\alpha A + \beta C)|^2 + |\alpha B^2|^2 + |\beta(B^2 - m^2)|^2 \quad (144)$$

It is apparent that the last two terms cannot vanish simultaneously, so that $V > 0$ and supersymmetry is broken.

Recently there have been analyses of high-redshift supernovae [120] and other astrophysical and cosmological data [121] that favour a non-zero cosmological constant, as would be suggested by such positive vacuum energy. Unfortunately, the model (142) and others like it are too much of a good thing. The “observed” cosmological constant, if it is real at all, would correspond to

$$\Lambda \lesssim 10^{-123} m_P^4 \quad (145)$$

whereas such a global model of supersymmetry breaking would correspond to

$$\Lambda \sim (1 \text{ TeV})^4 \sim 10^{-64} m_P^4, \quad (146)$$

a discrepancy by some 60 orders of magnitude! Even the QCD vacuum energy

$$E_{QCD} \sim (100 \text{ MeV})^4 \sim 10^{-80} m_P^4 \quad (147)$$

is much larger than the “observed” value (145). This discrepancy may be the biggest problem in theoretical physics, even bigger than the hierarchy problem. However, to address it requires a true quantum theory of gravity, as is discussed later in this lecture.

However, before doing so, let us briefly review the latest incarnation of global supersymmetry-breaking models, namely gauge-mediated or messenger models [66]. The basic idea is to hide the ugly origin of supersymmetry breaking in a hidden sector of the theory that is coupled to observable particles via an intermediate set of “messenger” particles that share some of the gauge interactions of the Standard Model, as seen in Fig. 32. Gauge interactions then mediate the supersymmetry breaking needed in the observable sector. These models were originally conceived in the early 1980’s [122] because neither the F -breaking scenario (142) nor D -breaking models fitted within the MSSM. They have recently been reincarnated [123] with the idea that supersymmetry breaking in the hidden sector might originate from non-perturbative phenomena, which are much better understood by now [66].

There have been two principal motivations for this reincarnation. One is that gauge mediation naturally imposes flavour universality in the observable sector [123]. All quarks (or leptons) with the same charge acquire universal soft supersymmetry-breaking scalar masses, avoiding any problems with flavour-changing neutral interactions [65] and reducing the effective number of parameters in the observable sector. A feature of gauge-mediated models is the appearance of a massless Goldstone fermion λ , which would acquire a small mass when gravity is taken into account, as discussed in the next section. This implies that the lightest neutralino χ is unstable: $\chi \rightarrow \lambda\gamma$ decay dominates.

This provides the second motivation for gauge-mediated models, which is the report by the CDF collaboration [124] of an apparent $\bar{p}p \rightarrow e^+e^-\gamma\gamma +$ missing p_T event. It has been

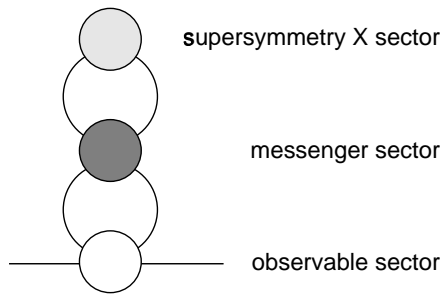


Fig. 32: Sketch of the principle of gauge-mediated (messenger) models: supersymmetry breaking in a hidden sector is communicated to the observable sector via gauge interactions [66].

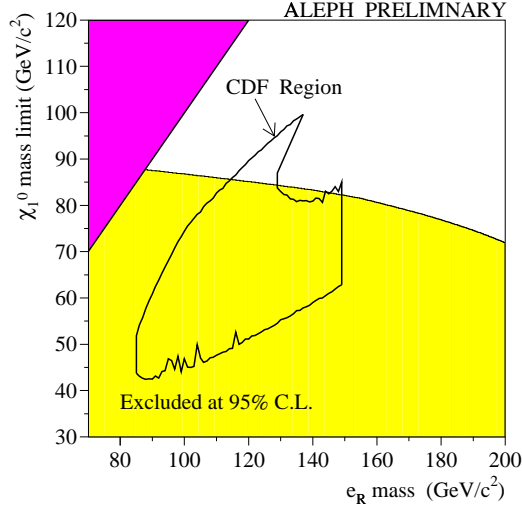


Fig. 33: Region of the $(m_{\tilde{e}}, m_{\chi})$ plane excluded by LEP [16] in models with a light gravitino, compared with the region favoured in the $\bar{p}p \rightarrow \tilde{e}\tilde{e} + X$ interpretation [125] of the CDF $\bar{p}p \rightarrow e^+e^-\gamma\gamma\cancel{p}_T + X$ event [124].

suggested [125] that this might be due to $\tilde{e}^+\tilde{e}^-$ pair production followed by $\tilde{e}^\pm \rightarrow e^\pm\chi$, $\chi \rightarrow \lambda\gamma$ decays, or due to $\chi^+\chi^-$ pair production followed by $\chi^\pm \rightarrow e^\pm\nu\chi$, $\chi \rightarrow \lambda\gamma$ decays. However, no other $\gamma\gamma$ + missing p_T events have been observed either at LEP [16] or at the FNAL Tevatron collider [126], and most of the parameter spaces for these interpretations have now been excluded, as seen in Fig. 33 [16]. Clearly, further experimental input from future collider runs is needed.

Some final comments on gauge-mediated models are in order. The first is that it has proved to be quite difficult to construct a model that is consistent with all the phenomenological constraints, has a desirable stable vacuum, etc. . The other is that, to the extent that the hidden scale $\Lambda_H \gg 1$ TeV, the apparent vacuum energy (cosmological constant) $\sim \Lambda_H^4 \gg 10^{-64} m_P^4$, worsening the discrepancy with the astrophysical upper limit (or observation) (145). However, in order to discuss this issue seriously, one needs a supersymmetric quantum theory of gravity, to which we now turn.

5.2 Local Supersymmetry and Supergravity

Why make a local theory of supersymmetry? One motivation is the analogy with gauge theories, in which bosonic symmetries are made local. Another is that local supersymmetry necessarily involves the introduction of gravity. Since both gravity and (surely!) supersymmetry exist, this

seems an inevitable step. It also leads to the possibility of unifying all the particle interactions including gravity, which was one of our original motivations for supersymmetry. Moreover, in the context of this Lecture, it is interesting that local supersymmetry (supergravity) admits an elegant mechanism for supersymmetry breaking [127], analogous to the Higgs mechanism in gauge theories, which allows us to address more seriously the possible existence of a cosmological constant.

The basic building block in a supergravity theory [45, 46] is the graviton supermultiplet of (51), which contains particles with helicities $(2, 3/2)$, the latter being the gravitino of spin $3/2$. Why is this required when one makes supersymmetry local?

We recall the basic global supersymmetry transformation laws (54) for bosons and fermions. Consider now the combination of two such global supersymmetry transformations:

$$[\delta_1, \delta_2] (\phi \text{ or } \psi) = -(\bar{\xi}_2 \gamma_\mu \xi_1) (i \partial_\mu) (\phi \text{ or } \psi) + \dots \quad (148)$$

The operator $(i \partial_\mu)$ corresponds to the momentum P_μ , and we see again that the combination of two global supersymmetry transformations is a translation. Consider now what happens when we consider local supersymmetry transformations characterized by a varying spinor $\xi(x)$. It is evident that the infinitesimal translation $\bar{\xi}_2 \gamma^\mu \xi_1$ in (148) is now x -dependent, and the previous global translation becomes a local coordinate transformation, as occurs in General Relativity.

How do we make the theory invariant under such local supersymmetry transformations? Consider again the simplest globally supersymmetric model containing a free spin-1/2 fermion and a free spin-0 boson (53), and make the local versions of the transformations (54). Following the same steps as in Lecture 2, we find that

$$\delta \mathcal{L} = \partial_\mu (\dots) + 2\bar{\psi} \gamma_\mu \not{\partial} S (\partial^\mu \xi(x)) + \text{herm. conj.} \quad (149)$$

In contrast to the global case, the action $A = \int d^4x \mathcal{L}$ is not invariant, because of the second term in (149). To cancel it out and restore invariance, we need more fields.

We proceed by analogy with gauge theories. In order to make the kinetic term $(i\bar{\psi}\not{\partial}\psi)$ invariant under gauge transformations $\psi \rightarrow e^{i\epsilon(x)}\psi$, we need to cancel a variation

$$-\bar{\psi} \partial_\mu \psi \partial^\mu \epsilon(x) \quad (150)$$

which is done by introducing a coupling to a gauge boson:

$$g\bar{\psi} \gamma_\mu \psi A^\mu(x) \quad (151)$$

and the corresponding transformation:

$$\delta A_\mu(x) = \frac{1}{g} \partial_\mu \epsilon(x) \quad (152)$$

In the supersymmetric case, we cancel the second term in (149) by a coupling:

$$\kappa \bar{\psi} \gamma_\mu \not{\partial} S \psi^\mu(x) \quad (153)$$

to a spin-3/2 spinor $\psi^\mu(x)$, representing a gauge fermion or gravitino, with the corresponding transformation:

$$\delta \psi^\mu = -\frac{2}{\kappa} \partial^\mu \xi(x) \quad (154)$$

where $\kappa \equiv 8\pi/m_P^2$.

For completeness, let us at least write down the Lagrangian for the graviton-gravitino supermultiplet:

$$L = -\frac{1}{2\kappa^2} \sqrt{-g}R - \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} \bar{\psi}_\mu \gamma_5 \gamma_\nu \mathcal{D}_\rho \psi_\sigma \quad (155)$$

where g denotes the determinant of the metric tensor:

$$g_{\mu\nu} = \epsilon_\mu^m \eta_{mn} \epsilon_\nu^u \quad (156)$$

where ϵ_μ^m is the vierbein and η_{mn} the Minkowski metric tensor, and \mathcal{D}_ρ is a covariant derivative

$$\mathcal{D}_\rho \equiv \partial_\rho + \frac{1}{4} \omega_\rho^{mn} [\gamma_m, \gamma_n] \quad (157)$$

where ω_ρ^{mn} is the spin connection. This is the simplest possible generally-covariant model of a spin-3/2 field. It is remarkable that it is invariant under the local supersymmetry transformations:

$$\begin{aligned} \delta \epsilon_\mu^m &= \frac{x}{2} \bar{\xi}(x) \gamma^m \psi_\mu(x), \\ \delta \omega_\mu^{mn} &= 0, \delta \psi_\mu = \frac{1}{x} \mathcal{D}_\mu \xi(x) \end{aligned} \quad (158)$$

just as the simplest possible (1/2, 0) theory (53) was globally supersymmetric, and also the action of an adjoint spin-1/2 field in a gauge theory.

It is also remarkable that supergravity admits an elegant analogue of the Higgs mechanism of spontaneous symmetry breaking [127]. Just as one combines the two polarization states of a massless gauge field with the single state of a massless Goldstone boson to obtain the three polarization states of a massive gauge boson, one may combine the two polarization states of a massless gravitino ψ_μ with the two polarization states of a massless Goldstone fermion λ to obtain the four polarization states of a massive spin-3/2 particle \tilde{G} . This super-Higgs mechanism corresponds to a spontaneous breakdown of local supersymmetry, since the massless graviton G has a different mass from the gravitino \tilde{G} :

$$m_G = 0 \neq m_{\tilde{G}}. \quad (159)$$

This is the only known consistent way of breaking local symmetry, just as the Higgs mechanism is the only way to generate $m_W \neq 0$.

Moreover, this can be achieved while keeping zero vacuum energy (cosmological constant), at least at the tree level. The reason for this is the appearance in local supersymmetry (supergravity) of a third term in the effective potential (141), which has a *negative* sign [127]. There is no time in these lectures to discuss this exciting feature in detail: the interested reader is referred to the original literature and the simplest example [128]. In this latter case, $\Lambda = V = 0$ for *any* value of the gravitino mass, for which reason it was named no-scale supergravity [129].

Again, there is no time to discuss here details of the coupling of supergravity to matter [127]. However, it is useful to have in mind the general features of the theory in the limit where $\kappa \rightarrow 0$, but the gravitino mass $m_{\tilde{G}} \equiv m_{3/2}$ remains fixed. One generally has non-zero gaugino masses $m_{1/2} \propto m_{3/2}$, and their universality is quite generic. One also has non-zero scalar masses $m_0 \propto m_{3/2}$, but their universality is much more problematic, and even violated in generic string models. It was this failing that partly refuelled the renewed interest in the gauge-mediated models mentioned in the previous section. A generic supergravity theory also yields non-universal trilinear soft supersymmetry-breaking couplings $A_\lambda \lambda \phi^3$: $A_\lambda \propto m_{3/2}$ and

bilinear scalar couplings $B_{\mu\mu}\phi^2 : B_{\mu} \propto m_{3/2}$. Therefore, supergravity may generate the full menagerie of soft supersymmetry-breaking terms:

$$-\frac{1}{2} \sum_a m_{1/2a} \tilde{V}_a \tilde{V}_a - \sum_i m_{0i}^2 |\phi_i|^2 - \left(\sum_{\lambda} A_{\lambda} \lambda \phi^3 + \text{h.c.} \right) - \left(\sum_{\mu} B_{\mu\mu} \phi^2 + \text{h.c.} \right) \quad (160)$$

Since these are generated at the supergravity scale near $m_P \sim 10^{19}$ GeV, the soft supersymmetry-breaking parameters are renormalized as discussed in Lecture 2. The analogous parameters in gauge-mediated models would also be renormalized, but to a different extent, because the mediation scale $\ll m_P$. This difference may provide a signature of such models, as discussed elsewhere [130].

Also renormalized is the vacuum energy (cosmological constant), which is a potential embarrassment. Loop corrections in a non-supersymmetric theory are quartically divergent, whereas those in a generic supergravity theory are only quadratically divergent, suggesting a contribution to the cosmological constant of order $m_{3/2}^2 m_P^2$, perhaps $O(10^{-32})m_P^4$! Particular models may have a one-loop quantum correction of order $m_{3/2}^4 = O(10^{-64})m_P^2$, but more magic (a new symmetry?) is needed to suppress the cosmological constant to the required level (145). This is one of the motivations for tackling string theory, which is our only candidate for a fundamental Theory of Everything including gravity.

5.3 Problems of Gravity

The greatest piece of unfinished business for twentieth-century physics is to reconcile general relativity with quantum mechanics. There are aspects of this problem, one being that of the cosmological constant, as discussed above. Another is that of perturbative quantum-gravity effects. Tree-level graviton exchange in $2 \rightarrow 2$ scattering, such as $e^+e^- \rightarrow e^+e^-$ at LEP, has an amplitude $A_G \sim E^2/m_P^2$, and hence a cross section

$$\sigma_G \sim E^2/m_P^4 \quad (161)$$

This is very small (negligible!) at LEP energies, reaching the unitarity limit only when $E \sim m_P$. However, when one calculates loop amplitudes involving gravitons, the rapid growth with energy (161) leads to uncontrollable, non-renormalizable divergences. These are of power type, and diverge faster and faster in higher orders of perturbation theory.

There are also non-perturbative problems in the quantization of gravity, that first arose in connection with black holes. From the pioneering work of Bekenstein and Hawking [131] on black-hole thermodynamics, we know that black holes have non-zero entropy S and temperature T , related to the Schwarzschild horizon radius. This means that the quantum description of a black hole should involve mixed states. The intuition underlying this feature is that information can be lost through the event horizon. Consider, for example, a pure quantum-mechanical pair state $|A, B\rangle \equiv \sum_i c_i |A_i\rangle |B_i\rangle$ prepared near the horizon, and what happens if one of the particles, say A , falls through the horizon while B escapes, as seen in Fig. 34. In this case,

$$\sum_i c_i |A_i B_i\rangle \rightarrow \sum_i |c_i|^2 |B_i\rangle \langle B_i| \quad (162)$$

and B emerges in a mixed state, as in Hawking's original treatment of the black-hole radiation that bears his name [131].

The problem is that conventional quantum mechanics does not permit the evolution of a pure initial state into a mixed final state. This is an issue both for the quantum particles discussed above and for the black hole itself. We could imagine having prepared the black hole

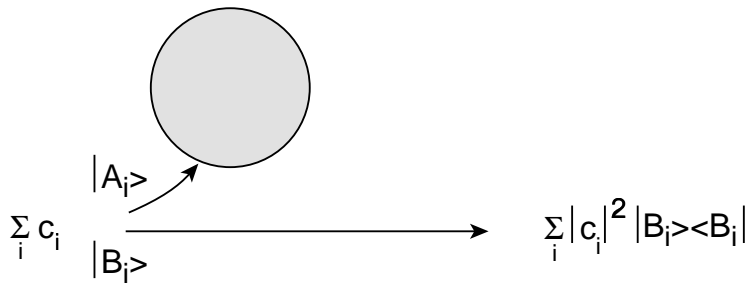


Fig. 34: If a pair of particles $|A\rangle |B\rangle$ is produced near the horizon of a black hole, and one of them ($|A\rangle$, say) falls in, the remaining particle $|B\rangle$ will appear to be in a mixed state.

by fusing massive or energetic particles in a pure initial state, e.g., by splitting a laser beam and then firing the sub-beams at each other as in a laser device for inertial nuclear fusion.

These problems point to a fundamental conflict between the proudest achievements of early twentieth-century physics, quantum mechanics and general relativity. One or the other should be modified, and perhaps both. Since quantum mechanics is sacred to field theorists, most particle physicists prefer to modify General Relativity by elevating it to string theory [30].⁹

5.4 Introduction to String Theory

At the level of perturbation theory, the divergence difficulties of quantum gravity can be related to the absence in a point-particle theory of a cutoff at short distances: for example,

$$\int^{\Lambda \rightarrow \infty} d^4k \left(\frac{1}{k^2} \right) \leftrightarrow \int_{1/\Lambda \rightarrow 0} d^4x \left(\frac{1}{x^6} \right) \sim \Lambda^2 \rightarrow \infty \quad (163)$$

Such divergences can be alleviated or removed if one replaces point particles by extended objects. The simplest possibility is to extend in just one dimension, leading us to a theory of strings. In such a theory, instead of point particles describing one-dimensional world lines, we have strings describing two-dimensional world sheets. Most popular have been closed loops of string, whose simplest world sheet is a tube. The “wiring diagrams” generated by the Feynman rules of conventional point-like particle theories become the “plumbing circuits” generated by the junctions and connections of these tubes of closed string. One could imagine generalizing this idea to higher-dimensional extended objects such as membranes describing world volumes, etc., and we return later to this option.

On a historical note, string models first arose from old-fashioned strong-interaction theory, before the advent of QCD. The lowest-lying hadronic states were joined by a very large number of excited states with increasing masses m and spins J :

$$J = \alpha' m^2 \quad (164)$$

where α' was called the “Regge slope”. One interpretation of this spectrum was of $\bar{q}q$ bound states in a linearly-rising potential, like an elastic string holding the constituents together, with tension $\mu = 1/\alpha'$. It was pointed out that such an infinitely (?) large set of resonances in the direct s -channel of a scattering process could be dual (equivalent) to the exchange of a similar infinite set in the crossed channel. Mathematically, this idea was expressed by the Veneziano [133] amplitude for $2 \rightarrow 2$ scattering, and its generalizations to $2 \rightarrow n$ particle

⁹It may be that this will eventually also require a modification of the *effective* quantum-mechanical space-time theory, even if the internal formulation of string theory is fully quantum-mechanical, but that is another story [132].

production processes. Then it was pointed out that these amplitudes could be derived formally from an underlying quantum theory of string [134]. However, this first incarnation of string theory was not able to accommodate the point-like partons seen inside hadrons at this time - the converse of the quantum-gravity motivation for string theory mentioned at the beginning of this section. Then along came QCD, which incorporated these point-like scaling properties and provided a qualitative understanding of confinement, which has now become quantitative with the advent of modern lattice calculations. Thus string theory languished as a candidate model of the strong interactions.

It was realized early on that unitarity required the existence of closed strings, even in an *a priori* open-string theory. Moreover, it was observed that the spectrum of a closed string included a massless spin-2 particle, which was an embarrassment for a theory of the strong interactions. However, this led to the idea [135] of reinterpreting string theory as a Theory of Everything, with this massless spin-2 state interpreted as the graviton and the string tension elevated to $\mu = O(m_P^2)$.

As already mentioned, one of the primary reasons for studying extended objects in connection with quantum gravity is the softening of divergences associated with short-distance behaviour. Since the string propagates on a world sheet, the basic formalism is two-dimensional. Accordingly, string vibrations may be described in terms of left- and right-moving waves:

$$\phi(r, t) \rightarrow \phi_L(r - t), \phi_R(r + t) \tag{165}$$

If the string has no boundary, as for a closed string, the left- and right-movers are independent. When quantized, they may be described by a two-dimensional field theory. Compared to a four-dimensional theory, it is relatively easy to make a two-dimensional field theory finite. In this case, it has conformal symmetry, which has an infinite-dimensional symmetry group in two dimensions. However, as you already know from gauge theories, one must be careful to ensure that this classical symmetry is not broken at the quantum level by anomalies. If the quantum string theory is to be consistent in a flat background space-time, the conformal anomaly fixes the number of left- and right-movers each to be equivalent to 26 free bosons if the theory has no supersymmetry, or 10 boson/fermion supermultiplets if the theory has $N = 1$ supersymmetry on the world sheet. There are other important quantum consistency conditions, and it was the demonstration by Green and Schwarz [136] that certain string theories are completely anomaly-free that opened the floodgates of theoretical interest in string theory as a Theory of Everything [30].

Among consistent string theories, one may enumerate the following. The *Bosonic String* exists in 26 dimensions, but this is not even its worst problem! It contains no fermionic matter degrees of freedom, and the flat-space vacuum is intrinsically unstable. *Superstrings* exist in 10 dimensions, have fermionic matter and also a stable flat-space vacuum. On the other hand, the ten-dimensional theory is left-right symmetric, and the incorporation of parity violation in four dimensions is not trivial. The *Heterotic String* [137] was originally formulated in 10 dimensions, with parity violation already incorporated, since the left- and right movers were treated differently. This theory also has a stable vacuum, but suffers from the disadvantage of having too many dimensions. *Four-Dimensional Heterotic Strings* may be obtained either by compactifying the six surplus dimensions: $10 = 4 + 6$ compact dimensions with size $R \sim 1/m_P$ [138], or by direct construction in four dimensions, replacing the missing dimensions by other internal degrees of freedom such as fermions [139] or group manifolds or ...? In this way it was possible to incorporate a GUT-like gauge group [99] or even something resembling the Standard Model [140].

What are the general features of such string models? First, they predict there are no more than 10 dimensions, which agrees with the observed number of 4! Secondly, they suggest that the rank of the four-dimensional gauge group should not be very large, in agreement with

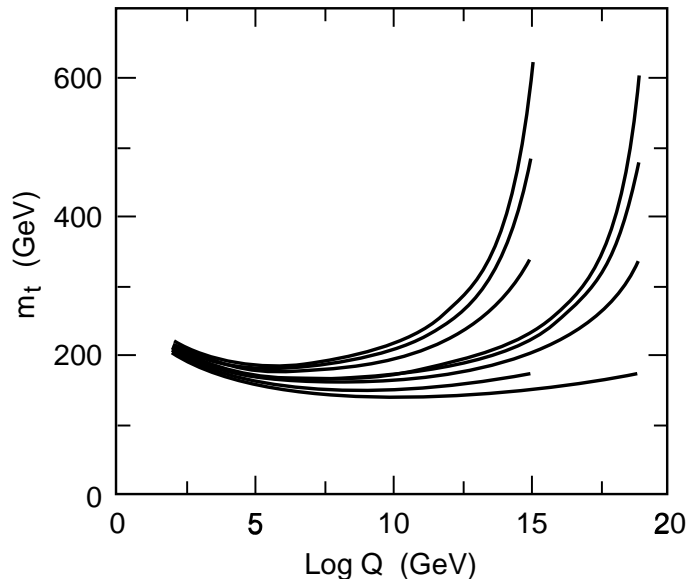


Fig. 35: The approximate infra-red fixed point of the renormalization-group equation for m_t means that a wide range of input Yukawa couplings at the GUT or string scale lead to similar physical values of $m_t \lesssim 190$ GeV.

the rank 4 of the Standard Model! Thirdly, the simplest four-dimensional string models do not accommodate large matter representations [141], such as an $\underline{8}$ of $SU(3)$ or a $\underline{3}$ of $SU(2)$, again in agreement with the known representation structure of the Standard Model! Fourthly, simple string models predict fairly successfully the mass of the top quark. This is because the maximum generic value of a Yukawa coupling λ_t is of the same order as the gauge coupling g . Applied to the top quark, this suggests that

$$m_t = \lambda_t \langle 0|H|0 \rangle = O(g) \times 250\text{GeV} \quad (166)$$

Moreover, the renormalization-group equation for λ_t exhibits an approximate infra-red fixed point, as seen in Fig. 35. This means that a large range of Yukawa coupling inputs at the Planck scale yield very similar physical values of $m_t \lesssim 190$ GeV. Fifthly, string theory makes a fairly successful prediction for the gauge unification scale in terms of m_P . If the intrinsic string coupling g_s is weak, one predicts [142]

$$M_{GUT} = O(g) \times \frac{m_P}{\sqrt{8\pi}} \simeq \text{few} \times 10^{17}\text{GeV} \quad (167)$$

where g is the gauge coupling, which is $\mathcal{O}(20)$ higher than the value calculated from the bottom up in Lecture 4 on the basis of LEP measurement of the gauge couplings. On the one hand, it is impressive that the bottom-up extrapolation over 14 decades agrees to within 10 % (on a logarithmic scale) with the top-down calculation (166). Nevertheless, it would be nice to obtain closer agreement, and this provides the major motivation for considering strongly-coupled string theory, which corresponds to a large internal dimension $l > m_{GUT}^{-1}$, as we discuss next.

5.5 Beyond String

Current developments involve going beyond string to consider higher-dimensional extended objects, such as generalized membranes with various numbers of internal dimensions. These may be obtained as solitons (non-perturbative classical solutions) of string theory, with masses

$$m \propto \frac{1}{g_s} \quad (168)$$

analogously to monopoles in gauge theory. It is evident from (168) that such membrane-solitons become light in the limit of strong string coupling $g_s \rightarrow \infty$.

It was observed some time ago that there should be a strong-coupling/weak-coupling duality [143] between elementary excitations and monopoles in supersymmetric gauge theories. These ideas have recently been confirmed in a spectacular solution of $N = 2$ supersymmetric gauge theory in four dimensions [144]. Similarly, it has recently been shown that there are analogous dualities in string theory, whereby solitons in some strongly-coupled string theory are equivalent to light string states in some other weakly-coupled string theory. Indeed, it appears that all string theories are related by such dualities. A particularity of this discovery is that the string coupling strength g_s is related to an extra dimension, in such a way that its size $R \rightarrow \infty$ as $g_s \rightarrow \infty$. This then leads to the idea of an underlying 11-dimensional framework called M theory [145] that reduces to the different string theories in different strong/weak-coupling limits, and reduces to eleven-dimensional supergravity in the low-energy limit.

A particular class of string solitons called D -branes [146] offers a promising approach to the black hole information paradox mentioned previously. According to this picture, black holes are viewed as solitonic balls of string, and their entropy simply counts the number of internal string states [132, 147]. These are in principle countable, so string theory may provide an accounting system for the information contained in black holes. Within this framework, the previously paradoxical process (162) becomes

$$|A, B \rangle + |BH \rangle \rightarrow |B' \rangle + |BH' \rangle \quad (169)$$

and the final state is pure if the initial state was. The apparent entropy of the final state in (162) is now interpreted as entanglement. The “lost” information is in principle encoded in the black-hole state, and this information could be extracted if we measured all properties of this ball of string.

In practice, we do not know how to recover this information from macroscopic black holes, so they appear to us as mixed states. What about microscopic black holes, namely fluctuations in the space-time background with $\Delta E = O(m_P)$, that last for a period $\Delta t = O(1/m_P)$ and have a size $\Delta x = O(1/m_P)$? Do these steal information from us [148], or do they give it back to us when they decay? Most people think there is no microscopic leakage of information in this way, but not all of us [132] are convinced. The neutral kaon system is among the most sensitive experimental areas [149, 150, 151] for testing this speculative possibility.

A final experimental comment concerns the magnitude of the extra dimension in M theory: LEP data suggest that it may be relatively large, with size $L_{11} \gg 1/m_{GUT} \simeq 1/10^{16} \text{ GeV} \gg 1/m_P$ [152]. Remember that the naïve string unification scale (167) is about 20 times larger than m_{GUT} . This may be traced to the fact that the gravitational interaction strength, although growing rapidly as a power of energy (161), is still much smaller than the gauge coupling strength at $E = m_{GUT}$. However, if an extra space-time dimension appears at an energy $E < m_{GUT}$, the gravitational interaction strength grows fast, as indicated in Fig. 36. Unification with gravity around 10^{16} GeV then becomes possible, *if* the gauge couplings do not also acquire a similar higher-dimensional kick. Thus we are led to the startling capacitor-plate framework for fundamental physics shown in Fig. 37.

Each plate is *a priori* ten-dimensional, and the bulk space between them is *a priori* eleven-dimensional. Six dimensions are compactified on a scale $L_6 \sim 1/m_{GUT}$, leaving a theory which is effectively five-dimensional in the bulk and four-dimensional on the walls. Conventional gauge interactions and observable matter particles are hypothesized to live on one capacitor plate, and there are other hidden gauge interactions and matter particles living on the other plate. The fifth dimension has a characteristic size which is estimated to be $O(10^{12} \text{ to } 10^{13} \text{ GeV})^{-1}$, and physics at large distances (smaller energies) looks effectively four-dimensional. Supersymmetry breaking

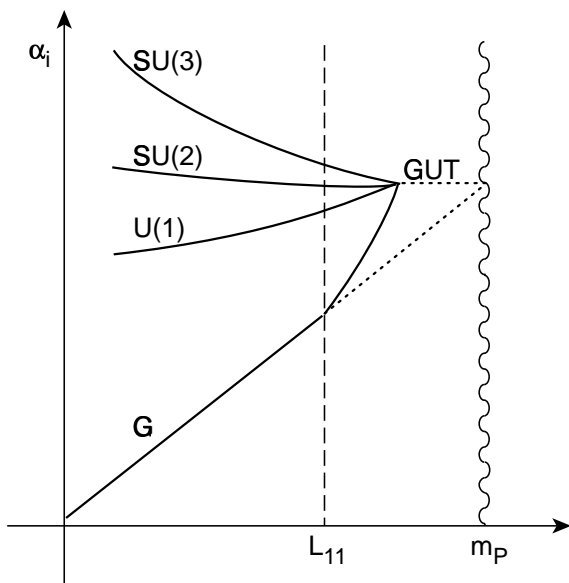


Fig. 36: Sketch of the possible evolution of the gauge couplings and the gravitational coupling G : if there is a large fifth dimension with size $\gg m_{GUT}^{-1}$, G may be unified with the gauge couplings at the GUT scale.

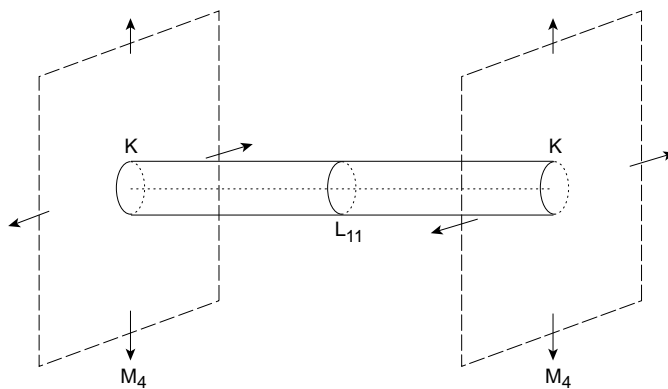


Fig. 37: The capacitor-plate scenario favoured in eleven-dimensional M theory. The eleventh dimension has a size $L_{11} \gg M_{GUT}^{-1}$, whereas dimensions 5, ..., 10 are compactified on a small manifold K with characteristic size $\sim M_{GUT}^{-1}$. the remaining four dimensions form (approximately) a flat Minkowski space M_4 .

is expected to originate on the hidden capacitor plate in this scenario, and to be transmitted to the observable wall by gravitational strength interactions in the bulk [153].

The phenomenological richness of this speculative M -theory approach is only beginning to be explored, and it remains to be seen whether it offers a realistic phenomenological description. However, it does embody all the available theoretical wisdom as well as offering the prospect of unifying all the observable gauge interactions with gravity at a single effective scale $\sim m_{GUT}$. As such, it constitutes our best contemporary guess about the Theory of Everything within and Beyond the Standard Model.

References

- [1] S.L. Glashow, *Nucl.Phys.* **22** (1961) 579;
S. Weinberg, *Phys.Rev.Lett.* **19** (1967) 1264;
A. Salam, Proc. 8th Nobel Symposium, Stockholm 1968, ed. N. Svartholm (Almqvist and Wiksells, Stockholm, 1968), p. 367.
- [2] Super-Kamiokande collaboration, Y. Fukuda et al., *Phys.Rev.Lett.* **81** (1998) 1562.
- [3] C. Quigg, *Gauge Theories of the Strong, Weak and Electromagnetic Interactions* (Benjamin-Cummings, Reading, 1983).
- [4] D. Ross and M. Veltman, *Nucl.Phys.* **B95** (1975) 135.
- [5] M. Grünewald and D. Karlen, talks at International Conference on High-Energy Physics, Vancouver 1998,
<http://www.cern.ch/LEPEWVG/misc>.
- [6] C. Bouchiat, J. Iliopoulos and Ph. Meyer, *Phys.Lett.* **138B** (1972) 652 and references therein.
- [7] H. Georgi, H. Quinn and S. Weinberg, *Phys.Rev.Lett.* **33** (1974) 451.
- [8] H. Georgi and S.L. Glashow, *Phys.Rev.Lett.* **32** (1974) 438.
- [9] S. Dimopoulos and H. Georgi, *Nucl.Phys.* **B193** (1981) 50;
S. Dimopoulos, S. Raby and F. Wilczek, *Phys.Rev.* **D24** (1981) 1681;
L. Ibáñez and G.G. Ross, *Phys.Lett.* **105B** (1981) 439.
- [10] J.S. Bell, *Nucl.Phys.* **60** (1973) 427;
C.H. Llewellyn Smith, *Phys.Lett.* **46B** (1973) 233;
J. Cornwall, D. Levin and G. Tiktopoulos, *Phys.Rev.Lett.* **30** (1973) 1268.
- [11] J. Ellis, M.K. Gaillard and D.V. Nanopoulos, *Nucl.Phys.* **B106** (1976) 292.
- [12] M. Veltman, *Acta Phys.Pol.* **B8** (1977) 475;
C. Vayonakis, *Lett.Nuovo Cimento* **17** (1976) 383.
- [13] B.W. Lee, C. Quigg and H.B. Thacker, *Phys.Rev.* **D16** (1977) 1519.
- [14] L. Durand and J. Lopez, *Phys.Rev.Lett.* **64** (1990) 1215 and *Phys.Rev.* **D45** (1992) 3112.
- [15] B.L. Ioffe and V.A. Khoze, *Sov.J.Part.Nucl.* **9** (1978) 50.

- [16] LEP Experiments Committee meeting, Nov. 12th, 1998,
<http://www.cern.ch/Committees/LEPC/minutes/LEPC50.html>.
 This source provides some preliminary updates of experimental plots reproduced in these notes, which are mainly taken from papers contributed to the International Conference on High-Energy Physics, Vancouver 1998:
<http://ichep.triumf.ca/main.asp>.
- [17] G. Altarelli, R. Kleiss and C. Verzegnassi, eds., *Physics at LEP 1*, CERN Report 89-08 (1989);
 D. Bardin, W. Hollik and G. Passarino, eds., *Reports of the Working Group on Precision Calculations for the Z Resonance*, CERN Report 95-03 (1995).
- [18] M. Davier and A. Hocker, *Phys.Lett.* **435** (1998) 427.
- [19] A. Sirlin, *Phys.Rev.* **D22** (1980) 971, *Nucl.Phys.* **B332** (1990) 20 and *Phys.Lett.* **B267** (1991) 240.
- [20] M. Veltman, *Nucl.Phys.* **123** (1977) 89;
 M.S. Chanowitz, M. Furman and I. Hinchliffe, *Phys.Lett.* **B78** (1978) 285.
- [21] M. Veltman, *Acta Phys.Pol.* **8** (1977) 475.
- [22] U. Amaldi et al., *Phys.Rev.* **D36** (1987) 1385
- [23] G. Costa et al., *Nucl.Phys.* **B297** (1988) 244.
- [24] Particle Data Group, C. Caso et al., *Eur.Phys.J.* **C3** (1998) 1.
- [25] G. Altarelli, T. Sjöstrand and F. Zwirner, eds., *Proceedings of the Workshop on Physics at LEP 2*, CERN Report 96-01 (1996).
- [26] J. Ellis, G.L. Fogli and E. Lisi, *Phys.Lett.* **B318** (1993) 148; *Phys.Lett.* **B333** (1994) 118;
Zeit.Phys. **C69** (1996) 627 and *Phys.Lett.* **B389** (1996) 321.
- [27] Y.F. Pirogov and O.V. Zenin, hep-ph/9808396.
- [28] P. Hasenfratz and J. Nager, *Zeit.Phys.* **C37** (1988) 477.
- [29] Y. Nir, lectures at this school.
- [30] M.B. Green, lectures at this school.
- [31] Y. Nambu, in: Z. Ajduk, S. Pokorski and A. Trautman, eds., *Proc. XI Int. Symp. on Elementary Particle Physics* (World Scientific, Singapore, 1989);
 A. Miranski, M. Tanabashi and K. Yamauraki, *Mod.Phys.Lett* **A4** (1989) 1043 and *Phys.Lett.* **B221** (1989) 177.
- [32] W.A. Bardeen, C.T. Hill and M. Lindner, *Phys.Rev.* **D41** (1990) 1647;
 For a review, see: E. Farhi and L. Susskind, *Phys.Rep.* **74C** (1981) 277.
- [33] S. Dimopoulos and L. Susskind, *Nucl.Phys.* **B155** (1979) 237;
 E. Eichten and K. Lane, *Phys.Lett.* **B90** (1980) 125.
- [34] M.E. Peskin and T. Takeuchi, *Phys.Rev.* **D46** (1992) 381.
- [35] G. Altarelli, R. Barbieri and S. Jadach, *Nucl.Phys.* **B369** (1992) 3; **376** (1992) 444 (E);
 G. Altarelli, R. Barbieri and F. Caravaglios, *Nucl.Phys.* **B405** (1993) 3.

- [36] S. Matsumoto, private communication.
- [37] G. Altarelli, R. Barbieri and F. Caravaglios, *Int.J.Mod.Phys.* **A13** (1998) 1031.
- [38] B. Holdom, *Phys.Lett.* **B105** (1985) 301;
T. Appelquist, D. Karabali and L.C.R. Wijewardhana, *Phys.Rev.Lett.* **57** (1986) 957;
M. Bando, T. Morozumi, H. So and K.Yamawaki, *Phys.Rev.Lett.* **59** (1987) 389;
T. Akiba and T. Yanagida, *Phys.Lett.* **B169** (1986) 432.
- [39] S. Coleman and J. Mandula, *Phys.Rev.* **159** (1967) 1251.
- [40] Y.A. Gol'fand and E.P. Likhtman, *Pis'ma Zh.Eksp.Teor.Fiz.* **13** (1971) 323.
- [41] P. Ramond, *Phys.Rev.* **D3** (1971) 2415;
A. Neveu and J.H. Schwarz, *Phys.Rev.* **D4** (1971) 1109.
- [42] D.V. Volkov and V.P. Akulov, *Phys.Lett.* **46B** (1973) 109.
- [43] J. Wess and B. Zumino, *Nucl.Phys.* **B70** (1974) 39.
- [44] J. Iliopoulos and B. Zumino, *Nucl.Phys.* **B76** (1974) 310;
S. Ferrara, J. Iliopoulos and B. Zumino, *Nucl.Phys.* **77** (1974) 413.
- [45] D.Z. Freedman, P. van Nieuwenhuizen and S. Ferrara, *Phys.Rev.* **D13** (1976) 3214.
- [46] S. Deser and B. Zumino, *Phys.Lett.* **62B** (1976) 335.
- [47] B. de Wit and H. Nicolai, *Nucl.Phys.* **B188** (1981) 98.
- [48] L. Maiani, *Proc. Summer School on Particle Physics*, Gif-sur-Yvette, 1979 (IN2P3, Paris, 1980) p. 3;
G't Hooft, in: G't Hooft et al., eds., *Recent Developments in Field Theories* (Plenum Press, New York, 1980);
E. Witten, *Nucl.Phys.* **B188** (1981) 513;
R.K. Kaul, *Phys.Lett.* **109B** (1982) 19.
- [49] P. Fayet and S. Ferrara, *Phys.Rep.* **32** (1977) 251.
- [50] R. Haag, J. Lopuszanski and M. Sohnius, *Nucl.Phys.* **B88** (1975) 257.
- [51] I. Antoniadis, J. Ellis and G. Leontaris, *Phys.Lett.* **B399** (1997) 92.
- [52] M. Peskin, in: N. Ellis and M. Neubert, eds., *Proc. 1996 European School of High-Energy Physics*, Carry-le-Rouet, France, CERN report 97-03, p.49, hep-ph/9705479.
- [53] P. Fayet, in: S. Ferrara, J. Ellis and P. van Nieuwenhuizen, eds., *Unification of the Fundamental Particle Interactions* (Plenum Press, New York, 1980), p. 587.
- [54] H.E. Haber and G.L. Kane, *Phys.Rep.* **117** (1985) 75.
- [55] See, for instance, the following papers and references therein:
G. Bhattacharyya, D. Choudhury and K. Sridhar, *Mod.Phys.Lett.* **A10** (1995) 1583 and *Phys.Lett.* **B355** (1995) 193;
C.E. Carlson, P. Roy and M. Sher, *Phys.Lett.* **B357** (1995) 2276;
M. Hirsch, H.V. Klapdor-Kleingrothaus and S.G. Kovalenko, *Phys.Rev.Lett.* **75** (1995) 17;
K.S. Babu and R.N. Mohapatra, *Phys.Rev.Lett.* **75** (1995) 2276;
A.Yu. Smirnov and F. Vissani, *Phys.Lett.* **B380** (1996) 317;

- D. Choudhury and P. Roy, *Phys.Lett.* **B378** (1996) 153;
M. Chaichian and K. Huitu, *Phys.Lett.* **B384** (1996) 157;
H. Nunokawa, A. Rossi and J.W.F. Valle, *Nucl.Phys.* **B482** (1996) 481.
- [56] G. Altarelli, J. Ellis, G. Giudice, S. Lola and M. Mangano, *Nucl.Phys.* **506** (1997) 3.
- [57] J. Ellis, Invited parallel session talk at the International Europhysics Conference on High-Energy Physics, Jerusalem, 1997, hep-ph/9712340.
- [58] H1 collaboration, C. Adloff et al., *Zeit.Phys.* **C74** (1997) 191;
ZEUS collaboration, J. Breitweg et al., *Zeit.Phys.* **C74** 207.
- [59] Y. Okada, M. Yamaguchi and T. Yanagida, *Progr.Theor.Phys.* **85** (1991) 1;
J. Ellis, G. Ridolfi and F. Zwirner, *Phys.Lett.* **B257** (1991) 83, *Phys.Lett.* **B262** (1991) 477;
H.E. Haber and R. Hempfling, *Phys.Rev.Lett.* **66** (1991) 1815;
R. Barbieri, M. Frigeni and F. Caravaglios, *Phys.Lett.* **B258** (1991) 167;
Y. Okada, M. Yamaguchi and T. Yanagida, *Phys.Lett.* **B262** (1991) 54.
- [60] J. Ellis, S. Kelley and D.V. Nanopoulos, *Phys.Lett.* **B249** (1990) 441.
- [61] S. Dimopoulos and H. Georgi, *Nucl.Phys.* **B193** (1981) 150.
- [62] L. Girardello and M. Grisaru, *Nucl.Phys.* **B194** (1982) 65.
- [63] K. Inoue, A. Kakuto, H. Komatsu and S. Takeshita, *Prog.Theor.Phys.* **68** (1982) 927.
- [64] I. Jack, D.R.T. Jones and A. Pickering, *Phys.Lett.* **B432** (1998) 114.
- [65] J. Ellis and D.V. Nanopoulos, *Phys.Lett.* **110B** (1982) 44.
- [66] G. Giudice and R. Rattazzi, hep-ph/9801271.
- [67] See, e.g., C. Kounnas, A.B. Lahanas, D.V. Nanopoulos and M. Quirós, *Phys.Lett.* **132B** (1983) 95.
- [68] L. Ibanez and G.G. Ross, *Phys.Lett.* **110B** (1982) 215.
- [69] M. Carena, M. Quiros and C.E.M. Wagner, *Nucl.Phys.* **B461** (1996) 407;
H.E. Haber, R. Hempfling and A.H. Hoang, *Zeit.Phys.* **C75** (1997) 539.
- [70] E. Richter-Was, D. Froidevaux, F. Gianotti, L. Poggioli, D. Cavalli and S. Resconi, *Int.J.Mod.Phys.* **A13** (1998) 1371.
- [71] ECFA/DESY LC Physics Working Group, E Accomando et al., *Phys.Rep.* **299** (1998) 1.
- [72] M. Carena, S. Mrenna and C.E.M. Wagner, hep-ph/9808312.
- [73] J. Ellis and S. Rudaz, *Phys.Lett.* **128B** (1983) 248.
- [74] J. Ellis, T. Falk, G. Ganis, K. Olive and M. Schmitt, *Phys.Rev.* **D58** (1998) 095002.
- [75] J. Ellis, J.S. Hagelin, D.V. Nanopoulos, K.A. Olive and M. Srednicki, *Nucl.Phys.* **B238** (1984) 453.
- [76] H. Goldberg, *Phys.Rev.Lett.* **50** (1983) 1419.
- [77] J. Rich, M. Spiro and J. Lloyd-Owen, *Phys.Rep.* **151** (1987) 239;
P.F. Smith, *Contemp.Phys.* **29** (1998) 159;
T.K. Hemmick et al., *Phys.Rev.* **D41** (1990) 2074.

- [78] J. Ellis, T. Falk, K. Olive and M. Schmitt, *Phys.Lett.* **B388** (1996) 97.
- [79] H.V. Klapdor-Kleingrothaus and Y. Ramachers, *Eur.Phys.J.* **A3** (1998) 85.
- [80] For a review, see: M.S. Turner, astro-ph/9808149.
- [81] J. Ellis, T. Falk, K. Olive and M. Schmitt, *Phys.Lett.* **B413** (1997) 355.
- [82] S. Abdullin and F. Charles, hep-ph/9811402.
- [83] J. Ellis, T. Falk and K. Olive, hep-ph/9810360.
- [84] I. Hinchliffe, F.E. Paige, M.D. Shapiro, J. Soderqvist and W. Yao, *Phys.Rev.* **D55** (1997) 5520;
LHCC Workshop on Supersymmetry, <http://www.cern.ch/Commitees/LHCC/SUSY96.html>.
- [85] See, for example:
A. Djouadi, G. Girardi, C. Verzegnassi, W. Hollik and F. Renard, *Nucl.Phys.* **B349** (1991) 2054;
M. Boulware and D. Finnell, *Phys.Rev.* **D44** (1991) 2054;
G. Altarelli, R. Barbieri and F. Caravaglios, *Phys.Lett.* **B314** 357;
D. Garcia and J. Sola, *Phys.Lett.* **B357** (1995) 349;
X. Wang, J. Lopez and D.V. Nanopoulos, *Phys.Rev.* **D52** (1995) 4116;
M. Shifman, *Mod.Phys.Lett.* **A10** (1995) 605;
G.L. Kane, R.G. Stuart and J.D. Wells, *Phys.Lett.* **B354** (1995) 350;
J. Erler and P. Langacker, *Phys.Rev.* **D52** (1995) 441;
P.H. Chankowski and S. Pokorski, *Nucl.Phys.* **B475** (1996) 3.
- [86] J. Ellis, J.L. Lopez and D.V. Nanopoulos, *Phys.Lett.* **B397** (1997) 88.
- [87] H. Georgi, H. Quinn and S. Weinberg, *Phys.Rev.Lett.* **33** (1974) 451.
- [88] M. Mangano, lectures at this school.
- [89] J. Ellis and D.V. Nanopoulos, *Nature* **292** (1981) 436.
- [90] M. Chanowitz, J. Ellis and M.K. Gaillard, *Nucl.Phys.* **B128** (1977) 506.
- [91] A.J. Buras, J. Ellis, M.K. Gaillard and D.V. Nanopoulos, *Nucl.Phys.* **B135** (1978) 66.
- [92] D.V. Nanopoulos and D.A. Ross, *Phys.Lett.* **118B** (1982) 99.
- [93] J. Ellis, S. Kelley and D.V. Nanopoulos, *Phys.Lett.* **260** (1991) 131;
U. Amaldi, W. de Boer and H. Furstenau, *Phys.Lett.* **B260** (1991) 447;
P. Langacker and M. Luo, *Phys.Rev.* **D44** (1991) 817.
- [94] J. Ellis, S. Kelley and D.V. Nanopoulos, *Nucl.Phys.* **B373** (1992) 55.
- [95] P. Langacker and N. Polonsky, *Phys.Rev.* **D47** (1993) 4028.
- [96] F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, *Nuovo Cimento* **104A** (1991) 1817;
F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, *Nuovo Cimento* **105A** (1992) 1210.
- [97] R. Barbieri and L.J. Hall, *Phys.Rev.Lett.* **68** (1992) 752;
J. Hisano, T. Moroi, K. Tobe and T. Yanagida, *Phys.Lett.* **B342** (1995) 138.
- [98] J. Ellis, M.K. Gaillard and D.V. Nanopoulos, *Phys.Lett.* **91B** (1980) 67.

- [99] I. Antoniadis, J. Ellis, J.S. Hagelin and D.V. Nanopoulos, *Phys.Lett.* **B194** (1987) 231 and **B231** (1989) 65.
- [100] J. Ellis and M.K. Gaillard, *Phys.Lett.* **88B** (1979) 315.
- [101] W. Marciano and A. Sirlin, *Phys.Rev.Lett.* **46** (1981) 163.
- [102] Super-Kamiokande collaboration, M. Shiozawa et al., *Phys.Rev.Lett.* **81** (1998) 3319.
- [103] J. Ellis, D.V. Nanopoulos and S. Rudaz, *Nucl.Phys.* **B202** (1982) 43;
S. Dimopoulos, S. Raby and F. Wilczek, *Phys.Lett.* **112B** (1982) 133.
- [104] S. Weinberg, *Phys.Rev.* **D26** (1982) 287,
N. Sakai and T. Yanagida, *Nucl.Phys.* **B197** (1982) 533.
- [105] J. Ellis, D.V. Nanopoulos and S. Rudaz, *Nucl.Phys.* **B202** (1982) 43.
- [106] J. Ellis, J.S. Hagelin, S. Kelley and D.V. Nanopoulos, *Nucl.Phys.* **B311** (1988) 1.
- [107] B.A. Campbell, J. Ellis and S. Rudaz, *Phys.Lett.* **141B** (1984) 229.
- [108] M. Girone for the ALEPH collaboration, parallel session talk presented at the International Europhysics Conference on High-Energy Physics, Jerusalem 1997, Imperial College preprint IC-HEP-97-17 (1997).
- [109] R. Barbieri, J. Ellis and M.K. Gaillard, *Phys.Lett.* **90B** (1980) 249.
- [110] T. Yanagida, Proc. Workshop on the Unified Theory and the Baryon Number in the Universe (KEK, Japan, 1979);
R. Slansky, Talk at the Sanibel Symposium, Caltech preprint CALT-68-709 (1979).
- [111] P. Litchfield, lectures at this school.
- [112] J.N. Bahcall, Invited talk at Neutrino 98, Takayama, astro-ph/9808162.
- [113] L. Wolfenstein, *Phys.Rev.* **D17** (1978) 2369;
S.P. Mikheyev and A.Yu. Smirnov, *Nuovo Cimento* **9C** (1986) 17.
- [114] For some recent work see,
S.F. King, hep-ph/9806440;
J.K. Elwood, N. Irges and P. Ramond, hep-ph/9807325;
V. Barger et al, hep-ph/9806328;
P. Osland and G. Vigdel, hep-ph/9806339;
A. Joshipura and A. Smirnov, hep-ph/9806376;
M. Tanimoto, hep-ph/9807283;
G. Lazarides and N. Vlachos, hep-ph/9807253;
V. Barger, T. Weiler and K. Whisnant, hep-ph/9807319;
J. Pati, hep-ph/9807315;
M. Gonzalez-Carcia et al, hep-ph/9807305;
G. Altarelli and F. Feruglio, hep-ph/9807353.
- [115] T. Kajita, for the Super-Kamiokande and Kamiokande collaborations, Invited talk at Neutrino 98, Takayama, hep-ex/9810001.
- [116] J. Ellis, G. Leontaris, S. Lola and D.V. Nanopoulos, hep-ph/9808251.
- [117] M. Tanimoto, *Phys.Lett.* **B360** (1995) 41.

- [118] L. O’Raifeartaigh, *Nucl.Phys.* **B96** (1975) 331.
- [119] P. Fayet and J. Iliopoulos, *Phys.Lett.* **51B** (1974) 461.
- [120] S. Perlmutter et al., astro-ph/9712212;
A.G. Riess et al., astro-ph/9805201.
- [121] N. Bahcall, astro-ph/9711062.
- [122] L.Alvarez-Gaume, M. Claudson and M.B. Wise, *Nucl.Phys.* **B207** (1982) 96;
J. Ellis, L. Ibanez and G.G. Ross, *Phys.Lett.* **113B** (1982) 283.
- [123] See, e.g., M. Dine and A. Nelson, *Phys.Rev.* **D48** (1993) 1277, **D51** (1995) 1362 and **D53** (1996) 2658.
- [124] CDF collaboration, F. Abe et al., *Phys.Rev.Lett.* **81** (1998) 1791.
- [125] See, for example:
D. Stump, M. Wiest and C.-P. Yuan, *Phys.Rev.* **D54** (1996) 1936;
S. Dimopoulos, M. Dine, A. Raby and S. Thomas, *Phys.Rev.Lett.* **76** (1996) 3494;
S. Ambrosanio, G. Kane, G. Kribs, S. Martin and S. Mrenna, *Phys.Rev.Lett.* **76** (1996) 3498 and *Phys.Rev.* **D54** (1996) 5395;
S. Dimopoulos, S. Thomas and J. Wells, *Phys.Rev.* **D54** (1996) 3283;
K. Babu, C. Kolda and F. Wilczek, *Phys.Rev.Lett.* **77** (1996) 3070;
J.L. Lopez and D.V. Nanopoulos, *Mod.Phys.Lett.* **A10** (1996) 2473 and *Phys.Rev.* **D55** (1997) 4450;
J.L. Lopez, D.V. Nanopoulos and A. Zichichi, *Phys.Rev.Lett.* **77** (1996) 5168 and *Phys.Rev.* **D55** (1997) 5813.
- [126] D0 collaboration, S. Abachi et al., *Phys.Rev.Lett.* **78** (1997) 2070.
- [127] E. Cremmer, B. Julia, J. Scherk, S. Ferrara, L. Girardello and P. Van Nieuwenhuizen, *Nucl.Phys.* **B147** (1979) 105.
- [128] E. Cremmer, S. Ferrara, C. Kounnas and D.V. Nanopoulos, *Phys.Lett.* **133B** (1983) 61.
- [129] J. Ellis, A.B. Lahanas, D.V. Nanopoulos and K.A. Tamvakis, *Phys.Lett.* **134B** (1984) 429.
- [130] A. Strumia, *Phys.Lett.* **B409** (1997) 213.
- [131] J. Bekenstein, *Phys.Rev.* **D12** (1975) 3077;
S. Hawking, *Comm.Math.Phys.* **43** (1975) 199.
- [132] J. Ellis, N.E. Mavromatos and D.V. Nanopoulos, *Mod.Phys.Lett.* **A10** (1995) 425 and references therein
- [133] G. Veneziano, *Nuovo Cimento* **57A** (1968) 190 and *Phys.Rep.* **C9** (1974) 199.
- [134] Y. Nambu, *Proc. Int. Conf. on Symmetries and Quark Models*, Wayne State University (1969);
P. Goddard, J. Goldstone, C.Rebbi and C. Thorn, *Nucl.Phys.* **B181** (1981) 502.
- [135] J. Scherk and J.H. Schwarz, *Nucl.Phys.* **B81** (1974) 118.
- [136] M.B. Green and J.H. Schwarz, *Phys.Lett.* **149B** (1984) 117 and **151B** (1985) 21.

- [137] D.J. Gross, J.A. Harvey, E. Martinec and R. Rohm, *Phys.Rev.Lett.* **54** (1985) 502 and *Nucl.Phys.* **B256** (1985) 253, **B267** (1985) 75.
- [138] P. Candelas, G. Horowitz, A. Strominger and E. Witten, *Nucl.Phys.* **B258** (1985) 46.
- [139] H. Kawai, D. Lewellen and S.H.H. Tye, *Nucl.Phys.* **B287** (1987) 1;
I. Antoniadis, C. Bachas and C. Kounnas, *Nucl.Phys.* **B289** (1987) 87;
I. Antoniadis and C. Bachas, *Nucl.Phys.* **B298** (1988) 586.
- [140] A. Faraggi, D.V. Nanopoulos and K. Yuan, *Nucl.Phys.* **B335** (1990) 347.
- [141] H. Dreiner, J. Lopez, D.V. Nanopoulos and D.B. Reiss, *Phys.Lett.* **B216** (1989) 283.
- [142] I. Antoniadis, J. Ellis, R. Lacaze and D.V. Nanopoulos, *Phys.Lett.* **B268** (1991) 188.
- [143] C. Montonen and D. Olive, *Phys.Lett.* **72B** (1977) 117.
- [144] N. Seiberg and E. Witten, *Nucl.Phys.* **B431** (1994) 484.
- [145] For a review, see: Miao Li, hep-th/9811019.
- [146] J. Polchinski, *Phys.Rev.Lett.* **75** (1995) 4724.
- [147] A. Strominger and C. Vafa, *Phys.Lett.* **B379** (1996) 99.
- [148] S. Hawking, *Commun.Math.Phys.* **87** (1983) 395 and *Phys.Rev.* **D37** (1988) 904.
- [149] J. Ellis, J.S. Hagelin, D.V. Nanopoulos and M. Srednicki, *Nucl.Phys.* **B241** (1984) 381.
- [150] J. Ellis, J. Lopez, N.E. Mavromatos and D.V. Nanopoulos, *Phys.Rev.* **D53** (1996) 3846.
- [151] CPLEAR collaboration, R. Adlet et al., and J. Ellis, J. Lopez, N.E. Mavromatos and D.V. Nanopoulos, *Phys.Lett.* **B364** (1995) 239.
- [152] P. Horava and E. Witten, *Nucl.Phys.* **B460** (1996) 506 and *Nucl.Phys.* **B475** (1996) 94;
P. Horava, *Phys.Rev.* **D54** (1996) 7561.
- [153] J. Ellis, Z. Lalak, S.Pokorski and W. Pokorski, hep-ph/9805377;
J. Ellis, Z. Lalak and W. Pokorski, hep-th/9811133.

COSMOLOGY AND PARTICLE PHYSICS

J.A. Peacock

Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK

Abstract

These lectures cover some of the basics of modern cosmology, assuming relatively little prior knowledge of the subject. They are organised into three main sections: (1) Models of the expanding universe (the Robertson-Walker metric, dynamics and the equation of state, the hot big bang, initial conditions and inflation); (2) Dark matter (astrophysical mass measurements, particle candidates for dark matter, constraints on dark matter from galaxy haloes, dark matter and cosmological perturbations); (3) Structure formation (statistics of cosmological density fields, generation of fluctuations via inflation, observations of large-scale structure, fluctuations in the microwave background).

1 THE ISOTROPIC UNIVERSE

1.1 The Robertson–Walker metric

Cosmological investigation began by considering the simplest possible mass distribution: one whose properties are **homogeneous** (constant density) and **isotropic** (the same in all directions). From this symmetry, the only allowed velocity field on a local scale is expansion (or contraction) with velocity proportional to distance:

$$\mathbf{v} = H\mathbf{r}. \quad (1)$$

Having chosen a model mass distribution, the next step is to solve the field equations to find the corresponding metric. Since our model is a particularly symmetric one, it is perhaps not too surprising that many of the features of the metric can be deduced from symmetry alone – and indeed will apply even if Einstein’s equations are replaced by something more complicated. These general arguments were put forward independently by H.P. Robertson and A.G. Walker in 1936.

Cosmological time The first point to note is that something suspiciously like a universal time exists in an isotropic universe. Consider a set of observers in different locations, all of whom are at rest with respect to the matter in their vicinity (these characters are usually termed **fundamental observers**). We can envisage them as each sitting on a different galaxy, and so receding from each other with the general expansion. We can define a global time coordinate t , which is the time measured by the clocks of these observers – *i.e.* t is the proper time measured by an observer at rest with respect to the local matter distribution. The coordinate is useful globally rather than locally because the clocks can be synchronized by the exchange of light signals between observers, who agree to set their clocks to a standard time when *e.g.* the universal homogeneous density reaches some given value. Using this time coordinate plus isotropy, we already have enough information to conclude that the metric must take the following form:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[f^2(r) dr^2 + g^2(r) d\psi^2 \right]. \quad (2)$$

Here, we have used the equivalence principle to say that the proper time interval between two distant events would look locally like special relativity to a fundamental observer on the spot: for them, $c^2 d\tau^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2$. Since we use the same time coordinate as they do, our only difficulty is in the spatial part of the metric: relating their dx *etc.* to spatial coordinates centred on us.

Because of spherical symmetry, the spatial part of the metric can be decomposed into a radial and a transverse part (in spherical polars, $d\psi^2 = d\theta^2 + \sin^2 \theta d\phi^2$). Distances have been decomposed into a product of a time-dependent **scale factor** $R(t)$ and a time-independent **comoving coordinate** r . The functions f and g are arbitrary; however, we can choose our radial coordinate such that either $f = 1$ or $g = r^2$, to make things look as much like Euclidean space as possible. Furthermore, we can determine the form of the remaining function from symmetry arguments.

To get some feeling for the general answer, it should help to think first about a simpler case: the metric on the surface of a sphere. A balloon being inflated is a common popular analogy for the expanding universe, and it will serve as a two-dimensional example of a space of constant curvature. If we call the polar angle in spherical polars r instead of the more usual θ , then the element of length on the surface of a sphere of radius R is

$$d\sigma^2 = R^2 (dr^2 + \sin^2 r d\phi^2). \quad (3)$$

It is possible to convert this to the metric for a 2-space of constant **negative curvature** by the device of considering an imaginary radius of curvature, $R \rightarrow iR$. If we simultaneously let $r \rightarrow ir$, we obtain

$$d\sigma^2 = R^2 (dr^2 + \sinh^2 r d\phi^2). \quad (4)$$

These two forms can be combined by defining a new radial coordinate that makes the transverse part of the metric look Euclidean:

$$d\sigma^2 = R^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\phi^2 \right), \quad (5)$$

where $k = +1$ for positive curvature and $k = -1$ for negative curvature.

An isotropic universe has the same form for the comoving spatial part of its metric as the surface of a sphere. This is no accident, since it is possible to define the equivalent of a sphere in higher numbers of dimensions, and the form of the metric is always the same. For example, a **3-sphere** embedded in four-dimensional Euclidean space would be defined as the coordinate relation $x^2 + y^2 + z^2 + w^2 = R^2$. Now define the equivalent of spherical polars and write $w = R \cos \alpha$, $z = R \sin \alpha \cos \beta$, $y = R \sin \alpha \sin \beta \cos \gamma$, $x = R \sin \alpha \sin \beta \sin \gamma$, where α , β and γ are three arbitrary angles. Differentiating with respect to the angles gives a four-dimensional vector (dx, dy, dz, dw) , and it is a straightforward exercise to show that the squared length of this vector is

$$|(dx, dy, dz, dw)|^2 = R^2 [d\alpha^2 + \sin^2 \alpha (d\beta^2 + \sin^2 \beta d\gamma^2)], \quad (6)$$

which is the Robertson–Walker metric for the case of positive spatial curvature. This $k = +1$ metric describes a **closed universe**, in which a traveller who sets off along a trajectory of fixed β and γ will eventually return to their starting point (when $\alpha = 2\pi$). In this respect, the positively curved 3D universe is identical to the case of the surface of a sphere: it is finite, but unbounded. By contrast, the $k = -1$ metric describes an **open universe** of infinite extent; as before, changing to negative spatial curvature replaces $\sin \alpha$ with $\sinh \alpha$, and α can be made as

large as we please without returning to the starting point. The $k = 0$ model describes a **flat universe**, which is also infinite in extent. This can be thought of as a limit of either of the $k = \pm 1$ cases, where the curvature scale R tends to infinity.

Notation and conventions The Robertson–Walker metric (which we shall often write in the shorthand **RW metric**) may be written in a number of different ways. The most compact forms are those where the comoving coordinates are *dimensionless*. Define the very useful function

$$S_k(r) = \begin{cases} \sin r & (k = 1) \\ \sinh r & (k = -1) \\ r & (k = 0), \end{cases} \quad (7)$$

and its cosine-like analogue, $C_k(r) \equiv \sqrt{1 - kS_k^2(r)}$. The metric can now be written in the preferred form that we shall use throughout:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left[dr^2 + S_k^2(r) d\psi^2 \right]. \quad (8)$$

The most common alternative is to use a different definition of comoving distance, $S_k(r) \rightarrow r$, so that the metric becomes

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left(\frac{dr^2}{1 - kr^2} + r^2 d\psi^2 \right). \quad (9)$$

There should of course be two different symbols for the different comoving radii, but each is often called r in the literature, so we have to learn to live with this ambiguity; the presence of terms like $S_k(r)$ or $1 - kr^2$ will usually indicate which convention is being used. Alternatively, one can make the scale factor dimensionless, defining

$$a(t) \equiv \frac{R(t)}{R_0}, \quad (10)$$

so that $a = 1$ at the present.

The redshift At small separations, where things are Euclidean, the proper separation of two fundamental observers is just $R(t) dr$, so that we obtain Hubble's law with

$$H = \frac{\dot{R}}{R}. \quad (11)$$

At large separations where spatial curvature becomes important, the concept of radial velocity becomes a little more slippery – but in any case how could one measure it directly in practice? At small separations, the recessional velocity gives the Doppler shift

$$\frac{\nu_{\text{emit}}}{\nu_{\text{obs}}} \equiv 1 + z \simeq 1 + \frac{v}{c}. \quad (12)$$

This defines the **redshift** z in terms of the shift of spectral lines. What is the equivalent of this relation at larger distances? Since photons travel on null geodesics of zero proper time, we see directly from the metric that

$$r = \int \frac{c dt}{R(t)}. \quad (13)$$

The comoving distance is constant, whereas the domain of integration in time extends from t_{emit} to t_{obs} ; these are the times of emission and reception of a photon. Photons that are emitted at later times will be received at later times, but these changes in t_{emit} and t_{obs} cannot alter the integral, since r is a comoving quantity. This requires the condition $dt_{\text{emit}}/dt_{\text{obs}} = R(t_{\text{emit}})/R(t_{\text{obs}})$, which means that events on distant galaxies time-dilate according to how much the universe has expanded since the photons we see now were emitted. Clearly (think of events separated by one period), this dilation also applies to frequency, and we therefore get

$$\boxed{\frac{\nu_{\text{emit}}}{\nu_{\text{obs}}} \equiv 1 + z = \frac{R(t_{\text{obs}})}{R(t_{\text{emit}})}}. \quad (14)$$

In terms of the normalized scale factor $a(t)$ we have simply $a(t) = (1+z)^{-1}$. Photon wavelengths therefore stretch with the universe, as is intuitively reasonable.

1.2 Dynamics of the expansion

Expansion and geometry The equation of motion for the scale factor can be obtained in a quasi-Newtonian fashion. Consider a sphere about some arbitrary point, and let the radius be $R(t)r$, where r is arbitrary. The motion of a point at the edge of the sphere will, in Newtonian gravity, be influenced only by the interior mass. We can therefore write down immediately a differential equation (**Friedmann's equation**) that expresses conservation of energy: $(\dot{R}r)^2/2 - GM/(Rr) = \text{constant}$. The Newtonian result that the gravitational field inside a uniform shell is zero does still hold in general relativity, and is known as **Birkhoff's theorem**. General relativity becomes even more vital in giving us the constant of integration in Friedmann's equation:

$$\boxed{\dot{R}^2 - \frac{8\pi G}{3}\rho R^2 = -kc^2}. \quad (15)$$

Note that this equation covers all contributions to ρ , *i.e.* those from matter, radiation and vacuum; it is independent of the equation of state.

For a given rate of expansion, there is thus a **critical density** that will yield $k = 0$, making the comoving part of the metric look Euclidean:

$$\boxed{\rho_c = \frac{3H^2}{8\pi G}}. \quad (16)$$

A universe with density above this critical value will be **spatially closed**, whereas a lower-density universe will be **spatially open**.

It is sometimes convenient to work with the time derivative of the Friedmann equation, because acceleration arguments in dynamics can often be more transparent than energy ones.

Differentiating with respect to time requires a knowledge of $\dot{\rho}$, but this can be eliminated by means of conservation of energy: $d[\rho c^2 R^3] = -pd[R^3]$. We then obtain

$$\ddot{R} = -4\pi GR(\rho c^2 + 3p)/3. \quad (17)$$

Both this equation and the Friedmann equation in fact arise as independent equations from different components of Einstein's equations for the RW metric.

Density parameters etc. The 'flat' universe with $k = 0$ arises for a particular **critical density**. We are therefore led to define a **density parameter** as the ratio of density to critical density:

$$\Omega \equiv \frac{\rho}{\rho_c} = \frac{8\pi G\rho}{3H^2}. \quad (18)$$

Since ρ and H change with time, this defines an epoch-dependent density parameter. The current value of the parameter should strictly be denoted by Ω_0 . Because this is such a common symbol, we shall keep the formulae uncluttered by normally dropping the subscript; the density parameter at other epochs will be denoted by $\Omega(z)$. The critical density therefore just depends on the rate at which the universe is expanding. If we now also define a dimensionless (current) Hubble parameter as

$$h \equiv \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}}, \quad (19)$$

then the current density of the universe may be expressed as

$$\begin{aligned} \rho_0 &= 1.88 \times 10^{-26} \Omega h^2 \text{ kg m}^{-3} \\ &= 2.78 \times 10^{11} \Omega h^2 M_{\odot} \text{ Mpc}^{-3}. \end{aligned} \quad (20)$$

A powerful approximate model for the energy content of the universe is to divide it into pressureless matter ($\rho \propto R^{-3}$), radiation ($\rho \propto R^{-4}$) and vacuum energy (ρ constant). The first two relations just say that the number density of particles is diluted by the expansion, with photons also having their energy reduced by the redshift; the third relation applies for Einstein's **cosmological constant**. In terms of observables, this means that the density is written as

$$\frac{8\pi G\rho}{3} = H_0^2(\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4}) \quad (21)$$

(introducing the normalized scale factor $a = R/R_0$). For some purposes, this separation is unnecessary, since the Friedmann equation treats all contributions to the density parameter equally:

$$\frac{kc^2}{H^2 R^2} = \Omega_m(a) + \Omega_r(a) + \Omega_v(a) - 1. \quad (22)$$

Thus, a flat $k = 0$ universe requires $\sum \Omega_i = 1$ at all times, whatever the form of the contributions to the density, even if the equation of state cannot be decomposed in this simple way.

Lastly, it is often necessary to know the present value of the scale factor, which may be read directly from the Friedmann equation:

$$R_0 = \frac{c}{H_0} [(\Omega - 1)/k]^{-1/2}. \quad (23)$$

The Hubble constant thus sets the **curvature length**, which becomes infinitely large as Ω approaches unity from either direction.

Solutions to the Friedmann equation The Friedmann equation may be solved most simply in ‘parametric’ form, by recasting it in terms of the conformal time $d\eta = c dt/R$ (denoting derivatives with respect to η by primes):

$$R'^2 = \frac{8\pi G}{3c^2} \rho R^4 - kR^2. \quad (24)$$

Because $H_0^2 R_0^2 = kc^2/(\Omega - 1)$, the Friedmann equation becomes

$$a'^2 = \frac{k}{(\Omega - 1)} \left[\Omega_r + \Omega_m a - (\Omega - 1)a^2 + \Omega_v a^4 \right], \quad (25)$$

which is straightforward to integrate provided $\Omega_v = 0$.

To the observer, the evolution of the scale factor is most directly characterised by the change with redshift of the Hubble parameter and the density parameter; the evolution of $H(z)$ and $\Omega(z)$ is given immediately by the Friedmann equation in the form $H^2 = 8\pi G\rho/3 - kc^2/R^2$. Inserting the above dependence of ρ on a gives

$$H^2(a) = H_0^2 \left[\Omega_v + \Omega_m a^{-3} + \Omega_r a^{-4} - (\Omega - 1)a^{-2} \right]. \quad (26)$$

This is a crucial equation, which can be used to obtain the relation between redshift and comoving distance. The radial equation of motion for a photon is $R dr = c dt = c dR/\dot{R} = c dR/(RH)$. With $R = R_0/(1+z)$, this gives

$$\begin{aligned} R_0 dr &= \frac{c}{H(z)} dz \\ &= \frac{c}{H_0} \left[(1 - \Omega)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4 \right]^{-1/2} dz. \end{aligned} \quad (27)$$

This relation is arguably the single most important equation in cosmology, since it shows how to relate comoving distance to the observables of redshift, Hubble constant and density parameters.

Lastly, using the expression for $H(z)$ with $\Omega(a) - 1 = kc^2/(H^2 R^2)$ gives the redshift dependence of the total density parameter:

$$\Omega(z) - 1 = \frac{\Omega - 1}{1 - \Omega + \Omega_v a^2 + \Omega_m a^{-1} + \Omega_r a^{-2}}. \quad (28)$$

This last equation is very important. It tells us that, at high redshift, all model universes apart from those with only vacuum energy will tend to look like the $\Omega = 1$ model. If $\Omega \neq 1$, then in

the distant past $\Omega(z)$ must have differed from unity by a tiny amount: the density and rate of expansion needed to have been finely balanced for the universe to expand to the present. This tuning of the initial conditions is called the **flatness problem** and is one of the motivations for the applications of quantum theory to the early universe.

Matter-dominated universe From the observed temperature of the microwave background (2.73 K) and the assumption of three species of neutrino at a slightly lower temperature (see below), we deduce that the total relativistic density parameter is $\Omega_r h^2 \simeq 4.2 \times 10^{-5}$, so at present it should be a good approximation to ignore radiation. However, the different redshift dependences of matter and radiation densities mean that this assumption fails at early times: $\rho_m/\rho_r \propto (1+z)^{-1}$. One of the critical epochs in cosmology is therefore the point at which these contributions were equal: the redshift of **matter–radiation equality**

$$1 + z_{\text{eq}} \simeq 23\,900 \Omega h^2. \quad (29)$$

At redshifts higher than this, the universal dynamics were dominated by the relativistic-particle content. By a coincidence discussed below, this epoch is close to another important event in cosmological history: **recombination**. Once the temperature falls below $\simeq 10^4$ K, ionized material can form neutral hydrogen. Observational astronomy is only possible from this point on, since Thomson scattering from electrons in ionized material prevents photon propagation. In practice, this limits the maximum redshift of observational interest to about 1000; unless Ω is very low or vacuum energy is important, a matter-dominated model is therefore a good approximation to reality.

Models with vacuum energy The solution of the Friedmann equation becomes more complicated if we allow a significant contribution from vacuum energy – *i.e.* a non-zero cosmological constant. Detailed discussions of the problem are given by Felten & Isaacman (1986) and Carroll, Press & Turner (1992); the most important features are outlined below.

The Friedmann equation itself is independent of the equation of state, and just says $H^2 R^2 = kc^2/(\Omega - 1)$, whatever the form of the contributions to Ω . In terms of the cosmological constant itself, we have

$$\Omega_v = \frac{8\pi G\rho_v}{3H^2} = \frac{\Lambda c^2}{3H^2}. \quad (30)$$

de Sitter space Before going on to the general case, it is worth looking at the endpoint of an outwards perturbation of Einstein’s static model, first studied by de Sitter and named after him. This universe is completely dominated by vacuum energy, and is clearly the limit of the unstable expansion, since the density of matter redshifts to zero while the vacuum energy remains constant. Consider again the Friedmann equation in its general form $\dot{R}^2 - 8\pi G\rho R^2/3 = -kc^2$: since the density is constant and R will increase without limit, the two terms on the lhs must eventually become almost exactly equal and the curvature term on the rhs will be negligible. Thus, even if $k \neq 0$, the universe will have a density that differs only infinitesimally from the critical, so that we can solve the equation by setting $k = 0$, in which case

$$R \propto \exp Ht, \quad H = \sqrt{\frac{8\pi G\rho_v}{3}} = \sqrt{\frac{\Lambda c^2}{3}}. \quad (31)$$

An interesting interpretation of this behaviour was promoted in the early days of cosmology by Eddington: the cosmological constant is what *caused* the expansion.

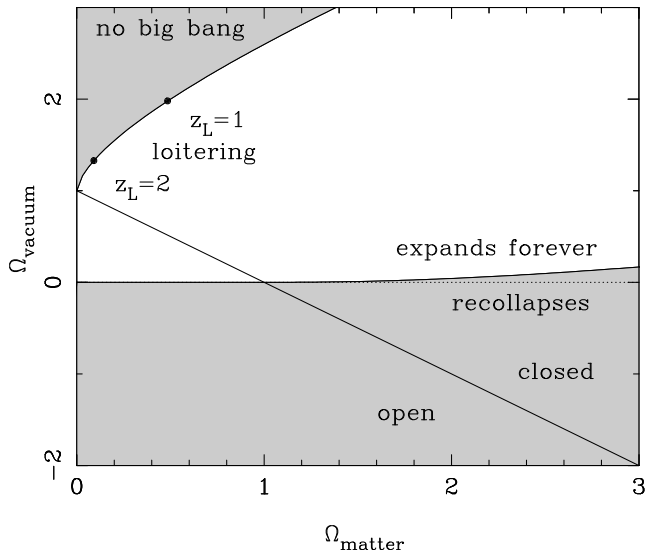


Fig. 1: This plot shows the different possibilities for the cosmological expansion as a function of matter density and vacuum energy. Models with total $\Omega > 1$ are always spatially closed (open for $\Omega < 1$), although closed models can still expand to infinity if $\Omega_v \neq 0$. If the cosmological constant is negative, recollapse always occurs; recollapse is also possible with a positive Ω_v if $\Omega_m \gg \Omega_v$. If $\Omega_v > 1$ and Ω_m is small, there is the possibility of a ‘loitering’ solution with some maximum redshift and infinite age (top left); for even larger values of vacuum energy, there is no big bang singularity.

Bouncing and loitering models Returning to the general case of models with a mixture of energy in the vacuum and normal components, we have to distinguish three cases. For models that start from a big bang (in which case radiation dominates completely at the earliest times), the universe will either recollapse or expand forever. The latter outcome becomes more likely for low densities of matter and radiation, but high vacuum density. It is however also possible to have models in which there is no big bang: the universe was collapsing in the distant past, but was slowed by the repulsion of a positive Λ term and underwent a ‘bounce’ to reach its present state of expansion. Working out the conditions for these different events is a matter of integrating the Friedmann equation. For the addition of Λ , this can only in general be done numerically. However, we can find the conditions for the different behaviours described above analytically, at least if we simplify things by ignoring radiation. The equation in the form of the time-dependent Hubble parameter looks like

$$\frac{H^2}{H_0^2} = \Omega_v(1 - a^{-2}) + \Omega_m(a^{-3} - a^{-2}) + a^{-2}, \quad (32)$$

and we are interested in the conditions under which the lhs vanishes, defining a turning point in the expansion. Setting the rhs to zero yields a cubic equation, and it is possible to give the conditions under which this has a solution (see Felten & Isaacman 1986). The main results of this analysis are summed up in figure 1. Since the radiation density is very small today, the main task of relativistic cosmology is to work out where on the $\Omega_{\text{matter}} - \Omega_{\text{vacuum}}$ plane the real universe lies. The existence of high-redshift objects rules out the bounce models, so that the idea of a hot big bang cannot be evaded.

Flat universe The most important model in cosmological research is that with $k = 0 \Rightarrow \Omega_{\text{total}} = 1$; when dominated by matter, this is often termed the **Einstein–de Sitter** model.

Paradoxically, this importance arises because it is an unstable state: as we have seen earlier, the universe will evolve away from $\Omega = 1$, given a slight perturbation. For the universe to have expanded by so many **e-foldings** (factors of e expansion) and yet still have $\Omega \sim 1$ implies that it was very close to being spatially flat at early times.

It now makes more sense to work throughout in terms of the normalized scale factor $a(t)$, so that the Friedmann equation for a matter–radiation mix is

$$\dot{a}^2 = H_0^2 \left(\Omega_m a^{-1} + \Omega_r a^{-2} \right), \quad (33)$$

which may be integrated to give the time as a function of scale factor:

$$H_0 t = \frac{2}{3\Omega_m^2} \left[\sqrt{\Omega_r + \Omega_m a} (\Omega_m a - 2\Omega_r) + 2\Omega_r^{3/2} \right]; \quad (34)$$

this goes to $\frac{2}{3}a^{3/2}$ for a matter-only model, and to $\frac{1}{2}a^2$ for radiation only.

One further way of presenting the model's dependence on time is via the density. Following the above, it is easy to show that

$$\begin{aligned} t &= \sqrt{\frac{1}{6\pi G\rho}} && \text{(matter domination)} \\ t &= \sqrt{\frac{3}{32\pi G\rho}} && \text{(radiation domination)}. \end{aligned} \quad (35)$$

Because Ω_r is so small, the deviations from a matter-only model are unimportant for $z \lesssim 1000$, and so the distance–redshift relation for the $k = 0$ matter plus radiation model is effectively just that of the $\Omega_m = 1$ Einstein–de Sitter model. An alternative $k = 0$ model of greater observational interest has a significant cosmological constant, so that $\Omega_m + \Omega_v = 1$ (radiation being neglected for simplicity). This may seem contrived, but once $k = 0$ has been established, it cannot change: individual contributions to Ω must adjust to keep in balance. The advantage of this model is that it is the only way of retaining the theoretical attractiveness of $k = 0$ while changing the age of the universe from the relation $H_0 t_0 = 2/3$, which characterises the Einstein–de Sitter model. Since much observational evidence indicates that $H_0 t_0 \simeq 1$, this model has received a good deal of interest in recent years. To keep things simple we shall neglect radiation, so that the Friedmann equation is

$$\dot{a}^2 = H_0^2 [\Omega_m a^{-1} + (1 - \Omega_m) a^2], \quad (36)$$

and the $t(a)$ relation is

$$H_0 t(a) = \int_0^a \frac{x \, dx}{\sqrt{\Omega_m x + (1 - \Omega_m) x^4}}. \quad (37)$$

The x^4 on the bottom looks like trouble, but it can be rendered tractable by the substitution $y = \sqrt{x^3 |\Omega_m - 1| / \Omega_m}$, which turns the integral into

$$H_0 t(a) = \frac{2}{3} \frac{S_k^{-1}(\sqrt{a^3 |\Omega_m - 1| / \Omega_m})}{\sqrt{|\Omega_m - 1|}}. \quad (38)$$

Here, k in S_k is used to mean \sin if $\Omega_m > 1$, otherwise \sinh ; these are still $k = 0$ models. This $t(a)$ relation is compared to models without vacuum energy in figure 2. Since there is nothing special about the current era, we can clearly also rewrite this expression as

$$H(a) t(a) = \frac{2}{3} \frac{S_k^{-1}(\sqrt{|\Omega_m(a) - 1| / \Omega_m(a)})}{\sqrt{|\Omega_m(a) - 1|}} \simeq \frac{2}{3} \Omega_m(a)^{-0.3}, \quad (39)$$

where we include a simple approximation that is accurate to a few % over the region of interest ($\Omega_m \gtrsim 0.1$). In the general case of significant Λ but $k \neq 0$, this expression still gives a very good approximation to the exact result, provided Ω_m is replaced by $0.7\Omega_m - 0.3\Omega_v + 0.3$ (Carroll, Press & Turner 1992).

Horizons For photons, the radial equation of motion is just $c dt = R dr$. How far can a photon get in a given time? The answer is clearly

$$\Delta r = \int_{t_0}^{t_1} \frac{c dt}{R(t)} = \Delta \eta, \quad (40)$$

i.e. just the interval of conformal time. What happens as $t_0 \rightarrow 0$ in this expression? We can replace dt by dR/\dot{R} , which the Friedmann equation says is $\propto dR/\sqrt{\rho R^2}$ at early times. Thus, this integral converges if $\rho R^2 \rightarrow \infty$ as $t_0 \rightarrow 0$, otherwise it diverges. Provided the equation of state is such that ρ changes faster than R^{-2} , light signals can only propagate a finite distance between the big bang and the present; there is then said to be a **particle horizon**. Such a horizon therefore exists in conventional big bang models, which are dominated by radiation at early times.

1.3 Observations in cosmology

We can now assemble some essential formulae for interpreting cosmological observations. Since we will mainly be considering the post-recombination epoch, these apply for a matter-dominated model only. Our observables are redshift, z , and angular difference between two points on the sky, $d\psi$. We write the metric in the form

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) [dr^2 + S_k^2(r) d\psi^2], \quad (41)$$

so that the *comoving* volume element is

$$dV = 4\pi [R_0 S_k(r)]^2 R_0 dr. \quad (42)$$

The *proper* transverse size of an object seen by us is its comoving size $d\psi S_k(r)$ times the scale factor at the time of emission:

$$d\ell = d\psi R_0 S_k(r) / (1 + z). \quad (43)$$

Probably the most important relation for observational cosmology is that between monochromatic flux density and luminosity. Start by assuming isotropic emission, so that the photons emitted by the source pass with a uniform flux density through any sphere surrounding the source. We can now make a shift of origin, and consider the RW metric as being centred on the source; however, because of homogeneity, the comoving distance between the source and the observer is the same as we would calculate when we place the origin at our location. The photons from the source are therefore passing through a sphere, on which we sit, of proper surface area $4\pi [R_0 S_k(r)]^2$. But redshift still affects the flux density in four further ways: photon energies and arrival rates are redshifted, reducing the flux density by a factor $(1+z)^2$; opposing this, the bandwidth $d\nu$ is reduced by a factor $1+z$, so the energy flux per unit bandwidth goes down by one power of $1+z$; finally, the observed photons at frequency ν_0 were emitted at

frequency $\nu_0(1+z)$, so the flux density is the luminosity at this frequency, divided by the total area, divided by $1+z$:

$$S_\nu(\nu_0) = \frac{L_\nu([1+z]\nu_0)}{4\pi R_0^2 S_k^2(r)(1+z)}. \quad (44)$$

A word about units: L_ν in this equation would be measured in units of W Hz^{-1} . Recognizing that emission is often not isotropic, it is common to consider instead the luminosity emitted into unit solid angle – in which case there would be no factor of 4π , and the units of L_ν would be $\text{W Hz}^{-1} \text{sr}^{-1}$.

The flux density received by a given observer can be expressed by definition as the product of the **specific intensity** I_ν (the flux density received from unit solid angle of the sky) and the solid angle subtended by the source: $S_\nu = I_\nu d\Omega$. Combining the angular size and flux-density relations thus gives the relativistic version of surface-brightness conservation. This is independent of cosmology:

$$I_\nu(\nu_0) = \frac{B_\nu([1+z]\nu_0)}{(1+z)^3}, \quad (45)$$

where B_ν is **surface brightness** (luminosity emitted into unit solid angle per unit area of source). We can integrate over ν_0 to obtain the corresponding total or **bolometric** formulae, which are needed *e.g.* for spectral-line emission:

$$S_{\text{tot}} = \frac{L_{\text{tot}}}{4\pi R_0^2 S_k^2(r)(1+z)^2}; \quad (46)$$

$$I_{\text{tot}} = \frac{B_{\text{tot}}}{(1+z)^4}. \quad (47)$$

The form of the above relations lead to the following definitions for particular kinds of distances:

<p>angular – diameter distance : $D_A = (1+z)^{-1} R_0 S_k(r)$</p> <p>luminosity distance : $D_L = (1+z) R_0 S_k(r)$.</p>	(48)
---	------

The last element needed for the analysis of observations is a relation between redshift and age for the object being studied. This brings in our earlier relation between time and comoving radius (consider a null geodesic traversed by a photon that arrives at the present):

$$c dt = R_0 dr / (1+z). \quad (49)$$

Distance–redshift relation The general relation between comoving distance and redshift was given earlier as

$$\begin{aligned} R_0 dr &= \frac{c}{H(z)} dz \\ &= \frac{c}{H_0} \left[(1-\Omega)(1+z)^2 + \Omega_v + \Omega_m(1+z)^3 + \Omega_r(1+z)^4 \right]^{-1/2} dz. \end{aligned} \quad (50)$$

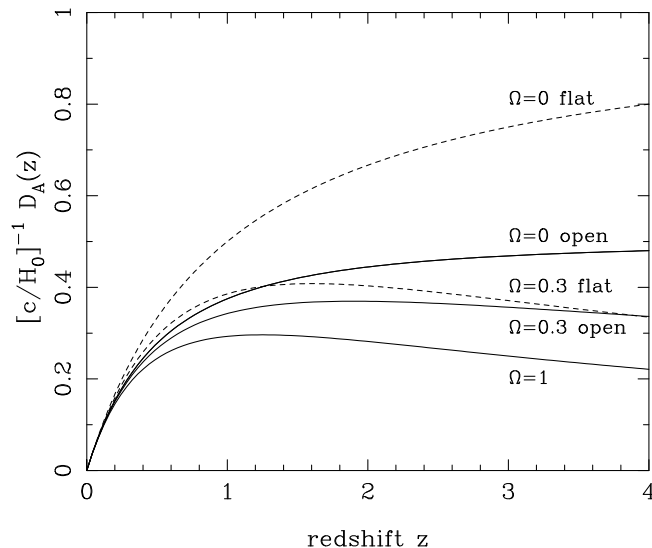


Fig. 2: A plot of dimensionless angular-diameter distance versus redshift for various cosmologies. Solid lines show models with zero vacuum energy; dashed lines show flat models with $\Omega_m + \Omega_v = 1$. In both cases, results for $\Omega_m = 1, 0.3, 0$ are shown; higher density results in lower distance at high z , due to gravitational focusing of light rays.

For a matter-dominated Friedmann model, this means that the distance of an object from which we receive photons today is

$$R_0 r = \frac{c}{H_0} \int_0^z \frac{dz'}{(1+z')\sqrt{1+\Omega z'}}. \quad (51)$$

Integrals of this form often arise when manipulating Friedmann models; they can usually be tackled by the substitution $u^2 = k(\Omega - 1)/[\Omega(1 + z)]$. This substitution produces **Mattig's formula** (1958), which is one of the single most useful equations in cosmology as far as observers are concerned:

$$R_0 S_k(r) = \frac{2c}{H_0} \frac{\Omega z + (\Omega - 2)[\sqrt{1 + \Omega z} - 1]}{\Omega^2(1 + z)}. \quad (52)$$

2 THE HOT BIG BANG

Adiabatic expansion What was the state of matter in the early phases of the big bang? Since the present-day expansion will cause the density to decline in the future, conditions in the past must have corresponded to high density – and thus to high temperature. We can deal with this quantitatively by looking at the thermodynamics of the fluids that make up a uniform cosmological model.

The expansion is clearly **adiathermal**, since the symmetry means that there can be no net heat flow through any surface. If the expansion is also reversible, then we can go one step further, because entropy change is defined in terms of the heat that flows during a reversible change. If no heat flows during a reversible change, then entropy must be conserved, and the expansion will be **adiabatic**. This can only be an approximation, since there will exist irreversible microscopic

processes. In practice, however, it will be shown below that the effects of these processes are overwhelmed by the entropy of thermal background radiation in the universe. It will therefore be an excellent approximation to treat the universe as if the matter content were a simple dissipationless fluid undergoing a reversible expansion. This means that, for a ratio of specific heats Γ , we get the usual adiabatic behaviour

$$T \propto R^{-3(\Gamma-1)}. \quad (53)$$

For radiation, $\Gamma = 4/3$ and we get just $T \propto 1/R$. A simple model for the energy content of the universe is to distinguish pressureless ‘dust-like’ matter (in the sense that $p \ll \rho c^2$) from relativistic ‘radiation-like’ matter (photons plus neutrinos). If these are assumed not to interact, then the energy densities scale as

$$\begin{aligned} \rho_m &\propto R^{-3} \\ \rho_r &\propto R^{-4} \end{aligned} \quad (54)$$

The universe must therefore have been **radiation dominated** at some time in the past, where the densities of matter and radiation cross over. To anticipate, we know that the current radiation density corresponds to thermal radiation with $T \simeq 2.73\text{K}$. We shall shortly show that one expects to find, in addition to this **cosmic microwave background** CMB (CMB), a background in neutrinos that has an energy density 0.68 times that from the photons (if the neutrinos are massless and therefore relativistic). If there are no other contributions to the energy density from relativistic particles, then the total effective radiation density is $\Omega_r h^2 \simeq 4.2 \times 10^{-5}$ and the redshift of **matter–radiation equality** is

$$1 + z_{\text{eq}} = 23\,900 \Omega h^2 (T/2.73\text{K})^{-4}. \quad (55)$$

The time of this change in the global equation of state is one of the key epochs in determining the appearance of the present-day universe.

Quantum gravity limit In principle, $T \rightarrow \infty$ as $R \rightarrow 0$, but there comes a point at which this extrapolation of classical physics breaks down. This is where the thermal energy of typical particles is such that their de Broglie wavelength is smaller than their Schwarzschild radius: quantum black holes clearly cause difficulties with the usual concept of background spacetime. Equating $2\pi\hbar/(mc)$ to $2Gm/c^2$ yields a characteristic mass for quantum gravity known as the **Planck mass**. This mass, and the corresponding length $\hbar/(m_p c)$ and time ℓ_p/c form the system of **Planck units**:

$$\begin{aligned} m_p &\equiv \sqrt{\frac{\hbar c}{G}} \simeq 10^{19} \text{GeV} \\ \ell_p &\equiv \sqrt{\frac{\hbar G}{c^3}} \simeq 10^{-35} \text{m} \\ t_p &\equiv \sqrt{\frac{\hbar G}{c^5}} \simeq 10^{-43} \text{s}. \end{aligned} \quad (56)$$

The Planck time therefore sets the origin of time for the classical phase of the big bang.

Collisionless equilibrium backgrounds We need the thermodynamics of a possibly relativistic perfect gas. We consider some box of volume $V = L^3$, and say that we will analyse the quantum

mechanics of particles in the box by taking the system to be periodic on scale L . Quantum fields in the box are expanded in plane waves, with allowed wavenumbers $k_x = n 2\pi/L$ etc.; these **harmonic boundary conditions** for the allowed eigenstates in the box lead to the density of states in k space:

$$dN = g \frac{V}{(2\pi)^3} d^3k \quad (57)$$

(where g is a degeneracy factor for spin etc.). This expression is nice because it is **extensive** ($N \propto V$) and hence the number density n is independent of V . The equilibrium **occupation number** for a quantum state of energy ϵ is given generally by

$$\langle f \rangle = \left[e^{(\epsilon - \mu)/kT} \pm 1 \right]^{-1} \quad (58)$$

(+ for fermions, – for bosons). Now, for a thermal radiation background, the **chemical potential**, μ is always zero. The reason for this is quite simple: μ appears in the first law of thermodynamics as the change in energy associated with a change in particle number, $dE = TdS - PdV + \mu dN$. So, as N adjusts to its equilibrium value, we expect that the system will be stationary with respect to small changes in N . More formally, the Helmholtz free energy $F = E - TS$ is minimized in equilibrium for a system at constant temperature and volume. Since $dF = -SdT - PdV + \mu dN$, $dF/dN = 0 \Rightarrow \mu = 0$. Thus, in terms of momentum space, the thermal equilibrium **background number density** of particles is

$$n = g \frac{1}{(2\pi\hbar)^3} \int_0^\infty \frac{4\pi p^2 dp}{e^{\epsilon(p)/kT} \pm 1}, \quad (59)$$

where $\epsilon = \sqrt{m^2 c^4 + p^2 c^2}$ and g is the degeneracy factor. There are two interesting limits of this expression.

- (1) Ultrarelativistic limit. For $kT \gg mc^2$ the particles behave as if they were massless, and we get

$$n = \left(\frac{kT}{c} \right)^3 \frac{4\pi g}{(2\pi\hbar)^3} \int_0^\infty \frac{y^2 dy}{e^y \pm 1}. \quad (60)$$

- (2) Non-relativistic limit. Here we can neglect the ± 1 in the occupation number, in which case

$$n = e^{-mc^2/kT} (2mkT)^{3/2} \frac{4\pi g}{(2\pi\hbar)^3} \int_0^\infty e^{-y^2} y^2 dy. \quad (61)$$

This shows us that the background ‘switches on’ at about $kT \sim mc^2$; at this energy, photons and other species in equilibrium will have sufficient energy to create particle-antiparticle pairs, which is how such an equilibrium background would be created. The point at which $kT \sim mc^2$ for some particle is known as a **threshold**.

Similar reasoning gives the energy density of the background, since it is only necessary to multiply the integrand by a factor $\epsilon(p)$ for the energy in each mode:

$$u = \rho c^2 = g \frac{1}{(2\pi\hbar)^3} \int_0^\infty \frac{4\pi p^2 dp}{e^{\epsilon(p)/kT} \pm 1} \epsilon(p). \quad (62)$$

In the same way, we can get the pressure from kinetic theory: $P = n\langle pv \rangle/3 = n\langle p^2 c^2/\epsilon \rangle/3$, where v is the particle velocity, and is related to its momentum and energy by $p = (\epsilon/c^2)v$. The pressure is therefore given by the following integral:

$$P = g \frac{1}{(2\pi\hbar)^3} \int_0^\infty \frac{4\pi p^2 dp}{e^{\epsilon/kT} \pm 1} \frac{p^2 c^2}{3\epsilon}. \quad (63)$$

Clearly, in the ultrarelativistic limit with $\epsilon \simeq pc$, the pressure obeys $P = \rho c^2/3$. In the nonrelativistic limit, the pressure is just $P = nkT$ (see below), whereas the density is dominated by the rest mass: $\rho = mn$, and therefore $P \ll \rho c^2/3$. At a threshold, the equation of state thus departs slightly from $P = \rho c^2/3$, even if the universe is radiation dominated on either side of the critical temperature.

Entropy of the background One quantity that is of considerable importance is the **entropy** of a thermal background. This may be derived in several ways. The most direct is to note that both energy and entropy are extensive quantities for a thermal background. Thus, writing the first law for $\mu = 0$ and using $\partial S/\partial V = S/V$ etc. for extensive quantities,

$$dE = TdS - PdV \quad \Rightarrow \quad \left(\frac{E}{V} dV + \frac{\partial E}{\partial T} dT \right) = \left(T \frac{S}{V} dV + T \frac{\partial S}{\partial T} dT \right) - PdV. \quad (64)$$

Equating the dV and dS parts gives the familiar $\partial E/\partial T = T \partial S/\partial T$ and

$$S = \frac{E + PV}{T} \quad (65)$$

Using the above integral for the pressure, the entropy is

$$S = \frac{4\pi gV}{(2\pi\hbar)^3} \int_0^\infty \frac{p^2 dp}{e^{\epsilon/kT} \pm 1} \left(\frac{\epsilon}{T} + \frac{p^2 c^2}{3\epsilon T} \right), \quad (66)$$

which becomes $S = 3.602Nk$ (bosons) or $4.202Nk$ (fermions) in the ultrarelativistic limit and $S = (mc^2/kT)Nk$ in the non-relativistic limit. So, for radiation, the entropy is just proportional to the number of particles. For this reason, the ratio of photon to baryon number densities n_γ/n_B is sometimes called the **entropy per baryon**.

Formulae for ultrarelativistic backgrounds We now summarize the most useful results from this discussion, which are the thermodynamic quantities for massless particles. These formulae are required time and time again in calculations of conditions in the early universe. Consider first bosons, such as the microwave background. Evaluating the dimensionless integrals encountered earlier (only possible numerically in the case of n) gives energy, number and entropy densities:

$$\begin{aligned} u &= g \frac{\pi^2}{30} kT \left(\frac{kT}{\hbar c} \right)^3 = 3P \\ \frac{s}{k} &= g \frac{2\pi^2}{45} \left(\frac{kT}{\hbar c} \right)^3 = 3.602n \end{aligned} \quad (67)$$

(remember that $g = 2$ for photons).

It is also expected that there will be a fermionic relic background of neutrinos left over from the big bang. Assume for now that the neutrinos are massless. In this case, the thermodynamic properties can be obtained from those of black-body radiation by the following trick:

$$\frac{1}{e^x + 1} = \frac{1}{e^x - 1} - \frac{2}{e^{2x} - 1}. \quad (68)$$

Thus, a gas of fermions looks like a mixture of bosons at two different temperatures. Knowing that boson number density and energy density scale as $n \propto T^3$ and $u \propto T^4$, we then get the corresponding fermionic results. The entropy requires just a little more care. Although we have said that entropy density is proportional to number density, in fact the entropy density for an ultrarelativistic gas was shown above to be $s = (4/3)u/T$, and so the fermionic factor is the same as for energy density:

$$\boxed{\begin{aligned} n_{\text{F}} &= \frac{3}{4} \frac{g_{\text{F}}}{g_{\text{B}}} n_{\text{B}} \\ u_{\text{F}} &= \frac{7}{8} \frac{g_{\text{F}}}{g_{\text{B}}} u_{\text{B}}. \\ s_{\text{F}} &= \frac{7}{8} \frac{g_{\text{F}}}{g_{\text{B}}} s_{\text{B}}. \end{aligned}} \quad (69)$$

Using these rules, it is usually possible to forget about the precise nature of the relativistic background in the universe and count bosonic degrees of freedom, given the effective degeneracy factor for u or s :

$$g_* \equiv \sum_{\text{bosons}} g_i + \frac{7}{8} \sum_{\text{fermions}} g_j, \quad (70)$$

although this definition needs to be modified if some species have different temperatures (see below).

Neutrino decoupling At the later stages of the big bang, energies are such that only light particles survive in equilibrium: γ , ν and the three leptons e , μ , τ . If neutrinos could be maintained in equilibrium, the lepton-antilepton pairs would annihilate as the temperature fell still further ($T_\tau = 10^{13.3}$ K, $T_\mu = 10^{12.1}$ K, $T_e = 10^{9.7}$ K), and the end result would be that the products of these annihilations would be shared among the only massless particles. However, in practice the weak reactions that maintain the neutrinos in thermal equilibrium ‘switch off’ at $T \simeq 10^{10}$ K. This **decoupling** is discussed in more detail below, but is a general cosmological phenomenon, which arises whenever the interaction timescales exceed the local Hubble time, leaving behind abundances of particles frozen at the values they had when last in thermal equilibrium. Two-body reaction rates scale proportional to density, times a cross-section that is often a declining function of energy, so that the interaction time changes at least as fast as R^{-3} . In contrast, the Hubble time changes no faster than R^{-2} (in the radiation era), so that there is inevitably a crossover. For neutrinos, this point occurs at a redshift of $\sim 10^{10}$, whereas the photons of the microwave background typically last interacted with matter at $z \simeq 1000$.

The effect of the electron-positron annihilation is therefore to enhance the numbers of photons relative to neutrinos. It is easy to see what quantitative effect this has: although we may talk loosely about the energy of e^\pm annihilation going into photons, what is actually conserved is the *entropy*. The entropy of an $e^\pm + \gamma$ gas is easily found by remembering that it is

proportional to the number density, and that all three particle species have $g = 2$ (polarization or spin). The total is then

$$s(\gamma + e^+ + e^-) = \frac{11}{4}s(\gamma). \quad (71)$$

Equating this to photon entropy at a new temperature gives the factor by which the photon temperature is enhanced with respect to that of the neutrinos. Equivalently, given the observed photon temperature today, we infer the existence of a neutrino background with a temperature

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma = 1.95 \text{ K}, \quad (72)$$

for $T_\gamma = 2.73 \text{ K}$. Although it is hard to see how such low energy neutrinos could ever be detected directly, their gravitation is certainly not negligible: they contribute an energy density that is a factor $(7/8) \times (4/11)^{4/3}$ times that of the photons (the fact that neutrinos have $g = 1$ whereas photons have $g = 2$ is cancelled by the fact that neutrinos and antineutrinos are distinguishable particles). For three neutrino species, this enhances the energy density in relativistic particles by a factor 1.68.

Massive neutrinos Although for many years the conventional wisdom was that neutrinos were massless, this assumption began to be increasingly challenged around the end of the 1970s. Theoretical progress in understanding the origin of masses in particle physics meant that it was no longer natural for the neutrino to be completely devoid of mass. Also, experimental evidence (Reines *et al.* 1980), which in fact turned out to be erroneous, seemed to imply a non-zero mass of $m \sim 10 \text{ eV}$ for the electron neutrino. The consequences of this for cosmology could be quite profound, as relic neutrinos are expected to be very abundant. The above section showed that $n(\nu + \bar{\nu}) = (3/4)n(\gamma; T = 1.95 \text{ K})$. That yields a total of 113 relic neutrinos in every cm^3 for each species.

The consequences of giving these particles a mass are easily worked out provided the mass is small enough. If this is the case, then the neutrinos were ultrarelativistic at decoupling and their statistics were those of massless particles. As the universe expands to $kT < m_\nu c^2$, the total number of neutrinos is preserved. Furthermore, their momentum redshifts as $p \propto 1/R$, so that the momentum-space distribution today will just be a redshifted version of the ultrarelativistic form. As discussed more fully below, the momentum-space distribution stays exactly what would have been expected for thermal-equilibrium neutrinos, even though they have long since decoupled. However, this illusion is broken once the temperature falls below $kT < m_\nu c^2$, because the effect of the rest-mass energy on the equilibrium occupation number causes the nonrelativistic momentum distribution to differ from the relativistic one. We therefore obtain the present-day mass density in neutrinos just by multiplying the zero-mass number density by m_ν , and the consequences for the cosmological density are easily worked out to be

$$\Omega h^2 = \frac{\sum m_i}{93.5 \text{ eV}}. \quad (73)$$

For a low Hubble parameter $h \simeq 0.5$, an average mass of only 8 eV will suffice to close the universe. In contrast, the current laboratory limits to the neutrino masses are

$$\begin{aligned} \nu_e &\lesssim 15 \text{ eV} \\ \nu_\mu &\lesssim 0.17 \text{ MeV} \\ \nu_\tau &\lesssim 24 \text{ MeV}. \end{aligned} \quad (74)$$

3 RELICS OF THE BIG BANG

The massive neutrino is the simplest example of a relic of the big bang: a particle that once existed in equilibrium, but which has decoupled and thus preserves a ‘snapshot’ of the properties of the universe at the time the particle was last in thermal equilibrium. The aim of this section is to give a little more detail on the processes that determine the final abundance of these relics.

3.1 Freeze-out

So far, we have used a simple argument, which decrees that the relic abundance becomes fixed when expansion and interaction timescales are equal. To do better than this, it is necessary to look at the differential equation that governs the abundance of particle species in the expanding universe. This is the **Boltzmann equation**, which considers the **phase-space density**: the joint probability density for finding a particle in a given volume element and in a given range of momentum, denoted by $f(\mathbf{x}, \mathbf{p})$. The general form of this equation is

$$\boxed{\frac{\partial f}{\partial t} + (\dot{\mathbf{x}} \cdot \nabla_{\mathbf{x}}) f + (\dot{\mathbf{p}} \cdot \nabla_{\mathbf{p}}) f = \dot{f}_c.} \quad (75)$$

The lhs is just the fluid-dynamical convective derivative of the phase-space density, generalized to 6D space. The rhs is the collisional term, and the equation therefore just says that groups of particles maintain their phase-space density as they stream through phase space, unless modified by collisions (**Liouville’s theorem**). The truth of this theorem is easily seen informally in one spatial dimension: a small square element $dx dv_x$ becomes sheared to a parallelogram of unchanged area, and so the phase-space density is unaltered.

The Boltzmann equation has been written with respect to a fixed system of laboratory coordinates, but it is quite easily adapted to the expanding universe. We should now interpret the particle velocities as being relative to a set of uniformly expanding observers. A particle that sets off from $r = 0$ with some velocity will effectively slow down as it tries to overtake distant receding observers. After time t , the particle will have travelled $x = vt$, and so encountered an observer with velocity $dv = Hx$. According to this observer, the particle’s momentum is now reduced by $dp = m dv = mHvt = Hpt$. There is therefore the appearance of a **Hubble drag** force:

$$\frac{\dot{p}}{p} = -H. \quad (76)$$

In the presence of density fluctuations, this needs to be supplemented by gravitational forces, which as usual manifest themselves through the affine connection. In a sense, most of cosmological theory comes down to solving the Boltzmann equation for photons plus neutrinos plus collisionless dark matter, coupled to the matter fluid via gravity in all cases and also by Thomson scattering in the case of photons. Since the interesting processes are operating at early times when the density fluctuations are small, this is an exercise in first-order relativistic perturbation theory. The technical difficulties in detail mean this will have to be omitted here (see Peebles 1980; Efstathiou 1990); when discussing perturbations, the main results can usually be understood in terms of a fluid approximation, and this approach is pursued below.

Things are much easier in the case of homogeneous backgrounds, where spatial derivatives can be neglected. The Boltzmann equation then has the simple form

$$\frac{\partial f}{\partial t} - Hp \frac{\partial f}{\partial p} = \dot{f}_c. \quad (77)$$

The collision term is usually dominated by particle–antiparticle annihilations (assuming for the moment that the numbers of each are identical, so that there is no significant asymmetry):

$$\dot{f}_c = - \int \langle \sigma v \rangle f \bar{f} d^3\bar{p}, \quad (78)$$

where $\langle \sigma v \rangle$ is the velocity-averaged product of the cross-section and the velocity. We can take $\langle \sigma v \rangle$ outside the integral even if it is not constant provided it is evaluated at some suitable average energy. Integrating over momentum then gives the moment equation for the number density,

$$\dot{n} + 3Hn = -\langle \sigma v \rangle n^2 + S, \quad (79)$$

where S is a source term added to represent the production of particles from thermal processes – effectively pair creation. This term is fixed by a thermodynamic equilibrium argument: for a non-expanding universe, n will be constant at the equilibrium value for that temperature, n_T , showing that

$$S = \langle \sigma v \rangle n_T^2. \quad (80)$$

If we define comoving number densities $N \equiv a^3 n$, the rate equation can be rewritten in the simple form

$$\boxed{\frac{d \ln N}{d \ln a} = -\frac{\Gamma}{H} \left[1 - \left(\frac{N_T}{N} \right)^2 \right]}, \quad (81)$$

where $\Gamma = n \langle \sigma v \rangle$ is the interaction rate experienced by the particles.

Unfortunately, this equation must be solved numerically. The main features are easy enough to see, however. Suppose first that the universe is sustaining a population in approximate thermal equilibrium, $N \simeq N_T$. If the population under study is relativistic, N_T does not change with time, because $n_T \propto T^3$ and $T \propto a^{-1}$. This means that it is possible to keep $N = N_T$ exactly, whatever Γ/H . It would however be grossly incorrect to conclude from this that the population stays in thermal equilibrium: if $\Gamma/H \ll 1$, a typical particle suffers no interactions even while the universe doubles in size, halving the temperature. A good example is the microwave background, whose photons last interacted with matter at $z \simeq 1000$. The CMB nevertheless still appears to be equilibrium black-body radiation because the number density of photons has fallen by the right amount to compensate for the redshifting of photon energy. This sounds like an incredible coincidence, but is in fact quite inevitable when looked at from the quantum-mechanical point of view. This says that the occupation number of a given mode, $= (\exp \hbar\omega/kT - 1)^{-1}$ for thermal radiation, is an adiabatic invariant that does not change as the universe expands – only the frequency alters, and thus the apparent temperature.

Now consider the opposite case, where the thermal solution would be nonrelativistic, with

$$N_T \propto T^{-3/2} \exp(-mc^2/kT). \quad (82)$$

If the background is to stay at the equilibrium value, the lhs of the rate equation must therefore be $\gg -1$. This is consistent if $\Gamma/H \gg 1$, because then the $(N_T/N)^2$ term on the rhs can still be close to unity. However, if $\Gamma/H \ll 1$, there must be a deviation from equilibrium. When N_T changes sufficiently fast with a , the actual abundance cannot keep up, so that the $(N_T/N)^2$ term on the rhs becomes negligible and $d \ln N/d \ln a \simeq -\Gamma/H$, which is $\ll 1$. There is therefore a critical time at which the reaction rate drops low enough that particles are simply

conserved as the universe expands – the population has **frozen out**. This provides a more detailed justification for the intuitive rule-of-thumb used above to define decoupling,

$$N(a \rightarrow \infty) = N(\Gamma/H = 1). \quad (83)$$

Exact numerical solutions of the rate equation almost always turn out very close to this simple rule (see chapter 5 of Kolb & Turner 1990).

3.2 Recombination and last scattering

One of the critical epochs in the evolution of the universe is reached when the temperature drops to the point ($T \sim 1000$ K) where it is thermodynamically favourable for the ionized plasma to form neutral atoms. This process is known as **recombination**: a complete misnomer, as the plasma has always been completely ionized up to this time.

There is a problem: highly excited atoms can be produced by a series of small transitions, but to reach the ground state requires the production of photons at least as energetic as the $2P \rightarrow 1S$ spacing (Lyman α , with $\lambda = 1216\text{\AA}$). Multiple absorption of these photons will cause reionization once they become abundant, so it would now appear that recombination can never occur at all (unlike a finite HII region, where the Ly α photons can escape; see *e.g.* Osterbrock 1974). There is a way out, however, using **two-photon emission**. The $2S \rightarrow 1S$ transition is strictly forbidden at first order and one can only conserve energy and angular momentum in the transition by emitting a *pair* of photons. This gives the mechanism we need for transferring the ionization energy into photons with $\lambda > \lambda_{\text{Ly}\alpha}$.

A highly stripped-down analysis of events simplifies the hydrogen atom to just two levels ($1S$ and $2S$). Any chain of recombinations that reaches the ground state can be ignored through the above argument: these reactions produce photons that are immediately re-absorbed elsewhere, so they have no effect on the ionization balance. The main chance of reaching the ground state comes through the recombinations that reach the $2S$ state, since some fraction of the atoms that reach that state will suffer two-photon decay before being re-excited. The rate equation for the fractional ionization is thus

$$\frac{d(nx)}{dt} = -R(nx)^2 \frac{\Lambda_{2\gamma}}{\Lambda_{2\gamma} + \Lambda_U(T)}, \quad (84)$$

where n is the number density of protons, x is the fractional ionization, R is the recombination coefficient ($R \simeq 3 \times 10^{-17} T^{-1/2} \text{m}^3 \text{s}^{-1}$), $\Lambda_{2\gamma}$ is the two-photon decay rate, and $\Lambda_U(T)$ is the stimulated transition rate upwards from the $2S$ state. This equation just says that recombinations are a two-body process, which create excited states that cascade down to the $2S$ level, from whence a competition between the upward and downward transition rates determines the fraction that make the downward transition. A fuller discussion (see chapter 6 of Peebles 1993) would include a number of other processes: depopulation of the ground state by inverse 2-photon absorption; redshifting of Ly alpha photons due to the universal expansion, which can prevent them being re-absorbed. However, at the redshifts of practical interest (1000 to 10), the simplified equation captures the main effect.

An important point about the rate equation is that it is only necessary to solve it once, and the results can then be scaled immediately to some other cosmological model. Consider the rhs: both R and $\Lambda_U(T)$ are functions of temperature, and thus of redshift only, so that any parameter dependence is carried just by n^2 , which scales $\propto (\Omega_B h^2)^2$, where Ω_B is the baryonic density parameter. Similarly, the lhs depends on $\Omega_B h^2$ through n ; the other parameter dependence comes if we convert time derivatives to derivatives with respect to redshift:

$$\frac{dt}{dz} \simeq -3.09 \times 10^{17} (\Omega h^2)^{-1/2} z^{-5/2} \text{ s}, \quad (85)$$

for a matter-dominated model at large redshift (Ω is the total density parameter). Putting these together, the fractional ionization must scale as

$$x(z) \propto \frac{(\Omega h^2)^{1/2}}{\Omega_B h^2}. \quad (86)$$

The last-scattering shell Putting in all the relevant processes, Jones & Wyse (1985) found the fractional ionization x near $z = 1000$ to be well approximated by

$$x(z) = 2.4 \times 10^{-3} \frac{(\Omega h^2)^{1/2}}{\Omega_B h^2} \left(\frac{z}{1000} \right)^{12.75}. \quad (87)$$

The scaling with Ω and h has a marvelous consequence. If we work out the optical depth to Thomson scattering, $\tau = \int n_e x \sigma_T dr_{\text{prop}}$, we find just

$$\tau(z) = 0.37 \left(\frac{z}{1000} \right)^{14.25}, \quad (88)$$

independent of cosmological parameters. The rate equation causes $x(z)$ to scale in just the right way that the optical depth is a completely robust quantity. Because τ changes rapidly with redshift, the distribution function for the redshift at which photons were last scattered, $e^{-\tau} d\tau/dz$, is sharply peaked, and is well fitted by a Gaussian of mean redshift 1065 and standard deviation in redshift 80. Thus, when we look at the sky, we can expect to see in all directions photons that originate from a **last-scattering surface** at $z \simeq 1065$. This independence of parameters is not quite exact in detail, however, and very accurate work needs to solve the evolution equations exactly (*e.g.* appendix C of Hu & Sugiyama 1995).

The microwave background In a famous piece of serendipity, the redshifted radiation from the last-scattering photosphere was detected as a 2.73 K microwave background by Penzias & Wilson (1965). Since the initial detection of the microwave background at $\lambda = 7.3$ cm, measurements of the spectrum have been made over an enormous range of wavelengths, from the depths of the Rayleigh–Jeans regime at 74 cm to well into the Wien tail at 0.5 mm. The most accurate measurements come from **COBE** – the NASA cosmic background explorer satellite. Early data showed the spectrum to be very close to a pure Planck function (Mather *et al.* 1990), and the final result verifies the lack of any distortion with breathtaking precision. The COBE temperature measurement and 95% confidence range of

$$T = 2.728 \pm 0.004 \text{ K} \quad (89)$$

improves significantly on the ground-based experiments. The lack of distortion in the shape of the spectrum is astonishing, and limits the chemical potential to $|\mu| < 9 \times 10^{-5}$ (Fixsen *et al.* 1996). These results also allow the limit $y \lesssim 1.5 \times 10^{-5}$ to be set on the Compton-scattering distortion parameter. These limits are so stringent that many competing cosmological models can be eliminated.

3.3 Primordial nucleosynthesis

At sufficiently early times, the temperature of the universe reaches the point where nuclear reactions can occur ($T \sim 10^9 \text{K}$). The abundance of light elements that results from these early reactions is fixed by an argument that can be outlined quite simply. In equilibrium, the numbers of neutrons and protons should vary as

$$\frac{N_n}{N_p} = e^{-\Delta mc^2/kT} \simeq e^{-1.5(10^{10} \text{K}/T)}. \quad (90)$$

The reason that neutrons exist today is that the timescale for the weak interactions needed to keep this equilibrium set up eventually becomes longer than the expansion timescale. The reactions thus rapidly cease, and the neutron–proton ratio undergoes **freeze-out** at some characteristic value. In practice this occurs at $N_n/N_p \simeq 1/6$. If most of the neutrons ended up in ${}^4\text{He}$, we would then expect 25% He by mass – which is very nearly what we see. One of the critical calculations in cosmology is therefore to calculate this freeze-out process in detail. As the following outline of the analysis will show, the result is due to a complex interplay of processes and the fact that there is a significant primordial abundance of anything other than hydrogen is a consequence of a number of coincidences.

Neutron freeze-out The only nuclear reactions that matter initially are the weak interactions that convert between protons and neutrons:



The main systematics of the relic abundances of the light elements can be understood by looking at the neutron to proton ratio and how it evolves.

The number density of neutrons obeys the kinetic equation

$$\frac{d n_n}{dt} = (\Lambda_{pe} + \Lambda_{p\nu}) n_p - (\Lambda_{ne} + \Lambda_{n\nu}) n_n, \quad (92)$$

where the rate coefficients Λ_i refer to the four possible processes given above. To these two-body processes should also be added the spontaneous decay of the neutron, which has the e -folding lifetime of

$$\boxed{\tau_n = 887 \pm 2 \text{ s}} \quad (93)$$

(according to the Particle Data Group). We neglect this for now, as it turns out that neutron freeze-out happens at slightly earlier times.

The quantum-field calculation of the rate coefficients is not difficult, because of the simple form of the Fermi Lagrangian for the weak interaction. Recall that this is proportional to the Fermi constant (G_F) times the fields of the various particles that participate, and that each field is to be thought of as a sum of creation and annihilation operators. This means that the matrix elements for all processes where various particles change their occupation numbers by one are the same, and the rate coefficients differ only by virtue of integration over states for the particles involved.

For example, the rate for neutron decay is

$$\tau_n^{-1} \propto G_F^2 \left(\frac{1}{[2\pi\hbar]^3} \right)^2 2 \int d^3 p_e d^3 p_\nu \delta(\epsilon_e + \epsilon_\nu - Q), \quad (94)$$

where Q is the neutron–proton energy difference. This expression can be more or less written down at sight. The delta function expresses conservation of energy and as usual arises automatically from integration over space, rather than having to be put in by hand. The factor of 2 before the integral allows for electron helicity, but there is no need to be concerned with the overall constant of proportionality. By performing the integral, this expression can be reduced to

$$\tau_n^{-1} \propto \frac{G_F^2}{2\pi^4} 1.636 m_e^5 \quad (\text{natural units}). \quad (95)$$

The reason why the overall constant of proportionality is not needed is that all other related weak processes have rates of the same form. The only difference is that, unlike the free decay in a vacuum just analysed, we need to include the probabilities that the initial and final states are occupied. For example, in $n + \nu \rightarrow p + e$, we need a factor n_ν for the initial neutrino state and $1 - n_e$ for the final electron state (effectively to allow for the fermionic equivalent of stimulated plus spontaneous emission; if the particles involved were bosons, this would become $1 + n$). The rate for this process is therefore

$$\Lambda_{n\nu} = [1.636 m_e^5 \tau_n]^{-1} \int_0^\infty \frac{p_e \epsilon_e p_\nu^2 dp_\nu}{[1 + \exp(p_\nu/kT)][1 + \exp(-\epsilon_e/kT)]}, \quad (96)$$

where $\epsilon_e = p_\nu + Q$. Very similar integrals can be written down for the other processes involved.

Since $Q \sim m_e c^2$, it is clear that at high temperatures $kT \gg m_e c^2$, all the rate coefficients will be of the same form; both p_e and ϵ_e can be replaced by p_ν in top and bottom of the integral, leaving a single rate coefficient to be determined by numerical integration:

$$\Lambda = 13.893 \tau_n^{-1} \left(\frac{kT}{m_e c^2} \right)^5 = \left(\frac{10^{10.135} \text{ K}}{T} \right)^{-5} \text{ s}^{-1}. \quad (97)$$

Since the number density of the thermal background of neutrinos and electrons is proportional to T^3 , this says that the effective cross-section for these weak interactions scales proportional to [energy]². The radiation-dominated era has

$$t = \sqrt{\frac{3}{32\pi G\rho}} = \left(\frac{10^{10.125} \text{ K}}{T} \right)^2 \text{ s}, \quad (98)$$

allowing for three massless neutrinos. The T^{-2} dependence of the expansion timescale is much slower than the interaction timescale, which changes as T^{-5} , so there is a quite sudden transition between thermal equilibrium and freeze-out, suggesting that weak interactions switch off, freezing the neutron abundance at a temperature of $T \simeq 10^{10.142} \text{ K}$. This implies an equilibrium neutron–proton ratio of

$$\frac{N_n}{N_p} = e^{-Q/kT} = e^{-10^{10.176} \text{ K}/T} \simeq 0.34. \quad (99)$$

This obviously cannot be a precisely correct result, because the freeze-out condition was calculated assuming a temperature well above the electron mass threshold, whereas it appears that freeze-out actually occurs at about this critical temperature. The rate Λ is in fact a little larger at the threshold than the high-temperature extrapolation would suggest, so the neutron abundance is in practice lower.

Neutrino freeze-out There is a yet more serious complication, because another important process that occurs around the same time is neutrino decoupling. The weak reaction that keeps neutrino numbers in equilibrium at this late time is $\nu + \bar{\nu} \leftrightarrow e^+ + e^-$. The rate for this process can be found in exactly the same way as above. At high energies, the result is identical, save only that the squared matrix element involved is smaller by a factor of about 5 because of the axial coupling of nucleons: the neutrino rate scales as G_F^2 , as opposed to $G_F^2(1 + 2g_A^2)$ for nucleons. This leads to the neutrinos decoupling at a slightly higher temperature:

$$\boxed{T(\nu \text{ decoupling}) \simeq 10^{10.5} \text{ K.}} \quad (100)$$

This is uncomfortably close to the electron mass threshold, but just sufficiently higher that it is not a bad approximation to say that all e^+e^- annihilations go into photons rather than neutrinos. During the time at which nucleons are decoupling, the neutrino and photon temperatures are therefore becoming different, and a detailed calculation must account for this. The resulting freeze-out temperature is very close to 10^{10} K, at which point the neutron-to-proton ratio is about 1:3.

Unfortunately, we are still not finished, because neutrons are not stable. It does not matter what abundance of them freezes out: unless they can be locked away in nuclei before $t = 887$ s, the relic abundance will decay freely to zero. The freeze-out point occurs at an age of a few seconds, so there are only a few e -foldings of expansion available in which to salvage some neutrons. So far, a remarkable sequence of coincidences has been assembled, in that the freeze-out of neutrinos and nucleons happens at about the same time as $e^+ - e^-$ annihilation, which is also a time of order τ_n . It may seem implausible that we can add one more – *i.e.* that nuclear reactions will become important at about the same time – but this is just what does happen.

Construction of nucleons This coincidence is not surprising, since the deuteron binding energy of 2.225 MeV is only 4.3 times larger than $m_e c^2$ and only 1.7 times larger than the neutron–proton mass difference. At higher temperatures, the strong interaction $n + p = \text{D} + \gamma$ is fast enough to produce deuterium, but thermal equilibrium favours a small deuterium fraction – *i.e.* typical photons are energetic enough to disrupt deuterium nuclei very easily. Furthermore, because of the large photon-to-baryon ratio, the photons can keep the deuterium from forming until the temperature has dropped well below the binding energy of the deuteron. The equilibrium abundance of deuterium is set in much the same way as the ionization abundance of hydrogen, and so obeys an equation that is identical (apart from spin-degeneracy factors) to the Saha equation for hydrogen ionization:

$$\frac{n_D}{n_p n_n} = \frac{3}{4} \frac{(2\pi\hbar)^3}{(2\pi kT m_p m_n / m_D)^{3/2}} \exp(\chi/kT), \quad (101)$$

where χ is the binding energy. As with hydrogen ionization, this defines an abrupt transition between the situation where deuterium is rare and where it dominates the equilibrium. The terms outside the exponential keep the deuterium density low until $kT \ll \chi$: the $n_D = n_n$ crossover occurs at

$$\boxed{T_{\text{deuteron}} \simeq 10^{8.9} \text{ K,}} \quad (102)$$

or a time of about 3 minutes. An exact integration of the weak-interaction kinetic equation for

the neutron abundance at $n_D = n_n$ (including free neutron decay, which is significant) gives

$$\boxed{\frac{n_n}{n_p} \simeq 0.163 (\Omega_B h^2)^{0.04} (N_\nu/3)^{0.2}} \quad (103)$$

(see *e.g.* chapter 4 of Kolb & Turner 1990; chapter 6 of Peebles 1993). The dependences on the baryon density and on the number of neutrino species are easily understood. A high baryon density means that the Saha equation gives a higher deuterium abundance, increasing the temperature at which nuclei finally form and giving a higher neutron abundance because fewer of them have decayed. The effect of extra neutrino species is to increase the overall rate of expansion, so that neutron freeze-out happens earlier, again raising the abundance.

The primordial helium abundance The argument so far has produced a universe consisting of just hydrogen and deuterium, but this is not realistic, as one would expect ${}^4\text{He}$ to be preferred on thermodynamic grounds, owing to its greater binding energy per nucleon (7 MeV, as opposed to 1.1 MeV for deuterium). In practice, the production of helium must await the synthesis of significant quantities of deuterium, which we have seen happens at a temperature roughly one-third that at which helium would be expected to dominate. What the thermodynamic argument does show, however, is that it is expected that the deuterium will be rapidly converted to helium once significant nucleosynthesis begins. This argument is what allows us to expect that the helium abundance can be calculated from the final n/p ratio. If all neutrons are incorporated into ${}^4\text{He}$, then the number density of hydrogen is set by the remaining protons: $n_H = n_p - n_n$. The mass fraction of helium, Y , is unity minus the hydrogen fraction, so that

$$\boxed{Y = 1 - \frac{n_p - n_n}{n_p + n_n} = 2 \left(1 + \frac{n_p}{n_n} \right)^{-1}} \quad (104)$$

For the earlier n/p ratio of 0.163, this gives $Y = 0.28$.

The ‘observed’ value of Y is in the region of $Y = 0.22$ to 0.23 (*e.g.* Pagel 1994), and there exists something of a difference of opinion on whether this is marvelously close agreement, or evidence for something seriously wrong with the standard model.

The number of particle generations Increasing the number of neutrino species widens the gap between theory and observation by $\Delta Y \simeq 0.01$ for each additional neutrino species. It is therefore clear that strong limits can be set on the number of unobserved species, and thus on the number of possible additional families in particle physics. For many years, these nucleosynthesis limits were stronger than those that existed from particle physics experiments. This changed in 1990, with a critical series of experiments carried out in **LEP**, which was the first experiment to produce Z^0 particles in large numbers. The particles are not seen directly, but their presence is inferred by detecting a peak in the energy-dependent cross-sections for producing pairs of leptons (l) or hadrons (h). The interpretation is that the peak is a ‘resonance’ due to the production of a Z^0 as an intermediate state, and that the energy of the peak measures the Z^0 mass:

$$e^+ + e^- \rightarrow Z^0 \rightarrow \begin{cases} l, \bar{l} \\ h, \bar{h}. \end{cases} \quad (105)$$

The width of the peak measures the Z^0 lifetime, through the uncertainty principle, and this gives a means of counting the numbers of neutrino species. The Z^0 can decay to pairs of neutrinos so

long as their rest mass sums to less than 91.2 GeV; more species increase the decay rate, and increase the Z^0 width, which measures the total decay rate.

Since 1990, these arguments have required N to be very close to 3 (see the Opal consortium, 1990); it is a matter of detailed argument over the helium data as to whether $N = 4$ was ruled out from cosmology prior to this. In any case, it is worth noting that these two routes do not measure exactly the same thing: both are sensitive only to relativistic particles, with upper mass scales of about 1 MeV and 100 GeV in the cosmological and accelerator cases respectively. If LEP had measured $N = 5$, that would have indicated extra species of rather massive neutrinos. The fact that both limits in fact agree is therefore good evidence for the correctness of the standard model, containing only three families.

Other light-element abundances The same thermodynamic arguments that say that helium should be favoured at temperatures around 0.1 MeV say that more massive nuclei would be preferred in equilibrium at lower temperatures. However, by the time helium synthesis is accomplished, the density and temperature are too low for significant synthesis of heavier nuclei to proceed: the lower density means that reactions tend to freeze out, even for a constant cross-section, and the need for penetration of the nuclear Coulomb barrier means that cross-sections decline rapidly as the temperature decreases.

Apart from helium, the main nuclear residue of the big bang is therefore those deuterium nuclei that escape being mopped up into helium, plus a trace of ^3He , which is produced en route to ^4He : $\text{D} + p \rightarrow ^3\text{He}$, followed by $^3\text{He} + n \rightarrow ^4\text{He}$ (the alternative route, of first $\text{D} + n$, then p also happens, but the intermediate tritium is not so strongly bound). There also exist extremely small fractions of other elements: ^7Li ($\sim 10^{-9}$ by mass) and ^7Be ($\sim 10^{-11}$). Unlike helium, the critical feature of these abundances is that they are rather sensitive to density. One of the major achievements of big bang cosmology is that it can account simultaneously for the abundances of H, ^2D , ^3He , ^4He and ^7Li – but *only for a low-density universe*. A proper understanding of the abundances really requires a numerical solution of the coupled rate equations for all the nuclear reactions, putting in the temperature-dependent cross-sections. This careful piece of numerical physics was first carried out impressively soon after the discovery of the microwave background, by Wagoner, Fowler & Hoyle (1967). At least the sense of the answer can be understood intuitively, however. We have seen that helium formation occurs at very nearly a fixed temperature, depending only weakly on density or neutrino species. The residual deuterium will therefore freeze out at about this temperature, leaving a number density fixed at whatever sets the reaction rate low enough to survive for a Hubble time. Since this density is a fixed quantity, the *proportion* of the baryonic density that survives as deuterium (or ^3He) should thus decline roughly as $1/(\text{density})$.

This provides a relatively sensitive means of weighing the baryonic content of the universe. A key event in the development of cosmology was thus the determination of the D/H ratio in the interstellar medium, carried out by the COPERNICUS UV satellite in the early 1970s (Rogerson & York 1973). This gave $\text{D}/\text{H} \simeq 2 \times 10^{-5}$, providing the first evidence for a low baryonic density, as follows. Figure 3 shows how the abundances of light elements vary with the cosmological density, according to detailed calculations. The baryonic density in these calculations is traditionally quoted in the field as the reciprocal of the entropy per baryon:

$$\eta \equiv (n_p + n_n)/n_\gamma = 2.74 \times 10^{-8} (T/2.73 \text{ K})^{-3} \Omega_{\text{B}} h^2. \quad (106)$$

Figure 3 shows that this deuterium abundance favours a low density, $\Omega_{\text{B}} h^2 \simeq 0.02$, and data on other elements give answers close to this. The constraint obtained from a comparison between

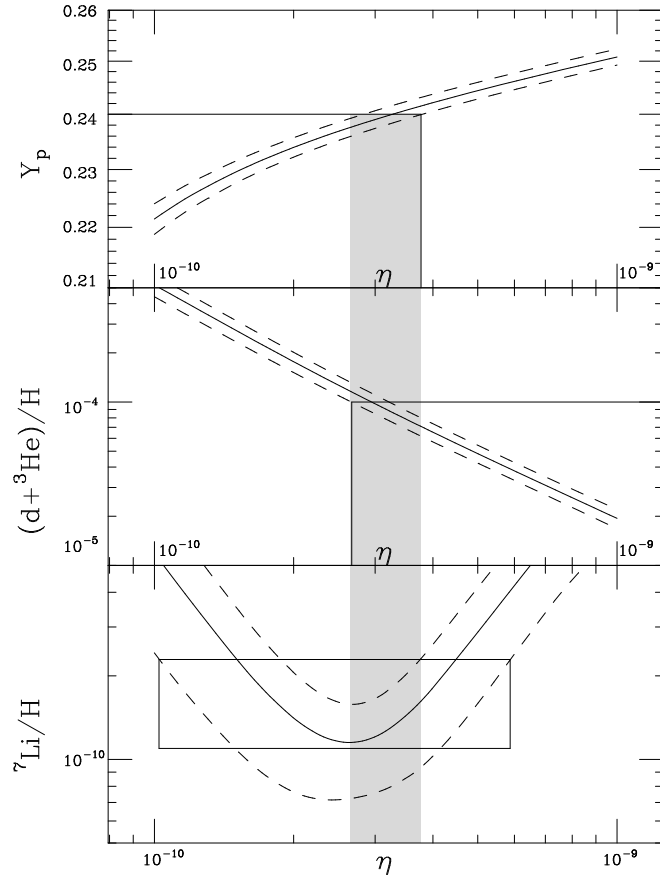


Fig. 3: The predicted primordial abundances of the light elements, as a function of the baryon-to-photon ratio η (Smith, Kawano & Malaney 1993). For a microwave-background temperature of 2.73 K, this is related to the baryonic density parameter via $\Omega_B h^2 = \eta/2.74 \times 10^{-8}$. Concordance with the data is found only for $\eta \simeq 3 \times 10^{-10}$, shown by the shaded strip.

nucleosynthesis predictions and observational data is rather tight:

$$0.010 \lesssim \Omega_B h^2 \lesssim 0.015 \quad (107)$$

(*e.g.* Walker *et al.* 1991; Smith, Kawano & Malaney 1993). This comparison is of course non-trivial, because we can only observe abundances in present-day stars and gas, rather than in primordial material. Nevertheless, making the best allowances possible for production or destruction of light elements in the course of stellar evolution, the conclusions obtained from different species are in remarkable agreement: baryons cannot close the universe. If $\Omega = 1$, the dark matter must be non-baryonic.

3.4 Baryogenesis

In discussing nucleosynthesis, we have taken it for granted that photons outnumber baryons in the universe by a large factor. Since baryons and antibaryons will annihilate at low temperatures, this is reasonable; but in that case why are there any baryons at all? A thermal background at high enough temperatures will contain equal numbers of protons and antiprotons, and this symmetry will be maintained as the particles annihilate. As usual, there would be some low

frozen-out abundance of both species at late times, but the universe would display a **matter-antimatter symmetry**. A recurring theme in cosmological debate has been to ask what happened to the antimatter. It appears that the universe began with a very slight asymmetry between matter and antimatter: at high temperatures there were $1 + O(10^{-9})$ protons for every antiproton. If baryon number is conserved, this imbalance cannot be altered once it is set in the initial conditions; but what generates it? One attractive idea is that the matter-antimatter asymmetry may have been set in the GUT era, before the universe cooled below the critical temperature of $\sim 10^{15}$ GeV.

The main idea to be exploited is that GUTs erase the distinction between baryons and leptons, treating them as different states of the same underlying unity. This raises the conceptual possibility of reactions that can generate a net baryon asymmetry, so long as the temperature is high enough that the GUT symmetry is not broken. The particles that will be involved are the gauge bosons of the GUT, since these are the ones that mediate the baryon \leftrightarrow lepton exchange reactions – *e.g.* the X & Y bosons of $SU(5)$. In particular, decay processes such as

$$X^{4/3} \rightarrow \begin{cases} e^+ + \bar{d} \\ 2u \end{cases} \quad (108)$$

are a promising direct source of baryon-number violation.

These mechanisms provide the first of three general **Sakharov conditions** for baryosynthesis (published in 1967, well before the invention of GUTs):

- (1) $\Delta B \neq 0$ reactions; (2) CP violation; (3) non-equilibrium conditions.

The second condition requires an asymmetry between particles and antiparticles. Recall what the symmetries C and P mean: a given observed reaction is possible (and will proceed with the same rate) if particles are replaced by antiparticles (C) or viewed in a mirror (P). Although P is violated by the weak interaction to the extent that only left-handed neutrinos are produced, the combined symmetry CP is obeyed in almost all cases. Consider the effect on the above X -boson decay if CP were to hold exactly: if a fraction f of decays produce $e^+ \bar{d}$, then decays of \bar{X} will produce the opposite baryon number via $e^- d$ the same fraction of the time, in which case no net asymmetry can be created. What we need is for the two fractions to be different, and such a process is observed in the laboratory in the form of the neutral kaons: both K_0 and \bar{K}_0 decay to either two or three pions, but with branching ratios that differ at the 10^{-3} level.

The third condition is necessary in order to prevent reverse reactions from erasing any baryon asymmetry as soon as it has been created. As in many cases in the expanding universe, the crucial physical results are contained in the ability of reactions to freeze out.

The challenge of baryosynthesis is to predict the observed asymmetry (in the form of a baryon-to-photon ratio $n_B/n_\gamma \sim 10^{-9}$). In principle, this can be done once the GUT model is given, and a connection can be made between laboratory measurements of CP violation and the baryon content of the universe. The simplest model for baryosynthesis would consist of a single Majorana particle, whose decays favour the production of baryons over antibaryons:

$$\begin{aligned} \Gamma(X \rightarrow \Delta B = +1) &= \frac{1}{2}(1 + \epsilon) \Gamma \\ \Gamma(X \rightarrow \Delta B = -1) &= \frac{1}{2}(1 - \epsilon) \Gamma. \end{aligned} \quad (109)$$

Here, $\Gamma = 1/\tau$ is the decay rate, and ϵ parameterizes the CP violation. The kinetic equations governing the effect of decays on the X number density and on baryon number B can be written down immediately following our earlier discussion of the Boltzmann equation:

$$\begin{aligned} \dot{n}_X + 3Hn_X &= -\Gamma(n_X - n_X^T) \\ \dot{n}_B + 3Hn_B &= \epsilon\Gamma(n_X - n_X^T). \end{aligned} \quad (110)$$

The terms involving the thermal-equilibrium X density, n_X^T , allow for inverse processes, and are deduced by asking what source term makes the lhs vanish in equilibrium, as before. What they say is that it is impossible for the X boson to decay when it is relativistic; we have to wait until $kT < mc^2$, so that n_X^T is suppressed. Note that the second equation explicitly makes clear the third Sakharov criterion for a violation of equilibrium. These equations are incomplete, as they neglect two-body processes that would contribute to the changing X abundance, such as $X + \bar{X}$ annihilation. The simplified form will apply after the X abundance has frozen out. In terms of comoving densities $N \equiv a^3 n$, the equation for baryon number is just

$$\dot{N}_B = -\epsilon \dot{N}_X \quad \Rightarrow \quad N_B = \epsilon(N_X^{\text{init}} - N_X), \quad (111)$$

so that the final baryon comoving density tends to $\epsilon N_X^{\text{init}}$ as the X 's decay away. The baryon-to-entropy ratio produced by this process is then

$$\frac{N_B}{s} \simeq \frac{\epsilon}{g_*} \exp\left(-\frac{m_X c^2}{kT_f}\right), \quad (112)$$

where the entropy density is defined here as g_* times the photon density, and the last term allows for the freeze-out suppression of the X density relative to massless backgrounds. Since the required ratio is $\sim 10^{-9}$ and $g_* \sim 100$ at early times, we need $\epsilon \gtrsim 10^{-7}$. Note that the freeze-out point cannot be very late: even for $\epsilon = 1$, $kT \gtrsim m_X c^2/16$ is needed.

Other mechanisms Baryosynthesis via GUT decay as above is the simplest mechanism, but there are other possibilities. First note that the above picture may well be inconsistent with an inflationary origin for the universe. Inflation generally involves a GUT-scale phase transition that leaves the universe reheated to a temperature somewhat below that of the GUT scale. Any pre-existing baryon asymmetry would be rendered irrelevant by the inflationary expansion, and things would not be hot enough afterwards for GUT processes to operate.

It is possible that baryosynthesis may proceed at still lower temperatures, since baryon non-conserving processes may even occur as part of the electroweak phase transition at $T \sim 200$ GeV. This is a surprise, since the electroweak Lagrangian contains no terms that would violate baryon number: leptons and hadrons are explicitly contained in different multiplets. This constraint may possibly be evaded by quantum tunnelling, but the exact extent to which baryon number violation may be realized in practice within the standard model is still a matter of debate (see *e.g.* section 6.8 of Kolb & Turner 1990; Grigoriev *et al.* 1992; Moore 1996; Ellis 1997).

4 INFLATIONARY COSMOLOGY

The standard isotropic cosmology is a very successful framework for interpreting observations, but prior to the early 1980s there were certain questions that had to be avoided. The initial conditions of the big bang appear to be odd in a number of ways; these puzzles are encapsulated in a set of classical ‘problems’, as follows.

The horizon problem Standard cosmology contains a particle horizon of comoving radius

$$r_H = \int_0^t \frac{c dt}{R(t)}, \quad (113)$$

which converges because $R \propto t^{1/2}$ in the early radiation-dominated phase. At late times, the integral is largely determined by the matter-dominated phase, for which

$$D_H = R_0 r_H \simeq \frac{6000}{\sqrt{\Omega_m}} h^{-1} \text{Mpc}. \quad (114)$$

The horizon at last scattering ($z \sim 1000$) was thus only ~ 100 Mpc in size, subtending an angle of about 1 degree. Why then are the large number of causally disconnected regions we see on the microwave sky all at the same temperature?

The flatness problem The $\Omega = 1$ universe is unstable:

$$[1 - 1/\Omega(z)] = f(z) [1 - 1/\Omega], \quad (115)$$

where $f(z) = (1+z)^{-1}$ in the matter-dominated era and $f(z) \propto (1+z)^{-2}$ for radiation domination, so that $f(z) \simeq (1+z_{\text{eq}})/(1+z)^2$ at early times. To get $\Omega \simeq 1$ today requires a **fine tuning of Ω** in the past, which becomes more and more precisely constrained as we increase the redshift at which the initial conditions are presumed to have been imposed. Ignoring annihilation effects, $1+z = T_{\text{init}}/2.7 \text{ K}$ and $1+z_{\text{eq}} \simeq 10^4$, so that the required fine tuning is

$$|\Omega(t_{\text{init}}) - 1| \lesssim 10^{-22} (E_{\text{init}}/\text{GeV})^2. \quad (116)$$

At the Planck epoch, which is the natural initial time, this requires a deviation of only 1 part in 10^{60} .

The expansion problem Even the most obvious fact of the cosmological expansion is unexplained. Although general relativity forbids a static universe, this is not enough to understand the expansion. As shown above, the gravitational dynamics of the cosmological scale factor $R(t)$ are just those of a cannonball travelling vertically in the Earth's gravity. Suppose we see a cannonball rising at a given time $t = t_0$: it may be true to say that it has $r = r_0$ and $v = v_0$ at this time because at a time Δt earlier it had $r = r - v_0 \Delta t$ and $v = v_0 - g \Delta t$, but this is hardly a satisfying explanation for the motion of a cannonball that was in fact fired by a cannon. Nevertheless, this is the only level of explanation that classical cosmology offers: the universe expands now because it did so in the past. Although it is not usually included in the list, one might thus with justice add an 'expansion problem' as perhaps the most fundamental in the catalogue of classical cosmological problems.

For many years, it was assumed that any solution to these difficulties would have to await a theory of quantum gravity. The classical singularity can be approached no closer than the Planck time of $\sim 10^{-43}$ s, and so the initial conditions for the classical evolution following this time must have emerged from behind the presently impenetrable barrier of the quantum gravity epoch. There remains a significant possibility that this policy of blaming everything on quantum gravity may be correct, but the great development of cosmology in the 1980s was the realization that the explanation of the initial-condition puzzles might involve physics at lower energies: 'only' 10^{15} GeV. Although this idea, now known as inflation, cannot be considered to be firmly established, the ability to treat gravity classically puts the discussion on a much less speculative foundation. What has emerged is a general picture of the early universe that has compelling simplicity, which moreover may be subject to observational verification. What follows is an outline of the main features of inflation; for more details see *e.g.* chapter 8 of Kolb & Turner (1990); Brandenberger (1990); Liddle & Lyth (1993).

Equation of state for inflation The list of problems with conventional cosmology provides a strong hint that the equation of state of the universe may have been very different at very early times. To solve the horizon problem and allow causal contact over the whole of the region observed at last scattering requires a universe that expands 'faster than light' near $t = 0$: $R \propto t^\alpha$, with $\alpha > 1$. If such a phase had existed, the integral for the comoving horizon would have diverged, and there would be no difficulty in understanding the overall homogeneity of the

universe – this could then be established by causal processes. Indeed, it is tempting to assert that the observed homogeneity *proves* that such causal contact must once have occurred. This phase of accelerated expansion is the most general feature of what has become known as the **inflationary universe**.

What condition does this place on the equation of state? In the integral for r_H , we can replace dt by dR/\dot{R} , which the Friedmann equation says is $\propto dR/\sqrt{\rho R^2}$ at early times. Thus, the horizon diverges provided the equation of state is such that ρR^2 vanishes or is finite as $R \rightarrow 0$. For a perfect fluid with $p \equiv (\Gamma - 1)\epsilon$ as the relation between pressure and energy density, we have the adiabatic dependence $p \propto R^{-3\Gamma}$, and the same dependence for ρ if the rest-mass density is negligible. A period of inflation therefore needs

$$\boxed{\Gamma < 2/3 \quad \Rightarrow \quad \rho c^2 + 3p < 0.} \quad (117)$$

An alternative way of seeing that this criterion is sensible is that the ‘active mass density’ $\rho + 3p/c^2$ then vanishes. Since this quantity forms the rhs of Poisson’s equation generalized to relativistic fluids, it is no surprise that the vanishing of $\rho + 3p/c^2$ allows a coasting solution with $R \propto t$.

Such a criterion can also solve the flatness problem. Consider the Friedmann equation,

$$\dot{R}^2 = \frac{8\pi G\rho R^2}{3} - kc^2. \quad (118)$$

As we have seen, the density term on the rhs must exceed the curvature term by a factor of at least 10^{60} at the Planck time, and yet a more natural initial condition might be to have the matter and curvature terms being of comparable order of magnitude. However, an inflationary phase in which ρR^2 increases as the universe expands can clearly make the curvature term relatively as small as required, provided inflation persists for sufficiently long.

de Sitter space and inflation We have seen that inflation will require an equation of state with negative pressure, and the only familiar example of this is the $p = -\rho c^2$ relation that applies for vacuum energy; in other words, we are led to consider inflation as happening in a universe dominated by a cosmological constant. As usual, any initial expansion will redshift away matter and radiation contributions to the density, leading to increasing dominance by the vacuum term. If the radiation and vacuum densities are initially of comparable magnitude, we quickly reach a state where the vacuum term dominates. The Friedmann equation in the vacuum-dominated case has three solutions:

$$R \propto \begin{cases} \sinh Ht & (k = -1) \\ \cosh Ht & (k = +1) \\ \exp Ht & (k = 0), \end{cases} \quad (119)$$

where $H = \sqrt{\Lambda c^2/3} = \sqrt{8\pi G\rho_{\text{vac}}/3}$; all solutions evolve towards the exponential $k = 0$ solution, known as **de Sitter space**. Note that H is not the Hubble parameter at an arbitrary time (unless $k = 0$), but it becomes so exponentially fast as the hyperbolic trigonometric functions tend to the exponential.

Because de Sitter space clearly has H^2 and ρ in the right ratio for $\Omega = 1$ (obvious, since $k = 0$), the density parameter in all models tends to unity as the Hubble parameter tends to H . If we assume that the initial conditions are not fine tuned (*i.e.* $\Omega = O(1)$ initially), then maintaining the expansion for a factor f produces

$$\Omega = 1 + O(f^{-2}). \quad (120)$$

This can solve the flatness problem, provided f is large enough. To obtain Ω of order unity today requires $|\Omega - 1| \lesssim 10^{-52}$ at the GUT epoch, and so

$$\boxed{\ln f \gtrsim 60} \tag{121}$$

e -foldings of expansion are needed; it will be proved below that this is also exactly the number needed to solve the horizon problem. It then seems almost inevitable that the process should go to completion and yield $\Omega = 1$ to measurable accuracy today. There is only a rather small range of e -foldings (60 ± 2 , say) around the critical value for which Ω today can be of order unity without it being equal to unity to within the tolerance set by density fluctuations ($\pm 10^{-5}$), and it would constitute an unattractive fine-tuning to require that the expansion hit this narrow window exactly.

This gives the first of two strong **predictions of inflation**: that the universe must be spatially flat

$$\boxed{\text{inflation} \Rightarrow k = 0.} \tag{122}$$

Note that this need not mean the Einstein–de Sitter model; the alternative possibility is that a vacuum contribution is significant in addition to matter, so that $\Omega_m + \Omega_v = 1$. Astrophysical difficulties in finding evidence for $\Omega_m = 1$ are thus one of the major motivations, through inflation, for taking the idea of a large cosmological constant seriously.

4.1 Inflation field dynamics

The general concept of inflation rests on being able to achieve a negative-pressure equation of state. This can be realized in a natural way by quantum fields in the early universe.

Quantum fields at high temperatures The critical fact we shall need from quantum field theory is that quantum fields can produce an energy density that mimics a cosmological constant. The discussion will be restricted to the case of a scalar field ϕ (complex in general, but often illustrated using the case of a single real field). The restriction to scalar fields is not simply for reasons of simplicity, but because the scalar sector of particle physics is relatively unexplored. While vector fields such as electromagnetism are well understood, it is expected in many theories of unification that additional scalar fields such as the Higgs field will exist. We now need to look at what these can do for cosmology.

The Lagrangian density for a scalar field is as usual of the form of a kinetic minus a potential term:

$$\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - V(\phi). \tag{123}$$

In familiar examples of quantum fields, the potential would be

$$V(\phi) = \frac{1}{2} m^2 \phi^2, \tag{124}$$

where m is the mass of the field in natural units. However, it will be better to keep the potential function general at this stage. As usual, Noether's theorem gives the energy–momentum tensor for the field as

$$T^{\mu\nu} = \partial^\mu \phi \partial^\nu \phi - g^{\mu\nu} \mathcal{L}. \tag{125}$$

From this, we can read off the energy density and pressure:

$$\begin{aligned}\rho &= \frac{1}{2}\dot{\phi}^2 + V(\phi) + \frac{1}{2}(\nabla\phi)^2 \\ p &= \frac{1}{2}\dot{\phi}^2 - V(\phi) - \frac{1}{6}(\nabla\phi)^2.\end{aligned}\tag{126}$$

If the field is constant both spatially and temporally, the equation of state is then $p = -\rho$, as required if the scalar field is to act as a cosmological constant; note that derivatives of the field spoil this identification.

If ϕ is a (complex) Higgs field, then the symmetry-breaking Mexican hat potential might be assumed:

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4.\tag{127}$$

At the classical level, such potentials determine where $|\phi|$ will be found in equilibrium: at the potential minimum. In quantum terms, this goes over to the **vacuum expectation value** $\langle 0|\phi|0\rangle$. However, these potentials do not include the inevitable fluctuations that will arise in thermal equilibrium. We know how to treat these in classical systems: at non-zero temperature a system of fixed volume will minimize not its potential energy, but the **Helmholtz free energy** $F = V - TS$, S being the entropy. The calculation of the entropy is technically complex, since it involves allowance for quantum interactions with a thermal bath of background particles. However, the main result can be justified, as follows. The effect of the thermal interaction must be to add an interaction term to the Lagrangian $\mathcal{L}_{\text{int}}(\phi, \psi)$, where ψ is a thermally fluctuating field that corresponds to the heat bath. In general, we would expect \mathcal{L}_{int} to have a quadratic dependence on $|\phi|$ around the origin: $\mathcal{L}_{\text{int}} \propto |\phi|^2$ (otherwise we would need to explain why the second derivative either vanishes or diverges); the coefficient of proportionality will be an effective mass² that depends on the thermal fluctuations in ψ . On dimensional grounds, this coefficient must be proportional to T^2 , although a more detailed analysis would be required to obtain the constant of proportionality.

There is thus a temperature-dependent **effective potential** that we have to minimize:

$$V_{\text{eff}}(\phi, T) = V(\phi, 0) + aT^2|\phi|^2.\tag{128}$$

The effect of this on the symmetry-breaking potential depends on the form of the zero-temperature $V(\phi)$. If the function is taken to be the simple Higgs form $V = -\mu^2 + \lambda\phi^4$, then the temperature-dependent part simply modifies the effective value of μ^2 : $\mu_{\text{eff}}^2 = \mu^2 - aT^2$. At very high temperatures, the potential will be parabolic, with a minimum at $|\phi| = 0$; below the critical temperature, $T_c = \mu/\sqrt{a}$, the ground state is at $|\phi| = [\mu_{\text{eff}}^2/(2\lambda)]^{1/2}$ and the symmetry is broken. At any given time, there is only a single minimum, and so this is a second-order phase transition.

It is easy enough to envisage more complicated behaviour, as illustrated in figure 4. This plots the potential

$$V_{\text{eff}}(\phi, T) = \lambda|\phi|^4 - b|\phi|^3 + aT^2|\phi|^2,\tag{129}$$

which displays two critical temperatures. At very high temperatures, the potential will have a parabolic minimum at $|\phi| = 0$; at T_1 , a second minimum appears in V_{eff} at $|\phi| \neq 0$, and this will be the global minimum for some $T_2 < T_1$. For $T < T_2$, the state at $|\phi| = 0$ is known as the **false vacuum**, whereas the global minimum is known as the **true vacuum**. For this particular form of potential, the second minimum around $\phi = 0$ always exists, so that there is a potential barrier preventing a transition to the false vacuum. This can be overcome by adding a small $-\mu^2|\phi|^2$ component to the potential, so that there will be a third critical temperature at which the curvature around the origin changes sign, leaving only one minimum in the potential.

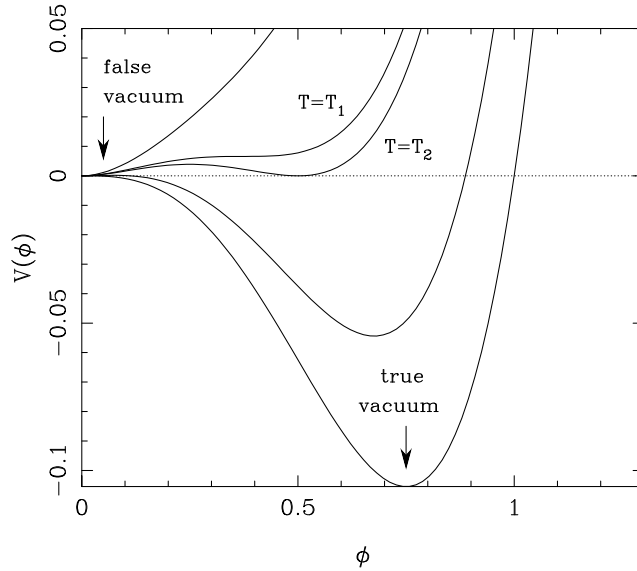


Fig. 4: The temperature-dependent effective potential $V = T^2|\phi|^2 - |\phi|^3 + |\phi|^4$, illustrated at several temperatures: $T^2 = 0.5, 9/32, 1/4, 0.1, 0$. For $T > T_1 = (9/32)^{1/2} \simeq 0.53$, only the false vacuum is available; for $T < T_2 = 1/2$ the true vacuum is energetically favoured and the potential approaches the zero-temperature form.

Alternatively, once the barrier is small enough, quantum tunnelling can take place and free ϕ to move. The universe is no longer trapped in the false vacuum and can make a first-order phase transition to the true vacuum state.

The crucial point to note for cosmology is that there is an energy-density difference between the two vacuum states:

$$\Delta V = \frac{\mu^4}{2\lambda} \quad (130)$$

If we say that the zero of energy is such that $V = 0$ in the true vacuum, this implies that the false-vacuum symmetric state displays an effective cosmological constant. On dimensional grounds, this must be an energy density $\sim m^4$ in natural units, where m is the energy at which the phase transition occurs. For GUTs, $m \simeq 10^{15}$ GeV; in laboratory units, this implies

$$\rho_{\text{vac}} = \frac{(10^{15}\text{GeV})^4}{\hbar^3 c^5} \simeq 10^{80} \text{ kg m}^{-3}. \quad (131)$$

The inevitability of such a colossal vacuum energy in models with GUT-scale symmetry breaking was the major motivation for the concept of inflation as originally envisaged by Guth (1981). At first sight, the overall package looks highly appealing, since the phase transition from false to true vacuum both terminates inflation and also reheats the universe to the GUT temperature, allowing the possibility that GUT-based reactions that violate baryon-number conservation can generate the observed matter/antimatter asymmetry. Because the transition is first-order, the original inflation model is known as **first-order inflation**.

However, while a workable inflationary cosmology will very probably deploy the three basic elements of vacuum-driven expansion, fluctuation generation and reheating, it has become clear that such a model must be more complex than Guth's initial proposal. To explain where the problems arise, we need to look in more detail at the functioning of the inflation mechanism.

Dynamics of the inflation field Treating the field classically (*i.e.* considering the expectation value $\langle\phi\rangle$), we get from energy–momentum conservation ($T_{;\nu}^{\mu\nu} = 0$) the equation of motion

$$\boxed{\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + dV/d\phi = 0.} \quad (132)$$

This can also be derived more easily by the direct route of writing down the action $S = \int \mathcal{L} \sqrt{-g} d^4x$ and applying the Euler–Lagrange equation that arises from a stationary action ($\sqrt{-g} = R^3(t)$ for an FRW model, which is the origin of the Hubble drag term $3H\dot{\phi}$).

The solution of the equation of motion becomes tractable if we both ignore spatial inhomogeneities in ϕ and make the **slow-rolling approximation** that $|\dot{\phi}|$ is negligible in comparison with $|3H\dot{\phi}|$ and $|dV/d\phi|$. Both these steps are required in order that inflation can happen; we have shown above that the vacuum equation of state only holds if in some sense ϕ changes slowly both spatially and temporally. Suppose there are characteristic temporal and spatial scales T and X for the scalar field; the conditions for inflation are that the negative-pressure equation of state from $V(\phi)$ must dominate the normal-pressure effects of time and space derivatives:

$$V \gg \phi^2/T^2, \quad V \gg \phi^2/X^2, \quad (133)$$

hence $|dV/d\phi| \sim V/\phi$ must be $\gg \phi/T^2 \sim \ddot{\phi}$. The $\ddot{\phi}$ term can therefore be neglected in the equation of motion, which then takes the slow-rolling form for homogeneous fields:

$$\boxed{3H\dot{\phi} = -dV/d\phi.} \quad (134)$$

The conditions for inflation can be cast into useful dimensionless forms. The basic condition $V \gg \dot{\phi}^2$ can now be rewritten using the slow-roll relation as

$$\boxed{\epsilon \equiv \frac{m_{\text{P}}^2}{16\pi} (V'/V)^2 \ll 1.} \quad (135)$$

Also, we can differentiate this expression to obtain the criterion $V'' \ll V'/m_{\text{P}}$. Using slow-roll once more gives $3H\dot{\phi}/m_{\text{P}}$ for the rhs, which is in turn $\ll 3H\sqrt{V}/m_{\text{P}}$ because $\dot{\phi}^2 \ll V$, giving finally

$$\boxed{\eta \equiv \frac{m_{\text{P}}^2}{8\pi} (V''/V) \ll 1} \quad (136)$$

(recall that for de Sitter space $H = \sqrt{8\pi GV(\phi)/3} \sim \sqrt{V}/m_{\text{P}}$ in natural units). These two criteria make perfect intuitive sense: the potential must be flat in the sense of having small derivatives if the field is to roll slowly enough for inflation to be possible.

Similar arguments can be made for the spatial parts. However, they are less critical: what matters is the value of $\nabla\phi = \nabla_{\text{comoving}}\phi/R$. Since R increases exponentially, these perturbations are damped away: assuming V is large enough for inflation to start in the first place, inhomogeneities rapidly become negligible. This ‘stretching’ of field gradients as we increase the cosmological horizon beyond the value predicted in classical cosmology also solves a related problem that was historically important in motivating the invention of inflation – the **monopole problem**. Monopoles are point-like topological defects that would be expected to arise in any

phase transition at around the GUT scale ($t \sim 10^{-35}$ s). If they form at approximately one per horizon volume at this time, then it follows that the present universe would contain $\Omega \gg 1$ in monopoles. This unpleasant conclusion is avoided if the horizon can be made much larger than the classical one at the end of inflation; the GUT fields have then been aligned over a vast scale, so that topological-defect formation becomes extremely rare.

Ending inflation Although spatial derivatives of the scalar field can thus be neglected, the same is not always true for time derivatives. Although they may be negligible initially, the relative importance of time derivatives increases as ϕ rolls down the potential and V approaches zero (leaving aside the subtle question of how we know that the minimum is indeed at zero energy). Even if the potential does not steepen, sooner or later we will have $\epsilon \simeq 1$ or $|\eta| \simeq 1$ and the inflationary phase will cease. Instead of rolling slowly ‘downhill’, the field will oscillate about the bottom of the potential, with the oscillations becoming damped by the $3H\dot{\phi}$ friction term (see figure 5). Eventually, we will be left with a stationary field that either continues to inflate without end, if $V(\phi = 0) > 0$, or which simply has zero density. This would be a most boring universe to inhabit, but fortunately there is a more realistic way in which inflation can end. We have neglected so far the couplings of the scalar field to matter fields. Such couplings will cause the rapid oscillatory phase to produce particles, leading to **reheating**. Thus, even if the minimum of $V(\phi)$ is at $V = 0$, the universe is left containing roughly the same energy density as it started with, but now in the form of normal matter and radiation – which starts the usual FRW phase, albeit with the desired special ‘initial’ conditions.

As well as being of interest for completing the picture of inflation, it is essential to realize that these closing stages of inflation are the *only* ones of observational relevance. Inflation might well continue for a huge number of e -foldings, all but the last few satisfying $\epsilon, \eta \ll 1$. However, the scales that left the de Sitter horizon at these early times are now vastly greater than our observable horizon, c/H_0 , which exceeds the de Sitter horizon by only a finite factor. If inflation terminated by reheating to the GUT temperature, then the expansion factor required to reach the present epoch is

$$a_{\text{GUT}}^{-1} \simeq E_{\text{GUT}}/E_\gamma. \quad (137)$$

The comoving horizon size at the end of inflation was therefore

$$d_{\text{H}}(t_{\text{GUT}}) \simeq a_{\text{GUT}}^{-1} [c/H_{\text{GUT}}] \simeq [E_{\text{P}}/E_\gamma] E_{\text{GUT}}^{-1}, \quad (138)$$

where the last expression in natural units uses $H \simeq \sqrt{V}/E_{\text{P}} \simeq E_{\text{GUT}}^2/E_{\text{P}}$. For a GUT energy of 10^{15} GeV, this is about 10 m. This is a sobering illustration of the magnitude of the horizon problem; if we relied on causal processes at the GUT era to produce homogeneity, then the universe would only be smooth in patches a few comoving metres across. To solve the problem, we need enough e -foldings of inflation to have stretched this GUT-scale horizon to the present horizon size

$$\boxed{N_{\text{obs}} = \ln \left[\frac{3000h^{-1} \text{ Mpc}}{(E_{\text{P}}/E_\gamma) E_{\text{GUT}}^{-1}} \right] \simeq 60.} \quad (139)$$

By construction, this is enough to solve the horizon problem, and it is also the number of e -foldings needed to solve the flatness problem. This is no coincidence, since we saw earlier that the criterion in this case was

$$N \gtrsim \frac{1}{2} \ln \left(\frac{a_{\text{eq}}}{a_{\text{GUT}}^2} \right). \quad (140)$$

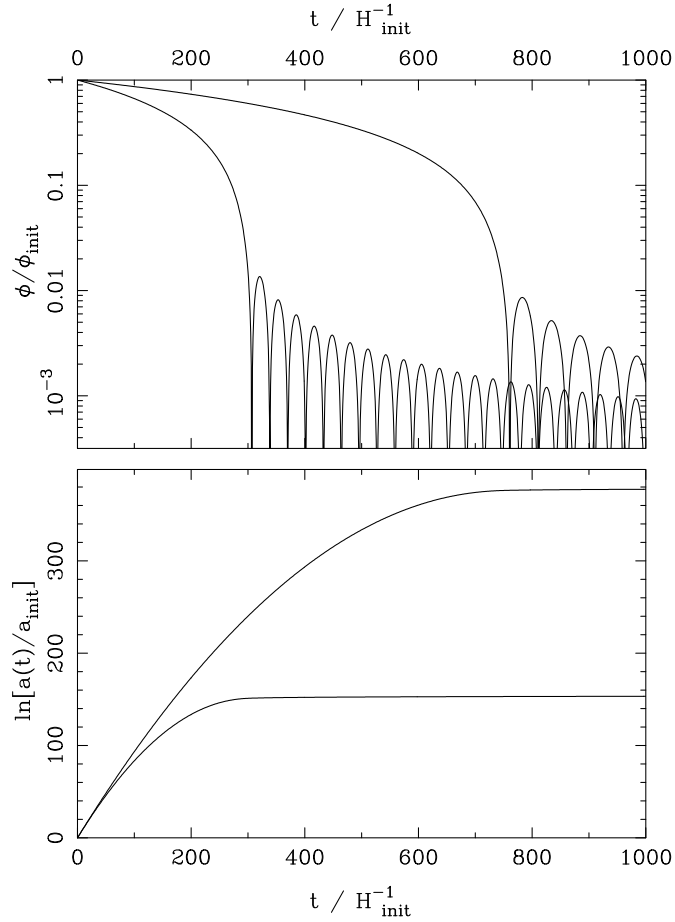


Fig. 5: A plot of the exact solution for the scalar field in a model with a $V \propto \phi^2$ potential. The top panel shows how the absolute value of ϕ falls smoothly with time during the inflationary phase, and then starts to oscillate when inflation ends. The bottom panel shows the evolution of the scale factor. We see the initial exponential behaviour flattening as the vacuum energy ceases to dominate. The two models shown have starting points of $W_i \equiv V_i/(\phi_i^2 H_i^2) = 0.002$ and 0.005 ; the former (upper lines in each panel) gives about 380 e -foldings of inflation, the latter (lower lines) only 150. According to the $\epsilon = \eta = 1$ criterion, inflation in these models ends at respectively $t = 730$ and 240 . The observationally relevant part of inflation is the last 60 e -foldings, and the behaviour of the scale factor is significantly non-exponential in this regime.

Now, $a_{\text{eq}} = \rho_\gamma / \rho$, and $\rho = 3H^2\Omega / (8\pi G)$. In natural units, this translates to $\rho \sim E_{\text{p}}^2 (c/H_0)^{-2}$, or $a_{\text{eq}}^{-1} \sim E_{\text{p}}^2 (c/H_0)^{-2} / E_\gamma^4$. The expression for N is then identical to that in the case of the horizon problem: the same number of e -foldings will always solve both.

Criteria for inflation Successful inflation in any of these models requires > 60 e -foldings of the expansion. The implications of this are easily calculated using the slow-roll equation, which gives the number of e -foldings between ϕ_1 and ϕ_2 as

$$N = \int H dt = -\frac{8\pi}{m_{\text{p}}^2} \int_{\phi_1}^{\phi_2} \frac{V}{V'} d\phi \quad (141)$$

For any potential that is relatively smooth, $V' \sim V/\phi$, and so we get $N \sim (\phi_{\text{start}}/m_{\text{p}})^2$, assuming that inflation terminates at a value of ϕ rather smaller than at the start. The criterion for successful inflation is thus that the initial value of the field exceeds the Planck scale:

$$\boxed{\phi_{\text{start}} \gg m_{\text{p}}.} \quad (142)$$

By the same argument, it is easily seen that this is also the criterion needed to make the slow-roll parameters ϵ and $\eta \ll 1$. To summarize, any model in which the potential is sufficiently flat that slow-roll inflation can commence will probably achieve the critical 60 e -foldings. Counterexamples can of course be constructed, but they have to be somewhat special cases.

It is interesting to review this conclusion for some of the specific inflation models listed above. Consider a mass-like potential $V = m^2\phi^2$. If inflation starts near the Planck scale, the fluctuations in V are $\sim m_{\text{p}}^4$ and these will drive ϕ_{start} to $\phi_{\text{start}} \gg m_{\text{p}}$ provided $m \ll m_{\text{p}}$; similarly, for $V = \lambda\phi^4$, the condition is weak coupling: $\lambda \ll 1$. Any field with a rather flat potential will thus tend to inflate, just because typical fluctuations leave it a long way from home in the form of the potential minimum. In a sense, inflation is realized by means of ‘inertial confinement’: there is nothing to prevent the scalar field from reaching the minimum of the potential – but it takes a long time to do so, and the universe has meanwhile inflated by a large factor.

4.2 Relic fluctuations from inflation

We have seen that de Sitter space contains a true event horizon, of proper size c/H . This suggests that there will be thermal fluctuations present, as with a black hole, for which the **Hawking temperature** is $kT_{\text{H}} = \hbar c / (4\pi r_{\text{s}})$. This analogy is close, but imperfect, and the characteristic temperature of de Sitter space is a factor 2 higher:

$$kT_{\text{deSitter}} = \frac{\hbar H}{2\pi}. \quad (143)$$

This existence of thermal fluctuations is one piece of intuitive motivation for expecting fluctuations in the quantum fields that are present in de Sitter space, but is not so useful in detail. In practice, we need a more basic calculation, which is to see how the zero-point fluctuations in small-scale quantum modes freeze out as classical density fluctuations once the modes have been inflated to super-horizon scales.

The details of this calculation are given below. However, we can immediately note that a natural prediction will be a spectrum of perturbations that are nearly *scale invariant*. This means that the metric fluctuations of spacetime receive equal levels of distortion from each decade of perturbation wavelength, and may be quantified in terms of the rms fluctuations, σ ,

in Newtonian gravitational potential, Φ ($c = 1$):

$$\delta_{\text{H}}^2 \equiv \Delta_{\Phi}^2 \equiv \frac{d\sigma^2(\Phi)}{d \ln k} = \text{constant}. \quad (144)$$

The notation δ_{H} arises because the potential perturbation is of the same order as the density fluctuation on the scale of the horizon at any given time.

It is commonly argued that the prediction of scale invariance arises because de Sitter space is invariant under time translation: there is no natural origin of time under exponential expansion. At a given time, the only length scale in the model is the horizon size c/H , so it is inevitable that the fluctuations that exist on this scale are the same at all times. After inflation ceases, the resulting fluctuations (at constant amplitude on the scale of the horizon) give us the Zeldovich spectrum **Zeldovich** or scale-invariant spectrum **scale-invariant** spectrum. The problem with this argument is that it ignores the issue of how the perturbations evolve while they are outside the horizon; we have only really calculated the amplitude for the last generation of fluctuations – *i.e.* those that are on the scale of the horizon at the time inflation ends. Fluctuations generated at earlier times will be inflated outside the de Sitter horizon, and will re-enter the FRW horizon at some time after inflation has ceased.

The evolution during this period is a topic where some care is needed, since the description of these large-scale perturbations is sensitive to the gauge freedom in general relativity. A technical discussion is given in *e.g.* Mukhanov, Feldman & Brandenberger (1992); for the present, we shall rely on simply motivating the inflationary result, which is that potential perturbations re-enter the horizon with the same amplitude they had on leaving. This may be made reasonable in two ways. Perturbations outside the horizon are immune to causal effects, so it is hard to see how any large-scale non-flatness in spacetime could ‘know’ whether it was supposed to grow or decline.

We therefore argue that the inflationary process produces a universe that is fractal-like in the sense that scale-invariant fluctuations correspond to a metric that has the same ‘wrinkliness’ per log length-scale. It then suffices to calculate that amplitude on one scale – *i.e.* the perturbations that are just leaving the horizon at the end of inflation, so that super-horizon evolution is not an issue. It is possible to alter this prediction of scale invariance only if the expansion is non-exponential; we have seen that such deviations plausibly do exist towards the end of inflation, so it is clear that exact scale invariance is not to be expected.

To anticipate the detailed treatment, the inflationary prediction is of a horizon-scale amplitude

$$\delta_{\text{H}} = \frac{H^2}{2\pi \dot{\phi}} \quad (145)$$

which can be understood as follows. Imagine that the main effect of fluctuations is to make different parts of the universe have fields that are perturbed by an amount $\delta\phi$. In other words, we are dealing with various copies of the same rolling behaviour $\phi(t)$, but viewed at different times

$$\delta t = \frac{\delta\phi}{\dot{\phi}}. \quad (146)$$

These universes will then finish inflation at different times, leading to a spread in energy densities (figure 6). The horizon-scale density amplitude is given by the different amounts that the

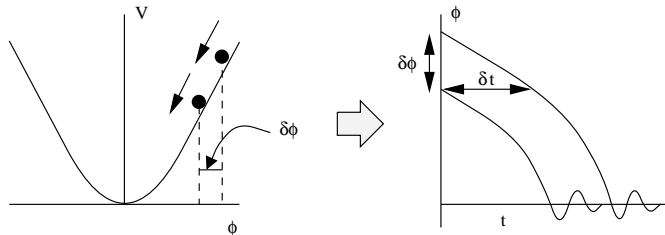


Fig. 6: This plot shows how fluctuations in the scalar field transform themselves into density fluctuations at the end of inflation. Different points of the universe inflate from points on the potential perturbed by a fluctuation $\delta\phi$, like two balls rolling from different starting points. Inflation finishes at times separated by δt in time for these two points, inducing a density fluctuation $\delta = H\delta t$.

universes have expanded following the end of inflation:

$$\delta_H \simeq H \delta t = \frac{H^2}{2\pi \dot{\phi}}, \quad (147)$$

where the last step uses the crucial input of quantum field theory, which says that the rms $\delta\phi$ is given by $H/2\pi$.

The fluctuation spectrum We now need to go over this vital result in rather more detail (see Liddle & Lyth 1993 for a particularly clear treatment). First, consider the equation of motion obeyed by perturbations in the inflaton field. The basic equation of motion is

$$\ddot{\phi} + 3H\dot{\phi} - \nabla^2\phi + V'(\phi) = 0, \quad (148)$$

and we seek the corresponding equation for the perturbation $\delta\phi$ obtained by starting inflation with slightly different values of ϕ in different places. Suppose this perturbation takes the form of a comoving plane-wave perturbation of comoving wavenumber k and amplitude A : $\delta\phi = A \exp(i\mathbf{k} \cdot \mathbf{x} - ikt/a)$. If the slow-roll conditions are also assumed, so that V' may be treated as a constant, then the perturbed field $\delta\phi$ obeys the first-order perturbation of the equation of motion for the main field:

$$[\ddot{\delta\phi}] + 3H[\dot{\delta\phi}] + (k/a)^2[\delta\phi] = 0, \quad (149)$$

which is a standard wave equation for a massless field evolving in an expanding universe.

Having seen that the inflaton perturbation behaves in this way, it is not much work to obtain the quantum fluctuations that result in the field at late times (*i.e.* on scales much larger than the de Sitter horizon). First consider the fluctuations in flat space: the field would be expanded as

$$\phi_k = \omega_k a_k + \omega_k^* a_k^\dagger, \quad (150)$$

and the field variance would be

$$\langle 0 | |\phi_k|^2 | 0 \rangle = |\omega_k|^2. \quad (151)$$

To solve the general problem, we only need to find how the amplitude ω_k changes as the universe expands. The idea is to start from the situation where we are well inside the horizon ($k/a \gg H$), in which case flat-space quantum theory will apply, and end at the point of interest outside the horizon (where $k/a \ll H$).

Returning now to the calculation, we want to know how the mode amplitude changes as the wavelength passes through the horizon. Initially, we have the standard result from flat-space quantum field theory, which can be rewritten in comoving units as

$$\omega_k = a^{-3/2} (2k/a)^{-1/2} e^{-ikt/a}. \quad (152)$$

The powers of the scale factor, $a(t)$, just allow for expanding the field in comoving wavenumbers k . The field amplitude contains a normalizing factor of $V^{-1/2}$, V being a proper volume; hence the $a^{-3/2}$ factor, if we use comoving $V = 1$. Another way of looking at this is that the proper number density of inflatons goes as a^{-3} as the universe expands. With this boundary condition, it is straightforward to check by substitution that the following expression satisfies the evolution equation:

$$\boxed{\omega_k = a^{-3/2} (2k/a)^{-1/2} e^{-ik/aH} (1 + iaH/k)} \quad (153)$$

(remember that H is a constant, so that $(d/dt)[aH] = H\dot{a} = aH^2$ etc.). At early times, when the horizon is much larger than the wavelength, $aH/k \ll 1$, and so ω_k is the flat-space result, except that the time dependence looks a little odd, being $\exp(-ik/aH)$. However, since $(d/dt)[k/aH] = -k/a$, we see that the oscillatory term has a leading dependence on t of the desired kt/a form. In the limit of very early times, the period of oscillation is $\ll H^{-1}$, so a is effectively a constant from the point of view of the epoch where quantum fluctuations dominate.

At the opposite extreme, $aH/k \gg 1$, the fluctuation amplitude becomes frozen out at the value

$$\langle 0 | |\phi_k|^2 | 0 \rangle = \frac{H^2}{2k^3}. \quad (154)$$

The initial quantum zero-point fluctuations in the field have been transcribed to a constant classical fluctuation that can eventually manifest itself as large-scale structure. The fluctuations in ϕ depend on k in such a way that the fluctuations per decade are constant:

$$\frac{d(\delta\phi)^2}{d \ln k} = \frac{4\pi k^3}{(2\pi)^3} \langle 0 | |\phi_k|^2 | 0 \rangle = \left(\frac{H}{2\pi}\right)^2 \quad (155)$$

(the factor $(2\pi)^{-3}$ comes from the Fourier transform; $4\pi k^2 dk = 4\pi k^3 d \ln k$ comes from the k -space volume element). This completes the argument. The rms value of fluctuations in ϕ can be used as above to deduce the power spectrum of mass fluctuations well after inflation is over. In terms of the variance per $\ln k$ in potential perturbations, the answer is

$$\boxed{\begin{aligned} \delta_H^2 &\equiv \Delta_{\Phi}^2(k) = \frac{H^4}{(2\pi\dot{\phi})^2} \\ H^2 &= \frac{8\pi}{3} \frac{V}{m_{\text{p}}^2} \\ 3H\dot{\phi} &= -V', \end{aligned}} \quad (156)$$

where we have also written once again the exact relation between H and V and the slow-roll condition, since manipulation of these three equations is often required in derivations.

This result calls for a number of comments. First, if H and $\dot{\phi}$ are both constant then the predicted spectrum is exactly scale invariant, with some characteristic inhomogeneity on

the scale of the horizon. As we have seen, exact de Sitter space with constant H will not be strictly correct for most inflationary potentials; nevertheless, in most cases the main points of the analysis still go through. The fluctuations in ϕ start as normal flat-space fluctuations (and so not specific to de Sitter space), which change their character as they are advected beyond the horizon and become frozen-out classical fluctuations. All that matters is that the Hubble parameter is roughly constant for the few e -foldings that are required for this transition to happen. If H does change with time, the number to use is the value at the time that a mode of given k crosses the horizon. Even if H were to be precisely constant, there remains the dependence on $\dot{\phi}$, which again will change as different scales cross the horizon. This means that different inflationary models display different characteristic deviations from a nearly scale-invariant spectrum, and this is discussed in more detail below.

Two other characteristics of the perturbations are more general: they will be Gaussian and adiabatic in nature. A Gaussian density field is one for which the joint probability distribution of the density at any given number of points is a multivariate Gaussian. The easiest way for this to arise in practice is for the density field to be constructed as a superposition of Fourier modes with independent random phases; the Gaussian property then follows from the central limit theorem. It is easy to see in the case of inflation that this requirement will be satisfied: the quantum commutation relations only apply to modes of the same k , so that modes of different wavelength behave independently and have independent zero-point fluctuations.

Gravity waves and tilt The density perturbations left behind as a residue of the quantum fluctuations in the inflaton field during inflation are an important relic of that epoch, but are not the only one. In principle, a further important test of the inflationary model is that it also predicts a background of gravitational waves, whose properties couple with those of the density fluctuations.

It is easy to see in principle how such waves arise. In linear theory, any quantum field is expanded in a similar way into a sum of oscillators with the usual creation and annihilation operators; the above analysis of quantum fluctuations in a scalar field is thus readily adapted to show that analogous fluctuations will be generated in other fields during inflation. In fact, the linearized contribution of a gravity wave, $h_{\mu\nu}$, to the Lagrangian looks like a scalar field $\phi = (m_{\text{P}}/4\sqrt{\pi}) h_{\mu\nu}$, the expected rms gravity-wave amplitude is

$$h_{\text{rms}} \sim H/m_{\text{P}}. \quad (157)$$

The fluctuations in ϕ are transmuted into density fluctuations, but gravity waves will survive to the present day, albeit redshifted.

This redshifting produces a break in the spectrum of waves. Prior to horizon entry, the gravity waves produce a scale-invariant spectrum of metric distortions, with amplitude h_{rms} per $\ln k$. These distortions are observable via the large-scale CMB anisotropies, where the tensor modes produce a spectrum with the same scale dependence as the Sachs–Wolfe gravitational redshift from scalar metric perturbations. In the scalar case, we have $\delta T/T \sim \phi/3c^2$, *i.e.* of order the Newtonian metric perturbation; similarly, the tensor effect is

$$\left(\frac{\delta T}{T}\right)_{\text{GW}} \sim h_{\text{rms}} \lesssim \delta_{\text{H}} \sim 10^{-5}, \quad (158)$$

where the second step follows because the tensor modes can constitute no more than 100% of the observed CMB anisotropy. The energy density of the waves is $\rho_{\text{GW}} \sim m_{\text{P}}^2 h^2 k^2$, where $k \sim H(a_{\text{entry}})$ is the proper wavenumber of the waves. At horizon entry, we therefore expect

$$\rho_{\text{GW}} \sim m_{\text{P}}^2 h_{\text{rms}}^2 H^2(a_{\text{entry}}). \quad (159)$$

After horizon entry, the waves redshift away like radiation, as a^{-4} , and generate a present-day energy spectrum per $\ln k$ that is constant for modes that entered the horizon while the universe was radiation dominated (because $a \propto t^{1/2} \Rightarrow H^2 a^4 = \text{constant}$). What is the density parameter of these waves? In natural units, $\Omega = (8\pi/3)\rho/(H^2 m_{\text{P}}^2)$, so $\Omega_{\text{GW}} \sim h_{\text{rms}}^2$ at the time of horizon entry, at which epoch the universe was radiation dominated with $\Omega_r = 1$ to an excellent approximation. Thereafter, the wave density maintains a constant ratio to the radiation density, since both redshift as a^{-4} , giving the present-day density as

$$\boxed{\Omega_{\text{GW}} \sim \Omega_r (H/m_{\text{P}})^2 \sim 10^{-4} V/m_{\text{P}}^4.} \quad (160)$$

The gravity-wave spectrum therefore displays a break between constant metric fluctuations on super-horizon scales and constant density fluctuations on small scales. An analogous break also exists in the spectrum of density perturbations in dark matter. If gravity waves make an important contribution to CMB anisotropies, we must have $h_{\text{rms}} \sim 10^{-5}$, and so $\Omega_{\text{GW}} \sim 10^{-14}$ is expected.

An alternative way of presenting the gravity-wave effect on the CMB anisotropies is via the ratio between the tensor effect of gravity waves and the normal scalar Sachs–Wolfe effect, as first analysed in a prescient paper by Starobinsky (1985). Denote the fractional temperature variance per natural logarithm of angular wavenumber by Δ^2 (constant for a scale-invariant spectrum). The tensor and scalar contributions are respectively

$$\Delta_{\text{T}}^2 \sim h_{\text{rms}}^2 \sim (H^2/m_{\text{P}}^2) \sim V/m_{\text{P}}^4. \quad (161)$$

$$\Delta_{\text{S}}^2 \sim \delta_{\text{H}}^2 \sim \frac{H^2}{\phi} \sim \frac{H^6}{(V')^2} \sim \frac{V^3}{m_{\text{P}}^6 V'^2}. \quad (162)$$

The ratio of the tensor and scalar contributions to the variance of microwave background anisotropies is therefore proportional to the inflationary parameter ϵ :

$$\frac{\Delta_{\text{T}}^2}{\Delta_{\text{S}}^2} \simeq 12.4 \epsilon, \quad (163)$$

inserting the exact coefficient from Starobinsky (1985). If it could be measured, the gravity-wave contribution to CMB anisotropies would therefore give a measure of ϵ , one of the dimensionless inflation parameters. The less ‘de Sitter-like’ the inflationary behaviour, the larger the relative gravitational-wave contribution.

Since deviations from exact exponential expansion also manifest themselves as density fluctuations with spectra that deviate from scale invariance, this suggests a potential test of inflation. Define the **tilt** of the fluctuation spectrum as follows:

$$\boxed{\text{tilt} \equiv 1 - n \equiv -\frac{d \ln \delta_{\text{H}}^2}{d \ln k}.} \quad (164)$$

We then want to express the tilt in terms of parameters of the inflationary potential, ϵ and η . These are of order unity when inflation terminates; ϵ and η must therefore be evaluated when the observed universe left the horizon, recalling that we only observe the last 60-odd e -foldings of inflation. The way to introduce scale dependence is to write the condition for a mode of given comoving wavenumber to cross the de Sitter horizon,

$$a/k = H^{-1}. \quad (165)$$

Since H is nearly constant during the inflationary evolution, we can replace $d/d \ln k$ by $d \ln a$, and use the slow-roll condition to obtain

$$\frac{d}{d \ln k} = a \frac{d}{da} = \frac{\dot{\phi}}{H} \frac{d}{d\phi} = -\frac{m_{\text{p}}^2}{8\pi} \frac{V'}{V} \frac{d}{d\phi}. \quad (166)$$

We can now work out the tilt, since the horizon-scale amplitude is

$$\delta_{\text{H}}^2 = \frac{H^4}{(2\pi\dot{\phi})^2} = \frac{128\pi}{3} \left(\frac{V^3}{m_{\text{p}}^6 V'^2} \right), \quad (167)$$

and derivatives of V can be expressed in terms of the dimensionless parameters ϵ and η . The tilt of the density perturbation spectrum is thus predicted to be

$$\boxed{1 - n = 6\epsilon - 2\eta} \quad (168)$$

In the section below on CMB anisotropies, we discuss whether this relation is observationally testable.

5 EVIDENCE FOR VACUUM ENERGY AT LATE TIMES

The idea of inflation is audacious, but undeniably speculative. However, once we accept the idea that quantum fields can generate an equation of state resembling a cosmological constant, we need not confine this mechanism to GUT-scale energies. There is no known mechanism that requires the minimum of $V(\phi)$ to lie exactly at zero energy, so it is quite plausible that there remains in the universe today some non-zero vacuum energy.

The most direct way of detecting vacuum energy has been the immense recent progress in the use of supernovae as standard candles. Type Ia SNe have been used as standard objects for around two decades, with an rms scatter in luminosity of 40%, and so a distance error of 20%. The big breakthrough came when it was realized that the intrinsic timescale of the SNe correlates with luminosity (brighter SNe last longer). Taking out this effect produces corrected standard candles that are capable of measuring distances to about 5% accuracy. Large search campaigns have made it possible to find of order 100 SNe over the range $0.1 \lesssim z \lesssim 1$, and two teams have used this strategy to make an empirical estimate of the cosmological distance-redshift relation.

The results of the *Supernova cosmology project* (e.g. Perlmutter et al. 1998) and the *High- z supernova search* (e.g. Riess et al. 1998) are highly consistent. Figure 7 shows the Hubble diagram from the latter team. The SNe magnitudes are K -corrected, so that their variation with redshift should be a direct measure of luminosity distance as a function of redshift.

We have seen above that this is written as the following integral, which must usually be evaluated numerically:

$$D_{\text{L}}(z) = (1+z)R_0 S_k(r) = (1+z) \frac{c}{H_0} |1-\Omega|^{-1/2} \times S_k \left[\int_0^z \frac{|1-\Omega|^{1/2} dz'}{\sqrt{(1-\Omega)(1+z')^2 + \Omega_v + \Omega_m(1+z')^3}} \right], \quad (169)$$

where $\Omega = \Omega_m + \Omega_v$, and S_k is sinh if $\Omega < 1$, otherwise sin. It is clear from figure 7 that the empirical distance-redshift relation is very different from the simplest inflationary prediction, which is the $\Omega = 1$ Einstein-de Sitter model; by redshift 0.6, the SNe are fainter than expected in this model by about 0.5 magnitudes. If this model fails, we can try adjusting Ω_m and Ω_v in an attempt to do better. Comparing each such model to the data yields the likelihood

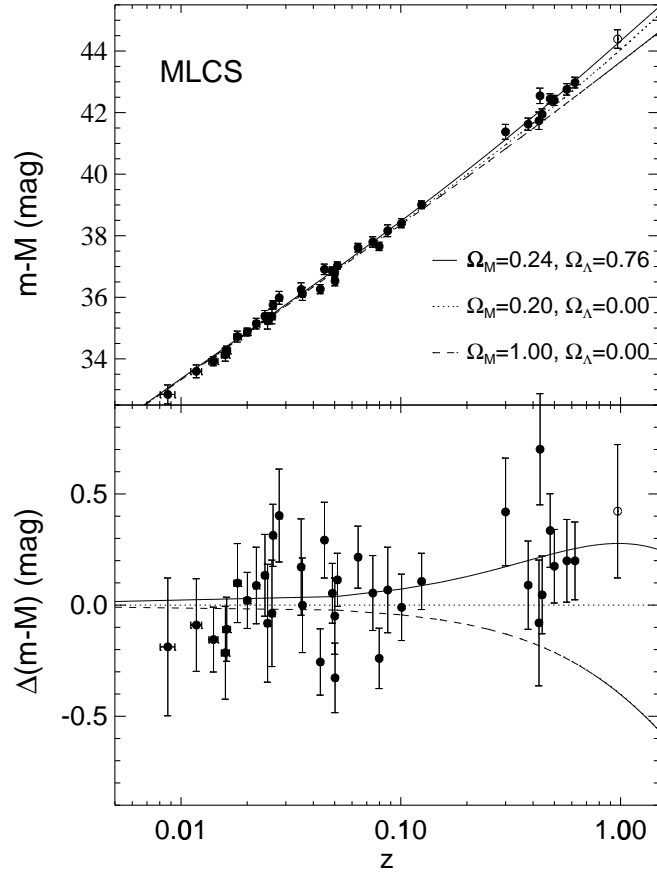


Fig. 7: The Hubble diagram produced by the High- z Supernova search team (Riess et al. 1998). The lower panel shows the data divided by a default model ($\Omega_m = 0.2$, $\Omega_v = 0$). The results lie clearly above this model, favouring a non-zero Λ . The lowest line is the Einstein-de Sitter model, which is in gross disagreement with observation.

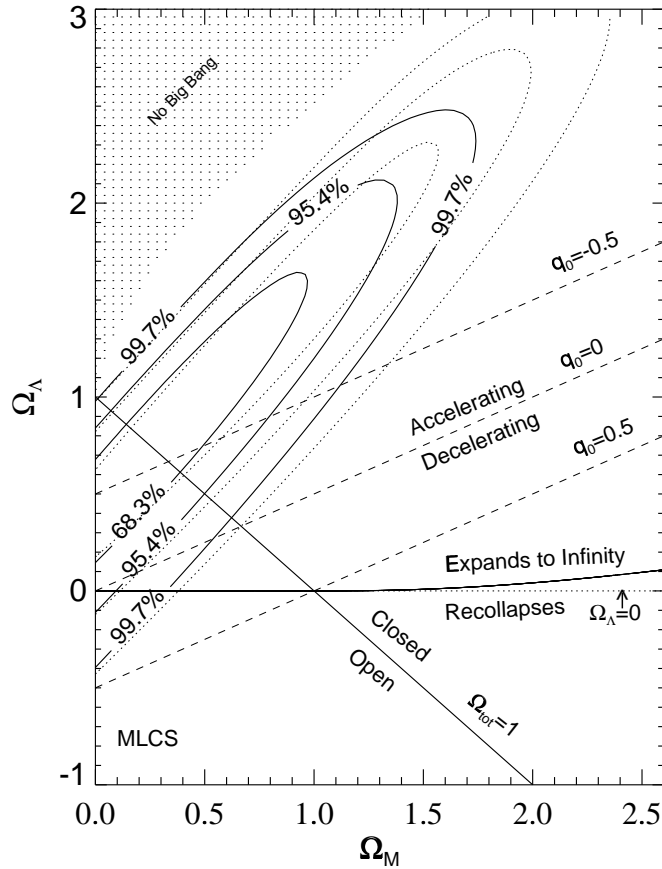


Fig. 8: Confidence contours on the Ω_v - Ω_m plane, according to Riess et al. (1998). Open models of all but the lowest densities are apparently ruled out, and nonzero Λ is strongly preferred. If we restrict ourselves to $k = 0$, then $\Omega_m \simeq 0.3$ is required. The constraints perpendicular to the $k = 0$ line are not very tight, but CMB data can help here in limiting the allowed degree of curvature.

contours shown in figure 8, which can be used in the standard way to set confidence limits on the cosmological parameters. The results very clearly require a low-density universe. For $\Lambda = 0$, a very low density is just barely acceptable, with $\Omega_m \lesssim 0.1$. However, the discussion of the CMB below shows that such a heavily open model is hard to sustain. The preferred model has $\Omega_v \simeq 1$; if we restrict ourselves to the inflationary $k = 0$, then the required parameters are very close to $(\Omega_m, \Omega_v) = (0.3, 0.7)$.

Cosmic coincidence This is an astonishing result – an observational detection of the physical reality of vacuum energy. The error bars continue to shrink, and no convincing systematic error has been suggested that could yield this result spuriously; this is one of the most important achievements of 20th-Century physics.

And yet, accepting the reality of vacuum energy raises a difficult question. If the universe contains a constant vacuum density and normal matter with $\rho \propto a^{-3}$, there is a unique epoch at which these two contributions cross over, and we seem to be living near to that time. This coincidence calls for some explanation. One might think of appealing to anthropic ideas, and these can limit Λ to some extent: if the universe became vacuum-dominated at $z > 1000$, gravitational instability as discussed in the next section would have been impossible – so that galaxies, stars and observers would not have been possible. On the other hand, Weinberg (1989)

argues that Λ could have been much larger than its actual value without making observers impossible. Efstathiou (1995) attempted to construct a probability distribution for Λ by taking this to be proportional to the number density of galaxies that result in a given model. However, there is no general agreement on how to set a probability measure for this problem.

It would be more satisfactory if we had some physical mechanism that guaranteed the coincidence, and one possibility has been suggested. We already have one coincidence, in that we live relatively close in time to the era of matter-radiation equality ($z \sim 10^3$, as opposed to $z \sim 10^{80}$ for the GUT era). What is required is a cosmological ‘constant’ that switches on around the equality era. Zlatev, Wang & Steinhardt (1998) have suggested how this might happen. The idea is to use the vacuum properties of a homogeneous scalar field as the physical origin of the negative-pressure term detected via SNe. This idea of a ‘rolling’ Λ was first explored by Ratra & Peebles (1988), and there has recently been a tendency towards use of the fanciful term ‘quintessence’. In any case, it is important to appreciate that the idea uses exactly the same physical elements that we discussed in the context of inflation: there is some $V(\phi)$, causing the expectation value of ϕ to obey the damped oscillator equation of motion, so the energy density and pressure are

$$\begin{aligned}\rho_\phi &= \dot{\phi}^2/2 + V \\ p_\phi &= \dot{\phi}^2/2 - V.\end{aligned}\tag{170}$$

This gives us two extreme equations of state: (i) vacuum-dominated, with $V \gg \dot{\phi}^2/2$, so that $p = -\rho$; (ii) kinetic-dominated, with $V \ll \dot{\phi}^2/2$, so that $p = \rho$. In the first case, we know that ρ does not alter as the universe expands, so the vacuum rapidly tends to dominate over normal matter. In the second case, the equation of state is the unusual $\Gamma = 2$, so we get the rapid behaviour $\rho \propto a^{-6}$. If a quintessence-dominated universe starts off with a large kinetic term relative to the potential, it may seem that things should always evolve in the direction of being potential-dominated. However, this ignores the detailed dynamics of the situation: for a suitable choice of potential, it is possible to have a **tracker field**, in which the kinetic and potential terms remain in a constant proportion, so that we can have $\rho \propto a^{-\alpha}$, where α can be anything we choose.

Putting this condition in the equation of motion shows that the potential is required to be exponential in form. More importantly, we can generalize to the case where the universe contains scalar field and ordinary matter. Suppose the latter dominates, and obeys $\rho_m \propto a^{-\alpha}$. It is then possible to have the scalar-field density obeying the same $\rho \propto a^{-\alpha}$ law, provided

$$V(\phi) = \frac{2}{\lambda^2}(6/\alpha - 1)\exp[-\lambda\phi].\tag{171}$$

The scalar-field density is $\rho_\phi = (\alpha/\lambda^2)\rho_{\text{total}}$ (see e.g. Liddle & Scherrer 1998). The impressive thing about this solution is that the quintessence density stays a fixed fraction of the total, whatever the overall equation of state: it automatically scales as a^{-4} at early times, switching to a^{-3} after matter-radiation equality.

This is not quite what we need, but it shows how the effect of the overall equation of state can affect the rolling field. Because of the $3H\dot{\phi}$ term in the equation of motion, ϕ ‘knows’ whether or not the universe is matter dominated. This suggests that a more complicated potential than the exponential may allow the arrival of matter domination to trigger the desired Λ -like behaviour. Zlatev, Wang & Steinhardt suggest two potentials which might achieve this:

$$V(\phi) = M^{4+\beta}\phi^{-\beta} \quad \text{or} \quad V(\phi) = M^4[\exp(m_P/\phi) - 1].\tag{172}$$

The evolution in these potentials may be described by $w(t)$, where $w = p/\rho$. We need $w \simeq 1/3$ in the radiation era, changing to $w \simeq -1$ today. The evolution in the inverse exponential potential

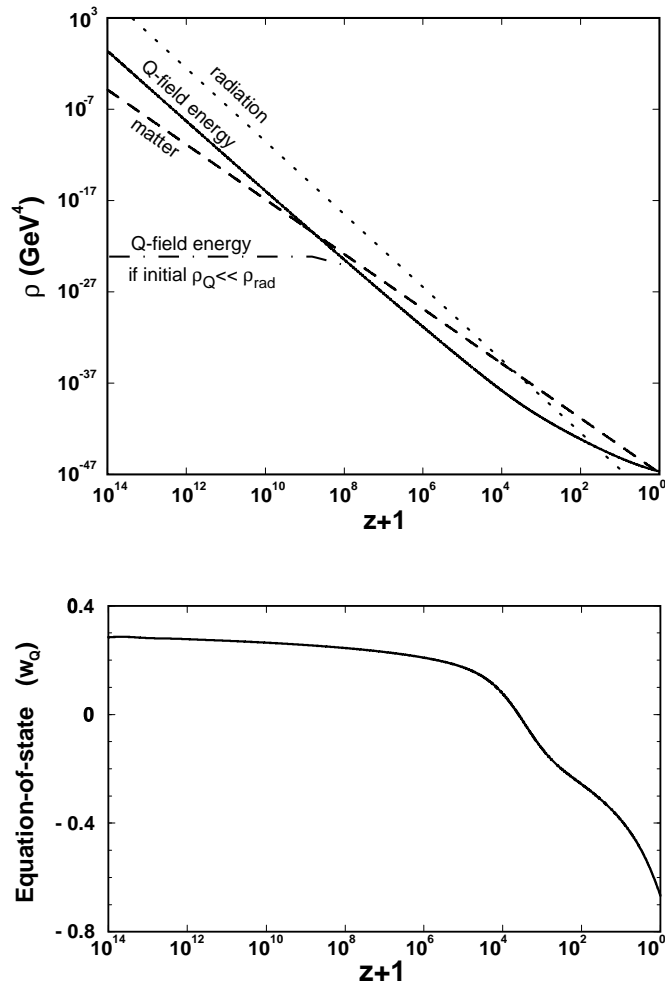


Fig. 9: This figure, taken from Zlatev, Wang & Steinhardt (1998), shows the evolution of the density in the ‘quintessence’ field (top panel), together with the effective equation of state of the quintessence vacuum (bottom panel), for the case of the inverse exponential potential. This allows vacuum energy to lurk at a few % of the total throughout the radiation era, but switching on a cosmological constant after the universe becomes matter dominated.

is shown in figure 9, demonstrating that the required behaviour can be found. However, a slight fine-tuning is still required, in that the trick only works for $M \sim 1$ meV, so there has to be an energy coincidence with the energy scale of matter-radiation equality.

So, the idea of tracker fields does not remove completely the puzzle concerning the level of present-day vacuum energy. In a sense, relegating the solution to a potential of unexplained form may seem a retrograde step. However, it is at least a testable step: the prediction of figure 9 is that $w \simeq -0.8$ today, so that the quintessence density scales as $\rho \propto a^{-0.6}$. This is a significant difference from the classical $w = -1$ vacuum energy, and it should be detectable as the SNe data improve. The existing data already require approximately $w < -0.5$, so there is the entrancing prospect that the equation of state for the vacuum will soon become the subject of experimental study.

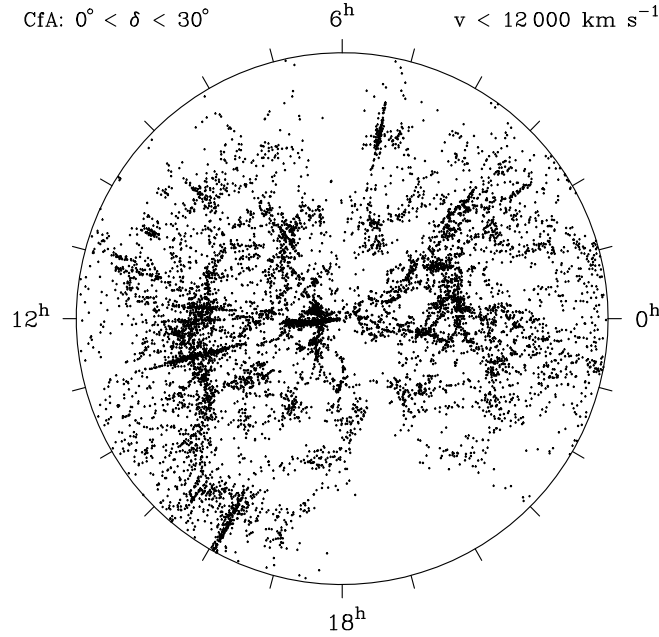


Fig. 10: One of the most dramatic pictures of the large-scale structure in the galaxy distribution is this slice made from the Harvard-Smithsonian Center for Astrophysics redshift survey to $B \simeq 15.5$. The survey coverage is not quite complete; as well as the holes due to the galactic plane around right ascensions 6^{h} and 19^{h} , the rich clusters are somewhat over-represented with respect to a true random sampling of the galaxy population. Nevertheless, this plot emphasizes nicely both the large-scale features such as the ‘great wall’ on the left, the totally empty void regions, and the radial ‘fingers of God’ caused by virialized motions in the clusters.

6 DYNAMICS OF STRUCTURE FORMATION

The overall properties of the universe are very close to being homogeneous; and yet telescopes reveal a wealth of detail on scales varying from single galaxies to **large-scale structures** of size exceeding 100 Mpc (see figure 10). The existence of these cosmological structures must be telling us something important about the initial conditions of the big bang, and about the physical processes that have operated subsequently.

The study of cosmological perturbations can be presented as a complicated exercise in linearized general relativity; fortunately, much of the essential physics can be extracted from a Newtonian approach. We start by writing down the fundamental equations governing fluid motion (non-relativistic for now):

$$\begin{aligned}
 \text{Euler : } \quad & \frac{D\mathbf{v}}{Dt} = -\frac{\nabla p}{\rho} - \nabla\Phi \\
 \text{energy : } \quad & \frac{D\rho}{Dt} = -\rho\nabla \cdot \mathbf{v} \\
 \text{Poisson : } \quad & \nabla^2\Phi = 4\pi G\rho,
 \end{aligned} \tag{173}$$

where $D/Dt = \partial/\partial t + \mathbf{v} \cdot \nabla$ is the usual convective derivative. We now produce the **linearized equations of motion** by collecting terms of first order in perturbations about a homogeneous background: $\rho = \rho_0 + \delta\rho$ etc. As an example, consider the energy equation:

$$[\partial/\partial t + (\mathbf{v}_0 + \delta\mathbf{v}) \cdot \nabla] (\rho_0 + \delta\rho) = -(\rho_0 + \delta\rho) \nabla \cdot (\mathbf{v}_0 + \delta\mathbf{v}). \tag{174}$$

For no perturbation, the zero-order equation is $(\partial/\partial t + \mathbf{v}_0 \cdot \nabla)\rho_0 = -\rho_0 \nabla \cdot \mathbf{v}_0$; since ρ_0 is homogeneous and $\mathbf{v}_0 = H\mathbf{x}$ is the Hubble expansion, this just says $\dot{\rho}_0 = -3H\rho_0$. Expanding the

full equation and subtracting the zeroth-order equation gives the equation for the perturbation:

$$(\partial/\partial t + \mathbf{v}_0 \cdot \nabla) \delta\rho + \delta\mathbf{v} \cdot \nabla(\rho_0 + \delta\rho) = -(\rho_0 + \delta\rho) \nabla \cdot \delta\mathbf{v} - \delta\rho \nabla \cdot \mathbf{v}_0. \quad (175)$$

Now, for sufficiently small perturbations, terms containing a product of perturbations such as $\delta\mathbf{v} \cdot \nabla \delta\rho$ must be negligible in comparison with the first-order terms. Remembering that ρ_0 is homogeneous leaves the linearized equation

$$[\partial/\partial t + \mathbf{v}_0 \cdot \nabla] \delta\rho = -\rho_0 \nabla \cdot \delta\mathbf{v} - \delta\rho \nabla \cdot \mathbf{v}_0. \quad (176)$$

It is straightforward to perform the same steps with the other equations; the results look simpler if we define the fractional density perturbation

$$\boxed{\delta \equiv \frac{\delta\rho}{\rho_0}}. \quad (177)$$

As above, when dealing with time derivatives of perturbed quantities, the full convective time derivative D/Dt can always be replaced by $d/dt \equiv \partial/\partial t + \mathbf{v}_0 \cdot \nabla$, which is the time derivative for an observer comoving with the unperturbed expansion of the universe. We then can write

$$\begin{aligned} \frac{d}{dt} \delta\mathbf{v} &= -\frac{\nabla \delta p}{\rho_0} - \nabla \delta\Phi - (\delta\mathbf{v} \cdot \nabla) \mathbf{v}_0 \\ \frac{d}{dt} \delta &= -\nabla \cdot \delta\mathbf{v} \\ \nabla^2 \delta\Phi &= 4\pi G \rho_0 \delta. \end{aligned} \quad (178)$$

There is now only one complicated term to be dealt with: $(\delta\mathbf{v} \cdot \nabla) \mathbf{v}_0$ on the rhs of the perturbed Euler equation. This is best attacked by writing it in components:

$$[(\delta\mathbf{v} \cdot \nabla) \mathbf{v}_0]_j = [\delta v]_i \nabla_i [v_0]_j = H [\delta v]_j, \quad (179)$$

where the last step follows because $\mathbf{v}_0 = H \mathbf{x}_0 \Rightarrow \nabla_i [v_0]_j = H \delta_{ij}$. This leaves a set of equations of motion that have no explicit dependence on the global expansion speed v_0 ; this is only present implicitly through the use of convective time derivatives d/dt .

These equations of motion are written in **Eulerian coordinates**: proper length units are used, and the Hubble expansion is explicitly present through the velocity \mathbf{v}_0 . The alternative approach is to use the comoving coordinates formed by dividing the Eulerian coordinates by the scale factor $a(t)$:

$$\boxed{\begin{aligned} \mathbf{x}(t) &= a(t) \mathbf{r}(t) \\ \delta\mathbf{v}(t) &= a(t) \mathbf{u}(t). \end{aligned}} \quad (180)$$

The next step is to translate spatial derivatives into comoving coordinates:

$$\nabla_x = \frac{1}{a} \nabla_r. \quad (181)$$

To keep the notation simple, subscripts on ∇ will normally be omitted hereafter, and spatial derivatives will be with respect to comoving coordinates. The linearized equations for conservation of momentum and matter as experienced by fundamental observers moving with the Hubble

flow then take the following simple forms in comoving units:

$$\boxed{\begin{aligned}\dot{\mathbf{u}} + 2\frac{\dot{a}}{a}\mathbf{u} &= \frac{\mathbf{g}}{a} - \frac{\nabla\delta p}{\rho_0} \\ \dot{\delta} &= -\nabla\cdot\mathbf{u},\end{aligned}} \quad (182)$$

where dots stand for d/dt . The peculiar gravitational acceleration $\nabla\delta\Phi/a$ is denoted by \mathbf{g} .

Before going on, it is useful to give an alternative derivation of these equations, this time working in comoving length units right from the start. First note that the comoving peculiar velocity \mathbf{u} is just the time derivative of the comoving coordinate \mathbf{r} :

$$\dot{\mathbf{x}} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}} = H\mathbf{x} + a\dot{\mathbf{r}}, \quad (183)$$

where the rhs must be equal to the Hubble flow $H\mathbf{x}$, plus the peculiar velocity $\delta\mathbf{v} = a\mathbf{u}$. In this equation, dots stand for exact convective time derivatives – *i.e.* time derivatives measured by an observer who follows a particle's trajectory – rather than partial time derivatives $\partial/\partial t$. This allows us to apply the continuity equation immediately in comoving coordinates, since this equation is simply a statement that particles are conserved, independent of the coordinates used. The exact equation is

$$\frac{D}{Dt}\rho_0(1+\delta) = -\rho_0(1+\delta)\nabla\cdot\mathbf{u}, \quad (184)$$

and this is easy to linearize because the background density ρ_0 is independent of time when comoving length units are used. This gives the first-order equation $\dot{\delta} = -\nabla\cdot\mathbf{u}$ immediately. The equation of motion follows from writing the Eulerian equation of motion as $\ddot{\mathbf{x}} = \mathbf{g}_0 + \mathbf{g}$, where $\mathbf{g} = \nabla\delta\Phi/a$ is the peculiar acceleration defined earlier, and \mathbf{g}_0 is the acceleration that acts on a particle in a homogeneous universe (neglecting pressure forces, for simplicity). Differentiating $\mathbf{x} = a\mathbf{r}$ twice gives

$$\ddot{\mathbf{x}} = a\dot{\mathbf{u}} + 2\dot{a}\mathbf{u} + \frac{\ddot{a}}{a}\mathbf{x} = \mathbf{g}_0 + \mathbf{g}. \quad (185)$$

The unperturbed equation corresponds to zero peculiar velocity and zero peculiar acceleration: $(\ddot{a}/a)\mathbf{x} = \mathbf{g}_0$; subtracting this gives the perturbed equation of motion $\mathbf{u} + 2(\dot{a}/a)\mathbf{u} = \mathbf{g}$, as before. This derivation is rather more direct than the previous route of working in Eulerian space. Also, it emphasizes that the equation of motion is exact, even though it happens to be linear in the perturbed quantities.

After doing all this, we still have three equations in the four variables δ , \mathbf{u} , $\delta\Phi$ and δp . The system needs an equation of state in order to be closed; this may be specified in terms of the sound speed

$$c_s^2 \equiv \frac{\partial p}{\partial \rho}. \quad (186)$$

Now think of a plane-wave disturbance $\delta \propto e^{-i\mathbf{k}\cdot\mathbf{r}}$, where \mathbf{k} is a comoving wavevector; in other words, suppose that the wavelength of a single Fourier mode stretches with the universe. All time dependence is carried by the amplitude of the wave, and so the spatial dependence can be factored out of time derivatives in the above equations (which would not be true with a constant comoving wavenumber k/a). An equation for the amplitude of δ can then be obtained by eliminating \mathbf{u} :

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \delta(4\pi G\rho_0 - c_s^2 k^2/a^2)}. \quad (187)$$

This equation is the one that governs the gravitational amplification of density perturbations.

There is a critical proper wavelength, known as the **Jeans length**, at which we switch from the possibility of exponential growth for long-wavelength modes to standing sound waves at short wavelengths. This critical length is

$$\boxed{\lambda_J = c_s \sqrt{\frac{\pi}{G\rho}}}, \quad (188)$$

and clearly delineates the scale at which sound waves can cross an object in about the time needed for gravitational free-fall collapse. When considering perturbations in an expanding background, things are more complex. Qualitatively, we expect to have no growth when the ‘driving term’ on the rhs is negative. However, owing to the expansion, λ_J will change with time, and so a given perturbation may switch between periods of growth and stasis.

Radiation-dominated universes At early enough times, the universe was radiation dominated ($c_s = c/\sqrt{3}$) and the analysis so far does not apply. It is common to resort to general relativity perturbation theory at this point. However, the fields are still weak, and so it is possible to generate the results we need by using special relativity fluid mechanics and Newtonian gravity with a relativistic source term. For simplicity, assume that accelerations due to pressure gradients are negligible in comparison with gravitational accelerations (*i.e.* restrict the analysis to $\lambda \gg \lambda_J$ from the start). The basic equations are then a simplified Euler equation and the full energy and gravitational equations:

$$\begin{aligned} \text{Euler : } & \frac{D\mathbf{v}}{Dt} = -\nabla\Phi \\ \text{energy : } & \frac{D}{Dt} (\rho + p/c^2) = \frac{\partial}{\partial t} (p/c^2) - (\rho + p/c^2) \nabla \cdot \mathbf{v} \\ \text{Poisson : } & \nabla^2\Phi = 4\pi G(\rho + 3p/c^2). \end{aligned} \quad (189)$$

For total radiation domination, $p = \rho c^2/3$, and it is easy to linearize these equations as before. The main differences come from factors of 2 and 4/3 due to the non-negligible contribution of the pressure. The result is a continuity equation $\nabla \cdot \mathbf{u} = -(3/4)\delta$, and the evolution equation for δ :

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = \frac{32\pi}{3}G\rho_0\delta}, \quad (190)$$

so the net result of all the relativistic corrections is a driving term on the rhs that is a factor 8/3 higher than in the matter-dominated case.

Solutions for $\delta(t)$ In both matter- and radiation-dominated universes with $\Omega = 1$, we have $\rho_0 \propto 1/t^2$:

$$\begin{aligned} \text{matter domination } (a \propto t^{2/3}) : & \quad 4\pi G\rho_0 = \frac{2}{3t^2} \\ \text{radiation domination } (a \propto t^{1/2}) : & \quad 32\pi G\rho_0/3 = \frac{1}{t^2}. \end{aligned} \quad (191)$$

Every term in the equation for δ is thus the product of derivatives of δ and powers of t , and a power-law solution is obviously possible. If we try $\delta \propto t^n$, then the result is $n = 2/3$ or -1

for matter domination; for radiation domination, this becomes $n = \pm 1$. For the growing mode, these can be combined rather conveniently using the **conformal time** $\eta \equiv \int dt/a$:

$$\boxed{\delta \propto \eta^2.} \quad (192)$$

Recall that η is proportional to the comoving size of the horizon.

The general case It is also interesting to think about the growth of matter perturbations in universes with nonzero vacuum energy, or even possibly some other exotic background with a peculiar equation of state. The differential equation for δ is as before, but $a(t)$ is altered. The way to deal with this is to treat a spherical perturbation as a small universe. Consider the Friedmann equation in the form

$$(\dot{a})^2 = \Omega_0^{\text{tot}} H_0^2 a^2 + K, \quad (193)$$

where $K = -kc^2/R_0^2$; this emphasizes that K is a constant of integration. A second constant of integration arises in the expression for time:

$$t = \int_0^a \dot{a}^{-1} da + C. \quad (194)$$

This lets us argue as before in the case of decaying modes: if a solution to the Friedmann equation is $a(t, K, C)$, then valid density perturbations are

$$\delta \propto \left(\frac{\partial \ln a}{\partial K} \right)_t \quad \text{or} \quad \left(\frac{\partial \ln a}{\partial C} \right)_t. \quad (195)$$

Since $\partial(\dot{a}^2)/\partial K = 1$, this gives the growing and decaying modes as

$$\boxed{\delta \propto \begin{cases} (\dot{a}/a) \int_0^a (\dot{a})^{-3} da & \text{(growing mode)} \\ (\dot{a}/a) & \text{(decaying mode)}. \end{cases}} \quad (196)$$

(Heath 1977; see also section 10 of Peebles 1980).

The equation for the growing mode requires numerical integration in general, with $\dot{a}(a)$ given by the Friedmann equation. A very good approximation to the answer is given by Carroll *et al.* (1992):

$$\boxed{\frac{\delta(z=0, \Omega)}{\delta(z=0, \Omega=1)} \simeq \frac{5}{2} \Omega_m \left[\Omega_m^{4/7} - \Omega_v + \left(1 + \frac{1}{2} \Omega_m\right) \left(1 + \frac{1}{70} \Omega_v\right) \right]^{-1}.} \quad (197)$$

This fitting formula for the growth suppression in low-density universes is an invaluable practical tool. For flat models with $\Omega_m + \Omega_v = 1$, it says that the growth suppression is less marked than for an open universe – approximately $\Omega^{0.23}$ as against $\Omega^{0.65}$ if $\Lambda = 0$. This reflects the more rapid variation of Ω_v with redshift; if the cosmological constant is important dynamically, this only became so very recently, and the universe spent more of its history in a nearly Einstein–de Sitter state by comparison with an open universe of the same Ω_m .

Mészáros effect What about the case of collisionless matter in a radiation background? The fluid treatment is not appropriate here, since the two species of particles can interpenetrate.

A particularly interesting limit is for perturbations well inside the horizon: the radiation can then be treated as a smooth, unclustered background that affects only the overall expansion rate. This is analogous to the effect of Λ , but an analytical solution does exist in this case. The perturbation equation is as before

$$\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi G\rho_m\delta, \quad (198)$$

but now $H^2 = 8\pi G(\rho_m + \rho_r)/3$. If we change variable to $y \equiv \rho_m/\rho_r = a/a_{\text{eq}}$, and use the Friedmann equation, then the growth equation becomes

$$\delta'' + \frac{2+3y}{2y(1+y)}\delta' - \frac{3}{2y(1+y)}\delta = 0 \quad (199)$$

(for $k = 0$, as appropriate for early times). It may be seen by inspection that a growing solution exists with $\delta'' = 0$:

$$\boxed{\delta \propto y + 2/3.} \quad (200)$$

It is also possible to derive the decaying mode. This is simple in the radiation-dominated case ($y \ll 1$): $\delta \propto -\ln y$ is easily seen to be an approximate solution in this limit.

What this says is that, at early times, the dominant energy of radiation drives the universe to expand so fast that the matter has no time to respond, and δ is frozen at a constant value. At late times, the radiation becomes negligible, and the growth increases smoothly to the Einstein–de Sitter $\delta \propto a$ behaviour (Mészáros 1974). The overall behaviour is therefore similar to the effects of pressure on a coupled fluid: for scales greater than the horizon, perturbations in matter and radiation can grow together, but this growth ceases once the perturbations enter the horizon. However, the explanations of these two phenomena are completely different. In the fluid case, the radiation pressure prevents the perturbations from collapsing further; in the collisionless case, the photons have free-streamed away, and the matter perturbation fails to collapse only because radiation domination ensures that the universe expands too quickly for the matter to have time to self-gravitate. Because matter perturbations enter the horizon (at $y = y_{\text{entry}}$) with $\dot{\delta} > 0$, δ is not frozen quite at the horizon-entry value, and continues to grow until this initial ‘velocity’ is redshifted away, giving a total boost factor of roughly $\ln y_{\text{entry}}$. This log factor may be seen below in the fitting formulae for the CDM power spectrum.

6.1 The peculiar velocity field

The equations for velocity-field perturbations were developed in section 5.2 as part of the machinery of analysing self-gravitating density fluctuations. There, the velocity field was eliminated, in order to concentrate on the behaviour of density perturbations. However, the peculiar velocity field is of great importance in cosmology, so it is convenient to give a summary that highlights the properties of velocity perturbations.

Consider first a galaxy that moves with some peculiar velocity in an otherwise uniform universe. Even though there is no peculiar gravitational acceleration acting, its velocity will decrease with time as the galaxy attempts to catch up with successively more distant (and therefore more rapidly receding) neighbours. If the proper peculiar velocity is v , then after time dt the galaxy will have moved a proper distance $x = v dt$ from its original location. Its near neighbours will now be galaxies with recessional velocities $Hx = Hv dt$, relative to which the peculiar velocity will have fallen to $v - Hx$. The equation of motion is therefore just

$$\dot{v} = -Hv = -\frac{\dot{a}}{a}v, \quad (201)$$

with the solution $v \propto a^{-1}$: peculiar velocities of nonrelativistic objects suffer redshifting by exactly the same factor as photon momenta. It is often convenient to express the peculiar velocity in terms of its comoving equivalent, $\mathbf{v} \equiv a \mathbf{u}$, for which the equation of motion becomes $\dot{\mathbf{u}} = -2H \mathbf{u}$. Thus, in the absence of peculiar accelerations and pressure forces, comoving peculiar velocities redshift away through the **Hubble drag** term $2H\mathbf{u}$.

If we now include the effects of peculiar acceleration, this simply adds the acceleration g on the right-hand side. This gives the equation of motion

$$\dot{\mathbf{u}} + \frac{2\dot{a}}{a} \mathbf{u} = -\frac{\mathbf{g}}{a}, \quad (202)$$

where $\mathbf{g} = \nabla\delta\Phi/a$ is the peculiar gravitational acceleration. Pressure terms have been neglected, so $\lambda \gg \lambda_j$. Remember that throughout we are using comoving length units, so that $\nabla_{\text{proper}} = \nabla/a$. This equation is the exact equation of motion for a single galaxy, so that the time derivative is $d/dt = \partial/\partial t + \mathbf{u} \cdot \nabla$. In linear theory, the second part of the time derivative can be neglected, and the equation then turns into one that describes the evolution of the linear peculiar velocity field at a fixed point in comoving coordinates.

The solutions for the peculiar velocity field can be decomposed into modes either parallel to \mathbf{g} or independent of \mathbf{g} (these are the homogeneous and inhomogeneous solutions to the equation of motion). The interpretation of these solutions is aided by knowing that the velocity field satisfies the **continuity equation**: $\dot{\rho} = -\nabla \cdot (\rho\mathbf{v})$ in proper units, which obviously takes the same form $\dot{\rho} = -\nabla \cdot (\rho\mathbf{u})$ if lengths and densities are in comoving units. If we express the density as $\rho = \rho_0(1 + \delta)$ (where in comoving units ρ_0 is just a number independent of time), the continuity equation takes the form

$$\dot{\delta} = -\nabla \cdot [(1 + \delta)\mathbf{u}], \quad (203)$$

which becomes just

$$\boxed{\nabla \cdot \mathbf{u} = -\dot{\delta}} \quad (204)$$

in linear theory when both δ and \mathbf{u} are small. This says that it is possible to have **vorticity modes** with $\nabla \cdot \mathbf{u} = 0$, for which $\dot{\delta}$ vanishes. We have already seen that δ either grows or decays as a power of time, so these modes require zero density perturbation, in which case the associated peculiar gravity also vanishes. These vorticity modes are thus the required homogeneous solutions, and they decay as $v = au \propto a^{-1}$, as with the kinematic analysis for a single particle. For any gravitational-instability theory, in which structure forms via the collapse of small perturbations laid down at very early times, it should therefore be a very good approximation to say that the linear velocity field must be curl-free.

For the growing modes, we want to try looking for a solution $\mathbf{u} = F(t)\mathbf{g}$. Then using continuity plus Gauss's theorem, $\nabla \cdot \mathbf{g} = 4\pi G a \rho \delta$, gives us

$$\delta\mathbf{v} = \frac{2f(\Omega)}{3H\Omega} \mathbf{g}, \quad (205)$$

where the function $f(\Omega) \equiv (a/\delta)d\delta/da$. A very good approximation to this (Peebles 1980) is $f \simeq \Omega^{0.6}$ (a result that is almost independent of Λ ; Lahav *et al.* 1991). Alternatively, we can work in Fourier terms. This is easy, as \mathbf{g} and \mathbf{k} are parallel, so that $\nabla \cdot \mathbf{u} = -i\mathbf{k} \cdot \mathbf{u} = -iku$. Thus, directly from the continuity equation,

$$\boxed{\delta\mathbf{v}_{\mathbf{k}} = -\frac{iHf(\Omega)a}{k} \delta_k \hat{\mathbf{k}}.} \quad (206)$$

The $1/k$ factor tells us that cosmological velocities come predominantly from larger-scale perturbations than those that dominate the density field. Deviations from the Hubble flow are therefore in principle a better probe of the inhomogeneity of the universe than large-scale clustering.

6.2 The Boltzmann equation

We now turn to the question of how to treat the matter source, without assuming that it is a fluid. The general approach that should be taken is to consider the phase-space distribution function $f(\mathbf{x}, \mathbf{p})$ – *i.e.* the product of the particle number density and the probability distribution for momentum. The equation that describes the evolution of f is the Boltzmann equation. The general relativistic form of the equation is

$$\left[p^\mu \frac{\partial}{\partial x^\mu} - \Gamma_{\alpha\beta}^\mu p^\alpha p^\beta \frac{\partial}{\partial p^\mu} \right] f = C. \quad (207)$$

This equation is exact for particles affected by gravitational forces and by collisions. The collision term on the rhs, C , has to contain all the appropriate scattering physics (Thomson scattering, in the case of a coupled system of electrons and photons); the gravitational forces are contained implicitly in the connection coefficients. What has to be done is to perturb this equation, using the perturbed metric coefficients, together with their equation of motion derived from the Einstein equations. Although this is easily said, the detailed algebra of the calculation consumes many pages, and it will not be reproduced here (see Peebles 1980; Efstathiou 1990; Bond 1997). The result is a system of coupled differential equations for the distribution functions of the non-fluid components (photons, neutrinos, plus possibly collisionless dark matter), together with the density and pressure of the collisional baryon fluid. Remembering to include gravitational waves, the whole system has to be integrated numerically, starting with a single Fourier mode of wavelength much greater than the horizon scale, and evolving to the present. Finally, the present-day perturbations to observational quantities such as density and radiation specific intensity are constructed by adding together modes of all wavelengths (which evolve independently in the linear approximation).

This, then, is a brief summary of the professional approach to cosmological perturbations. A modern cosmological Boltzmann code, such as that described by Seljak & Zaldarriaga (1996), is a large and sophisticated piece of machinery, which is the final outcome of decades of intellectual effort. Although heroic analytical efforts have been made in an attempt to find alternative methods of calculation (*e.g.* Hu & Sugiyama 1995), results of high precision demand the full approach. For a non-specialist, the best that can be done is to attempt to use simple approximate physical arguments in order to understand the main features of the results; on large scales, this approach is usually quantitatively successful.

6.3 Transfer functions

Real power spectra result from modifications of any primordial power by a variety of processes: growth under self-gravitation; the effects of pressure; dissipative processes. In general, modes of short wavelength have their amplitudes reduced relative to those of long wavelength in this way. The overall effect is encapsulated in the **transfer function**, which gives the ratio of the late-time amplitude of a mode to its initial value:

$$T_k \equiv \frac{\delta_k(z=0)}{\delta_k(z) D(z)}, \quad (208)$$

where $D(z)$ is the linear growth factor between redshift z and the present. The normalization redshift is arbitrary, so long as it refers to a time before any scale of interest has entered the

horizon. Once we possess the transfer function, it is a most valuable tool. The evolution of linear perturbations back to last scattering obeys the simple growth laws summarized above, and it is easy to see how structure in the universe will have changed during the matter-dominated epoch.

There are in essence two ways in which the power spectrum that exists at early times may differ from that which emerges at the present, both of which correspond to a reduction of small-scale fluctuations:

(1) Jeans mass effects. Prior to matter–radiation equality, we have already seen that perturbations inside the horizon are prevented from growing by radiation pressure. Once z_{eq} is reached, one of two things can happen. If collisionless dark matter dominates, perturbations on all scales can grow. If baryonic gas dominates, the Jeans length remains approximately constant, as follows: The sound speed, $c_s^2 = \partial p / \partial \rho$, may be found by thinking about the response of matter and radiation to small adiabatic compressions:

$$\delta p = (4/9)\rho_r c^2 (\delta V/V), \quad \delta \rho = [\rho_m + (4/3)\rho_r](\delta V/V), \quad (209)$$

implying

$$c_s^2 = c^2 \left(3 + \frac{9}{4} \frac{\rho_m}{\rho_r} \right)^{-1} = c^2 \left[3 + \frac{9}{4} \left(\frac{1 + z_{\text{rad}}}{1 + z} \right) \right]^{-1}. \quad (210)$$

Here, z_{rad} is the redshift of equality between matter and photons; $1 + z_{\text{rad}} = 1.68(1 + z_{\text{eq}})$ because of the neutrino contribution. At $z \ll z_{\text{rad}}$, we therefore have $c_s \propto \sqrt{1 + z}$. Since $\rho = (1 + z)^3 3\Omega_B H_0^2 / (8\pi G)$, the *comoving* Jeans length is constant at

$$\lambda_J = \frac{c}{H_0} \left(\frac{32\pi^2}{27\Omega_B(1 + z_{\text{rad}})} \right)^{1/2} = 50 (\Omega_B h^2)^{-1} \text{ Mpc}. \quad (211)$$

Thus, in either case, one of the critical length scales for the power spectrum will be the horizon distance at z_{eq} ($= 23900\Omega h^2$ for $T = 2.73$ K, counting neutrinos as radiation). In the matter-dominated approximation, we get

$$d_H = \frac{2c}{H_0} (\Omega z)^{-1/2} \Rightarrow d_{\text{eq}} = 39 (\Omega h^2)^{-1} \text{ Mpc}. \quad (212)$$

The exact distance–redshift relation is

$$R_0 dr = \frac{c}{H_0} \frac{dz}{(1 + z) \sqrt{1 + \Omega_m z + (1 + z)^2 \Omega_r}}, \quad (213)$$

from which it follows that the correct answer for the horizon size including radiation is a factor $\sqrt{2} - 1$ smaller: $d_{\text{eq}} = 16.0 (\Omega h^2)^{-1}$ Mpc.

It is easy from the above to see the approximate scaling that must be obeyed by transfer functions. Consider the adiabatic case first. Perturbations with $kd_{\text{eq}} \ll 1$ always undergo growth as $\delta \propto d_H^2$. Perturbations with larger k enter the horizon when $d_H \simeq 1/k$; they are then frozen until z_{eq} , at which point they can grow again. The missing growth factor is just the square of the change in d_H during this period, which is $\propto k^2$. The approximate limits of an adiabatic transfer function would therefore be

$$T_k \simeq \begin{cases} 1 & (kd_{\text{eq}} \ll 1) \\ [kd_{\text{eq}}]^{-2} & (kd_{\text{eq}} \gg 1). \end{cases} \quad (214)$$

For isocurvature perturbations, the situation is the opposite. Consider a perturbation of short wavelength: once it comes well inside the horizon, the photons disperse, and so all the perturbation to the entropy density (which must be conserved) is carried by the matter perturbation. The perturbation thus enters the horizon with the original amplitude δ_i . Thereafter, it grows in the same way as an isothermal perturbation. This means there are two regimes, for perturbations that enter the horizon before and after matter–radiation equality. The former match onto the Mészáros solution, and keep their amplitudes constant until they start to grow after a_{eq} . The present-day amplitude for these is $\delta/\delta_i = (3/2)[1/a_{\text{eq}}]$. Perturbations that enter after matter–radiation equality start to grow immediately, so that their present amplitude is $\delta/\delta_i \simeq 1/a_{\text{entry}}$. Entry occurs when $kd_{\text{H}} \simeq 1$, and the horizon evolves as $d_{\text{H}} = (2c/H_0)a^{1/2}$ (assuming $\Omega = 1$). Putting these arguments together, the isocurvature transfer function relative to δ_i is

$$T_k \simeq \begin{cases} (2/15) [kc/H_0]^2 & (kd_{\text{eq}} \ll 1) \\ (3/2) a_{\text{eq}}^{-1} & (kd_{\text{eq}} \gg 1) \end{cases} \quad (215)$$

(a more sophisticated argument is required to obtain the exact factor 2/15 in the long-wavelength limit; see Efstathiou 1990). Since this goes to a constant at high k , it is also common to quote the transfer function relative to this value. This means that $T_k < 1$ at $kd_{\text{eq}} \lesssim 1$, and so the isocurvature transfer function is the mirror image of the adiabatic case: one falls where the other rises (see figure 11).

(2) Damping. In addition to having their growth retarded, very small-scale perturbations will be erased entirely, which can happen in one of two ways. For collisionless dark matter, perturbations are erased simply by **free streaming**: random particle velocities cause blobs to disperse. At early times ($kT > mc^2$), the particles will travel at c , and so any perturbation that has entered the horizon will be damped. This process switches off when the particles become non-relativistic; for massive particles, this happens long before z_{eq} (resulting in **cold dark matter**; CDM CDM). For massive neutrinos, on the other hand, it happens *at* z_{eq} : only perturbations on very large scales survive in the case of **hot dark matter** HDM (HDM). In a purely baryonic universe, the corresponding process is called **Silk damping**: the mean free path of photons due to scattering by the plasma is non-zero, and so radiation can diffuse out of a perturbation, convecting the plasma with it. The typical distance of a random walk in terms of the diffusion coefficient D is $x \simeq \sqrt{Dt}$, which gives a damping length of

$$\lambda_{\text{S}} \simeq \sqrt{\lambda d_{\text{H}}}, \quad (216)$$

the geometric mean of the horizon size and the mean free path. Since $\lambda = 1/(n\sigma_{\text{T}}) = 44.3(1+z)^{-3}(\Omega_{\text{B}}h^2)^{-1}$ proper Gpc, we obtain a comoving damping length of

$$\lambda_{\text{S}} = 16.3(1+z)^{-5/4}(\Omega_{\text{B}}^2\Omega h^6)^{-1/4} \text{ Gpc}. \quad (217)$$

This becomes close to the Jeans length by the time of last scattering, $1+z \simeq 1000$.

It is invaluable in practice to have some accurate analytic formulae that fit the numerical results for transfer functions. We give below results for some common models of particular interest (illustrated in figure 11, along with other cases where a fitting formula is impractical). For the models with collisionless dark matter, $\Omega_{\text{B}} \ll \Omega$ is assumed, so that all lengths scale with the horizon size at matter–radiation equality, leading to the definition

$$q \equiv \frac{k}{\Omega h^2 \text{Mpc}^{-1}}. \quad (218)$$

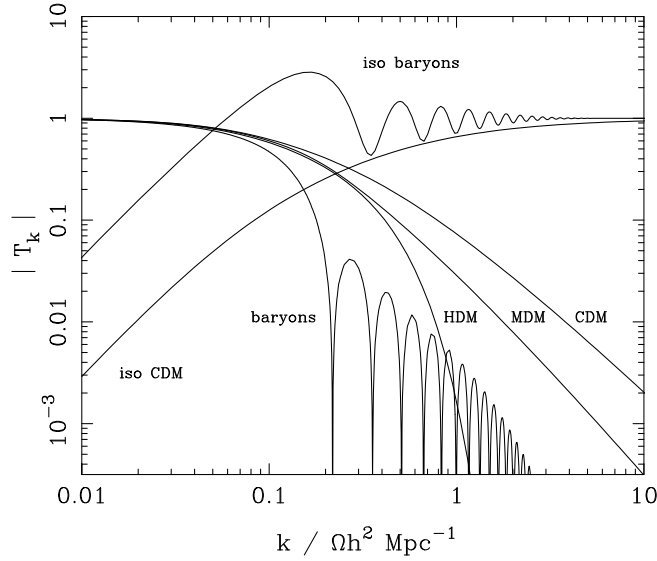


Fig. 11: A plot of transfer functions for various models. For adiabatic models, $T_k \rightarrow 1$ at small k , whereas the opposite is true for isocurvature models. A number of possible matter contents are illustrated: pure baryons; pure CDM; pure HDM; MDM (30% HDM, 70% CDM). For dark-matter models, the characteristic wavenumber scales proportional to Ωh^2 . The scaling for baryonic models does not obey this exactly; the plotted cases correspond to $\Omega = 1$, $h = 0.5$.

We consider the following cases: (1) Adiabatic CDM; (2) Adiabatic massive neutrinos (1 massive, 2 massless); (3) Isocurvature CDM; these expressions come from Bardeen *et al.* (1986; BBKS). Since the characteristic length-scale in the transfer function depends on the horizon size at matter-radiation equality, the temperature of the CMB enters. In the above formulae, it is assumed to be exactly 2.7 K; for other values, the characteristic wavenumbers scale $\propto T^{-2}$. For these purposes massless neutrinos count as radiation, and three species of these contribute a total density that is 0.68 that of the photons.

$$\begin{aligned}
 (1) \quad T_k &= \frac{\ln(1 + 2.34q)}{2.34q} \left[1 + 3.89q + (16.1q)^2 + (5.46q)^3 + (6.71q)^4 \right]^{-1/4} \\
 (2) \quad T_k &= \exp(-3.9q - 2.1q^2) \\
 (3) \quad T_k &= (5.6q)^2 \left(1 + \left[15.0q + (0.9q)^{3/2} + (5.6q)^2 \right]^{1.24} \right)^{-1/1.24}
 \end{aligned}
 \tag{219}$$

The case of **mixed dark matter** (MDM:MDM a mixture of massive neutrinos and CDM) is more complex. See Pogosyan & Starobinsky (1995) for a fit in this case.

The above expressions assume pure dark matter, which is unrealistic. At least for CDM models, a non-zero baryonic density lowers the apparent dark-matter density parameter. We can define an apparent shape parameter Γ for the transfer function:

$$\boxed{q \equiv (k/h \text{ Mpc}^{-1})/\Gamma,}
 \tag{220}$$

and $\Gamma = \Omega h$ in a model with zero baryon content. This parameter was originally defined by Efstathiou, Bond & White (1992), in terms of a CDM model with $\Omega_B = 0.03$. Peacock & Dodds (1994) showed that the effect of increasing Ω_B was to preserve the CDM-style spectrum shape,

but to shift to lower values of Γ . This shift was generalized to models with $\Omega \neq 1$ by Sugiyama (1995):

$$\Gamma = \Omega h \exp[-\Omega_{\text{B}}(1 + \sqrt{2h}/\Omega)]. \quad (221)$$

This formula fails if the baryon content is too large, and the transfer function develops oscillations (see Eisenstein & Hu 1998 for a more accurate approximation in this case).

6.4 The spherical model

An overdense sphere is a very useful nonlinear model, as it behaves in exactly the same way as a closed sub-universe. The density perturbation need not be a uniform sphere: any spherically symmetric perturbation will clearly evolve at a given radius in the same way as a uniform sphere containing the same amount of mass. In what follows, therefore, density refers to the *mean* density inside a given sphere. The equations of motion are the same as for the scale factor, and we can therefore write down the cycloid solution immediately. For a matter-dominated universe, the relation between the proper radius of the sphere and time is

$$\begin{aligned} r &= A(1 - \cos \theta) \\ t &= B(\theta - \sin \theta), \end{aligned} \quad (222)$$

and $A^3 = GMB^2$, just from $\ddot{r} = -GM/r^2$. Expanding these relations up to order θ^5 gives $r(t)$ for small t :

$$r \simeq \frac{A}{2} \left(\frac{6t}{B}\right)^{2/3} \left[1 - \frac{1}{20} \left(\frac{6t}{B}\right)^{2/3}\right], \quad (223)$$

and we can identify the density perturbation within the sphere:

$$\delta \simeq \frac{3}{20} \left(\frac{6t}{B}\right)^{2/3}. \quad (224)$$

This all agrees with what we knew already: at early times the sphere expands with the $a \propto t^{2/3}$ Hubble flow and density perturbations grow proportional to a .

We can now see how linear theory breaks down as the perturbation evolves. There are three interesting epochs in the final stages of its development, which we can read directly from the above solutions. Here, to keep things simple, we compare only with linear theory for an $\Omega = 1$ background.

- (1) **Turnround.** The sphere breaks away from the general expansion and reaches a maximum radius at $\theta = \pi$, $t = \pi B$. At this point, the true density enhancement with respect to the background is just $[A(6t/B)^{2/3}/2]^3/r^3 = 9\pi^2/16 \simeq 5.55$.
- (2) **Collapse.** If only gravity operates, then the sphere will collapse to a singularity at $\theta = 2\pi$. This occurs when $\delta_{\text{lin}} = (3/20)(12\pi)^{2/3} \simeq 1.69$.
- (3) **Virialization.** Consider the time at which the sphere has collapsed by a factor 2 from maximum expansion. At this point, it has kinetic energy K related to potential energy V by $V = -2K$. This is the condition for equilibrium, according to the **virial theorem**. For this reason, many workers take this epoch as indicating the sort of density contrast to be expected as the endpoint of gravitational collapse. This occurs at $\theta = 3\pi/2$, and the corresponding density enhancement is $(9\pi + 6)^2/8 \simeq 147$, with $\delta_{\text{lin}} \simeq 1.58$. Some authors prefer to assume that this virialized size is eventually achieved only at collapse, in which case the contrast becomes $(6\pi)^2/2 \simeq 178$.

These calculations are the basis for a common ‘rule of thumb’, whereby one assumes that linear theory applies until δ_{lin} is equal to some δ_c a little greater than unity, at which point virialization is deemed to have occurred. Although the above only applies for $\Omega = 1$, analogous results can be worked out from the full $\delta_{\text{lin}}(z, \Omega)$ and $t(z, \Omega)$ relations; $\delta_{\text{lin}} \simeq 1$ is a good criterion for collapse for any value of Ω likely to be of practical relevance. The full density contrast at virialization may be approximated by

$$1 + \delta_{\text{vir}} \simeq 178 \Omega^{-0.7} \quad (225)$$

(although flat Λ -dominated models show less dependence on Ω ; Eke *et al.* 1996).

7 COSMOLOGICAL DENSITY FIELDS

The next step is to see how the above theoretical ideas can be confronted with statistical measures of the observed matter distribution, and to summarize what is known about the dimensionless **density perturbation field**

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x}) - \langle \rho \rangle}{\langle \rho \rangle}. \quad (226)$$

This quantity need not be assumed to be small. Indeed, some of the most interesting issues arise in understanding the evolution of the density field to large values of δ .

A critical feature of the δ field is that it inhabits a universe that is isotropic and homogeneous in its large-scale properties. This suggests that the statistical properties of δ should also be homogeneous, even though it is a field that describes inhomogeneities. This statement sounds contradictory, and yet it makes perfect sense if there exists an **ensemble of universes**. The concept of an ensemble is used every time we apply probability theory to an event such as tossing a coin: we imagine an infinite sequence of repeated trials, half of which result in heads, half in tails. To say that the probability of heads is 1/2 means that the coin lands heads up in half the members of this ensemble of universes. The analogy of coin tossing in cosmology is that the density at a given point in space will have different values in each member of the ensemble, with some overall variance $\langle \delta^2 \rangle$ between members of the ensemble. Statistical homogeneity of the δ field then means that this variance must be independent of position. The actual field found in a given member of the ensemble is a **realization** of the statistical process.

There are two problems with this line of argument: (i) we have no evidence that the ensemble exists; (ii) in any case, we only get to observe one realization, so how is the variance $\langle \delta^2 \rangle$ to be measured? The first objection applies to coin tossing, and may be evaded if we understand the physics that generates the statistical process – we only need to *imagine* tossing the coin many times, and we do not actually need to perform the exercise. The best that can be done in answering the second objection is to look at widely separated parts of space, since the δ fields there should be causally unconnected; this is therefore as good as taking measurements from two different member of the ensemble. In other words, if we measure the variance $\langle \delta^2 \rangle$ by averaging over a sufficiently large volume, the results would be expected to approach the true ensemble variance, and the averaging operator $\langle \dots \rangle$ is often used without being specific about which kind of average is intended. Fields that satisfy this property, whereby

$$\text{volume average} \quad \leftrightarrow \quad \text{ensemble average} \quad (227)$$

are termed **ergodic**. Giving a formal proof of ergodicity for a random process is not always easy (Adler 1981); in cosmology it is perhaps best regarded as a common-sense axiom.

7.1 Fourier analysis of density fluctuations

It is often convenient to consider building up a general field by the superposition of many modes. For a flat comoving geometry, the natural tool for achieving this is via Fourier analysis. How do we make a Fourier expansion of the density field in an infinite universe? If the field were periodic within some box of side L , then we would just have a sum over wave modes:

$$F(\mathbf{x}) = \sum F_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}. \quad (228)$$

The requirement of periodicity restricts the allowed wavenumbers to **harmonic boundary conditions**

$$k_x = n \frac{2\pi}{L}, \quad n = 1, 2, \dots, \quad (229)$$

with similar expressions for k_y and k_z . Now, if we let the box become arbitrarily large, then the sum will go over to an integral that incorporates the density of states in k -space, exactly as in statistical mechanics. The Fourier relations in n dimensions are thus

$$\begin{aligned} F(x) &= \left(\frac{L}{2\pi}\right)^n \int F_k(k) \exp(-i\mathbf{k}\cdot\mathbf{x}) d^n k \\ F_k(k) &= \left(\frac{1}{L}\right)^n \int F(x) \exp(i\mathbf{k}\cdot\mathbf{x}) d^n x. \end{aligned} \quad (230)$$

Correlation functions and power spectra As an immediate example of the Fourier machinery in action, consider the important quantity

$$\xi(\mathbf{r}) \equiv \langle \delta(\mathbf{x}) \delta(\mathbf{x} + \mathbf{r}) \rangle, \quad (231)$$

which is the autocorrelation function of the density field – usually referred to simply as the **correlation function**. The angle brackets indicate an averaging over the normalization volume V . Now express δ as a sum and note that δ is real, so that we can replace one of the two δ 's by its complex conjugate, obtaining

$$\xi = \left\langle \sum_{\mathbf{k}} \sum_{\mathbf{k}'} \delta_{\mathbf{k}} \delta_{\mathbf{k}'}^* e^{i(\mathbf{k}' - \mathbf{k})\cdot\mathbf{x}} e^{-i\mathbf{k}\cdot\mathbf{r}} \right\rangle. \quad (232)$$

Alternatively, this sum can be obtained without replacing $\langle \delta\delta \rangle$ by $\langle \delta\delta^* \rangle$, from the relation between modes with opposite wavevectors that holds for any real field: $\delta_{\mathbf{k}}(-\mathbf{k}) = \delta_{\mathbf{k}}^*(\mathbf{k})$. Now, by the periodic boundary conditions, all the cross terms with $\mathbf{k}' \neq \mathbf{k}$ average to zero. Expressing the remaining sum as an integral, we have

$$\xi(\mathbf{r}) = \frac{V}{(2\pi)^3} \int |\delta_{\mathbf{k}}|^2 e^{-i\mathbf{k}\cdot\mathbf{r}} d^3 k. \quad (233)$$

In short, the correlation function is the Fourier transform of the **power spectrum**. This relation has been obtained by volume averaging, so it applies to the specific mode amplitudes and correlation function measured in any given realization of the density field. Taking ensemble

averages of each side, the relation clearly also holds for the ensemble average power and correlations – which are really the quantities that cosmological studies aim to measure. We shall hereafter often use the alternative notation

$$\boxed{P(k) \equiv \langle |\delta_k|^2 \rangle} \quad (234)$$

for the ensemble-average power. The distinction between the ensemble average and the actual power measured in a realization is clarified below in the section on Gaussian fields.

In an isotropic universe, the density perturbation spectrum cannot contain a preferred direction, and so we must have an **isotropic power spectrum**: $\langle |\delta_{\mathbf{k}}|^2(\mathbf{k}) \rangle = |\delta_k|^2(k)$. The angular part of the k -space integral can therefore be performed immediately: introduce spherical polars with the polar axis along \mathbf{k} , and use the reality of ξ so that $e^{-i\mathbf{k}\cdot\mathbf{x}} \rightarrow \cos(kr \cos \theta)$. In three dimensions, this yields

$$\xi(r) = \frac{V}{(2\pi)^3} \int P(k) \frac{\sin kr}{kr} 4\pi k^2 dk. \quad (235)$$

The 2D analogue of this formula is

$$\xi(r) = \frac{A}{(2\pi)^2} \int P(k) J_0(kr) 2\pi k dk. \quad (236)$$

We shall usually express the power spectrum in dimensionless form, as the variance per $\ln k$ ($\Delta^2(k) = d\langle \delta^2 \rangle / d \ln k \propto k^3 P[k]$):

$$\boxed{\Delta^2(k) \equiv \frac{V}{(2\pi)^3} 4\pi k^3 P(k) = \frac{2}{\pi} k^3 \int_0^\infty \xi(r) \frac{\sin kr}{kr} r^2 dr.} \quad (237)$$

This gives a more easily visualizable meaning to the power spectrum than does the quantity $VP(k)$, which has dimensions of volume: $\Delta^2(k) = 1$ means that there are order-unity density fluctuations from modes in the logarithmic bin around wavenumber k . $\Delta^2(k)$ is therefore the natural choice for a Fourier-space counterpart to the dimensionless quantity $\xi(r)$.

Power-law spectra The above shows that the power spectrum is a central quantity in cosmology, but how can we predict its functional form? For decades, this was thought to be impossible, and so a minimal set of assumptions was investigated. In the absence of a physical theory, we should not assume that the spectrum contains any preferred length scale, otherwise we should then be compelled to explain this feature. Consequently, the spectrum must be a featureless power law:

$$\boxed{\langle |\delta_k|^2 \rangle \propto k^n} \quad (238)$$

The index n governs the balance between large- and small-scale power. The meaning of different values of n can be seen by imagining the results of filtering the density field by passing over it a box of some characteristic comoving size x and averaging the density over the box. This will filter out waves with $k \gtrsim 1/x$, leaving a variance $\langle \delta^2 \rangle \propto \int_0^{1/x} k^n 4\pi k^2 dk \propto x^{-(n+3)}$. Hence, in terms of a mass $M \propto x^3$, we have

$$\delta_{\text{rms}} \propto M^{-(n+3)/6}. \quad (239)$$

Similarly, a power-law spectrum implies a power-law correlation function. If $\xi(r) = (r/r_0)^{-\gamma}$, with $\gamma = n + 3$, the corresponding 3D power spectrum is

$$\Delta^2(k) = \frac{2}{\pi} (kr_0)^\gamma \Gamma(2 - \gamma) \sin \frac{(2 - \gamma)\pi}{2} \equiv \beta(kr_0)^\gamma \quad (240)$$

($= 0.903(kr_0)^{1.8}$ if $\gamma = 1.8$). This expression is only valid for $n < 0$ ($\gamma < 3$); for larger values of n , ξ must become negative at large r (because $P(0)$ must vanish, implying $\int_0^\infty \xi(r) r^2 dr = 0$). A cutoff in the spectrum at large k is needed to obtain physically sensible results.

The Zeldovich spectrum Most important of all is the **scale-invariant spectrum**, which corresponds to the value $n = 1$, *i.e.* $\Delta^2 \propto k^4$. To see how the name arises, consider a perturbation $\delta\Phi$ in the gravitational potential:

$$\nabla^2 \delta\Phi = 4\pi G \rho_0 \delta \quad \Rightarrow \quad \delta\Phi_k = -4\pi G \rho_0 \delta_k / k^2. \quad (241)$$

The two powers of k pulled down by ∇^2 mean that, if $\Delta^2 \propto k^4$ for the power spectrum of density fluctuations, then Δ_{Φ}^2 is a constant. Since potential perturbations govern the flatness of spacetime, this says that the scale-invariant spectrum corresponds to a metric that is a **fractal**: spacetime has the same degree of ‘wrinkliness’ on each resolution scale. The total curvature fluctuations diverge, but only logarithmically at either extreme of wavelength.

Another way of looking at this spectrum is in terms of perturbation growth balancing the scale dependence of δ : $\delta \propto x^{-(n+3)/2}$. We know that δ viewed on a given comoving scale will increase with the size of the horizon: $\delta \propto r_{\text{H}}^2$. At an arbitrary time, though, the only natural length provided by the universe (in the absence of non-gravitational effects) is the horizon itself:

$$\delta(r_{\text{H}}) \propto r_{\text{H}}^2 r_{\text{H}}^{-(n+3)/2} = r_{\text{H}}^{-(n-1)/2}. \quad (242)$$

Thus, if $n = 1$, the growth of both r_{H} and δ with time cancels out so that the universe always looks the same when viewed on the scale of the horizon; such a universe is self-similar in the sense of always appearing the same under the magnification of cosmological expansion. This spectrum is often known as the **Zeldovich spectrum** (sometimes hyphenated with Harrison and Peebles, who invented it independently).

Filtering and moments A common concept in the manipulation of cosmological density fields is that of **filtering**, where the density field is convolved with some **window function**: $\delta \rightarrow \delta * f$. Many observable results can be expressed in this form. Some common 3D filter functions are Gaussian filter top-hat filter

$$\begin{aligned} \text{Gaussian : } f &= \frac{V}{(2\pi)^{3/2} R_{\text{G}}^3} e^{-r^2/2R_{\text{G}}^2} \Rightarrow f_k = e^{-k^2 R_{\text{G}}^2/2} \\ \text{top-hat : } f &= \frac{3V}{4\pi R_{\text{T}}^3} \quad (r < R_{\text{T}}) \Rightarrow f_k = \frac{3}{y^3} (\sin y - y \cos y) \quad (y \equiv k R_{\text{T}}). \end{aligned} \quad (243)$$

Note the factor of V in the definition of f ; this is needed to cancel the $1/V$ in the definition of convolution. For some power spectra, the difference in these filter functions at large k is unimportant, and we can relate them by equating the expansions near $k = 0$, where $1 - |f_k|^2 \propto k^2$. This equality requires

$$R_{\text{T}} = \sqrt{5} R_{\text{G}}. \quad (244)$$

We are often interested not in the convolved field itself, but in its variance, for use as a statistic (*e.g.* to measure the rms fluctuations in the number of objects in a cell). By the

convolution theorem, this means we are interested in a **moment** of the power spectrum times the squared filter transform. We shall generally use the following notation:

$$\sigma_n^2 \equiv \frac{V}{(2\pi)^3} \int P(k) |f_k|^2 k^{2n} d^3k; \quad (245)$$

the filtered variance is thus σ_0^2 (often denoted by just σ^2). Clustering results are often published in the form of **cell variances** of δ as a function of scale, σ^2 , using either cubical cells of side ℓ (Efstathiou *et al.* 1990) or Gaussian spheres of radius R_G (Saunders *et al.* 1991). For a power-law spectrum ($\Delta^2 \propto k^{n+3}$), we have for the Gaussian sphere

$$\sigma^2 = \Delta^2 \left(k = \left[\frac{1}{2} \left(\frac{n+1}{2} \right)! \right]^{1/(n+3)} R_G^{-1} \right). \quad (246)$$

For $n \lesssim 0$, this formula also gives a good approximation to the case of cubical cells, with $R_G \rightarrow \ell/\sqrt{12}$. The result is rather insensitive to assumptions about the power spectrum, and just says that the variance in a cell is mainly probing waves with $\lambda \simeq 2\ell$.

Moments may also be expressed in terms of the correlation function over the sample volume:

$$\sigma^2 = \iint \xi(|\mathbf{x} - \mathbf{x}'|) f(\mathbf{x}) f(\mathbf{x}') d^3x d^3x'. \quad (247)$$

To prove this, it is easiest to start from the definition of σ^2 as an integral over the power spectrum times $|f_k|^2$, write out the Fourier representations of P and f_k and use $\int \exp[i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}' + \mathbf{r})] d^3k = (2\pi)^3 \delta_D^{(3)}(\mathbf{x} - \mathbf{x}' + \mathbf{r})$. Finally, it is also sometimes convenient to express things in terms of derivatives of the correlation function at zero lag. Odd derivatives vanish, but even derivatives give

$$\xi^{(2n)}(0) = (-1)^n \frac{\sigma_n^2}{2n+1}. \quad (248)$$

Normalization For scale-invariant spectra, a natural amplitude measure is the variance in gravitational potential per unit $\ln k$, which is a constant, independent of scale:

$$\epsilon^2 \equiv \frac{\Delta_\Phi^2}{c^4} = \frac{9}{4} \left(\frac{ck}{H_0} \right)^{-4} \Delta^2(k). \quad (249)$$

Two further commonly encountered measures relate to the clustering field around 10 Mpc. One is σ_8 , the rms density variation when smoothed with a **top-hat filter** (sphere of uniform weight) of radius $8h^{-1}$ Mpc; this is observed to be very close to unity. The other is an integral over the correlation function: J_3

$$J_3 \equiv \int_0^r \xi(y) y^2 dy = \int \Delta^2(k) W(k) \frac{dk}{k}, \quad (250)$$

where $W(k) = (\sin kr - kr \cos kr)/k^3$. The canonical value of this is $J_3(10 h^{-1} \text{ Mpc}) = 277h^{-3} \text{ Mpc}^3$ (from the CfA survey; see Davis & Peebles 1983). It is sometimes more usual to use instead the dimensionless **volume-averaged correlation function** $\bar{\xi}$:

$$\bar{\xi}(r) = \frac{3}{4\pi r^3} \int_0^r \xi(x) 4\pi x^2 dx = \frac{3}{r^3} J_3(r). \quad (251)$$



Fig. 12: The N -point correlation functions of a density field consisting of a set of particles are calculated by looking at a set of cells of volume dV (so small that they effectively only ever contain 0 or 1 particles). The Poisson probability that two cells at separation r_{12} are both occupied is $\rho_0^2 dV_1 dV_2$; with clustering, this is modified by a factor $1 + \xi(r_{12})$, where ξ is the two-point correlation function. Similarly, the probability of finding a triplet of occupied cells is a factor $1 + \xi(r_{12}, r_{13}, r_{23})$ times the random probability; this defines the three-point correlation function.

The canonical value then becomes $\bar{\xi}(10 h^{-1} \text{ Mpc}) = 0.83$; this measure is clearly very close in content to $\sigma_8 = 1$.

A point to beware of is that the normalization of a theory is often quoted in terms of a value of these parameters extrapolated according to *linear* time evolution. Since the observed values are clearly nonlinear, there is no reason why theory and observation should match exactly. Even more confusingly, it is quite common in the literature to find the linear value of σ_8 called $1/b$, where b is a **bias parameter**. The implication is that $b \neq 1$ means that light does not follow mass; this may well be true in reality, but with this definition, nonlinearities will produce $b \neq 1$ even in models where mass traces light exactly. Use of this convention is not recommended.

7.2 N-point correlations

An alternative definition of the autocorrelation function is as the **two-point correlation function**, which gives the excess probability for finding a neighbour a distance r from a given galaxy (see figure 12). By regarding this as the probability of finding a pair with one object in each of the volume elements dV_1 and dV_2 ,

$$dP = \rho_0^2 [1 + \xi(r)] dV_1 dV_2, \quad (252)$$

this is easily seen to be equivalent to the autocorrelation definition of ξ : $\xi = \langle \delta(x_1)\delta(x_2) \rangle$. A related quantity is the **cross-correlation function**. Here, one considers two different classes of object (a and b, say), and the cross-correlation function ξ_{ab} is defined as the (symmetric) probability of finding a pair in which dV_1 is occupied by an object from the first catalogue and dV_2 by one from the second:

$$dP = \rho_a \rho_b [1 + \xi_{ab}(r)] dV_1 dV_2. \quad (253)$$

In terms of density fields, clearly $\xi_{ab} = \langle \delta_a(x_1)\delta_b(x_2) \rangle$. Cross-correlations give information about the density profile around objects; for example, ξ_{gc} between galaxies and clusters measures the average galaxy density profile around clusters (at least out to radii where clusters overlap).

7.3 Gaussian density fields

Apart from statistical isotropy of the fluctuation field, there is another reasonable assumption we might make: that the phases of the different Fourier modes δ_k are uncorrelated and random. This corresponds to treating the initial disturbances as some form of random noise, analogous

to Johnson noise in electrical circuits; indeed, many mathematical tools that have become invaluable in cosmology were first established with applications to communication circuits in mind (*e.g.* Rice 1954). The random-phase approximation has a powerful consequence, which derives from the **central limit theorem**: loosely, the sum of a large number of independent random variables will tend to be normally distributed. This will be true not just for the field δ ; all quantities that are derived from linear sums over waves (such as field derivatives) will be have a joint Normal distribution. The result is a **Gaussian random field**, whose properties are characterised entirely by its power spectrum.

Clustering of peaks Another important calculation that can be performed with density peaks is to estimate the clustering of cosmological objects. Peaks have some inbuilt clustering as a result of the statistics of the linear density field: they are ‘born clustered’. For galaxies, this clustering amplitude is greatly altered by the subsequent dynamical evolution of the density field, but this is not true for clusters of galaxies, which are the largest nonlinear systems at the current epoch. We recognize clusters simply because they are the most spectacularly large galaxy systems to have undergone gravitational collapse; this has an important consequence, as first realized by Kaiser (1984). The requirement that these systems have become nonlinear by the present means that they must have been associated with particularly high peaks in the initial conditions. If we thus confine ourselves to peaks above some **density threshold** in ν , the statistical correlations can be very strong – especially for the richer clusters corresponding to high peaks.

The main effect is easy to work out, using the **peak–background split**. Here, one conceptually decomposes the density field into short-wavelength terms, which generate the peaks, plus terms of much longer wavelength, which modulate the peak number density. Consider the large-wavelength field as if it were some extra perturbation δ_+ ; if we select all peaks above a threshold ν in the final field, this corresponds to taking all peaks above $\delta = \nu\sigma_0 - \delta_+$ in the initial field. This varying effective threshold will now produce more peaks in the regions of high δ_+ , leading to amplification of the clustering pattern. For high peaks, $P(> \nu) \propto \nu^2 e^{-\nu^2/2}$; the exponential is the most important term, leading to a perturbation $\delta P/P \simeq \nu(\delta_+/\sigma_0)$. Hence, we obtain the high-peak amplification factor for the correlation function:

$$\xi_{\text{pk}}(r) \simeq \frac{\nu^2}{\sigma_0^2} \xi_{\text{mass}}(r). \quad (254)$$

It is important to realize that the process as described need have nothing to do with biased galaxy formation; it works perfectly well if galaxy light traces mass exactly in the universe. Clusters occur at special places in the mass distribution, so there is no reason to expect their correlations to be the same as those of the mass field.

In more detail, the exact clustering of peaks is just an extension of the calculation of the number density of peaks. We want to find the density of peaks of height ν_2 at a distance r from a peak of height ν_1 . This involves a 6×6 covariance matrix for the fields and first and second derivatives even in 1D (20×20 in 3D). Moreover, most of the elements in this matrix are non-zero, so that the analytical calculation of ξ is sadly not feasible (see Lumsden, Heavens & Peacock 1989). However, a closely related calculation is easier to solve: the correlations of **thresholded regions**. Assume that objects form with unit probability in all regions whose density exceeds some threshold value, so that we need to deal with the correlation function of a

modified density field that is constant above the threshold and zero elsewhere. This is

$$1 + \xi_{>\nu}(r) = \frac{1}{[P(>\nu)]^2} \int_{\nu}^{\infty} \int_{\nu}^{\infty} \frac{dx dy}{2\pi[1 - \psi^2(r)]^{1/2}} \times \exp\left(-\frac{x^2 + y^2 - 2xy\psi(r)}{2[1 - \psi^2(r)]}\right) \quad (255)$$

(Kaiser 1984). For high thresholds, this should be very close to the correlation function of peaks. The complete solution of this equation is given by Jensen & Szalay (1986) (see also Kashlinsky 1991 for the extension to the cross-correlation of fields above different thresholds). A good approximation, which extends Kaiser's original result, is

$$1 + \xi_{>\nu} \simeq \exp\left(1 + \frac{\nu^2}{\sigma_0^2} \xi_{\text{mass}}\right). \quad (256)$$

There remains the question of the inclusion of dynamics into the above treatment. As the density field evolves, density peaks will move from their initial locations, and the clustering will alter. The general problem is rather nasty (see Bardeen *et al.* 1986), but things are relatively straightforward in the linear regime where the mass fluctuations are small. If the statistical enhancement of correlations produces a fractional perturbation in the numbers of thresholded objects of $\delta_{\text{statistical}} = f \delta_{\text{mass}}$, then the effect of allowing weak dynamical evolution is just

$$\delta_{\text{obs}} = \delta_{\text{statistical}} + \delta_{\text{mass}}. \quad (257)$$

To see this, think of density perturbations arising as in the Zeldovich approximation, via objects moving closer together. Density peaks will be convected with the flow and compressed in number density in the same way as for any other particle. Thus, the effective enhancement ends up as $f \rightarrow f + 1$. We can deduce the value of f for Kaiser's model by looking at the expression for the correlation function in the limit of small correlations: $f \simeq \nu/\sigma_0$. So, for large-scale correlations of high peaks, we expect

$$\xi_{\text{pk}} \simeq \left(1 + \frac{\nu}{\sigma_0}\right)^2 \xi_{\text{mass}}. \quad (258)$$

This idea of obtaining enhanced correlations by means of a threshold in density has been highly influential in cosmology. As well as the original application to clusters, attempts have also been made to use this mechanism to explain why galaxies might have clustering properties that differ from those of the mass.

Application to galaxy clusters This is the class of object that forms the main application of the peak clustering method. In order to model these systems as density peaks, it is necessary to specify a filter radius and a threshold; once we choose a filter radius to select cluster-sized fluctuations, the threshold is then fixed mainly by the number density (although altering the power-spectrum model also has a slight influence through γ). For Gaussian filtering, the conventional choice of R_f for clusters is $5h^{-1}$ Mpc. For $h = 1/2$ cold dark matter with $\gamma = 0.74$ on this scale, the required threshold is $\nu = 2.81$. These figures seem quite reasonable: Abell clusters are the rare high peaks of the mass distribution, and collapsed only recently. The reason for setting any threshold at all is the requirement of gravitational collapse by the present, so it is inevitable that $\nu \sim 1$.

The observations of the spatial correlations of clusters are somewhat controversial. The correlation function found by most workers is consistent with a scaled version of the galaxy

function, $\xi = (r/r_0)^{-1.8}$, but values of r_0 vary. The original value found by Bahcall & Soneira (1983) was $25 h^{-1}$ Mpc, but later work favoured values in the range $15 - 20 h^{-1}$ Mpc (Sutherland 1988; Dalton *et al.* 1992; Peacock & West 1992). The enhancement with respect to ξ for galaxies is thus a factor $\simeq 10$. Since σ_0 is close to unity for this smoothing, the simple asymptotic scaling would imply a threshold $\nu \simeq 3$, which is reasonable for moderately rare peaks.

7.4 Nonlinear clustering evolution

Observations of galaxy clustering extend into the highly nonlinear regime, $\xi \lesssim 10^4$, so it is essential to understand how this nonlinear clustering relates to the linear-theory initial conditions. A useful trick for dealing with this problem is to think of the density field under full nonlinear evolution as consisting of a set of collapsed, virialized clusters. What is the density profile of one of these objects? At least at separations smaller than the clump separation, the density profile of the clusters is directly related to the correlation function, since this just measures the number density of neighbours of a given galaxy. For a very steep cluster profile, $\rho \propto r^{-\epsilon}$, most galaxies will lie near the centres of clusters, and the correlation function will be a power law, $\xi(r) \propto r^{-\gamma}$, with $\gamma = \epsilon$. In general, because the correlation function is the convolution of the density field with itself, the two slopes differ. In the limit that clusters do not overlap, the relation is $\gamma = 2\epsilon - 3$ (for $3/2 < \epsilon < 3$; see Peebles 1974 or McClelland & Silk 1977). In any case, the critical point is that the correlation function may be thought of as arising directly from the density profiles of clumps in the density field.

In this picture, it is easy to see how ξ will evolve with redshift, since clusters are virialized objects that do not expand. The hypothesis of **stable clustering** states that, although the separation of clusters will alter as the universe expands, their internal density structure will stay constant with time. This hypothesis clearly breaks down in the outer regions of clusters, where the density contrast is small and linear theory applies, but it should be applicable to small-scale clustering. Regarding ξ as a density profile, its small-scale shape should therefore be fixed in *proper* coordinates, and its amplitude should scale as $(1+z)^{-3}$ owing to the changing mean density of unclustered galaxies, which dilute the clustering at high redshift. Thus, with $\xi \propto r^{-\gamma}$, we obtain the comoving evolution

$$\xi(r, z) \propto (1+z)^{\gamma-3} \quad (\text{nonlinear}). \quad (259)$$

Since the observed $\gamma \simeq 1.8$, this implies slower evolution than is expected in the linear regime:

$$\xi(r, z) \propto (1+z)^{-2} g(\Omega) \quad (\text{linear}). \quad (260)$$

This argument does not so far give a relation between the nonlinear slope γ and the index n of the linear spectrum. However, the linear and nonlinear regimes match at the scale of quasilinearity, *i.e.* $\xi(r_0) = 1$; each regime must make the same prediction for how this break point evolves. The linear and nonlinear predictions for the evolution of r_0 are respectively $r_0 \propto (1+z)^{-2/(n+3)}$ and $r_0 \propto (1+z)^{-(3-\gamma)/\gamma}$, so that $\gamma = (3n+9)/(n+5)$. In terms of an effective index $\gamma = 3 + n_{\text{NL}}$, this becomes

$$n_{\text{NL}} = -\frac{6}{5+n}. \quad (261)$$

The power spectrum resulting from power-law initial conditions will evolve self-similarly with this index. Note the narrow range predicted: $-2 < n_{\text{NL}} < -1$ for $-2 < n < +1$, with an $n = -2$ spectrum having the same shape in both linear and nonlinear regimes.

Indications from the angular clustering of faint galaxies (Efstathiou *et al.* 1991) and directly from redshift surveys (Le Fèvre *et al.* 1996) are that the observed clustering of galaxies evolves at about the linear-theory rate, rather more rapidly than the scaling solution would indicate. However, any interpretation of such data needs to assume that galaxies are unbiased tracers of the mass, whereas the observed high amplitude of clustering of quasars at $z \simeq 1$ ($r_0 \simeq 7 h^{-1}$ Mpc; see Shanks *et al.* 1987, Shanks & Boyle 1994) warns that at least some high-redshift objects have clustering that is apparently not due to gravity alone.

For many years it was thought that only these limiting cases of extreme linearity or nonlinearity could be dealt with analytically, but in a marvelous piece of alchemy, Hamilton *et al.* (1991; HKLM) suggested a general way of understanding the linear \leftrightarrow nonlinear mapping. The conceptual basis of their method can be understood with reference to the spherical collapse model. For $\Omega = 1$, a spherical clump virializes at a density contrast of order 100 when the linear contrast is of order unity. The trick now is to think about the density contrast in two distinct ways. To make a connection with the statistics of the density field, the correlation function $\xi(r)$ may be taken as giving a typical clump profile. What matters for collapse is that the integrated overdensity within a given radius reaches a critical value, so one should work with the volume-averaged correlation function $\bar{\xi}(r)$:

$$\bar{\xi}(R) \equiv \frac{3}{4\pi R^3} \int_0^R \xi(r) 4\pi r^2 dr. \quad (262)$$

A density contrast of $1 + \delta$ can also be thought of as arising through collapse by a factor $(1 + \delta)^{1/3}$ in radius, which suggests that a given nonlinear correlation $\bar{\xi}_{\text{NL}}(r_{\text{NL}})$ should be thought of as resulting from linear correlations on a linear scale:

$$r_{\text{L}} = [1 + \bar{\xi}_{\text{NL}}(r_{\text{NL}})]^{1/3} r_{\text{NL}}. \quad (263)$$

This is the first part of the **HKLM procedure**. Having performed this translation of scales, the second step is to conjecture that the nonlinear correlations are a universal function of the linear ones:

$$\bar{\xi}_{\text{NL}}(r_{\text{NL}}) = f_{\text{NL}}(\bar{\xi}_{\text{L}}(r_{\text{L}})). \quad (264)$$

The asymptotics of the function can be deduced readily. For small arguments $x \ll 1$, $f_{\text{NL}}(x) \simeq x$; the spherical collapse argument suggests $f_{\text{NL}}(1) \simeq 10^2$. Following collapse, $\bar{\xi}_{\text{NL}}$ depends on scale factor as a^3 (stable clustering), whereas $\bar{\xi}_{\text{L}} \propto a^2$; the large- x limit is therefore $f_{\text{NL}}(x) \propto x^{3/2}$. HKLM deduced from numerical experiments a numerical fit that interpolated between these two regimes, in a manner that empirically showed negligible dependence on the power spectrum.

However, these equations are often difficult to use stably for numerical evaluation; it is better to work directly in terms of power spectra. The key idea here is that $\bar{\xi}(r)$ can often be thought of as measuring the power at some effective wavenumber: it is obtained as an integral of the product of $\Delta^2(k)$, which is often a rapidly rising function, and a window function that cuts off rapidly at $k \gtrsim 1/r$:

$$\bar{\xi}(r) = \Delta^2(k_{\text{eff}}), \quad k_{\text{eff}} \simeq 2/r, \quad (265)$$

where n is the effective power-law index of the power spectrum. This approximation for the effective wavenumber is within 20 per cent of the exact answer over the range $-2 < n < 0$. In most circumstances, it is therefore an excellent approximation to use the HKLM formulae directly to scale wavenumbers and powers:

$$\begin{aligned} \Delta_{\text{NL}}^2(k_{\text{NL}}) &= f_{\text{NL}}(\Delta_{\text{L}}^2(k_{\text{L}})) \\ k_{\text{L}} &= [1 + \Delta_{\text{NL}}^2(k_{\text{NL}})]^{-1/3} k_{\text{NL}}. \end{aligned} \quad (266)$$

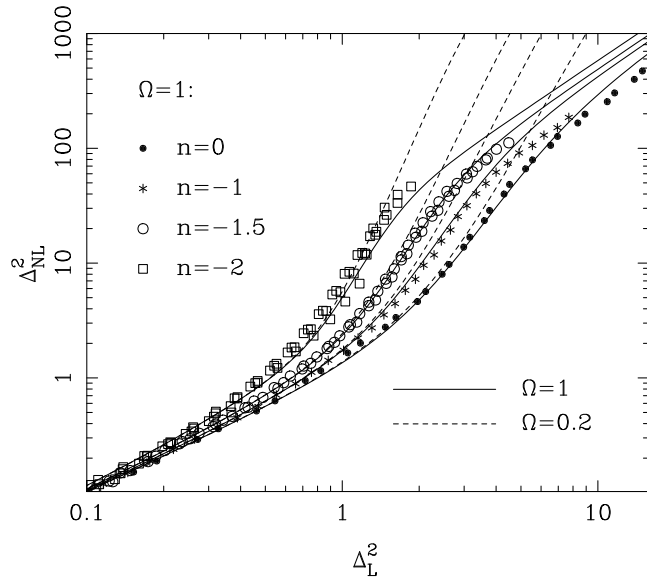


Fig. 13: The generalization of the HKLM function relating nonlinear power to linear power, for the cases $n = 0$, -1 , -1.5 and -2 . Data points are shown for the case $\Omega = 1$ only, with the corresponding fitting formulae shown as solid lines. This diagram clearly displays three regimes: (i) linear ($\Delta_{\text{NL}}^2 \lesssim 1$); (ii) quasilinear ($1 \lesssim \Delta_{\text{NL}}^2 \lesssim 100$); (iii) stable-clustering ($\Delta_{\text{NL}}^2 \gtrsim 100$). For a given linear power, the nonlinear power increases for more negative n . There is also a greater nonlinear response in the case of an open universe with $\Omega = 0.2$, indicated by the dashed lines. The fitting formula is shown for models with zero vacuum energy only, but what matters in general is the Ω -dependent linear growth suppression factor $g(\Omega)$.

What about models with $\Omega \neq 1$? The argument that leads to the $f_{\text{NL}}(x) \propto x^{3/2}$ asymptote in the nonlinear transformation is just that linear and nonlinear correlations behave as a^2 and a^3 respectively following collapse. If collapse occurs at high redshift, then $\Omega = 1$ may be assumed at that time, and the nonlinear correlations still obey the a^3 scaling to low redshift. All that has changed is that the linear growth is suppressed by some Ω -dependent factor $g(\Omega)$. According to Carroll, Press & Turner (1992), the required factor may be approximated almost exactly by

$$g(\Omega) = \frac{5}{2}\Omega_m \left[\Omega_m^{4/7} - \Omega_v + \left(1 + \frac{1}{2}\Omega_m\right)\left(1 + \frac{1}{70}\Omega_v\right) \right]^{-1}. \quad (267)$$

where we have distinguished matter (m) and vacuum (v) contributions to the density parameter explicitly. It then follows that the large- x asymptote of the nonlinear function is

$$f_{\text{NL}}(x) \propto [g(\Omega)]^{-3} x^{3/2}. \quad (268)$$

This says that the amplitude of highly nonlinear clustering is greater for low-density universes.

The suggestion of HKLM was that f_{NL} might be independent of the form of the linear spectrum, but Jain, Mo & White (1995) showed that this is not true, especially when the linear spectrum is rather flat ($n \lesssim -1.5$). Peacock & Dodds (1996) suggested that the HKLM method should be generalized by using the following fitting formula for the n -dependent nonlinear function (strictly, the one that applies to the power spectrum, rather than to ξ):

$$f_{\text{NL}}(x) = x \left[\frac{1 + B\beta x + [Ax]^{\alpha\beta}}{1 + ([Ax]^{\alpha} g^3(\Omega) / [Vx^{1/2}])^{\beta}} \right]^{1/\beta}. \quad (269)$$

B describes a second-order deviation from linear growth; A and α parameterize the power law that dominates the function in the quasilinear regime; V is the virialization parameter that gives

the amplitude of the $f_{\text{NL}}(x) \propto x^{3/2}$ asymptote; β softens the transition between these regimes. An excellent fit to N -body data (illustrated in figure 13) is given by the following spectrum dependence of the expansion coefficients:

$$\begin{aligned}
A &= 0.542 (1 + n/3)^{-0.685} \\
B &= 0.097 (1 + n/3)^{-0.224} \\
\alpha &= 3.235 (1 + n/3)^{-0.236} \\
\beta &= 0.659 (1 + n/3)^{-0.356} \\
V &= 11.54 (1 + n/3)^{-0.371}.
\end{aligned}
\tag{270}$$

The more general case of curved spectra can be dealt with very well by using the tangent spectral index at each linear wavenumber:

$$n_{\text{eff}} \equiv \frac{d \ln P}{d \ln k}, \tag{271}$$

although evaluating n_{eff} at a wavenumber of $k/2$ gives even better results.

Evolution of clustering The discussion of nonlinear evolution has revealed that in practice the regime $1 \lesssim \Delta^2 \lesssim 100$ is dominated by the steep **quasilinear transition** where $f_{\text{NL}}(x) \propto x^\alpha$, $\alpha \simeq 3.5$ – 4.5 . This turns out to predict a rate of evolution that is very different from the extremes of linear evolution or stable clustering. For $\Delta^2 \gg 1$, the transitional spectrum scales as

$$\begin{aligned}
k_{\text{NL}} &\simeq [\Delta_{\text{NL}}^2]^{1/3} k_{\text{L}} \\
\Delta_{\text{NL}}^2 &\propto [D^2(a) \Delta_{\text{L}}^2]^{1+\alpha},
\end{aligned}
\tag{272}$$

where $D(a)$ is the linear growth law for density perturbations. For a power-law linear spectrum, $\Delta^2 \propto k^{3+n}$, this predicts a quasilinear power law

$$\Delta_{\text{NL}}^2 \propto D^{(6-2\gamma)(1+\alpha)/3} k_{\text{NL}}^\gamma, \tag{273}$$

where the nonlinear power-law index depends as follows on the slope of the linear spectrum:

$$\gamma = \frac{3(3+n)(1+\alpha)}{3+(3+n)(1+\alpha)}. \tag{274}$$

For the observed index of $\gamma \simeq 1.8$, this would require $n \simeq -2.2$, very different from the $n = 0$ that would give $\gamma = 1.8$ in the virialized regime.

We can now summarize the rate of evolution of clustering in the three different regimes:

linear : $\xi(r, z) \propto [D(z)]^2$ quasilinear : $\xi(r, z) \propto [D(z)]^{(6-2\gamma)(1+\alpha)/3}$ nonlinear : $\xi(r, z) \propto (1+z)^{-(3-\gamma)}$,	$\tag{275}$
--	-------------

where γ is the power-law slope in the relevant regime. For $\alpha = 4$, $\gamma = 1.7$, this gives $\xi \propto D^{4.3}$ for the quasilinear evolution; this is more than twice as fast as the linear evolution, and over three times the rate of stable-clustering evolution if $\Omega = 1$, so that $D(z) = 1/(1+z)$. The conclusion is that clustering in the regime where most data exist is expected to evolve very rapidly with redshift, unless Ω is low. We discuss below whether this effect has been seen.

7.5 Real-space clustering

It is possible to avoid the complications of redshift space. One can deal with pure two-dimensional projected clustering, as discussed in the next section. Alternatively, peculiar velocities may be dealt with by using the correlation function evaluated explicitly as a 2D function of transverse (r_\perp) and radial (r_\parallel) separation. Integrating along the redshift axis then gives the **projected correlation function**, which is independent of the velocities

$$w_p(r_\perp) \equiv \int_{-\infty}^{\infty} \xi(r_\perp, r_\parallel) dr_\parallel = 2 \int_{r_\perp}^{\infty} \xi(r) \frac{r dr}{(r^2 - r_\perp^2)^{1/2}}. \quad (276)$$

In principle, this statistic can be used to recover the real-space correlation function by using the inverse relation for the **Abel integral equation**:

$$\xi(r) = -\frac{1}{\pi} \int_r^{\infty} w'_p(y) \frac{dy}{(y^2 - r^2)^{1/2}}. \quad (277)$$

An alternative notation for the projected correlation function is $\Xi(r_\perp)$ (Saunders, Rowan-Robinson & Lawrence 1992). Note that the projected correlation function is not dimensionless, but has dimensions of length. The quantity $\Xi(r_\perp)/r_\perp$ is more convenient to use in practice as the projected analogue of $\xi(r)$.

The reason that $w_p(r_\perp)$ is independent of redshift-space distortions is that peculiar velocities simply move pairs of points in r_\parallel , but not in r_\perp , and the expected pair count is just proportional to $2\pi r_\perp dr_\perp dr_\parallel$. Suppose we ignore the linear-theory velocities (which are more easily treated in Fourier space as above), and just consider the effect of a small-scale velocity dispersion. The correlation function is then convolved in the radial direction:

$$\begin{aligned} \xi(r_\perp, r_\parallel) &= \int_{-\infty}^{\infty} \xi_{\text{true}}(r_\perp, r) f(r_\parallel - r) dr \\ &= \frac{r_0^\gamma}{\sqrt{2\pi} \sigma_v} \int_{-\infty}^{\infty} [r_\perp^2 + (r_\parallel - x)^2]^{-\gamma/2} e^{-x^2/2\sigma_v^2} dx, \end{aligned} \quad (278)$$

where the latter expression applies for power-law clustering and a Gaussian velocity dispersion. Looking at the function in the redshift direction thus allows the pairwise velocity dispersion to be estimated; this is the origin of the above estimate of σ_p . See Fisher (1995) for more discussion of this method.

Sometimes these complications are neglected, and correlations are calculated in redshift space assuming isotropy. The result is a small increase in scalelength, as power on small scales is transferred to separations of order the velocity smearing. The result is a scale length around $7h^{-1}$ Mpc for the redshift-space $\xi(s)$ as opposed to the $5h^{-1}$ Mpc that applies for $\xi(r)$.

Projection on the sky A more common situation is where we lack any distance data; we then deal with a projection on the sky of a magnitude-limited set of galaxies at different depths. The statistic that is observable is the angular correlation function, $w(\theta)$, or its angular power spectrum Δ_θ^2 . If the sky were flat, the relation between these would be the usual **Hankel transform** pair:

$$\begin{aligned} w(\theta) &= \int_0^\infty \Delta_\theta^2 J_0(K\theta) dK/K \\ \Delta_\theta^2 &= K^2 \int_0^\infty w(\theta) J_0(K\theta) \theta d\theta. \end{aligned} \quad (279)$$

For power-law clustering, $w(\theta) = (\theta/\theta_0)^{-\epsilon}$, this gives

$$\Delta_\theta^2(K) = (K\theta_0)^\epsilon 2^{1-\epsilon} \frac{\Gamma(1-\epsilon/2)}{\Gamma(\epsilon/2)}, \quad (280)$$

which is equal to $0.77(K\theta_0)^\epsilon$ for $\epsilon = 0.8$. At large angles, these relations are not quite correct. We should really expand the sky distribution in **spherical harmonics**:

$$\delta(\hat{\mathbf{q}}) = \sum a_\ell^m Y_{\ell m}(\hat{\mathbf{q}}), \quad (281)$$

where $\hat{\mathbf{q}}$ is a unit vector that specifies direction on the sky. The functions $Y_{\ell m}$ are the eigenfunctions of the angular part of the ∇^2 operator: $Y_{\ell m}(\theta, \phi) \propto \exp(im\phi)P_\ell^m(\cos\theta)$, where P_ℓ^m are the **associated Legendre polynomials** (see *e.g.* section 6.8 of Press *et al.* 1992). Since the spherical harmonics satisfy the orthonormality relation $\int Y_{\ell m} Y_{\ell' m'}^* d^2q = \delta_{\ell\ell'}\delta_{mm'}$, the inverse relation is

$$a_\ell^m = \int \delta(\hat{\mathbf{q}}) Y_{\ell m}^* d^2q. \quad (282)$$

The analogues of the Fourier relations for the correlation function and power spectrum are

$$w(\theta) = \frac{1}{4\pi} \sum_\ell \sum_{m=-\ell}^{m=+\ell} |a_\ell^m|^2 P_\ell(\cos\theta)$$

$$|a_\ell^m|^2 = 2\pi \int_{-1}^1 w(\theta) P_\ell(\cos\theta) d\cos\theta.$$

(283)

For small θ and large ℓ , these go over to a form that looks like a flat sky, as follows. Consider the asymptotic forms for the Legendre polynomials and the J_0 Bessel function:

$$P_\ell(\cos\theta) \simeq \sqrt{\frac{2}{\pi\ell\sin\theta}} \cos\left[\left(\ell + \frac{1}{2}\right)\theta - \frac{1}{4}\pi\right]$$

$$J_0(z) \simeq \sqrt{\frac{2}{\pi z}} \cos\left[z - \frac{1}{4}\pi\right], \quad (284)$$

for respectively $\ell \rightarrow \infty$, $z \rightarrow \infty$; see chapters 8 & 9 of Abramowitz & Stegun 1965. This shows that, for $\ell \gg 1$, we can approximate the small-angle correlation function in the usual way in terms of an angular power spectrum Δ_θ^2 and angular wavenumber K :

$$w(\theta) = \int_0^\infty \Delta_\theta^2(K) J_0(K\theta) \frac{dK}{K}, \quad \Delta_\theta^2\left(K = \ell + \frac{1}{2}\right) = \frac{2\ell + 1}{8\pi} \sum_m |a_\ell^m|^2. \quad (285)$$

An important relation is that between the angular and spatial power spectra. In outline, this is derived as follows. The perturbation seen on the sky is

$$\delta(\hat{\mathbf{q}}) = \int_0^\infty \delta(\mathbf{y}) y^2 \phi(y) dy, \quad (286)$$

where $\phi(y)$ is the **selection function**, normalized such that $\int y^2 \phi(y) dy = 1$, and y is comoving distance. The function ϕ is the comoving density of objects in the survey, which is given by the integrated luminosity function down to the luminosity limit corresponding to the limiting flux of the survey seen at different redshifts; a flat universe ($\Omega = 1$) is assumed for now. Now write down the Fourier expansion of δ . The plane waves may be related to spherical harmonics via

the expansion of a plane wave in **spherical Bessel functions** $j_\ell(x) = (\pi/2x)^{1/2} J_{n+1/2}(x)$ (see chapter 10 of Abramowitz & Stegun 1965 or section 6.7 of Press *et al.* 1992):

$$e^{ikr \cos \theta} = \sum_0^\infty (2\ell + 1) i^\ell P_\ell(\cos \theta) j_\ell(kr), \quad (287)$$

plus the spherical harmonic addition theorem

$$P_\ell(\cos \theta) = \frac{4\pi}{2\ell + 1} \sum_{m=-\ell}^{m=+\ell} Y_{\ell m}^*(\hat{\mathbf{q}}) Y_{\ell m}(\hat{\mathbf{q}}'); \quad \hat{\mathbf{q}} \cdot \hat{\mathbf{q}}' = \cos \theta. \quad (288)$$

These relations allow us to take the angular correlation function $w(\theta) = \langle \delta(\hat{\mathbf{q}}) \delta(\hat{\mathbf{q}}') \rangle$ and transform it to give the angular power spectrum coefficients. The actual manipulations involved are not as intimidating as they may appear, but they are left as an exercise and we simply quote the final result (Peebles 1973):

$$\boxed{\langle |a_\ell^m|^2 \rangle = 4\pi \int \Delta^2(k) \frac{dk}{k} \left[\int y^2 \phi(y) j_\ell(ky) dy \right]^2.} \quad (289)$$

What is the analogue of this formula for small angles? Rather than manipulating large- ℓ Bessel functions, it is easier to start again from the correlation function. By writing as above the overdensity observed at a particular direction on the sky as a radial integral over the spatial overdensity, with a weighting of $y^2 \phi(y)$, we see that the angular correlation function is

$$\langle \delta(\hat{\mathbf{q}}_1) \delta(\hat{\mathbf{q}}_2) \rangle = \iint \langle \delta(\mathbf{y}_1) \delta(\mathbf{y}_2) \rangle y_1^2 y_2^2 \phi(y_1) \phi(y_2) dy_1 dy_2. \quad (290)$$

We now change variables to the mean and difference of the radii, $y \equiv (y_1 + y_2)/2$; $x \equiv (y_1 - y_2)$. If the depth of the survey is larger than any correlation length, we only get a signal when $y_1 \simeq y_2 \simeq y$. If the selection function is a slowly varying function, so that the thickness of the shell being observed is also of order the depth, the integration range on x may be taken as being infinite. For small angles, we then obtain **Limber's equation**:

$$\boxed{w(\theta) = \int_0^\infty y^4 \phi^2 dy \int_{-\infty}^\infty \xi \left(\sqrt{x^2 + y^2 \theta^2} \right) dx} \quad (291)$$

(see sections 51 and 56 of Peebles 1980). Theory usually supplies a prediction about the linear density field in the form of the power spectrum, and so it is convenient to recast Limber's equation:

$$w(\theta) = \int_0^\infty y^4 \phi^2 dy \int_0^\infty \pi \Delta^2(k) J_0(ky\theta) dk/k^2. \quad (292)$$

The form $\phi \propto y^{-1/2} \exp[-(y/y^*)^2]$ is often taken as a reasonable approximation to the Schechter function, and this gives

$$w(\theta) = \frac{\pi}{2\Gamma^2\left(\frac{5}{4}\right)} \int_0^\infty \Delta^2(k) e^{-(k\theta y^*)^2/8} \left[1 - \frac{1}{8}(k\theta y^*)^2 \right] \frac{dk}{k^2 y^*}. \quad (293)$$

The power-spectrum version of Limber's equation is already in the form required for the relation to the angular power spectrum ($w = \int \Delta_\theta^2 J_0(K\theta) dK/K$), and so we obtain the direct small-angle relation between spatial and angular power spectra:

$$\Delta_\theta^2 = \frac{\pi}{K} \int \Delta^2(K/y) y^5 \phi^2(y) dy. \quad (294)$$

This is just a convolution in log space, and is considerably simpler to evaluate and interpret than the $w - \xi$ version of Limber's equation.

Finally, note that it is not difficult to make allowance for spatial curvature in the above discussion. Write the Robertson–Walker metric in the form

$$c^2 d\tau^2 = c^2 dt^2 - R^2 \left[\frac{dr^2}{1 - kr^2} + r^2 \theta^2 \right]; \quad (295)$$

for $k = 0$, the notation $y = R_0 r$ was used for comoving distance, where $R_0 = (c/H_0)|1 - \Omega|^{-1/2}$. The radial increment of comoving distance was $dx = R_0 dr$, and the comoving distance between two objects was $(dx^2 + y^2 \theta^2)^{1/2}$. To maintain this version of Pythagoras's theorem, we clearly need to keep the definition of y and redefine radial distance: $dx = R_0 dr C(y)$, where $C(y) = [1 - k(y/R_0)^2]^{-1/2}$. The factor $C(y)$ appears in the non-Euclidean comoving volume element, $dV \propto y^2 C(y) dy$, so that we now require the normalization equation for ϕ to be

$$\int_0^\infty y^2 \phi(y) C(y) dy = 1. \quad (296)$$

The full version of Limber's equation therefore gains two powers of $C(y)$, but one of these is lost in converting between $R_0 dr$ and dx :

$$w(\theta) = \int_0^\infty [C(y)]^2 y^4 \phi^2 dy \int_{-\infty}^\infty \xi \left(\sqrt{x^2 + y^2 \theta^2} \right) \frac{dx}{C(y)}. \quad (297)$$

The net effect is therefore to replace $\phi^2(y)$ by $C(y)\phi^2(y)$, so that the full power-spectrum equation is

$$\Delta_\theta^2 = \frac{\pi}{K} \int \Delta^2(K/y) C(y) y^5 \phi^2(y) dy. \quad (298)$$

It is also straightforward to allow for evolution. The power version of Limber's equation is really just telling us that the angular power from a number of different radial shells adds incoherently, so we just need to use the actual evolved power at that redshift. These integral equations can be inverted numerically to obtain the real-space 3D clustering results from observations of 2D clustering; see Baugh & Efstathiou (1993; 1994).

7.6 Measuring the clustering spectrum

The history of attempts to quantify galaxy clustering goes back to Hubble's demonstration that the distribution of galaxies on the sky was non-uniform. The major post-war landmarks were the angular analysis of the Lick catalogue, described in Peebles (1980), and the analysis of the CfA redshift survey (Davis & Peebles 1983). It has taken some time to obtain data on samples that greatly exceed these in depth, but several pieces of work appeared around the start of the 1990s that clarified many of the discrepancies between different surveys and which paint a relatively consistent picture of large-scale structure. Perhaps the most significant of these surveys have been the **APM survey** (automatic plate-measuring machine survey; see Maddox

et al. 1990 and Maddox *et al.* 1996), the **CfA survey** (Center for Astrophysics survey; see Huchra *et al.* 1990) and the **LCRS** (Las Campanas redshift survey; see Shectman *et al.* 1996), together with a variety of surveys based on galaxies selected at $60\mu\text{m}$ by the infrared astronomy satellite, **IRASIRAS** satellite (Saunders *et al.* 1991; Fisher *et al.* 1993). These surveys are sensitive to rather different galaxy populations: the APM, CfA and LCRS surveys select in blue light and are sensitive to stellar populations of intermediate age; the IRAS emission originates from hot dust, associated with bursts of active star formation. A compilation of clustering results for a variety of tracers was given by Peacock & Dodds (1994); the results are shown in figure 14.

There is a wide range of power measured, ranging over perhaps a factor 20 between the real-space APM galaxies and the rich Abell clusters. Are these measurements all consistent with one Gaussian power spectrum for mass fluctuations? Corrections for redshift-space distortions and nonlinearities can be applied to these data to reconstruct the linear mass fluctuations, subject to an unknown degree of bias. The simplest assumption for this is that the bias is a linear response of the galaxy-formation process, and may be taken as independent of scale:

$$\Delta_{\text{tracer}}^2 = b^2 \Delta_{\text{mass}}^2. \quad (299)$$

There thus exist five free parameters that can be adjusted to optimize the agreement between the various estimates of the linear power spectrum; these are Ω and the four bias parameters for Abell clusters, radio galaxies, optical galaxies and IRAS galaxies (b_A, b_R, b_O, b_I); however, only two of these really matter: Ω and some measure of the overall level of fluctuations. For now, we take the IRAS bias parameter to play this latter role. Once these two are specified, the other bias parameters are well determined, principally from the linear data at small k , and have the approximate ratios

$$b_A : b_R : b_O : b_I = 4.5 : 1.9 : 1.3 : 1 \quad (300)$$

(Peacock & Dodds 1994). The reasons why different galaxy tracers may show different strengths of clustering are discussed above. Rich clusters are inevitably biased with respect to the mass, simply through the statistics of rare high-density regions. Massive ellipticals such as radio galaxies share some of this bias through the effect of morphological segregation, which says that the E/S0 fraction rises in clusters to almost 100%, by comparison with a mean of 20%. At high overdensities, the fraction of optical galaxies that are IRAS galaxies declines by a factor $\simeq 3-5$ (Strauss *et al.* 1992), reflecting the fact that IRAS galaxies are mainly spirals. Generally, the analysis of galaxies of a given type assumes that the luminosity function is independent of environment so that bias is independent of luminosity. This is not precisely true, and the amplitude of ξ does appear to rise slightly for galaxies of luminosity above several times L^* . (Valls-Gabaud *et al.* 1989; Loveday *et al.* 1995; Benoist *et al.* 1996). However, for the bulk of the galaxies in a given population, it is a good approximation to say that **luminosity segregation** can be neglected.

The various reconstructions of the linear power spectrum for the case $\Omega = b_I = 1$ are shown superimposed in figure 14, and display an impressive degree of agreement. This argues very strongly that what we measure from large-scale galaxy clustering has a direct relation to mass fluctuations, rather than being an optical illusion caused by non-uniform galaxy-formation efficiency (Bower *et al.* 1993). If effects other than gravity were dominant, the shape of spectrum inferred from clusters would have a very different shape at large scales, contrary to observation.

Large-scale power-spectrum data and models It is interesting to ask whether the power spectrum contains any features or whether it is consistent with a single smooth curve. A convenient description is in terms of the CDM power spectrum, which is $\Delta^2(k) \propto k^{n+3} T_k^2$. The CDM

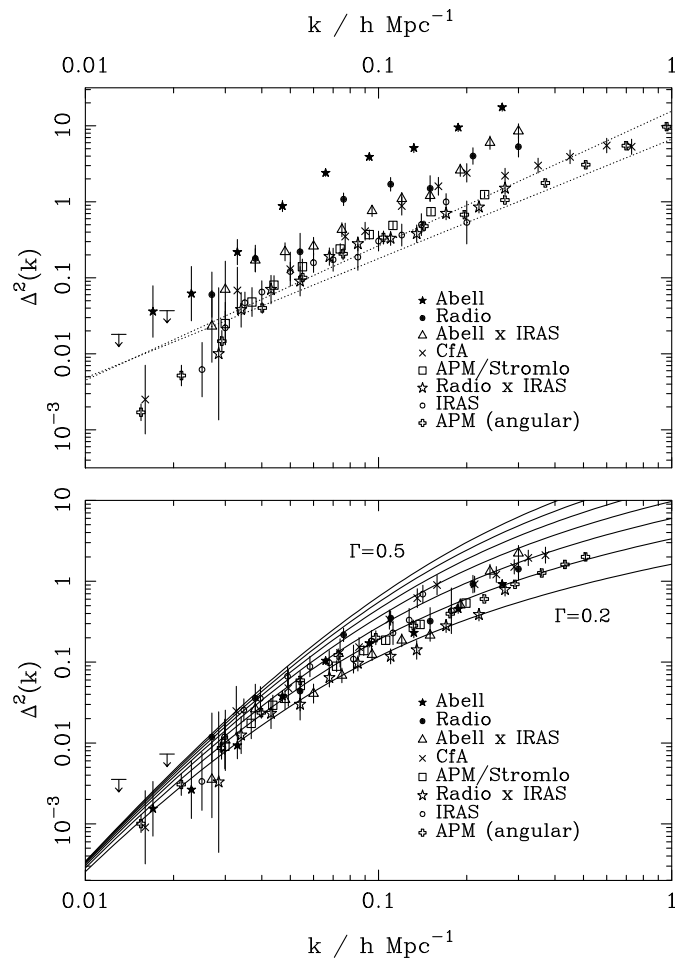


Fig. 14: A compilation of power-spectrum data, adapted from Peacock & Dodds (1994). The upper panel shows raw power-spectrum data in the form $\Delta^2 \equiv d\sigma^2/d \ln k$; all data with the exception of the APM power spectrum are in redshift space. The two dotted lines shown for reference are the transforms of the canonical real-space correlation functions for optical and IRAS galaxies ($r_0 = 5 h^{-1} \text{ Mpc}$ and $3.78 h^{-1} \text{ Mpc}$ and slopes of 1.8 and 1.57 respectively). The lower panel shows the results of correcting these datasets for different degrees of bias and for nonlinear evolution. There is an excellent degree of agreement, particularly in the detection of a break around $k = 0.03h$. The data are compared to various CDM models, assuming scale-invariant initial conditions, with the same large-wavelength normalization. Values of the fitting parameter $\Gamma = 0.5, 0.45, \dots, 0.25, 0.2$ are shown. The best-fit model has $\Gamma = 0.25$.

spectrum is very commonly used as a basis for comparison with cosmological observations, and it is essential to realize that this can be done in two ways. The CDM physics can be accepted, in which case the best-fitting value of Γ constrains Ω , h and Ω_B . Alternatively, the CDM spectrum can be used as a completely empirical fitting formula, which is assumed to approximate some different set of physics. For example, the MDM model includes CDM and an admixture of massive neutrinos; over a limited range of k , this will appear similar to a CDM spectrum, but with an effective value of Γ that is very much less than Ωh , because of the way in which neutrinos remove small-scale power in this case.

The normalization of the spectrum is specified by the rms variation in the fractional density contrast, averaged over $8 h^{-1}$ Mpc spheres; for CDM-like spectra, this measures power at an effective wavenumber well approximated by

$$\sigma_8^2 = \Delta^2(k_{\text{eff}}), \quad k_{\text{eff}}/h \text{ Mpc}^{-1} = 0.172 + 0.011 [\ln(\Gamma/0.34)]^2. \quad (301)$$

Fitting this spectrum to the large-scale linearised data of figure 14 requires the parameters

$$\Gamma \simeq 0.25 + 0.3(1/n - 1), \quad (302)$$

in agreement with many previous arguments suggesting that an apparently low-density model is needed; the linear transfer function does not bend sharply enough at the break wavenumber if a ‘standard’ high-density $\Gamma = 0.5$ model is adopted. For any reasonable values of h and baryon density, a high-density CDM model is not viable. Even a high degree of ‘tilt’ in the primordial spectrum (Cen *et al.* 1992) does not help change this conclusion unless n is set so low that major difficulties result when attempting to account for microwave-background anisotropies.

An important general lesson can also be drawn from the lack of large-amplitude features in the power spectrum. This is a strong indication that collisionless matter is deeply implicated in forming large-scale structure. Purely baryonic models contain large bumps in the power spectrum around the Jeans’ length prior to recombination ($k \sim 0.03\Omega h^2 \text{ Mpc}^{-1}$), whether the initial conditions are isocurvature or adiabatic. It is hard to see how such features can be reconciled with the data, beyond a ‘visibility’ in the region of 20%.

These ideas can be illustrated with a simple empirical model. Consider a spectrum in the form of a break between two power laws:

$$\Delta^2(k) = \frac{(k/k_0)^\alpha}{1 + (k/k_1)^{\alpha-\beta}}. \quad (303)$$

As shown in figure 16, the nonlinear power that results from this linear spectrum matches the data very nicely, if we choose the parameters

$$\begin{aligned} k_0 &= 0.3 h \text{ Mpc}^{-1} \\ k_1 &= 0.05 h \text{ Mpc}^{-1} \\ \alpha &= 0.8 \\ \beta &= 4.0. \end{aligned} \quad (304)$$

A value of $\beta = 4$ corresponds to a scale-invariant spectrum at large wavelengths, whereas the effective small-scale index is $n = -2.2$. We consider below the physical ways in which a spectrum of this shape might arise.

Finally, it is important to note that bias of either sign may have to be considered. A high-density universe with the cluster-normalized value of σ_8 predicts clustering well below that observed. However, the opposite is true for low Ω . If we simply take the APM power spectrum and ignore nonlinear corrections, the apparent value of σ_8 is about 0.9. Contrast this with the

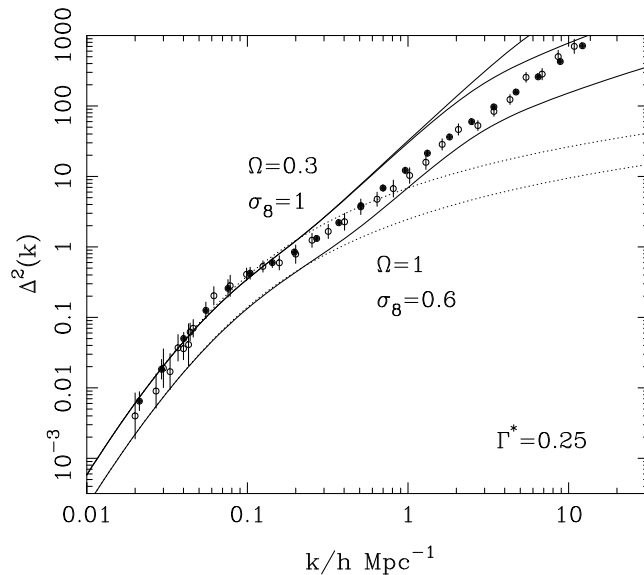


Fig. 15: The clustering data for optical galaxies, compared to three models with $\Gamma = 0.25$: $\Omega = 1$, $\sigma_8 = 0.6$; $\Omega_m = 0.3$, $\Omega_v = 0$, $\sigma_8 = 1$; $\Omega_m = 0.3$, $\Omega_v = 0.7$, $\sigma_8 = 1$. The linear spectra are shown dotted; solid lines denote evolved nonlinear spectra. All these models are chosen with a normalization that is approximately correct for the rich-cluster abundance and large-scale peculiar velocities. In all cases, the shape of the predicted spectrum fails to match observation. The high-density model would require a bias that is not a monotonic function of scale, whereas the low-density models exceed the observed small-scale clustering (figure adapted from Peacock 1997).

prediction of the cluster-normalization formula, which requires $\sigma_8 = 1.4$ for $\Omega = 0.2$, or 2.1 for $\Omega = 0.1$. Thus, low-density models inevitably require significant **antibias**, and we would have to consider the possibility that galaxy formation was suppressed in high-density regions. Bias in this sense has one advantage over positive bias, since it will tend to make the predicted small-scale spectrum less steep, which figure 15 suggests may be required in order to match the data. However, as discussed above, it is implausible that the scale dependence of the bias will be very extreme; a model that matches the data at $k \simeq 1 h \text{ Mpc}^{-1}$ will probably significantly undershoot the ‘bump’ at $k \simeq 0.1 h \text{ Mpc}^{-1}$. This large-scale feature therefore has a critical importance in the interpretation of large-scale structure. If it is correct, then the simplest CDM models fail and must be replaced by something more complicated. However, if future observations should yield lower power values at this point, then a low-density CDM model with antibias would provide a model for large-scale structure that is attractive in many ways (Jing *et al.* 1998).

7.7 Peculiar velocity fields

A research topic that has assumed increasing importance since about 1986 is the subject of deviations from the Hubble flow. Although the relation $v = Hr$ is a good approximation, it has long been known that individual galaxies have random velocities of a few hundred km s^{-1} superimposed on the general expansion. An exciting development has been the realization that these peculiar velocities display large-scale coherence in the form of **bulk flows** or **streaming flows**, which gives us the chance to probe very large-scale density fluctuations in the universe, and perhaps even to measure its mean density. Detailed reviews of these developments are given by Dekel (1994) and Strauss & Willick (1995).

In linear perturbation theory, the peculiar velocity is parallel to the peculiar gravitational

acceleration $\mathbf{g} = \nabla\delta\Phi/a$:

$$\delta\mathbf{v} = \frac{2f(\Omega)}{3H\Omega}\mathbf{g}, \quad (305)$$

where

$$f(\Omega) \equiv \left(\frac{a}{\delta}\right) \frac{d\delta}{da} \simeq \Omega^{0.6}. \quad (306)$$

Alternatively, we can work in Fourier terms, where

$$\delta\mathbf{v}_{\mathbf{k}} = -\frac{iHf(\Omega)a}{k}\delta_k\hat{\mathbf{k}}. \quad (307)$$

In either case, the linear peculiar-velocity field satisfies the continuity relation

$$\nabla \cdot \mathbf{v} = -Hf(\Omega)\delta \quad (308)$$

(where the divergence is in terms of proper coordinates). Since the fractional density perturbation (denoted by δ) is unobservable directly, one makes a connection with the galaxy distribution via the linear bias parameter

$$\delta_{\text{light}} = b\delta_{\text{mass}}. \quad (309)$$

The combination

$$\beta \equiv \Omega^{0.6}/b \quad (310)$$

can therefore be measured in principle, given the observed velocity field plus a large deep redshift survey from which the density perturbation field can be estimated.

Cosmological dipoles The well-determined absolute motion of the Earth with respect to the microwave background provides one of the possible general methods of estimating the cosmological density parameter Ω . Given a galaxy redshift survey, an estimate can be made of the gravitational acceleration of the local group produced by large-scale galaxy clustering. In perturbation theory, the relation between \mathbf{v} and \mathbf{g} can then be used to derive the peculiar velocity at a point in terms of the surrounding density field:

$$\mathbf{v} = \frac{H_0}{4\pi}\Omega^{0.6}\int\frac{\delta}{r^2}\hat{\mathbf{r}}d^3r \quad (311)$$

(Peebles 1980).

Weighing the universe The power of velocity fields is that they sample scales large enough that density perturbations are fully in the linear regime. In combination with large redshift surveys to define the spatial distribution of light, this has allowed not only a test of the assumption that large-scale clustering reflects gravitational instability but also a much more powerful extension of velocity-based methods for estimating the global density. We showed above how a knowledge of the density field surrounding the local group could be used to estimate the motion of the local group and hence predict the CMB dipole. Given a deep enough redshift survey, it is possible to use the same method to predict the peculiar velocity for any point in our local region. If we know the galaxy density perturbation field δ_g , then the peculiar velocity of a point is given by

$$\mathbf{v} = \beta \frac{H_0}{4\pi} \int \frac{\delta_g}{r^2} \hat{\mathbf{r}} d^3r, \quad (312)$$

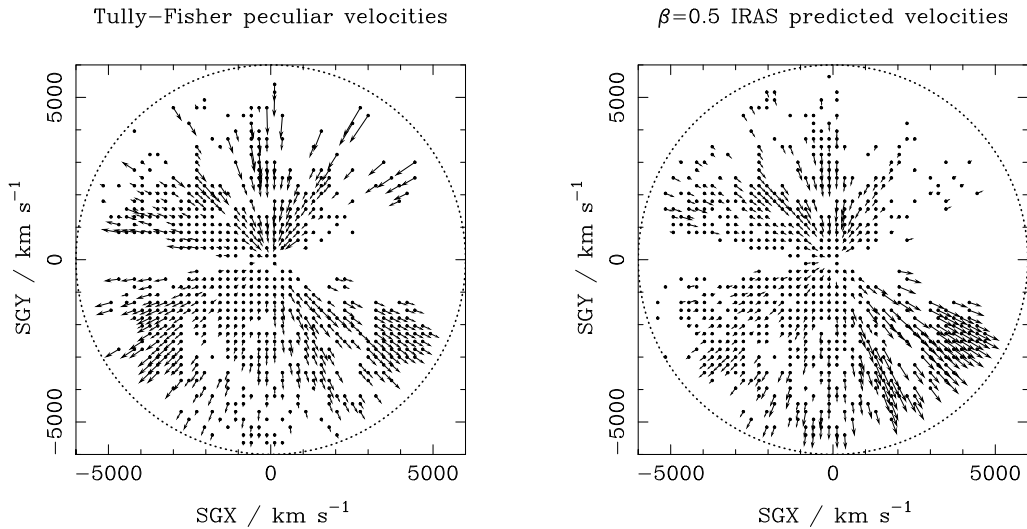


Fig. 16: A comparison of peculiar velocities inferred from the infrared Tully–Fisher method with predictions from the IRAS galaxy density field, for galaxies within 30° of the supergalactic plane and within 6000 km s^{-1} , assuming $\beta\Omega^{0.6}/b = 0.5$. The data of Davis, Nusser & Willick (1996) have been averaged onto a grid for clarity in regions of high galaxy density.

where the \mathbf{r} coordinate system is centred on the point of interest. The only subtlety is that δ_g is not observed directly in real space, but is deduced in redshift space. In practice, this can be corrected in an iterative way: δ_g is used to predict \mathbf{v} , the galaxy redshifts are corrected for the peculiar velocities, and the exercise is repeated until a stable real-space estimate of δ_g is obtained. Both this correction and the final prediction for \mathbf{v} depend on β , and so it should be possible to estimate β from a comparison between the predicted velocity field and the peculiar velocities derived using distance indicators such as the Tully–Fisher method. Figure 16 shows a recent result from one of these studies (Davis, Nusser & Willick 1997). Since most galaxies are concentrated towards the supergalactic plane defined by the local supercluster, it makes sense to plot the velocity vectors projected onto this plane. Since thousands of galaxies are involved, the velocity field is shown averaged onto a grid. The right-hand panel shows the field predicted from the gravity field due to the local density distribution.

This figure shows that, within 5000 km s^{-1} , the velocity field appears to be dominated by a few distinct regions in which the flow is nearly coherent. Of these features, the one that has received the most attention is the outward flow seen near $\text{SGX} = -4000 \text{ km s}^{-1}$, $\text{SGY} = -1000 \text{ km s}^{-1}$. This suggests the existence of a single large mass concentration at somewhat larger radii, which has been dubbed the **great attractor** (Dressler *et al.* 1987). Popular discussions of this object have sometimes given the impression of some mysterious concentration of mass that is detected only through its gravitational attraction. However, it should be clear from figure 16 that this is implausible; the overwhelming impression is that the observed and predicted velocity fields follow each other with reasonable fidelity, strongly suggesting that it should be possible to see the great attractor. This is not a totally straightforward process, since the long-range nature of gravity leaves some ambiguity over the distance at which the mass responsible for the peculiar velocities may lie. However, it is clear that the region of sky towards the great attractor contains many particularly rich superclusters, so there is no shortage of candidates (see *e.g.* Hudson 1993). The general agreement between the observed flows and the predictions of gravitational instability is enormously encouraging; furthermore, the amplitude of the prediction scales with β , and $\beta = 0.5$ seems to give a reasonable overall match (Davis,

Nusser & Willick 1996; Willick *et al.* 1997; see figure 16). This is very satisfying, as it agrees with the determinations from clustering anisotropy discussed earlier.

Could we make the comparison the other way around, and predict the density from the velocities? At first sight, this seems impossible, since observations only reveal *radial* components of velocity. However, in linear theory, vorticity perturbations become negligible relative to the growing mode for times sufficiently long after the perturbations are created. It is therefore very tempting to make the assumption that the linear velocity field should be completely irrotational at the present epoch. Furthermore, **Kelvin's circulation theorem** (see section 8 of Landau & Lifshitz 1959) guarantees that the flow will remain irrotational even in the presence of non-linearities, provided these are not so large as to cause dissipative processes. Dissipation does of course operate on the smallest scales (galaxies rotate, after all), but this should not affect the large-scale motions. We are therefore driven to write

$$\mathbf{v} = -\nabla\psi. \quad (313)$$

The problem is now solved in principle: the **velocity potential** ψ can be estimated by integrating the peculiar velocities along radial lines of sight. The unobservable transverse components can then be recovered by differentiation of the potential. In practice, this is a nontrivial problem, given that we are dealing with a limited number of galaxies, each of which has a rather noisy velocity estimate, of 20% precision at best. Nevertheless, by averaging over large numbers of galaxies to produce a smoothed representation of the radial velocity field on a grid, it is possible to use this method. The practical application goes by the name of **POTENT** (Bertschinger *et al.* 1990).

8 COSMIC BACKGROUND FLUCTUATIONS

8.1 Mechanisms for primary fluctuations

At the last-scattering redshift ($z \simeq 1000$), gravitational instability theory says that fractional density perturbations $\delta \gtrsim 10^{-3}$ must have existed in order for galaxies and clusters to have formed by the present. A long-standing challenge in cosmology has been to detect the corresponding fluctuations in brightness temperature of the cosmic microwave background (CMB) radiation, and it took over 25 years of ever more stringent upper limits before the first detections were obtained, in 1992. The study of CMB fluctuations has subsequently blossomed into a critical tool for pinning down cosmological models.

This can be a difficult subject; the treatment given here is intended to be the simplest possible. For technical details see *e.g.* Bond (1997), Efstathiou (1990), Hu & Sugiyama (1995), Seljak & Zaldarriaga (1996); for a more general overview, see White, Scott & Silk (1994) or Partridge (1995). The exact calculation of CMB anisotropies is complicated because of the increasing photon mean free path at recombination: a fluid treatment is no longer fully adequate. For full accuracy, the Boltzmann equation must be solved to follow the evolution of the photon distribution function. A convenient means for achieving this is provided by the public domain **CMBFAST** code (Seljak & Zaldarriaga 1996). Fortunately, these exact results can usually be understood via a more intuitive treatment, which is quantitatively correct on large and intermediate scales. This is effectively what would be called local thermodynamic equilibrium in stellar structure: imagine that the photons we see each originated in a region of space in which the radiation field was a Planck function of a given characteristic temperature. The observed brightness temperature field can then be thought of as arising from a superposition of these fluctuations in thermodynamic temperature.

We distinguish **primary anisotropies** (those that arise due to effects at the time of recombination) from **secondary anisotropies**, which are generated by scattering along the

line of sight. There are three basic primary effects, illustrated in figure 17, which are important on respectively large, intermediate and small angular scales:

- (1) Gravitational (Sachs–Wolfe) perturbations. Photons from high-density regions at last scattering have to climb out of potential wells, and are thus redshifted.
- (2) Intrinsic (adiabatic) perturbations. In high-density regions, the coupling of matter and radiation can compress the radiation also, giving a higher temperature.
- (3) Velocity (Doppler) perturbations. The plasma has a non-zero velocity at recombination, which leads to Doppler shifts in frequency and hence brightness temperature.

To make quantitative progress, the next step is to see how to predict the size of these effects in terms of the spectrum of mass fluctuations.

The temperature power spectrum The statistical treatment of CMB fluctuations is very similar to that of spatial density fluctuations. We have a 2D field of random fluctuations in brightness temperature, and this can be analysed by the same tools that are used in the case of 2D galaxy clustering.

Suppose that the fractional temperature perturbations on a patch of sky of side L are Fourier expanded:

$$\begin{aligned}\frac{\delta T}{T}(\mathbf{X}) &= \frac{L^2}{(2\pi)^2} \int T_K \exp(-i\mathbf{K} \cdot \mathbf{X}) d^2K \\ T_K(\mathbf{K}) &= \frac{1}{L^2} \int \frac{\delta T}{T}(\mathbf{X}) \exp(i\mathbf{K} \cdot \mathbf{X}) d^2X,\end{aligned}\tag{314}$$

where \mathbf{X} is a 2D position vector on the sky, and \mathbf{K} is a 2D wavevector. This is only a valid procedure if the patch of sky under consideration is small enough to be considered flat; we give the full machinery below. We will normally take the units of length to be angle on the sky, although they could also in principle be h^{-1} Mpc at a given redshift. The relation between angle and comoving distance on the last-scattering sphere requires the comoving angular-diameter distance to the last-scattering sphere; because of its high redshift, this is effectively identical to the horizon size at the present epoch, R_H :

$$\begin{aligned}R_H &= \frac{2c}{\Omega_m H_0} \quad (\text{open}) \\ R_H &\simeq \frac{2c}{\Omega_m^{0.4} H_0} \quad (\text{flat});\end{aligned}\tag{315}$$

the latter approximation for models with $\Omega_m + \Omega_v = 1$ is due to Vittorio & Silk (1991).

As with the density field, it is convenient to define a dimensionless power spectrum of fractional temperature fluctuations,

$$\mathcal{T}^2 \equiv \frac{L^2}{(2\pi)^2} 2\pi K^2 |T_K|^2,\tag{316}$$

so that \mathcal{T}^2 is the fractional variance in temperature from modes in unit range of $\ln K$. The corresponding dimensionless spatial statistic is the two-point correlation function

$$C(\theta) = \left\langle \frac{\delta T}{T}(\psi) \frac{\delta T}{T}(\psi + \theta) \right\rangle,\tag{317}$$

which is the Fourier transform of the power spectrum, as usual:

$$C(\theta) = \int \mathcal{T}^2(K) J_0(K\theta) \frac{dK}{K}. \quad (318)$$

Here, the Bessel function comes from the angular part of the Fourier transform:

$$\int \exp(ix \cos \phi) d\phi = 2\pi J_0(x). \quad (319)$$

Now, in order to predict the observed anisotropy of the microwave background, the problem we must solve is to integrate the temperature perturbation field through the **last-scattering shell**. In order to do this, we assume that the sky is flat; we also neglect curvature of the 3-space, although this is only strictly valid for flat models with $k = 0$. Both these restrictions mean that the results are not valid for very large angles. Now, introducing the Fourier expansion of the 3D temperature perturbation field (with coefficients T_k^{3D}) we can construct the observed 2D temperature perturbation field by integrating over k space and optical depth:

$$\frac{\delta T}{T} = \frac{V}{(2\pi)^3} \iint T_k^{3D} e^{-i\mathbf{k}\cdot\mathbf{r}} d^3k e^{-\tau} d\tau. \quad (320)$$

A further simplification is possible if we approximate $e^{-\tau} d\tau$ by a Gaussian in comoving radius:

$$\exp(-\tau) d\tau \propto \exp[-(r - r_{\text{LS}})^2/2\sigma_r^2] dr. \quad (321)$$

This says that we observe radiation from a last-scattering shell centred at comoving distance r_{LS} (which is very nearly identical to r_{H} , since the redshift is so high), with a thickness σ_r . The section on recombination showed that the appropriate value of σ_r is approximately

$$\sigma_r = 7(\Omega h^2)^{-1/2} \text{ Mpc}. \quad (322)$$

An intuitively useful way of thinking about the integral for the observed temperature perturbation is as a two-stage process: produce a temperature field that is convolved in the radial direction, and then say that we observe a single shell that slices through this convolved field at the radius of last scattering. If the observed CMB is a slice in the (x, y) plane, the effect of the last-scattering convolution in the z direction is $T_k^{3D} \rightarrow T_k^{3D} \exp[-k_z^2 \sigma_r^2/2]$ (z will briefly denote the Cartesian coordinate in the redshift direction, not redshift itself). As a result of this radial convolution and the angular dependence of the Doppler scattering term, the temperature spatial power spectrum is anisotropic. Nevertheless, we can still write down 2D and 3D Fourier-transform expressions for the correlation function in the plane $z = 0$ (taking the origin to be in the centre of the last-scattering shell):

$$\begin{aligned} C_{3D} &= \int \frac{\mathcal{T}_{3D}^2}{4\pi k^3} e^{-i\mathbf{k}\cdot\mathbf{x}} dk_x dk_y dk_z \\ C_{2D} &= \int \frac{\mathcal{T}_{2D}^2}{2\pi K^2} e^{-i\mathbf{K}\cdot\mathbf{x}} dK_x dK_y. \end{aligned} \quad (323)$$

Note the distinction between k and K – wavenumbers in 3D and 2D respectively. The definition of \mathcal{T}_{3D}^2 as the dimensionless power spectrum of spatial variations in temperature is analogous to the 3D spatial power spectrum:

$$\mathcal{T}_{3D}^2 = \frac{V}{(2\pi)^3} 4\pi k^3 |T_k^{3D}|^2 \quad (324)$$

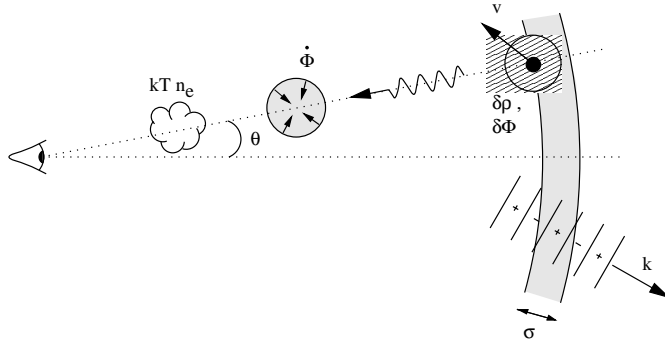


Fig. 17: Illustrating the physical mechanisms that cause CMB anisotropies. The shaded arc on the right represents the last-scattering shell; an inhomogeneity on this shell affects the CMB through its potential, adiabatic and Doppler perturbations. Further perturbations are added along the line of sight by time-varying potentials (Rees–Sciama effect) and by electron scattering from hot gas (Sunyaev–Zeldovich effect). The density field at last scattering can be Fourier analysed into modes of wavevector \mathbf{k} . These spatial perturbation modes have a contribution that is in general damped by averaging over the shell of last scattering. Short-wavelength modes are more heavily affected (i) because more of them fit inside the scattering shell, and (ii) because their wavevectors point more nearly radially for a given projected wavelength.

(the 2D equivalent was written above just as \mathcal{T}^2 , but sometimes it will be convenient for clarity to add an explicit subscript 2D). Equating the two expressions for $C(\mathbf{x})$ gives the usual expression relating 2D and 3D power spectra, which we shall write in the slightly different form

$$\mathcal{T}_{2\text{D}}^2(K) = K^2 \int_0^\infty \mathcal{T}_{3\text{D}}^2(\sqrt{K^2 + w^2}) e^{-w^2 \sigma_r^2} \frac{dw}{(w^2 + K^2)^{3/2}}, \quad (325)$$

This simple expression gives the 2D spectrum as a projection, to which all modes with wavelength shorter than the projected wavelength of interest contribute; short-wavelength modes that run nearly towards the observer have a much longer apparent wavelength on the sky; see figure 17. The integral will generally be dominated by the contribution around $w = 0$, unless $\mathcal{T}_{3\text{D}}^2$ is a very rapidly increasing function, in which case what matters will be the small-scale cutoff governed by the width of the last-scattering shell.

The 2D power spectrum is thus a smeared version of the 3D one: any feature that appears at a particular wavenumber in 3D will cause a corresponding feature at the same wavenumber in 2D. A particularly simple converse to this rule arises when there are *no* features: the 3D power spectrum is scale-invariant ($\mathcal{T}_{3\text{D}}^2 = \text{constant}$). In this case, for scales large enough that we can neglect the radial smearing from the last-scattering shell,

$$\mathcal{T}_{2\text{D}}^2 = \mathcal{T}_{3\text{D}}^2 \quad (326)$$

so that the pattern on the CMB sky is scale invariant also. To apply the above machinery for a general spectrum, we now need quantitative expressions for the spatial temperature anisotropies.

Sachs–wolfe effect This is the dominant large-scale effect, and arises from potential perturbations at last scattering. These have two effects: (i) they redshift the photons we see, so that an overdensity *cools* the background as the photons climb out, $\delta T/T = \delta\Phi/c^2$; (ii) they cause time dilation at the last-scattering surface, so that we seem to be looking at a younger (and hence *hotter*) universe where there is an overdensity. The time dilation is $\delta t/t = \delta\Phi/c^2$; since

the time dependence of the scale factor is $a \propto t^{2/3}$ and $T \propto 1/a$, this produces the counterterm $\delta T/T = -(2/3)\delta\Phi/c^2$. The net effect is thus one-third of the gravitational redshift:

$$\boxed{\frac{\delta T}{T} = \frac{\delta\Phi}{3c^2}} \quad (327)$$

This effect was originally derived by Sachs & Wolfe (1967) and bears their name SW effect (SW effect). It is common to see the first argument alone, with the factor 1/3 attributed to some additional complicated effect of general relativity. However, in weak fields, general relativistic effects should already be incorporated within the concept of gravitational time dilation; the above argument shows that this is indeed all that is required to explain the full result.

To relate to density perturbations, use Poisson's equation $\nabla^2\delta\Phi_k = 4\pi G\rho\delta_k$. The effect of ∇^2 is to pull down a factor of $-k^2/a^2$ (a^2 because k is a comoving wavenumber). Eliminating ρ in terms of Ω and z_{LS} gives

$$T_k = -\frac{\Omega(1+z_{\text{LS}})}{2} \left(\frac{H_0}{c}\right)^2 \frac{\delta_k(z_{\text{LS}})}{k^2}. \quad (328)$$

Doppler source term The effect here is just the Doppler effect from the scattering of photons by moving plasma:

$$\boxed{\frac{\delta T}{T} = \frac{\delta\mathbf{v} \cdot \hat{\mathbf{r}}}{c}} \quad (329)$$

Using the standard expression for the linear peculiar velocity, the corresponding k -space result is

$$T_k = -i\sqrt{\Omega(1+z_{\text{LS}})} \left(\frac{H_0}{c}\right) \frac{\delta_k(z_{\text{LS}})}{k} \hat{\mathbf{k}} \cdot \hat{\mathbf{r}}. \quad (330)$$

Adiabatic source term This is the simplest of the three effects mentioned earlier:

$$\boxed{T_k = \frac{\delta_k(z_{\text{LS}})}{3}}, \quad (331)$$

because $\delta n_\gamma/n_\gamma = \delta\rho/\rho$ and $n_\gamma \propto T^3$. However, this simplicity conceals a paradox. Last scattering occurs only when the universe recombines, which occurs at roughly a fixed temperature: $kT \sim \chi$, the ionization potential of hydrogen. Surely, then, we should just be looking back to a surface of constant temperature? Hot and cold spots should normalize themselves away, so that the last-scattering sphere appears uniform. The solution is that a denser spot recombines *later*: it is therefore less redshifted and appears hotter. In algebraic terms, the observed temperature perturbation is

$$\left(\frac{\delta T}{T}\right)_{\text{obs}} = -\frac{\delta z}{1+z} = \frac{\delta\rho}{\rho}, \quad (332)$$

where the last expression assumes linear growth, $\delta \propto (1+z)^{-1}$. Thus, even though a more correct picture for the temperature anisotropies seen on the sky is of a crinkled surface at

constant temperature, thinking of hot and cold spots gives the right answer. Any observable cross-talk between density perturbations and delayed recombination is confined to effects of order higher than linear.

We now draw the above results together to form the spatial power spectrum of CMB fluctuations in terms of the power spectrum of mass fluctuations at last scattering:

$$\boxed{\mathcal{T}_{3\text{D}}^2 = \left[(f_{\text{A}} + f_{\text{SW}})^2(k) + f_{\text{V}}^2(k)\mu^2 \right] \Delta_k^2(z_{\text{LS}}).} \quad (333)$$

There is no cross term between the adiabatic and Sachs–Wolfe terms proportional to δ and the Doppler term proportional to $i\delta$: $|a\delta + ib\delta|^2 = (a\delta + ib\delta)(a\delta^* - ib\delta^*)$. The dimensionless factors can be written most simply as

$$\begin{aligned} f_{\text{SW}} &= -\frac{2}{(kD_{\text{LS}})^2} \\ f_{\text{V}} &= \frac{2}{kD_{\text{LS}}} \\ f_{\text{A}} &= 1/3, \end{aligned} \quad (334)$$

where

$$D_{\text{LS}} = \frac{2c}{\Omega_m^{1/2} H_0} (1 + z_{\text{LS}})^{-1/2} = 184(\Omega h^2)^{-1/2} \text{ Mpc} \quad (335)$$

is the comoving horizon size at last scattering (a result that is independent of whether there is a cosmological constant).

We can see immediately from these expressions the relative importance of the various effects on different scales. The Sachs–Wolfe effect dominates for wavelengths $\gtrsim 1h^{-1}$ Gpc; Doppler effects then take over but are almost immediately dominated by adiabatic effects on the smallest scales.

Small-scale fluctuations The above expressions apply to perturbations for which only gravity has been important up till last scattering, *i.e.* those larger than the horizon at z_{eq} . For smaller wavelengths, a variety of additional physical processes act on the radiation perturbations, generally reducing the predicted anisotropies. An accurate treatment of these effects is not really possible without a more complicated analysis, as is easily seen by considering the thickness of the last-scattering shell, $\sigma_r = 7(\Omega h^2)^{-1/2}$ Mpc. This clearly has to be of the same order of magnitude as the photon mean free path at this time; on any smaller scales, a fluid approximation for the radiation is inadequate and a proper solution of the Boltzmann equation is needed. Nevertheless, some qualitative insight into the small-scale processes is possible. The radiation fluctuations will be damped relative to the baryon fluid by photon diffusion, characterised by the Silk-damping scale, $\lambda_{\text{S}} = 2.7(\Omega\Omega_{\text{B}}h^6)^{-1/4}$ Mpc. Below the horizon scale at z_{eq} , $16(\Omega h^2)^{-1}$ Mpc, there is also the possibility that dark-matter perturbations can grow while the baryon fluid is still held back by radiation pressure, which results in adiabatic radiation fluctuations that are less than would be predicted from the dark-matter spectrum alone. In principle, this suggests a suppression factor of $(1 + z_{\text{eq}})/(1 + z_{\text{LS}})$, or roughly a factor 10. In detail, the effect is an oscillating function of scale, since we have seen that baryonic perturbations oscillate as sound waves when they come inside the horizon:

$$\delta_b \propto (3c_{\text{S}})^{1/4} \exp\left(\pm i \int kc_{\text{S}} d\tau\right); \quad (336)$$

here, τ stands for conformal time. There is thus an oscillating signal in the CMB, depending on the exact phase of these waves at the time of last scattering. These oscillations in the fluid of baryons plus radiation cause a set of **acoustic peaks** in the small-scale power spectrum of the CMB fluctuations (see below).

It is clear that small-scale CMB anisotropies are a complex area, because of the near-coincidence between z_{eq} and z_{LS} , and between σ_r , $r_{\text{H}}(z_{\text{eq}})$ and λ_{S} . To some extent, these complications can be ignored, because the finite thickness of the last-scattering shell smears out small-scale perturbations in any case. However, the damping is exponential in $(k\mu)^2$ and so modes with low μ receive little damping; averaging over all directions gives a reduction in power that goes only $\propto k^{-1}$. In the absence of the other effects listed above, small-scale adiabatic fluctuations would still dominate the anisotropy pattern.

Large-scale fluctuations The flat-space formalism becomes inadequate for very large angles; the proper basis functions to use are the spherical harmonics:

$$\frac{\delta T}{T}(\hat{\mathbf{q}}) = \sum a_{\ell}^m Y_{\ell m}(\hat{\mathbf{q}}), \quad (337)$$

where $\hat{\mathbf{q}}$ is a unit vector that specifies direction on the sky. Since the spherical harmonics satisfy the orthonormality relation $\int Y_{\ell m} Y_{\ell' m'}^* d^2 q = \delta_{\ell\ell'} \delta_{mm'}$, the inverse relation is

$$a_{\ell}^m = \int \frac{\delta T}{T} Y_{\ell m}^* d^2 q. \quad (338)$$

The analogues of the Fourier relations for the correlation function and power spectrum are

$$\begin{aligned} C(\theta) &= \frac{1}{4\pi} \sum_{\ell} \sum_{m=-\ell}^{m=+\ell} |a_{\ell}^m|^2 P_{\ell}(\cos \theta) \\ |a_{\ell}^m|^2 &= 2\pi \int_{-1}^1 C(\theta) P_{\ell}(\cos \theta) d \cos \theta. \end{aligned} \quad (339)$$

These are exact relations, governing the actual correlation structure of the observed sky. However, the sky we see is only one of infinitely many possible realizations of the statistical process that yields the temperature perturbations; as with the density field, we are more interested in the **ensemble average power**. A common notation is to define C_{ℓ} as the expectation value of $|a_{\ell}^m|^2$:

$$C(\theta) = \frac{1}{4\pi} \sum_{\ell} (2\ell + 1) C_{\ell} P_{\ell}(\cos \theta), \quad C_{\ell} \equiv \langle |a_{\ell}^m|^2 \rangle, \quad (340)$$

where now $C(\theta)$ is the ensemble-averaged correlation. For small θ and large ℓ , the exact form reduces to a Fourier expansion:

$$C(\theta) = \int_0^{\infty} \mathcal{T}^2(K) J_0(K\theta) \frac{dK}{K}, \quad \mathcal{T}^2(K = \ell + \frac{1}{2}) = \frac{(\ell + \frac{1}{2})(2\ell + 1)}{4\pi} C_{\ell}. \quad (341)$$

The effect of filtering the microwave sky with the beam of a telescope may be expressed as a multiplication of the C_{ℓ} , as with convolution in Fourier space:

$$C_{\text{S}}(\theta) = \frac{1}{4\pi} \sum_{\ell} (2\ell + 1) W_{\ell}^2 C_{\ell} P_{\ell}(\cos \theta). \quad (342)$$

When the telescope beam is narrow in angular terms, the Fourier limit can be used to deduce the appropriate ℓ -dependent filter function. For example, for a Gaussian beam of **FWHM** (full-width to half maximum) 2.35σ , the filter function is $W_\ell = \exp(-\ell^2\sigma^2/2)$.

For the large-scale temperature anisotropy, we have already seen that what matters is the Sachs–Wolfe effect, for which we have derived the spatial anisotropy power spectrum. The spherical harmonic coefficients for a spherical slice through such a field can be deduced using the results for large-angle galaxy clustering, in the limit of a selection function that goes to a delta function in radius:

$$C_\ell^{\text{SW}} = 16\pi \int (kD_{\text{LS}})^{-4} \Delta_k^2(z_{\text{LS}}) j_\ell^2(kR_{\text{H}}) \frac{dk}{k}, \quad (343)$$

where the j_ℓ are **spherical Bessel functions** (see chapter 10 of Abramowitz & Stegun 1965). This formula, derived by Peebles (1982), strictly applies only to spatially flat models, since the Fourier expansion of the density field is invalid in an open model. Nevertheless, since the curvature radius R_0 subtends an angle of $\Omega/[2(1-\Omega)^{1/2}]$, even the lowest few multipoles are not seriously affected by this point, provided $\Omega \gtrsim 0.1$.

For simple mass spectra, the integral for the C_ℓ can be performed analytically. The case of most practical interest is a scale-invariant spectrum ($\Delta_k^2 \propto k^4$), for which the integral scales as

$$C_\ell = \frac{6}{\ell(\ell+1)} C_2 \quad (344)$$

(see equation 6.574.2 of Gradshteyn & Ryzhik 1980). The direct relation between the mass fluctuation spectrum and the multipole coefficients of CMB fluctuations mean that either can be used as a measure of the normalization of the spectrum. One measure that has become common is to work in terms of the amplitude of the quadrupole ($\ell = 2$), by means of the rms temperature fluctuation Q_{rms} produced just by the $\ell = 2$ term(s) in the spherical harmonic expansion:

$$Q_{\text{rms}}^2 = \frac{1}{4\pi} \sum_{m=-2}^{m=+2} |a_2^m|^2; \quad (345)$$

Unfortunately, although the quadrupole is the largest-scale intrinsic anisotropy signal (the intrinsic dipole is unobservable, owing to the Earth’s motion), it is not a good choice as a reference point, for several reasons. First, the large-scale temperature pattern is subject to corruption by emission from the Milky Way, and it is better to work at galactic latitudes $|b| \gtrsim 20^\circ$; second, the intrinsic quadrupole is badly affected by **cosmic variance**. The C_ℓ coefficients are the average of $|a_\ell^m|^2$ over an ensemble, and so the Q_{rms}^2 value seen by a given observer is distributed like χ^2 with five degrees of freedom. A more useful quantity is the ensemble-averaged quadrupole, since this relates directly to the power spectrum:

$$Q_{\text{rms-ps}}^2 \equiv \frac{5}{4\pi} C_2. \quad (346)$$

However, in practice power is measured over a range of multipoles, centred at $\ell > 2$, so that the value at $\ell = 2$ is really an extrapolation that assumes a specific index for the spectrum. The most common choice is a scale-invariant spectrum, and so the clumsy quantity $Q_{\text{rms-ps}|n=1}$ is used as a way of expressing the normalization of a scale-invariant spectrum. More generally, following

our discussion of small-scale anisotropies, it makes sense to define a broad-band measure of the ‘power per log ℓ ’:

$$\mathcal{T}^2(\ell) = \frac{\ell(\ell+1)}{2\pi} C_\ell, \quad (347)$$

so that \mathcal{T}^2 is constant for a scale-invariant spectrum. This is a close relation of another measure that is sometimes encountered: $Q^2(\ell) = (5/12)\mathcal{T}^2(\ell)$ is an obvious generalization of the Q notation for the quadrupole amplitude. Finally, whichever measure is adopted, there is still the choice of units. The temperature fluctuation $\Delta T/T$ is dimensionless, but anisotropy experiments generally measure ΔT directly, independent of the mean temperature. It is therefore common practice to quote numbers like Q in units of μK .

8.2 Characteristics of CMB anisotropies

We are now in a position to understand the characteristic angular structure of CMB fluctuations. The change-over from scale-invariant Sachs–Wolfe fluctuations to fluctuations dominated by Doppler scattering has been shown to occur at $k \simeq D_{\text{LS}}$. This is one critical angle (call it θ_1); its definition is $\theta_1 = D_{\text{LS}}/R_{\text{H}}$, and for a matter-only model it takes the value

$$\theta_1 = 1.8 \Omega^{1/2} \text{ degrees}. \quad (348)$$

For flat low-density models with significant vacuum density, R_{H} is smaller; θ_1 and all subsequent angles would then be larger by about a factor $\Omega^{-0.6}$ (*i.e.* θ_1 is roughly independent of Ω in flat Λ -dominated models).

The second dominant scale is the scale of last-scattering smearing set by $\sigma_r = 7(\Omega h^2)^{-1/2}$ Mpc. This subtends an angle

$$\theta_2 = 4 \Omega^{1/2} \text{ arcmin}. \quad (349)$$

Finally, a characteristic scale in many density power spectra is set by the horizon at z_{eq} . This is $16(\Omega h^2)^{-1}$ Mpc and subtends

$$\theta_3 = 9h^{-1} \text{ arcmin}, \quad (350)$$

independent of Ω . This is quite close to θ_2 , so that alterations in the transfer function are an effect of secondary importance in most models.

We therefore expect that all scale-invariant models will have similar CMB power spectra: a flat Sachs–Wolfe portion down to $K \simeq 1 \text{ degree}^{-1}$, followed by a bump where Doppler and adiabatic effects come in, which turns over on arcminute scales through damping and smearing. This is illustrated well in figure 18, which shows some detailed calculations of 2D power spectra, generated with the CMBFAST package. From these plots, the key feature of the anisotropy spectrum is clearly the peak at $\ell \sim 100$. This is often referred to as the **Doppler peak**, but it is not so clear that this name is accurate. Our simplified analysis suggests that Sachs–Wolfe anisotropy should dominate for $\theta > \theta_1$, with Doppler and adiabatic terms becoming of comparable importance at θ_1 , and adiabatic effects dominating at smaller scales. There are various effects that cause the simple estimate of adiabatic effects to be too large, but they clearly cannot be neglected for $\theta < \theta_1$. A better name, which is starting to gain currency, is the **acoustic peak**. In any case, it is clear that the peak is the key diagnostic feature of the CMB anisotropy spectrum: its height above the SW ‘plateau’ is sensitive to Ω_{B} and its angular location depends on Ω and Λ . It is therefore no surprise that many experiments are

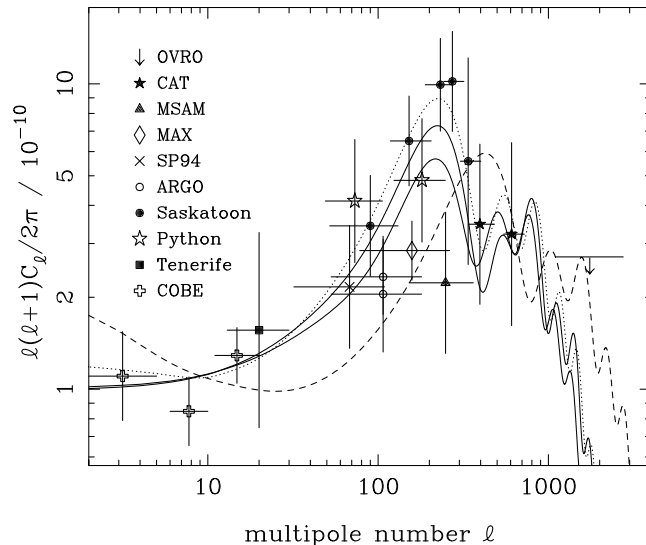


Fig. 18: Angular power spectra $\mathcal{T}^2(\ell) = \ell(\ell + 1)C_\ell/2\pi$ for the CMB, plotted against angular wavenumber ℓ in radians⁻¹. The experimental data are an updated version of the compilation described in White, Scott & Silk (1994), communicated by M. White; see also Hancock *et al.* (1997). Various model predictions for adiabatic scale-invariant CDM fluctuations are shown. The two solid lines correspond to $(\Omega, \Omega_B, h) = (1, 0.05, 0.5)$ and $(1, 0.1, 0.5)$, with the higher Ω_B increasing power by about 20% at the peak. The dotted line shows a flat Λ -dominated model with $(\Omega, \Omega_B, h) = (0.3, 0.05, 0.65)$; the dashed line shows an open model with the same parameters. Note the very similar shapes of all the curves. The normalisation has been set to the large-scale amplitude, and so any dependence on Ω is quite modest. The main effects are that open models shift the peak to the right, and that the height of the peak increases with Ω_B and h .

currently attempting accurate measurements of this feature. Furthermore, it is apparent that sufficiently accurate experiments will be able to detect higher ‘harmonics’ of the peak, in the form of smaller oscillations of amplitude perhaps 20% in power, around $\ell \simeq 500\text{--}1000$. These features arise because the matter–radiation fluid undergoes small-scale oscillations, the phase of which at last scattering depends on wavelength, since the density oscillation varies roughly as $\delta \propto \exp(ic_S k\tau)$. Accurate measurement of these oscillations would pin down the sound speed at last scattering, and help give an independent measurement of the baryon density.

Ω dependence and normalization It is not uncommon to encounter the claim that the level of CMB fluctuations is inconsistent with a low-density universe, and in particular that a high density in collisionless dark matter is required. In fact, this statement is something of a fallacy, and it is worth examining the issue of density dependence in some detail.

Suppose we perform calculations assuming some mass power spectrum and $\Omega = 1$. If we now change Ω while keeping the shape and normalization of the power spectrum fixed, there are two effects: the power spectrum is translated both horizontally and vertically. The horizontal translation is quite simple: the main angles of importance scale as $\Omega^{1/2}$ so the CMB pattern shifts to smaller scales as Ω is reduced (unless Λ is important, in which case the shift is almost negligible; see above). To predict the vertical shift, it will suffice to consider the Sachs–Wolfe portion of the spectrum. This is

$$\mathcal{T}_{\text{SW}}^2 = \frac{4}{(kD_{\text{LS}})^4} \Delta^2(z_{\text{LS}}). \quad (351)$$

To relate δ at last scattering to its present value, we need the Ω -dependent growth-suppression factor for density perturbations:

$$\delta(z_{\text{LS}}) \simeq \frac{\delta_0}{1 + z_{\text{LS}}} [g(\Omega)]^{-1} \quad (352)$$

i.e. there is less growth in low-density universes. Including the Ω dependence of D_{LS} gives

$$\mathcal{T}_{\text{sw}}^2 = \frac{1}{4} \left(\frac{ck}{H_0} \right)^{-4} \Delta_0^2 \Omega^2 [g(\Omega)]^{-2}. \quad (353)$$

The approximate power-law dependence of g is $\Omega^{0.65}$ for open models or $\Omega^{0.23}$ for flat models, so it appears that low-density universes predict lower fluctuations. This is clearly contrary to the common idea that low-density universes are ruled out owing to the freeze-out of density-perturbation growth requiring higher fluctuations at last scattering. The fractional density fluctuations are indeed higher, but the potential fluctuations that are observable depend on the *total* density fluctuation, which is lower.

Predictions from galaxy clustering However, Δ_0^2 here is the *mass* power spectrum, and the normalization we deduce from the light will generally depend on Ω . The effect this has depends on the scale at which we normalize. One common approach is to use σ_8 , the rms density contrast in spheres of radius $8 h^{-1}$ Mpc, since this is closely related to the abundance of rich clusters of galaxies. Alternatively, the amplitude of large-scale peculiar velocities measures the fractional density fluctuation on a somewhat larger scale – in both cases with a dependence of approximately $\delta_0 \propto \Omega^{-0.6}$. Note that neither of these determinations use galaxy clustering data at all: the present-day potential fluctuations are measurable directly, and yield $\delta_0 \Omega^{0.6}$. What we deduce from galaxy clustering, conversely, is the biased quantity $b\delta_0$, and so galaxy clustering observations allow the parameter $\beta \equiv \Omega^{0.6}/b$ to be measured (as well as the shape of the spectrum, of course). The requirement for larger matter fluctuations in the case of lower Ω now makes the density dependence of the Sachs–Wolfe effect very weak: roughly $(\Delta_{\text{sw}}^2)^{1/2} \propto \Omega^{-0.25}$ (open) or $\Omega^{0.17}$ (flat).

How then can we constrain Ω from CMB observations? One possible route arises because the transfer function on small scales is rather sensitive to the total density, since modes with wavelengths below $r_{\text{H}}(z_{\text{eq}}) = 16(\Omega h^2)^{-1}$ Mpc have their amplitudes reduced. For a given primordial amplitude, this reduction clearly increases as Ω decreases, and is particularly severe for pure baryon universes where the small-scale power is removed by Silk damping. In the extreme case, low-density universes can have very little power on 8-Mpc wavelengths, and so normalizing on this scale gives silly answers. The number σ_8 measures the total rms density fluctuation after filtering with a sphere, and the only way this number can be large in models with a damping cutoff is to set the amplitude of 100-Mpc modes high. This is the real reason why low-density universes have often been associated with very large CMB fluctuations. In any case, now that we have clustering data on 100-Mpc scales, it makes much more sense to fix the normalization there, in which case the predicted amplitude of large-scale temperature fluctuations loses almost all dependence on Ω , as discussed above.

Following this discussion, it should be clear that it was possible to make relatively clear predictions of the likely level of CMB anisotropies, even in advance of the first detections. What was required was a measurement of the typical depth of large-scale potential wells in the universe, and many lines of argument pointed inevitably to numbers of order 10^{-5} . This was already clear from the existence of massive clusters of galaxies with velocity dispersions of up to 1000 km s^{-1} :

$$v^2 \sim \frac{GM}{r} \quad \Rightarrow \quad \frac{\Phi}{c^2} \sim \frac{v^2}{c^2}, \quad (354)$$

so the potential well of a cluster is of order 10^{-5} deep. More exactly, the abundance of rich clusters is determined by the amplitude σ_8 , which measures $[\Delta^2(k)]^{1/2}$ at an effective wavenumber of very nearly $0.17 h \text{ Mpc}^{-1}$. If we assume that this is a large enough scale that what we are measuring is the amplitude of any scale-invariant spectrum, then the earlier expression for the temperature power spectrum gives

$$\sqrt{\mathcal{T}_{\text{sw}}^2} \simeq 10^{-5.7} \Omega \sigma_8 [g(\Omega)]^{-1}. \quad (355)$$

There were thus strong grounds to expect that large-scale fluctuations would be present at about the 10^{-5} level, and it was a significant boost to the credibility of the gravitational-instability model that such fluctuations were eventually seen.

8.3 Observations of CMB anisotropies

Prior to April 1992, no CMB fluctuations had been detected, but existing limits were close to the interesting 10^{-5} level. Continued non-detection of fluctuations at a much lower level would have forced a fundamental overhaul of our ideas about cosmological structure formation, so this period was a critical time for cosmology.

April 23 1992 saw the announcement by the COBE DMR team of the first detection of CMB fluctuations (Smoot *et al.* 1992). COBE is an acronym for NASA's **cosmic background explorer** satellite; launched in November 1989, this carried several experiments to probe the large-scale radiation field over the wavelength range $1\mu\text{m}$ to 1cm . The one concerned with the CMB at $\lambda \gtrsim 1\text{mm}$ was the **differential microwave radiometer**. The DMRDMR experiment made a map of the sky with an angular resolution set by its 7° FWHM beam. This resolution means that only the low-order multipoles are accessible; DMR thus probes pure Sachs–Wolfe, and so it is easy to relate the DMR detection of sky fluctuations to a limit on the power spectrum.

In the case of the COBE measurements, the simplest and most robust datum in the initial detection reports was just the sky variance (convolved to 10° FWHM resolution, in order to suppress the noise a little), *i.e.* $C_s(0)$ with $\sigma = 4.25^\circ$. The expected result for this for pure Sachs–Wolfe anisotropies can be predicted to almost perfect accuracy by using a small-angle approximation:

$$C_s(0) = \frac{\Omega^2}{g^2(\Omega)} \int 4[k(2c/H_0)]^{-4} \Delta^2(k) W^2(kR_H) \frac{dk}{k} \quad (356)$$

$$W^2(y) = [1 - j_0^2(y) - 3j_1^2(y)] F(y\sigma)/(y\sigma),$$

where a Gaussian beam of FWHM 2.35σ is assumed, and

$$F(x) \equiv \exp(-x^2) \int_0^x \exp(t^2) dt \quad (357)$$

is Dawson's integral. The terms involving Bessel functions correspond to the subtraction of the unobservable monopole and dipole terms. The window function is relatively sharply peaked and so the COBE variance essentially picks out the power at a given characteristic scale, which is well approximated as follows (σ in radians):

$$C_s(0) = 1.665 \frac{\Omega^2}{g^2(\Omega)} 4[k_s(2c/H_0)]^{-4} \Delta^2(k_s) \quad (358)$$

$$k_s R_H = \frac{0.54}{\sigma} + 2.19(n-1)$$

(Peacock & Dodds 1994). The original reported value was $C^{1/2}(0) = 1.10 \pm 0.18 \times 10^{-5}$ (Smoot *et al.* 1992). For scale-invariant spectra, this corresponds to an rms quadrupole $Q_{\text{rms-ps}} =$

$15.0 \pm 2.5 \mu\text{K}$. The final results from 4 years of data are consistent with the initial announcement: a scale-invariant fitted $Q_{\text{rms-ps}} = 18.0 \pm 1.8 \mu\text{K}$. Only weak constraints on the spectrum index can be set, but the results are certainly consistent with the scale-invariant prejudice: $n = 1.2 \pm 0.3$ (Bennett *et al.* 1996).

This detection required a large number of systematic effects to be eliminated in order to be certain that the reported signal was indeed cosmological. The COBE experiment had the advantage of being a satellite in a stable environment, with no atmospheric fluctuations to contend with. By observing the same piece of sky many times from different parts of its orbit, any signals that related to low-level interference from the Earth or Sun could be eliminated. The most serious remaining problem was astronomical foregrounds – principally emission from the Milky Way. The DMR experiment observed with three different frequency channels: 31.5, 53 and 90 GHz, each of which had two independent receivers. At these wavelengths, the main contaminants are galactic synchrotron and bremsstrahlung emission (brightness temperatures varying roughly as $\nu^{-2.8}$ and ν^{-2} , respectively). A constant-temperature cosmological signal can thus be picked up by averaging the channels so as to make the galactic signal vanish (and analysing only high galactic latitudes for good measure). Even after the systematics have been dealt with, the individual DMR receivers had significant thermal noise; in the 4-year data set, the full-resolution combined map suffers a noise of around $30 \mu\text{K}$ at any given position. Since the sky map has 1844 pixels, the excess cosmological variance can be detected with huge significance, but some have criticised COBE on the grounds that it failed to detect individual structures in the CMB. Such comments are unwarranted, since all that any CMB experiment can do is to measure multipole components of the temperature perturbation field down to some limit. COBE was unable to measure individual multipole coefficients to its full resolution ($\ell \simeq 20$), but did perfectly well to $\ell \simeq 10$: the hot and cold spots on the CMB sky at 20° resolution were correctly identified even in the first-year data.

Small-scale experiments Eventually, the CMB sky will be revisited by satellite experiments with resolutions well below 1 degree. In the meantime, experiments that seek to improve on the COBE map have to work either from the ground or from balloons. In either case, they are forced to work with restricted patches of the sky, and have to contend with variable atmospheric emission. As a result, multiple-beam experiments designed to remove atmospheric emission are the norm. The simplest strategy is to switch rapidly between either two beams separated by an angle θ (chopping) or between three beam positions in a line, each separated by θ (chopping plus nodding). It is then possible to form combinations of these signals that reduce the atmospheric contribution: $T_2 - T_1$ in the two-beam case or $T_2 - (T_1 + T_3)/2$ in the three-beam case. The first case gives a signal insensitive to the mean atmosphere, whereas the second also cancels any contribution from a constant gradient in atmospheric emission. Squaring these expressions and taking expectation values shows that the rms fluctuations measured in such experiments are

$$\begin{aligned} \Delta T/T &= \sqrt{2[C(0) - C(\theta)]} \quad (\text{two-beam}) \\ &= \sqrt{2[C(0) - C(\theta)] - \frac{1}{2}[C(0) - C(2\theta)]} \quad (\text{three-beam}) \end{aligned} \quad (359)$$

Putting in the relation between power spectrum and $C(\theta)$, we see that three-beam experiments produce an effective 2D window function,

$$\begin{aligned} \langle (\delta T/T)^2 \rangle &= \int \mathcal{T}^2 W_K^2 dK/K \\ W_K^2 &= \left[\frac{3}{2} - 2J_0(K\theta) + \frac{1}{2}J_0(2K\theta) \right] e^{-K^2\theta_s^2}, \end{aligned} \quad (360)$$

where $2.35\theta_s$ is the beam FWHM and θ is the beam throw. The filter function peaks at some effective wavelength that is very close to 2θ . In general, all such experiments can be viewed in

this way as observing the CMB sky with some effective window function,

$$\boxed{\langle (\delta T/T)^2 \rangle = \frac{1}{4\pi} \sum_{\ell} (2\ell + 1) W_{\ell}^2 C_{\ell}} \quad (361)$$

(see Partridge 1995 for more complex observing strategies).

8.4 Conclusions and outlook

Having reviewed the physical mechanisms that cause anisotropies in the microwave background, and summarized the observational situation, it is time to ask what conclusions can be drawn. In order to narrow the field of possibilities, the discussion will concentrate on models with primordial fluctuations that are adiabatic and Gaussian. As well as being the simplest models, they will also turn out to be in reasonably good agreement with observation. Isocurvature models suffer from the high amplitude of the large-scale perturbations, and do not become any more attractive when modelled in detail (Hu, Bunn & Sugiyama 1995). Topological defects were for a long time hard to assess, since accurate predictions of their CMB properties were difficult to make. Recent progress does, however, indicate that these theories may have difficulty matching the main details of CMB anisotropies, even as they are presently known (Pen, Seljak & Turok 1997).

Inflationary predictions Matching the CMB sky with what we know of mass inhomogeneities today is important for physical cosmology, since we have seen that the anisotropies depend on the cosmological parameters Ω , Λ , Ω_{B} and h . As if this were not already potentially a rich enough prize, CMB anisotropies also offer the chance to probe the very earliest phases of the big bang, and to test whether the expanding universe really did begin with an inflationary phase. Let us recall what predictions inflation makes for the fluctuation spectrum. Inflation is driven by a scalar field ϕ , with a potential $V(\phi)$. As well as the characteristic energy density of inflation, V , this can be characterised by two dimensionless parameters

$$\begin{aligned} \epsilon &\equiv \frac{m_{\text{P}}^2}{16\pi} (V'/V)^2 \\ \eta &\equiv \frac{m_{\text{P}}^2}{8\pi} (V''/V), \end{aligned} \quad (362)$$

where m_{P} is the Planck mass, $V' = dV/d\phi$, and all quantities are evaluated towards the end of inflation, when the present large-scale structure modes were comparable in size to the inflationary horizon. Prior to transfer-function effects, the primordial fluctuation spectrum is specified by a horizon-scale amplitude (extrapolated to the present) δ_{H} and a slope n :

$$\Delta^2(k) = \delta_{\text{H}}^2 \left(\frac{ck}{H_0} \right)^{3+n}. \quad (363)$$

The inflationary predictions for these numbers are

$$\begin{aligned} \delta_{\text{H}} &\sim \frac{V^{1/2}}{m_{\text{P}}^2 \epsilon^{1/2}} \\ n &= 1 - 6\epsilon + 2\eta, \end{aligned} \quad (364)$$

which leaves us in the unsatisfactory position of having two observables and three parameters.

The critical ingredient for testing inflation by making further predictions is the possibility that, in addition to scalar modes, the CMB could also be affected by gravitational waves (following the original insight of Starobinsky 1985). We therefore distinguish explicitly between scalar and tensor contributions to the CMB fluctuations by using appropriate subscripts. The former category are those described by the Sachs–Wolfe effect, and are gravitational potential fluctuations that relate directly to mass fluctuations. The relative amplitude of tensor and scalar contributions depended on the inflationary parameter ϵ alone:

$$\frac{C_\ell^T}{C_\ell^S} \simeq 12.4\epsilon \simeq 6(1 - n). \quad (365)$$

The second relation to the **tilt** (which is defined to be $1 - n$) is less general, as it assumes a polynomial-like potential, so that η is related to ϵ . If we make this assumption, inflation can be tested by measuring the tilt and the tensor contribution. For simple models, this test should be feasible: $V = \lambda\phi^4$ implies $n \simeq 0.95$ and $C_\ell^T/C_\ell^S \simeq 0.3$. To be safe, we need one further observation, and this is potentially provided by the spectrum of C_ℓ^T . Suppose we write separate power-law index definitions for the scalar and tensor anisotropies:

$$C_\ell^S \propto \ell^{n_S-3}, \quad C_\ell^T \propto \ell^{n_T-3}. \quad (366)$$

From the discussion of the Sachs–Wolfe effect, we know that, on large scales, the scalar index is the same as index in the matter power spectrum: $n_S = n = 1 - 6\epsilon + 2\eta$. By the same method, it is easily shown that $n_T = 1 - 2\epsilon$ (although different definitions of n_T are in use in the literature; the convention here is that $n = 1$ always corresponds to a constant $\mathcal{T}^2(\ell)$). Finally, then, we can write the **inflationary consistency equation**:

$$\boxed{\frac{C_\ell^T}{C_\ell^S} = 6.2(1 - n_T)}. \quad (367)$$

The slope of the scalar perturbation spectrum is the only quantity that contains η , and so n_S is not involved in a consistency equation, since there is no independent measure of η with which to compare it.

From the point of view of an inflationary purist, the scalar spectrum is therefore an annoying distraction from the important business of measuring the tensor contribution to the CMB anisotropies. A certain degree of degeneracy exists here (see Bond *et al.* 1994), since the tensor contribution has no acoustic peak; C_ℓ^T is roughly constant up to the horizon scale and then falls. A spectrum with a large tensor contribution therefore closely resembles a scalar-only spectrum with smaller Ω_b (and hence a relatively lower peak). One way in which this degeneracy may be lifted is through polarization of the CMB fluctuations. A nonzero polarization is inevitable because the electrons at last scattering experience an anisotropic radiation field. Thomson scattering from an anisotropic source will yield polarization, and the practical size of the fractional polarization P is of the order of the quadrupole radiation anisotropy at last scattering: $P \gtrsim 1\%$. Furthermore, the polarization signature of tensor perturbations differs from that of scalar perturbations (*e.g.* Seljak 1997; Hu & White 1997); the different contributions to the total unpolarized C_ℓ can in principle be disentangled, allowing the inflationary test to be carried out.

Implications of large-scale anisotropies Despite the above discussion, it will be convenient to compare the present-day mass distribution with the CMB data by considering only scalar perturbations at first. Possible complications due to tensor contributions can be brought in a little

later. The best and cleanest anisotropy measurements are those due to COBE, and we have seen above how the large-scale Sachs–Wolfe anisotropy can be calculated. It is possible to argue in both directions, either using the mass spectrum to predict the CMB, or *vice versa*; we shall start with the latter route. Górski *et al.* (1995), Bunn, Scott & White (1995), and White & Bunn (1995) discuss the large-scale normalization from the 2-year COBE data in the context of CDM-like models. The final 4-year COBE data favour very slightly lower results, and we scale to these in what follows. For scale-invariant spectra and $\Omega = 1$, the best normalization is

$$\text{COBE} \quad \Rightarrow \quad \Delta^2(k) = \left(\frac{k}{0.0737 h \text{ Mpc}^{-1}} \right)^4, \quad (368)$$

which is equivalent to $Q_{\text{rms-ps}} = 18.0 \mu\text{K}$, or $\delta_{\text{H}} = 2.05 \times 10^{-5}$.

For low-density models, the earlier discussion suggests that the power spectrum should depend on Ω and the growth factor g as $P \propto g^2/\Omega^2$. Because of the time dependence of the gravitational potential (integrated Sachs–Wolfe effect) and because of spatial curvature, this expression is not exact, although it captures the main effect. From the data of White & Bunn (1995), a better approximation is

$$\Delta^2(k) \propto \frac{g^2}{\Omega^2} g^{0.7}. \quad (369)$$

This applies for low- Ω models both with and without vacuum energy, with a maximum error of 2% in density fluctuation provided $\Omega > 0.2$. Since the rough power-law dependence of g is $g(\Omega) \simeq \Omega^{0.65}$ and $\Omega^{0.23}$ for open and flat models respectively, we see that the implied density fluctuation amplitude scales approximately as $\Omega^{-0.12}$ and $\Omega^{-0.69}$ respectively for these two cases. The dependence is weak for open models, but vacuum energy implies much larger fluctuations for low Ω .

What if we consider a **tilted spectrum**? To see the effect of $n \neq 1$, we need to know the effective k at which COBE determines the spectrum. We saw earlier that the Sachs–Wolfe contribution to the rms sky fluctuations filtered with a Gaussian beam of FWHM 2.35σ effectively measured the power at $kR_{\text{H}} = (0.54/\sigma) + 2.19(n - 1)$. For 10° resolution, $\sigma = 0.0742$, and so the effective wavenumber is approximately

$$k_{\text{eff}}^{\text{COBE}} = 0.0012 h \text{ Mpc}^{-1} \times \begin{cases} \Omega^{1.0} & (\text{open}) \\ \Omega^{0.4} & (\text{flat}), \end{cases} \quad (370)$$

ignoring the small n -dependent correction. This is a scale at least 20 times beyond the largest wavelength on which large-scale structure is reliably measured, and so the effects of tilt will be substantial. We will adopt the following measure of power on the largest reliable scales:

$$\Delta_{\text{opt}}^2(k = 0.02 h \text{ Mpc}^{-1}) \simeq 0.005 \pm 0.0015. \quad (371)$$

Furthermore, we know from redshift-space distortions and the value of σ_8 inferred from the cluster abundance that the corresponding number for the mass must be lower if $\Omega = 1$. As before, $b \simeq 1.6$ seems the best guess for $\Omega = 1$. Since σ_8 from clusters scales as $\Omega^{-0.56}$, this suggests that $\Delta_{\text{mass}}^2(k = 0.02 h \text{ Mpc}^{-1}) \simeq 0.002\Omega^{-1.1}$. We can compare this number with the COBE prediction, scaling the COBE-predicted amplitude with Ω as above and pivoting the power-law spectrum about k_{eff} ; clearly there will be a unique value of n that matches prediction and observation. The only problem is that, although $k = 0.02 h \text{ Mpc}^{-1}$ is a very large scale, the power measured there is not quite the primordial value. Two physical models that fit the shape of the large-scale clustering spectrum were discussed above: (i) $\Gamma = 0.25$ CDM; (ii) $\Gamma = 0.4$, $f_\nu = 0.3$ MDM. At $k = 0.02 h \text{ Mpc}^{-1}$, the transfer functions for these models are 0.69 and

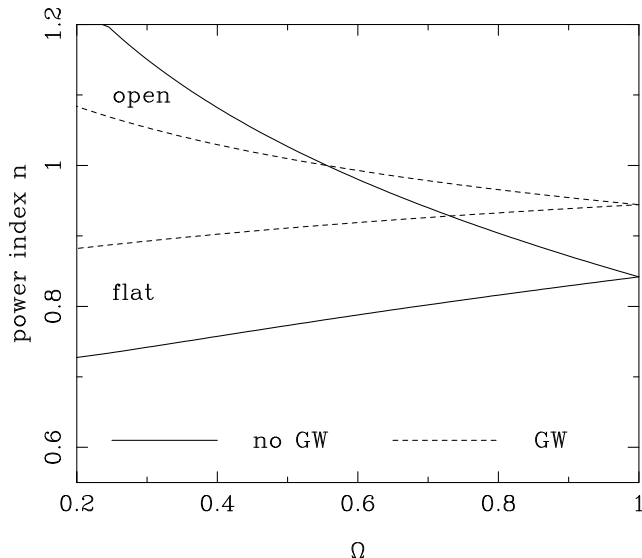


Fig. 19: A plot of the power-law index of the spectrum needed to reconcile the COBE level of CMB fluctuations with the inferred present-day mass fluctuations on the largest scales, $k = 0.02 h \text{ Mpc}^{-1}$, allowing for the effects of the transfer function as described in the text. Both open models and flat Λ -dominated models are shown. Solid lines denote models without gravity waves; dashed lines show the effect of adding gravity waves with the usual coupling to tilt. A scale-invariant spectrum ($n = 1$) requires an open universe with $\Omega \simeq 0.6$. For lower densities, Λ -dominated flat models with gravity waves come closest to simple inflationary predictions ($n \simeq 0.9$).

0.81 respectively. We therefore adopt a mean of 0.75, and scale the inferred primordial mass fluctuations upwards by $1/(0.75)^2$.

Figure 19 shows the values of n that are required to reconcile this measure of primordial large-scale structure with COBE. Gravity waves are treated in two distinct ways: in the first case they are ignored; in the second they are added in with the above inflationary coupling to tilt, $C_\ell^T/C_\ell^S = 6(1 - n)$. Figure 19 has several interesting features:

- (1) For $\Omega = 1$, a significant tilt is needed: $n \simeq 0.84$ without tensors, rising to $n = 0.95$ if they are included.
- (2) Going to low-density Λ -dominated models requires a greater degree of tilt. Even though the inferred mass fluctuations today increase for low Ω , the Ω dependence of the CMB fluctuations in Λ models is even stronger.
- (3) Conversely, the weak CMB dependence on Ω for open models means that the required tilt changes rapidly with Ω . A scale-invariant spectrum with no tensor contribution is consistent with the data if $\Omega \simeq 0.6$.

Some of these results look more attractive than others. High degrees of tilt are not expected in simple models of inflation (*e.g.* $n = 0.95$ for $V = \lambda\phi^4$). Moreover, large tilt causes problems with the CMB on smaller scales. COBE normalization corresponds to $\ell \simeq 20$, whereas we have seen that there is mounting evidence for a peak at $\ell \simeq 200$. The existence of tilt thus reduces the amplitude of this peak relative to COBE by a factor 10^{1-n} , which is equal to 1.4 for $\Omega = 1$ without gravity waves. Things are just as bad if gravity waves are included: the tilt is less, but the gravity-wave component has no peak, so that the reduction of the relative height of the peak is again a factor 1.4. Looking at figure 18, we see that $n = 1$ models with a reasonable baryon

content in fact get the height of the peak about right; to allow the tilted variants, the baryon content would have to be boosted above what is allowed from nucleosynthesis.

Although the preference for $n < 1$ is thus a negative feature of flat models, it may not be fatal, since the inferred tilt depends on the accuracy of the large-scale clustering measurements. Changing the assumed large-scale power by a factor 1.5 changes n by 0.14 without gravity waves, or 0.06 with gravity waves. If flat models are to survive, the true power at $k = 0.02 h \text{ Mpc}^{-1}$ would thus need to be larger by a factor of order 1.5, and current data do not exclude this possibility.

Implications of small-scale anisotropies Stronger diagnostics for Ω and Λ come from the intermediate-scale and small-scale CMB anisotropies. The location of the peak at $\ell \simeq 200$ is sensitive to Ω , since it measures directly the angular size of the horizon at last scattering, which scales as $\ell \propto \Omega^{-1/2}$ for open models. The cutoff at $\ell \simeq 1000$ caused by last-scattering smearing also moves to higher ℓ for low Ω ; if Ω were small enough, the smearing cutoff would be carried to large ℓ , where it would be inconsistent with the upper limits to anisotropies on 10-arcminute scales. For flat models with $\Lambda \neq 0$, the Ω dependence is much weaker, which is one possible way of detecting Λ , should other arguments favour $\Omega < 1$.

This tendency for open models to violate the upper limits to arcminute-scale anisotropies is a long-standing problem, which allowed Bond & Efstathiou (1984) to deduce the following limit on CDM universes:

$$\Omega \gtrsim 0.3h^{-4/3} \quad (372)$$

(in the absence of reionization, with a spectrum normalization that was independent of Ω , thus not allowing for the possibility of bias).

Coda The study of anisotropies in the CMB is presently one of the most exciting observational areas in cosmology, as a plethora of experiments map out the anisotropy spectrum over a wide range of scales. The fact that these detections are at the $\lesssim 10^{-5}$ level makes an amusing contrast with the early days of the subject, when fluctuations of order 10^{-3} were expected, based on the simplistic formula ‘ $\delta T/T = (1/3)\delta\rho/\rho$ and I must make galaxies by $z = 3$ ’. Today, we have a much more sophisticated appreciation of the scales that are accessible to observation, plus much improved data on the inhomogeneity of the local universe.

The COBE detection and smaller-scale measurements are enormously encouraging indications of the overall correctness of the picture of structure formation via gravitational instability, but they leave open many possibilities. These will be constrained by the information which resides in the $\ell \gtrsim 100$ peaks in the anisotropy spectrum. At present, all we can say is that there are hints of a acoustic peak in the spectrum at $\ell \simeq 200$ and a sharp fall by $\ell \simeq 1000$. If confirmed, these facts would make it very difficult to sustain the idea of an open universe. As we saw earlier, the supernovae Hubble diagram strongly favours a low-density universe, if we consider only $k = 0$ models. We therefore need to consider a ‘standard model’ in which the majority of the energy density is in the form of vacuum energy: either a classical cosmological constant, or ‘quintessence’, where the scalar field continues to roll.

Definitive measurements of the CMB fluctuation spectrum will require a new generation of experiments, which are expected to yield results in the first decade of the 21st Century. As well as accurate large-scale mapping, these probes will measure the fine-scale anisotropy down to $\ell \sim 1000$. Measurements of the higher harmonics of the acoustic oscillations will be sensitive to the fine details of the physical conditions at last scattering: these will either rule out all the standard range of models – or determine the nature of dark matter and measure very accurately

the main cosmological parameters, if the present framework is correct (Bond, Efstathiou & Tegmark 1997; Zaldarriaga, Spergel & Seljak 1997).

Finally, the deepest attraction offered by CMB studies is the possibility of testing inflation. We have seen that one characteristic prediction of inflation is the existence of tensor anisotropies, and have discussed how these may in principle be detected via their contribution to the CMB anisotropy spectrum, and also through the polarization of CMB fluctuations. These will be challenging observations, but ones whose importance it would be difficult to overstate. The detection of the inflationary background of gravitational waves would give us experimental evidence on the nature of physics at almost the Planck energy. It is astonishing to realize that this might be accomplished within a mere 100 years from the first discovery of the expansion of the universe. The present is undoubtedly a golden age for cosmology, but perhaps the best is yet to come.

REFERENCES

- Abramowitz M., Stegun I.A. (1965) *Handbook of Mathematical Functions*, Dover
- Adler R.J. (1981) *The Geometry of Random Fields*, Wiley
- Bahcall N.A., Soneira R.M. (1983) *Astrophys. J.*, **270**, 20
- Baugh C.M., Efstathiou G. (1993) *Mon. Not. R. Astr. Soc.*, **265**, 145
- Baugh C.M., Efstathiou G. (1994) *Mon. Not. R. Astr. Soc.*, **267**, 323
- Bardeen J.M., Bond J.R., Kaiser N., Szalay A.S. (1986) *Astrophys. J.*, **304**, 15
- Bennett C.L. *et al.* (1996) *Astrophys. J.*, **464**, L1
- Benoist C., Maurogordato S., da Costa L.N., Cappi A., Schaeffer R. (1996) *Astrophys. J.*, **472**, 452
- Bertschinger E., Dekel A., Faber S.M., Dressler A., Burstein D. (1990) *Astrophys. J.*, **364**, 370
- Bond J.R., Efstathiou G. (1984) *Astrophys. J.*, **285**, L45
- Bond J.R. (1997) in *Cosmology and Large-scale Structure*, proc. 60th Les Houches School, eds R. Schaeffer, J. Silk, M. Spiro & J. Zinn-Justin, Elsevier, p469
- Bond J.R., Crittenden R., Davis R.L., Efstathiou G., Steinhardt P.J. (1994) *Phys. Rev. Lett.*, **72**, 13
- Bond J.R., Efstathiou G., Tegmark M. (1997) *Mon. Not. R. Astr. Soc.*, **291**, L33
- Bower R.G., Coles P., Frenk C.S., White S.D.M. (1993) *Astrophys. J.*, **405**, 403
- Brandenberger R.H. (1990) in *Physics of the Early Universe*, proc 36th Scottish Universities Summer School in Physics, eds Peacock J.A., Heavens A.F., Davies A.T., Adam Hilger, p281
- Bunn E.F., Scott D., White M. (1995) *Astrophys. J.*, **441**, 9
- Burbidge E.M., Burbidge G.R., Fowler W.A., Hoyle F. (1957) *Rev. Mod. Phys.*, **29**, 547
- Carroll S.M., Press W.H., Turner E.L. (1992) *Ann. Rev. Astr. Astrophys.*, **30**, 499
- Cen R., Gnedin N.Y., Kofman L.A., Ostriker J.P. (1992) *Astrophys. J.*, **399**, L11
- Collins P.D.B., Martin A.D., Squires E.J. (1989) *Particle Physics and Cosmology*, Wiley
- Dalton G.B., Efstathiou G., Maddox S.J., Sutherland, W. (1992) *Astrophys. J.*, **390**, L1
- Davis M., Peebles P.J.E. (1983) *Astrophys. J.*, **267**, 465
- Davis M., Nusser A., Willick J.A. (1996) *Astrophys. J.*, **473**, 22
- Dekel A. (1994) *Ann. Rev. Astr. Astrophys.*, **32**, 371
- Dressler A., Lynden-Bell D., Burstein D., Davies R.L., Faber S.M., Terlevich R.J., Wegner G. (1987) *Astrophys. J.*, **313**, 42
- Efstathiou G. (1990) in *Physics of the Early Universe*, proc 36th Scottish Universities Summer School in Physics, eds Peacock J.A., Heavens A.F., Davies A.T., Adam Hilger, p361
- Efstathiou G. (1995) *Mon. Not. R. Astr. Soc.*, **274**, L73
- Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R.S., Frenk

- C.S. (1990) *Mon. Not. R. Astr. Soc.*, **247**, 10
- Efstathiou G., Bernstein G., Katz N., Tyson T., Guhathakurta P. (1991) *Astrophys. J.*, **380**, 47
- Efstathiou G., Bond J.R., White S.D.M. (1992) *Mon. Not. R. Astr. Soc.*, **258**, 1P
- Eisenstein D.J., Hu W. (1998) *Astrophys. J.*, **496**, 605
- Eke V.R., Cole S., Frenk C.S. (1996) *Mon. Not. R. Astr. Soc.*, **282**, 263
- Ellis J. (1997) in *Cosmology and Large-scale Structure*, proc. 60th Les Houches School, eds R. Schaeffer, J. Silk, M. Spiro & J. Zinn-Justin, Elsevier, p715
- Felten J.E., Isaacman R. (1986) *Rev. Mod. Phys.*, **58**, 689
- Fisher K.B. (1995) *Astrophys. J.*, **448**, 494
- Fisher K.B., Davis M., Strauss M.A., Yahil A., Huchra J.P. (1993) *Astrophys. J.*, **402**, 42
- Fixsen D.J., Cheng E.S., Gales J.M., Mather J.C., Shafer R.A., Wright E.L. (1996) *Astrophys. J.*, **473**, 576
- Górski K.M., Ratra B., Sugiyama N., Banday A.J. (1995) *Astrophys. J.*, **444**, L65
- Gradshteyn I.S., Ryzhik I.M. (1980) *Table of Integrals, Series and Products*, Academic Press
- Grigoriev D., Shaposhnikov M., Turok N. (1992) *Phys. Lett.*, **275B**, 395
- Guth A.H. (1981) *Phys. Rev. D*, **23**, 347
- Hamilton A.J.S., Kumar P., Lu E., Matthews A. (1991) *Astrophys. J.*, **374**, L1
- Hancock S. *et al.* (1997) *Mon. Not. R. Astr. Soc.*, **289**, 505
- Heath D. (1977) *Mon. Not. R. Astr. Soc.*, **179**, 351
- Hu W., Bunn E.F., Sugiyama N. (1995) *Astrophys. J.*, **447**, L59
- Hu W., Sugiyama N. (1995) *Astrophys. J.*, **444**, 489
- Hu W., White M. (1997) *New Astronomy*, **2**, 323
- Huchra J.P., Geller M.J., de Lapparant V., Corwin H.G. (1990) *Astrophys. J. Suppl.*, **72**, 433
- Hudson M.J. (1993) *Mon. Not. R. Astr. Soc.*, **265**, 43
- Jain B., Mo H.J., White S.D.M. (1995) *Mon. Not. R. Astr. Soc.*, **276**, L25
- Jensen L.G., Szalay A.S. (1986) *Astrophys. J.*, **305**, L5
- Jing Y.P., Mo H.J., Börner G. (1998) *Astrophys. J.*, **494**, 1
- Jones B.J.T., Wyse R.F.G. (1985) *Astr. Astrophys.*, **149**, 144
- Kaiser N. (1984) *Astrophys. J.*, **284**, L9
- Kashlinsky A. (1991) *Astrophys. J.*, **376**, L5
- Kolb E.W., Turner M.S. (1990) *The Early Universe*, Addison-Wesley
- Lahav O., Lilje P.B., Primack J.R., Rees M.J. (1991) *Mon. Not. R. Astr. Soc.*, **251**, 128
- Landau L.D., Lifshitz E.M. (1959) *Fluid Mechanics*, Pergamon
- Le Fèvre O. *et al.* (1996) *Astrophys. J.*, **461**, 534
- Liddle A.R., Lyth D. (1993) *Phys. Reports*, **231**, 1
- Liddle A.R., Scherrer R.J. (1998) astro-ph/9809272
- Loveday J., Maddox S.J., Efstathiou G., Peterson B.A. (1995) *Astrophys. J.*, **442**, 457
- Lumsden S.L., Heavens A.F., Peacock J.A. (1989) *Mon. Not. R. Astr. Soc.*, **238**, 293
- McClelland J., Silk J. (1977) *Astrophys. J.*, **217**, 331
- Maddox S.J., Efstathiou G., Sutherland W.J., Loveday J. (1990) *Mon. Not. R. Astr. Soc.*, **247**, 1P
- Maddox S.J., Efstathiou G., Sutherland W.J. (1996) *Mon. Not. R. Astr. Soc.*, **283**, 1227
- Mather J.C. *et al.* (1990) *Astrophys. J.*, **354**, L37
- Mészáros P. (1974) *Astr. Astrophys.*, **37**, 225
- Moore G. (1996) *Nucl. Phys.*, **B480**, 689
- Mukhanov V.F., Feldman H.A., Brandenberger R.H. (1992) *Phys. Reports*, **215**, 203
- Opal Consortium (1990) *Phys. Lett.*, **B240**, 497
- Osterbrock D.E. (1974) *Astrophysics of Gaseous Nebulae*, Freeman
- Pagel B.E.J. (1994) in *The Formation and Evolution of Galaxies*, eds C. Muñoz-Tuñón, F. Sánchez, Cambridge University Press, p151

- Partridge R.B. (1995) *3K: The Cosmic Microwave Background*, Cambridge University Press
- Peacock J.A. (1997) *Mon. Not. R. Astr. Soc.*, **284**, 885
- Peacock J.A., Dodds S.J. (1994) *Mon. Not. R. Astr. Soc.*, **267**, 1020
- Peacock J.A., Dodds S.J. (1996) *Mon. Not. R. Astr. Soc.*, **280**, L19
- Peacock J.A., West M.J. (1992) *Mon. Not. R. Astr. Soc.*, **259**, 494
- Peebles P.J.E. (1973) *Astrophys. J.*, **185**, 413
- Peebles P.J.E. (1974) *Astrophys. J.*, **32**, 197
- Peebles P.J.E. (1982) *Astrophys. J.*, **263**, L1
- Peebles P.J.E. (1980) *The Large-scale Structure of the Universe*, Princeton University Press
- Peebles P.J.E. (1993) *Principles of Physical Cosmology*, Princeton University Press
- Pen U.-L., Seljak U., Turok N. (1997) *Phys. Rev. Lett.*, **79**, 1611
- Penzias A.A., Wilson R.W. (1965) *Astrophys. J.*, **142**, 419
- Perlmutter S. *et al.* (1998) astro-ph/9812133
- Pogosyan D., Starobinsky A.A. (1995) *Astrophys. J.*, **447**, 465
- Press W.H., Schechter P. (1974) *Astrophys. J.*, **187**, 425
- Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. (1992) *Numerical Recipes (2nd edition)*, Cambridge University Press
- Ratra B., Peebles P.J.E. (1988) *Phys. Rev. D*, **37**, 3406
- Reines F., Sobel H., Pasierb E. (1980) *Phys. Rev. Lett.*, **45**, 1307
- Rice S.O. (1954) in *Selected Papers on Noise and Stochastic Processes*, ed. Wax N., Dover, p133
- Riess A.G. *et al.* (1998) *Astr. J.*, **116**, 1009
- Rogerson J.B., York D.G. (1973) *Astrophys. J.*, **186**, L95
- Sachs R.K., Wolfe A.M. (1967) *Astrophys. J.*, **147**, 73
- Sakharov A.D. (1967) *JETP Lett.*, **5**, 24
- Saunders W., Frenk C., Rowan-Robinson M., Efstathiou G., Lawrence A., Kaiser N., Ellis R., Crawford J., Xia X.-Y., Parry I. (1991) *Nature*, **349**, 32
- Saunders W., Rowan-Robinson M., Lawrence A. (1992) *Mon. Not. R. Astr. Soc.*, **258**, 134
- Seljak U., Zaldarriaga M. (1996) *Astrophys. J.*, **469**, 437
- Seljak U. (1997) *Astrophys. J.*, **482**, 6
- Shanks T., Fong R., Boyle B.J., Peterson B.A. (1987) *Mon. Not. R. Astr. Soc.*, **227**, 739
- Shanks T., Boyle B.J. (1994) *Mon. Not. R. Astr. Soc.*, **271**, 753
- Shectman S.A., Landy S.D., Oemler A., Tucker D.L., Lin H., Kirshner R.P., Schechter P.L., (1996) *Astrophys. J.*, **470**, 172
- Smith M.S., Kawano L.H., Malaney R.A. (1993) *Astrophys. J. Suppl.*, **85**, 219
- Smoot G.F. *et al.* (1992) *Astrophys. J.*, **396**, L1
- Starobinsky A.A. (1985) *Sov. Astr. Lett.*, **11**, 133
- Strauss M.A., Davis M., Yahil Y., Huchra J.P. (1992) *Astrophys. J.*, **385**, 421
- Strauss M.A., Willick J.A. (1995) *Phys. Reports*, **261**, 271
- Sugiyama N. (1995) *Astrophys. J. Suppl.*, **100**, 281
- Sutherland W.J. (1988) *Mon. Not. R. Astr. Soc.*, **234**, 159
- Valls-Gabaud D., Alimi J.-M., Blanchard A. (1989) *Nature*, **341**, 215
- Vittorio N., Silk J. (1991) *Astrophys. J.*, **385**, L9
- Walker T.P., Steigman G., Kang H.-S., Schramm D.M., Olive K.A. (1991) *Astrophys. J.*, **376**, 51
- Weinberg S. (1989) *Rev. Mod. Phys.*, **61**, 1
- White M., Scott D., Silk J. (1994) *Ann. Rev. Astr. Astrophys.*, **32**, 319
- White M., Bunn E.F. (1995) *Astrophys. J.*, **450**, 477
- Willick J.A., Strauss M.A., Dekel A., Kollatt T. (1997) *Astrophys. J.*, **486**, 629
- Zaldarriaga M., Spergel D.N., Seljak U. (1997) *Astrophys. J.*, **488**, 1
- Zlatev I., Wang L., Steinhardt P.J. (1998) astro-ph/9807002

THE DETECTION OF GRAVITATIONAL WAVES.

S. Rowan and J. Hough

Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ

Abstract

Gravitational waves, one of the more exotic predictions of Einstein's General Theory of Relativity, may, after 80 years of controversy over their existence, be detected within the next decade. Sources such as coalescing compact binary systems, stellar collapses and pulsars are all possible candidates for detection. The most promising design of gravitational wave detector, offering the possibility of very high sensitivities over a wide range of frequency, uses test masses a long distance apart and freely suspended as pendulums on earth or in drag free craft in space; laser interferometry provides a means of sensing the motion of the masses produced as they interact with a gravitational wave. The main theme of this paper will be a review of the mechanical and optical principles used in the various long baseline systems being built around the world - LIGO (USA), VIRGO (Italy/France), TAMA300 (Japan) and GEO600 (UK/Germany) - and in LISA, a proposed space-borne interferometer.

1. INTRODUCTION

Gravitational Waves, predicted by General Relativity to result from the non-symmetrical acceleration of mass, may be directly detected within the next decade. Construction of a worldwide system of the largest optical interferometers ever to be built on earth (LIGO Project, USA [1], VIRGO Project, Italy/France [2], GEO 600 Project, Germany/UK [3] and the TAMA 300 Project, Japan [4]) is proceeding vigorously, and this detector array should have the capability of detecting gravitational wave signals from violent astrophysical events in the Universe.

Some early relativists were sceptical about the existence of gravitational waves however their reality is no longer in doubt. Indeed the evolution of the orbit of the binary pulsar, PSR 1913 +16, can only be explained if angular momentum and energy is carried away from this system by gravitational waves [5], and the 1993 Nobel Prize in Physics was awarded to Hulse and Taylor for their experimental observations and subsequent interpretations of this system [6,7].

Gravitational waves are produced when matter is accelerated in an asymmetrical way; but due to the nature of the gravitational interaction, significant levels of radiation are produced only when very large masses are accelerated in very strong gravitational fields. Such a situation cannot be found on earth but is found in a variety of astrophysical systems.

Gravitational wave signals are expected over a wide range of frequencies; from $\sim 10^{-17}$ Hz in the case of ripples in the cosmological background to $\sim 10^3$ Hz from the formation of neutron stars in supernova explosions. The most predictable sources are binary star systems. However there are many sources of much greater astrophysical interest associated with black hole interactions and coalescences, neutron star coalescences, stellar collapses to neutron stars and black holes (supernova explosions), pulsars, and the physics of the early Universe. For a full discussion of sources refer to the material contained in reference [8].

Why is there currently such interest worldwide in the detection of gravitational waves? Partly because observation of the velocity and polarisation states of the signals will allow a direct experimental check of the wave predictions of General Relativity; but more importantly because the detection of the signals will allow access to an as yet untapped source of information about astrophysical processes including these mentioned above. It is interesting to note that the gravitational wave signal from a coalescing compact binary star system has a relatively simple form and the distance to the source can be obtained from a combination of its signal strength and its evolution in time; if the redshift at that distance is found, Hubble's Constant - the value for which has been a source of lively debate for many years - may then be determined with a high degree of accuracy [9].

2. DETECTION OF GRAVITATIONAL WAVES

Gravitational waves are most simply thought of as ripples in the curvature of space-time, their effect being to change the separation of adjacent masses on earth or in space; this tidal effect is the basis of all present detectors. The problem for the experimental physicist is that the predicted magnitudes of the strains in space caused by gravitational waves are of the order of 10^{-21} or lower [6]. Indeed current theoretical models suggest that in order to detect a few events per year – from coalescing Neutron star binary systems for example – sensitivity close to 10^{-22} is required. Signal strengths at the earth for a number of sources are shown in Fig. 1.

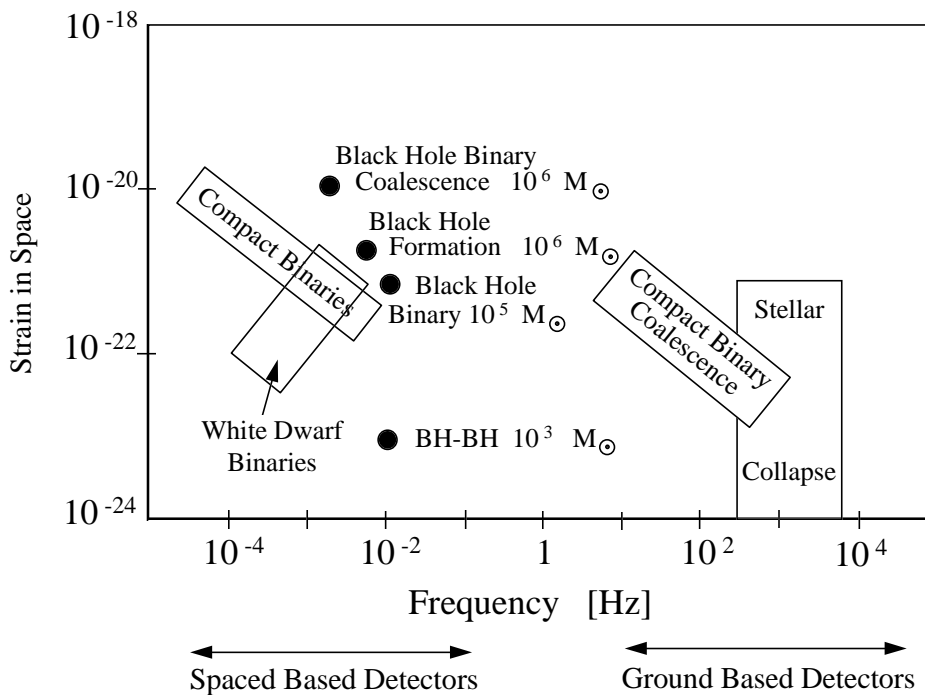


Fig. 1. Some possible sources for ground based and space-borne detectors

The small signal levels mean that limiting noise sources resulting from the thermal motion of molecules in the detector (thermal noise), from seismic or other mechanical disturbances, and from noise associated with the detector readout, whether electronic or optical, must be reduced to a very low level. For signals above ~ 10 Hz ground based experiments are possible, but for lower frequencies

where local fluctuating gravitational gradients and seismic noise on earth become a problem, it is best to consider developing detectors to be used in space [10].

2.1 Initial Detectors and their Development

The earliest experiments in the field were ground based and were carried out by Joseph Weber of the University of Maryland about 30 years ago. Having looked for evidence of excitation of the normal modes of the earth by very low frequency gravitational waves [11], Weber then moved on to look for tidal strains in aluminium bars which were at room temperature and were well isolated from ground vibrations and acoustic noise in the laboratory [12,13]. The bars were resonant at 1600Hz, a frequency where the energy spectrum of the signals from collapsing stars was predicted to peak. Despite the fact that Weber observed coincident excitations of his detectors placed up to 1000km apart, at a rate of approximately one event per day, his results were not substantiated by similar experiments carried out in several other laboratories in the USA, Germany, Britain and Russia. It seems unlikely that Weber was observing gravitational wave signals because, although his detectors were very sensitive, being able to detect strains of the order of 10^{-15} over millisecond timescales, their sensitivity was far away from what was predicted to be required theoretically. Development of Weber bar type detectors has continued with the emphasis being on cooling to reduce the noise levels, and currently systems at the Universities of Rome [14], Louisiana [15] and Perth (Western Australia) [16] are achieving sensitivity levels better than 10^{-18} for millisecond pulses. Bar detectors have a disadvantage, however, of being sensitive only to signals that have significant spectral energy in a narrow band around their resonant frequency.

2.2 Long Baseline Detectors on Earth

An alternative and very flexible design of gravitational wave detector offers the possibility of very high sensitivities over a wide range of frequency [1-4].

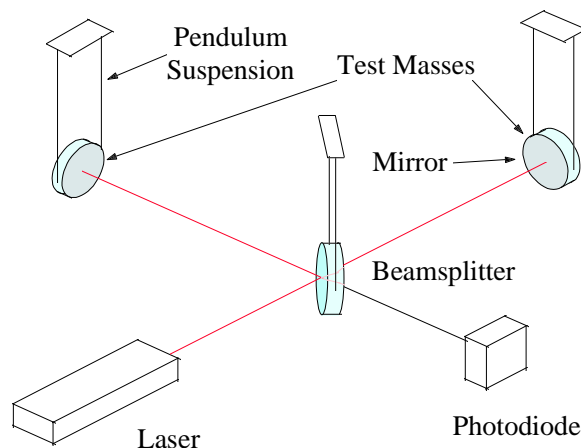


Fig.2 Schematic of gravitational wave detector using laser interferometry

This uses test masses a long distance apart and freely suspended as pendulums to isolate against seismic noise and reduce the effects of thermal noise; laser interferometry provides a means of sensing the motion of the masses produced as they interact with a gravitational wave (Fig.2). This technique is based on the Michelson interferometer and is particularly suited to the detection of gravitational waves as they have a quadrupole nature. Waves propagating perpendicular to the plane of the interferometer will result in one arm of the interferometer being increased in length while the other arm is decreased and vice versa. This results in a small change in the interference pattern of the light observed at the interferometer output. A typical design specification to allow a reasonable probability for detecting sources is to achieve a noise floor in strain smaller than $2 \times 10^{-23}/\sqrt{\text{Hz}}$. The distance between test masses possible on earth is limited to a few km by geographical and cost factors. If we assume an arm length of 3km the above specification sets the requirement that the residual motion of each test mass is smaller than $3 \times 10^{-20} \text{ m}/\sqrt{\text{Hz}}$ over the operating range of the detector, which may be from $\sim 10\text{Hz}$ to a few kHz, and requires that the optical detection system at the output of the interferometer must be good enough to detect such small motions.

2.3 Noise Sources which limit the Sensitivity of Interferometric Gravitational Wave Detectors

Fundamentally it should be possible to build systems using laser interferometry to monitor strains in space which are only limited by the Heisenberg Uncertainty Principle; however there are other practical issues which must be taken into account. Fluctuating gravitational gradients pose one limitation to the interferometer sensitivity achievable at low frequencies and it is the level of noise from this source which dictates that experiments to look for sub-Hz gravitational wave signals have to be carried out in space [17,18]. In general for ground based detectors the most important limitations to sensitivity are a result of seismic and other ground-borne mechanical noise, thermal noise associated with the test masses and their suspensions and shot noise in the photocurrent from the photodiode which detects the interference pattern. The significance of each of these sources will be briefly reviewed.

2.3.1 Seismic Noise

Seismic noise at a reasonably quiet site on the earth follows a spectrum in all three dimensions close to $10^{-7}/f^2 \text{ mHz}^{-1/2}$ and thus if the motion of each test mass has to be less than $3 \times 10^{-20} \text{ m}/\sqrt{\text{Hz}}$ at a frequency such as 30Hz then the level of seismic isolation required at 30Hz in the horizontal direction is greater than 10^9 . Since there is liable to be some coupling of vertical noise through to horizontal a significant level of isolation has to be provided in the vertical direction also. Thus the suspension systems for the test masses must be carefully designed to have this order of seismic isolation. Such isolation can be provided in a relatively simple way; e.g. by suspending each test mass as the last stage of a multiple pendulum system which itself is hung from a plate mounted on passive ‘rubber’ isolation mounts or on an active (electro-mechanical) anti-vibration system. The multiple pendulum system also has to be soft in the vertical direction to allow vertical isolation.

2.3.2 Thermal Noise

Of the many noise sources which may degrade the sensitivity of ground based laser interferometric gravitational wave detectors, thermal noise associated with the mirror masses and the last stage of their suspensions is likely to be the most significant at the lower end of the operating range of the detector [19]. The operating range of the detector lies between the resonances of the test masses and their pendulum suspensions, and thus it is the thermal noise in the tails of the resonant modes which is important. In order to keep the off-resonance thermal noise as low as possible the mechanical loss

factors of the material of the test masses and of the fibres or wires used to suspend the test masses need to be kept low. This is achieved if the mechanical quality factors of the masses and pendulum resonances are as high as possible. Indeed, to achieve the level of sensitivity discussed above the quality factor of the test masses (~15-20kg) must be $\sim 3 \times 10^7$ and the quality factor of the pendulum resonances should be greater than 10^8 . Discussions relevant to this are given in [20]. These are very high quality factors and they put significant constraints on the choice of material for the test masses and the suspending fibres as well as demanding that very low loss jointing techniques of the fibres to the masses be used. One viable solution is to use fused silica masses hung by fused silica fibres [21,22] although the use of other materials such as sapphire may be possible [23,20].

2.3.3 Photoelectron Shot Noise

As mentioned earlier it is very important that the system used for sensing the optical fringe movement on the output of the interferometer can sense strains in space of $2 \times 10^{-23}/\sqrt{\text{Hz}}$ or differences in the lengths of the two arms of less than 10^{-19} m, a minute displacement compared to the wavelength of light $\sim 10^{-6}$ m. The limitation is set by shot noise in the detected photocurrent and from consideration of the number of photoelectrons measured in a time τ it can be shown that the detectable strain sensitivity depends on the level of laser power (I_0) of wavelength λ used to illuminate the interferometer of arm length L , and on the time τ , such that:

$$\text{Detectable strain in time } \tau = 1/L [\lambda hc/4\pi^2 I_0 \tau]^{0.5}$$

$$\text{or Detectable strain}/\sqrt{\text{Hz}} = 1/L[\lambda hc/2\pi^2 I_0]^{0.5}$$

where c is the velocity of light.

Thus achievement of the required strain sensitivity level requires a laser, operating at a wavelength of 10^{-6} m, to provide 3×10^6 W power at the input to the interferometer. This is a formidable requirement.

However the situation can be helped greatly if a multipass arrangement is used in the arms of the interferometer as this multiplies up the apparent movement by the number of bounces. The multiple beams can either be separate as in an optical delay line, or may lie on top of each other as in a Fabry-Perot resonant cavity as shown in Fig. 3.

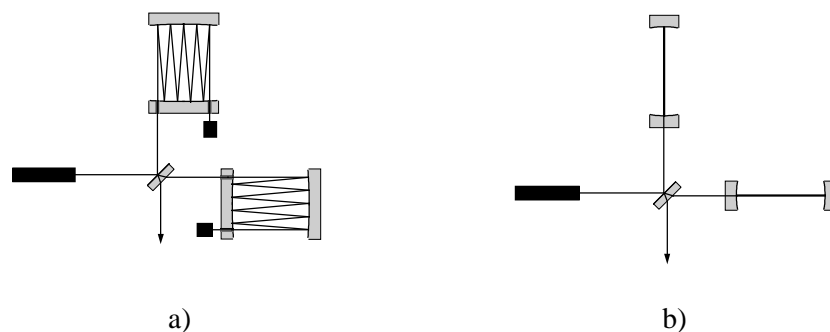


Fig.3. Michelson Interferometers with (a) delay lines and (b) Fabry-Perot cavities in the arms of the interferometer.

The number of bounces is set by the fact that the light should not stay in the arms longer than the characteristic timescale of the signal - otherwise some cancellation of the detected signal may

occur. Thus if signals of characteristic timescale 1msec are to be searched for the number of bounces should not exceed 50 for an arm length of 3km. With 50 bounces the required laser power is reduced to 1.2×10^3 W, still a formidable requirement.

It can be shown that the optimum signal to noise ratio in a Michelson interferometer is obtained when the arm lengths are such that the output light is very close to a minimum. (This is not intuitively obvious and is discussed more fully in [24]). In this situation if the mirrors are of very low optical loss, nearly all of the light, which is supplied to the interferometer, is reflected back towards the laser. In other words the laser is not properly impedance matched to the interferometer. The impedance matching can be improved by placing another mirror of carefully chosen transmission - a power recycling mirror - between the laser and the interferometer so that a resonant cavity is formed between this mirror and the rest of the interferometer and no light comes back towards the laser [25]. There is then a power build up inside the interferometer, a build up which can be high enough to create the required kilowatt of laser light at the beamsplitter from ~ 10 W or so out of the laser. This level of laser power is currently achievable from large single frequency Nd:YAG systems [26,27,28]. To enhance the sensitivity of the detector further and to allow some narrowing of the detection bandwidth, which may be valuable in searches for continuous wave sources of gravitational radiation, another technique known as signal recycling can be implemented [29,30,31]. This relies on the fact that sidebands created on the light by gravitational wave signals interacting with the arms do not interfere destructively and so do appear at the output of the interferometer. If a mirror of suitably chosen reflectivity is put at the output of the system as shown in Fig 4, then the sidebands can be recycled back into the interferometer where they resonate and hence the signal size over a given bandwidth (set by the mirror reflectivity) is enhanced.

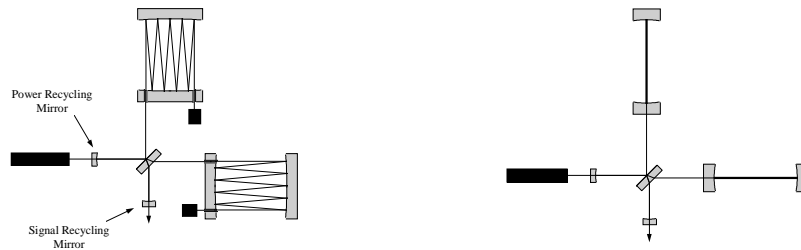


Fig.4. The implementation of power and signal recycling on the interferometers in the previous figure.

2.4 Long Baseline Detectors under Construction

Prototype detectors using laser interferometry have been constructed by various research groups around the world - at the Max-Planck-Institut für Quantenoptik in Garching [32], at the University of Glasgow [33], at California Institute of Technology [34], at the Massachusetts Institute of Technology [35], at the Institute of Space and Astronautical Science in Tokyo [36] and at the astronomical observatory in Tokyo [37]. These detectors have arm lengths varying from 10m to 100m and have or had either multibeam delay lines or resonant Fabry-Perot cavities in their arms. Several years ago the sensitivities of some of these detectors reached a level - better than 10^{-18} for millisecond pulses - where it was sensible to decide to build detectors of much longer baseline which should be capable of reaching the performance required to have a real possibility of detecting gravitational waves. It should be noted that in order to improve the confidence level of any detection and to obtain the location of the source a number of interferometers are required worldwide. Thus an international network of gravitational wave detectors is now under construction. The American LIGO project comprises the

building of two detector systems with arms of 4km length, one in Hanford, Washington State, and one in Livingston, Louisiana. The vacuum system, laser and input optics and first suspension system for the detector in Hanford are now installed and the vacuum system is in place in Louisiana. The French/Italian VIRGO detector of 3km arm length at Cascina near Pisa is at the stage where the central buildings are close to completion and vacuum tanks to house the interferometry are being installed. It should be noted that this detector which uses five-stage multi-pendulum systems for the suspension of its test masses is specially designed to be able to operate down to approximately 10Hz. The TAMA 300 detector, which has arms of length 300m, is at a relatively advanced stage of construction at the Tokyo Astronomical Observatory. This detector is being built mainly underground; the vacuum system is complete and initial operation with light in the arms has started. All the systems mentioned above are designed to use resonant cavities in the arms of the detectors, use standard wire sling techniques for suspending the test masses, and are to be illuminated by infra-red light from a Nd:YAG laser.

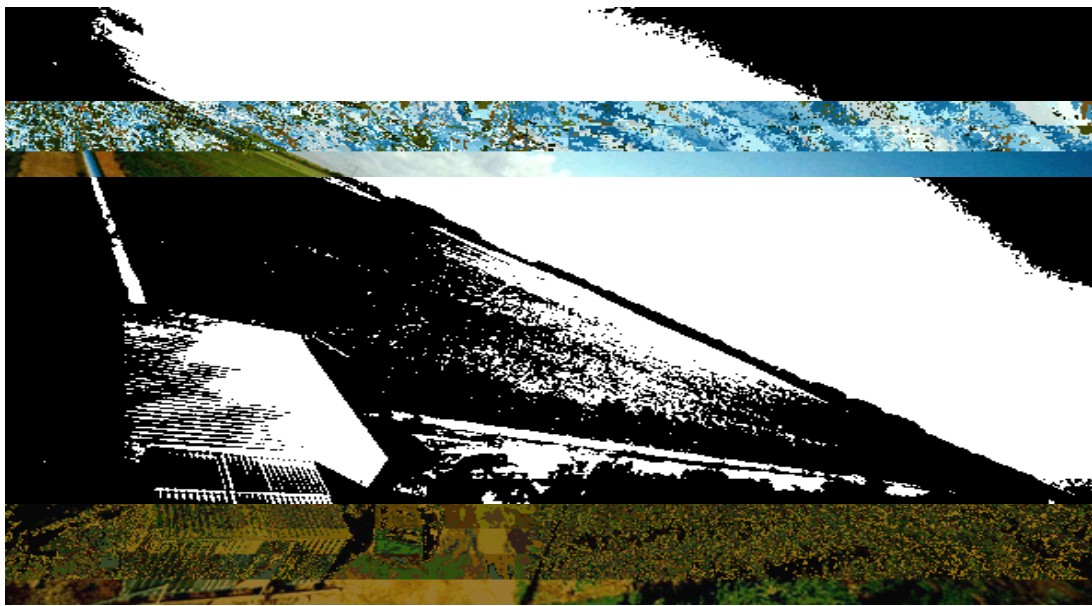


Fig. 5 A birds eye view of the GEO 600 detector, sited in Ruthe, near Hannover in Germany

The German/British project - GEO 600 - for a 600m detector near Hanover is somewhat different. It makes use of a four pass delay line system with advanced optical signal enhancement techniques, utilises very low loss fused silica suspensions for the test masses, and should have a sensitivity at frequencies above a few hundred Hz comparable to the first phases of VIRGO and LIGO when they are in operation at the beginning of next century. Illumination is again to be by infra-red light, provided by a 10W single frequency YAG laser. Construction is advancing well with the necessary buildings and vacuum pipes for the arms being in place. Installation of the suspensions for the optical elements is now beginning. A birds-eye view of the site showing the central building and the directions of the two arms is shown in Fig. 5.

This detector is based on the fundamental research carried out over many years at the Max-Planck-Institut für Quantenoptik, Garching, the University of Glasgow, the University of Wales, Cardiff, and more recently the University of Hanover and the Albert Einstein Institut, Potsdam. In two years time initial operation of the detector is expected to commence and during the following years we can

expect some very interesting coincidence searches for gravitational waves, at a sensitivity level of approximately 10^{-21} for pulses of several milliseconds duration.

2.5 Longer Baseline Detectors in Space

Searches for gravitational wave sources at low frequency have been underway for a number of years using their possible effect on the phase of the radio signals from the millisecond pulsars PSR 1937+21 and PSR 1855+09 [38] and their possible effect on the Doppler shift of radar signals transponded back from space craft such as Ulysses [39]. The former experiment, sensitive to signals of frequency of the order of one cycle per year and lower, has been used to set a limit on the gravitational wave background in this frequency regime corresponding to an energy density per logarithmic frequency interval of less than 6×10^{-8} of the closure density of the Universe. The latter experiments have searched for continuous wave signals in the mHz region and have set a limit of approximately 3×10^{-15} for such signals. However perhaps the most interesting sources of gravitational waves - those resulting from black hole formation and coalescence - lie in the region of 10^{-4} Hz to 10^{-1} Hz and a detector whose strain sensitivity is approximately 10^{-23} over relevant timescales is required to search for these. The most promising way of looking for such signals is to fly a laser interferometer in space i.e. to launch a number of drag free space craft into orbit and to compare the distances between test masses in these craft along arms making significant angles with each other using laser interferometry. Two such experiments have been proposed. The first, LISA [8] is being proposed by an American/European team; it consists of an array of 3 drag free spacecraft at the vertices of an equilateral triangle of length of side 5×10^6 km, and this cluster is placed in an Earth-like orbit at a distance of 1 AU from the Sun, and 20 degrees behind the Earth. Proof masses inside the spacecraft (two in each spacecraft) form the end points of three separate but not independent interferometers. Each single two-arm Michelson type interferometer is formed from a vertex (actually consisting of the proof masses in a 'central' spacecraft), and the masses in two remote spacecraft as indicated in Fig. 6.

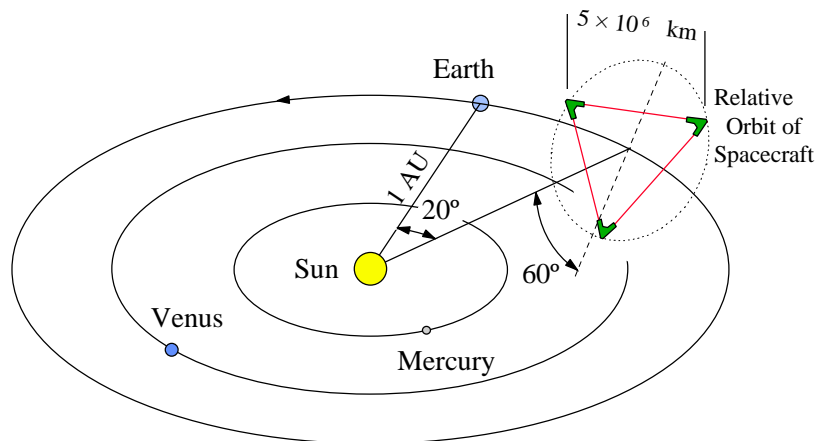


Fig.6 The proposed LISA detector

The three-interferometer configuration provides redundancy against component failure, gives better detection probability, and allows the determination of polarisation of the incoming radiation. The spacecraft in which they are accommodated shields each pair of proof masses from external disturbances (e.g. solar radiation pressure). Drag free control servos enable the spacecraft to follow the proof masses to a high level of precision, the drag compensation being effected using proportional

electric thrusters. Illumination of the interferometers is by highly stabilised laser light from Nd:YAG lasers at a wavelength of 1.064 microns. For each interferometer - consisting of a central spacecraft and two distant spacecraft - the two lasers in the central spacecraft are phase locked together so they effectively behave as a single laser. The lasers in the end spacecraft are phase locked to the incoming light, and thus act as amplifying mirrors.

LISA, in a slightly earlier form with six spacecraft, two at each vertex, was adopted by ESA as a Cornerstone project in their post Horizon 2000 programme. However the timescale of this programme is somewhat long and following suggestions from the LISA Team group at JILA the lower cost three spacecraft version has recently been studied at JPL and could allow the possibility of an earlier launch as a NASA led or ESA led collaborative medium scale mission.

The second experiment, OMEGA [40], calls for three craft to be placed in a geocentric orbit, the arm length in this case being 10^9 m and this is currently being proposed to NASA as a MIDEX mission.

3. CONCLUSION

A large amount of effort worldwide is now being invested in the development of both ground and spaced based searches for gravitational radiation and we are entering a new era where the signals from neutron star and black hole interactions will widen our understanding of the Universe. Beyond this, however, there is the very exciting prospect that gravitational wave astronomy will be like radio astronomy and X-ray astronomy and will allow the discovery of very active sources currently unknown to us.

ACKNOWLEDGEMENTS

We wish to thank G. Cagnoli for assistance with the preparation of this paper and PPARC and the University of Glasgow for support.

REFERENCES

- [1] B.C. Barish, "LIGO: Status and Prospects" Proc. of Conference on Gravitational Wave Detection, Tokyo, Eds. K. Tsubono, M.K. Fujimoto, K. Kuroda, Universal Academy Press Inc, Tokyo (1997) 163-173
- [2] J.-Y. Vinet et al, Proc. of Gravitation and Cosmology, ICGC-95 Conference, Pune, India 13-19 Dec 1995 Astrophysics and Space Science Library, Kluwer Academic Publishers 211 (1997) 89-93
- [3] J. Hough et al "GEO 600: Current status and some aspects of the design" Proc. of Conference on Gravitational Wave Detection, Tokyo, Eds. K. Tsubono, M.K. Fujimoto, K. Kuroda, Universal Academy Press Inc, Tokyo (1997) 175-182
- [4] K. Tsubono and the TAMA collaboration, "TAMA project" Proc. of Conference on Gravitational Wave Detection, Tokyo, Eds. K. Tsubono, M.K. Fujimoto, K. Kuroda, Universal Academy Press Inc, Tokyo (1997) 183-191
- [5] C.M. Will, "Theory and experiment in gravitational physics" Revised edition, Cambridge University Press, Cambridge (1983)
- [6] R.A. Hulse Rev. Mod. Phys. 66 (1994) 699

- [7] J.H. Taylor *Rev. Mod. Phys.* 66 (1994) 711.
- [8] “First International LISA Symposium”, *Classical and Quantum Gravity* 14, 6, (1997); “Gravitational Waves, Sources and Detectors”, Edoardo Amaldi Foundation Series Volume 2, Ed. I. Ciufolini and F. Fidecaro, World Scientific, (1997)
- [9] B.F. Schutz, “Determining the nature of the Hubble constant” *Nature* 323 (1986) 310
- [10] K. Danzmann et al, “LISA: Laser Interferometer Space Antenna for Gravitational Wave Measurements” *Class. Quant. Grav.* 13 11A (1996) A247-50
- [11] R.L. Forward, D. Zipoy, J. Weber, S. Smith, H. Benioff, 189 *Nature* 473 (1961)
- [12] J. Weber, *Phys. Rev. Lett.* 22 (1969) 1320
- [13] J. Weber, *Phys. Rev. Lett* 25 (1970) 180
- [14] P. Astone, M. Bassan, P. Bonifazi, P. Carelli, E. Coccia, C. Cosmelli, V. Fafone, S. Frasca, A. Marini, G. Mazzitelli, Y. Minenkov, I. Modena, G. Modestino, A. Moleti, G.V. Palletino, M.A. Papa, G. Pizzella, P. Rapagnani, F. Ricci, F. Ronga, R. Terenzi, M. Visco, L. Votano, International Conference on Low Temperature Physics, Prague, Czech Republic 8-14 Aug 1996, *Czech. J. Phys.* 46 suppl. pt S5 (1996) 2907-8
- [15] E. Amaldi, O. Aguiar, M. Bassan, P. Bonifazi, P. Carelli, M.G. Castellano, G. Cavallari, E. Coccia, C. Cosmelli, W.M. Fairbank, S. Frasca, V. Foglietti, R. Habel, W.O. Hamilton, J. Henderson, W. Johnson, K.R. Lane, A.G. Mann, M.S. McAshan, P.F. Michelson, I. Modena, G.V. Palletino, G. Pizzella, J.C. Price, R. Rapagnani, F. Ricci, N. Solomonson, T.R. Stevenson, R.C. Taber, B.-X. Xu, “First gravity wave coincident experiment between resonant cryogenic detectors: Louisiana-Rome-Stanford” *Astron. and Astro.* 216 1-2 (1989) 325-32
- [16] I.S. Heng, D.G. Blair, E.N. Ivanov, M.E. Tobar “Long term operation of a niobium resonant bar gravitational wave antenna” *Phys. Lett. A* 218 3-6 (1996) 190-6
- [17] R.E. Spero in “Science Underground” ed. M.M. Nieto et al. AIP New York, (1983)
- [18] P.R. Saulson, “Terrestrial gravitational noise on a gravitational wave antenna” *Phys. Rev. D* 30 (1984) 732
- [19] P.R. Saulson, “Thermal noise in mechanical experiments” *Phys. Rev. D* 42 (1990) 2437-2445
- [20] S. Rowan, S.M. Twyford and J. Hough, “The design of low loss suspensions for advanced gravitational wave detectors” *Proc. 2nd E. Amaldi conf. on gravitational waves, CERN, Geneva July (1996)* In press
- [21] V. B. Braginsky, V.P. Mitrofanov, K.V. Tokmakov, “Energy dissipation in the pendulum mode of the test mass suspension of a gravitational wave antenna” *Phys. Lett. A* 218 (1996) 164-166
- [22] S. Rowan, S.M. Twyford, R. Hutchins, J. Kovalik, J.E. Logan, A.C. McLaren, N.A. Robertson, J. Hough, “Q factor measurements on prototype fused quartz pendulum suspensions for use in gravitational wave detectors” *Phys. Lett. A* 233 (1997) 303-308

- [23] L. Ju, M. Notcutt, D. Blair, F. Bondu, C.N. Zhao, "Sapphire beamsplitters and test masses for advanced laser interferometric gravitational wave detectors" *Phys. Lett. A* 218 3-6 (1996) 197-206
- [24] W.A. Edelstein, J. Hough, J.R. Pugh, W. Martin, "Limits to the measurement of displacement in an interferometric gravitational radiation detector" *J. Phys. E (Scientific Instruments)* 11 7 (1978) 710
- [25] R.W.P. Drever, J. Hough, A.J. Munley, S.-A. Lee, R. Spero, S.E. Whitcomb, H. Ward, G.M. Ford, M. Hereld, N.A. Robertson, I. Kerr, J.R. Pugh, G.P. Newton, B. Meers, E.D. Brooks III, Y. Gursel, in "Quantum optics, experimental gravitation and measurement theory" Eds. P. Meystre, M.O. Scully, Plenum Press, New York, (1983) 503-514
- [26] O. Cregut, C.N. Man, D. Shoemaker, A. Brillat, A. Menhert, P. Peuser, N.P. Schmitt, P. Zeller, K. Wallermoth, "18W single-frequency operation of an injection-locked, CW, Nd:YAG laser, *Phys. Lett. A* 140 6 (1989) 294-8
- [27] D. Golla, I. Freitag, H. Zellmer, W. Schone, I. Kropke, H. Welling, "15W single-frequency operation of a CW diode laser-pumped Nd:YAG ring laser" *Opt. Comm.* 98 1-3 (1993) 86-90
- [28] R.J. Shine Jr., A.J. Alfrey, R.L. Byer, "40W TEM₀₀-mode, diode-laser-pumped Nd:YAG miniature-slab laser", *Opt. Lett.* 20 5 (1995) 459-61
- [29] B.J. Meers, "Recycling in laser-interferometric gravitational-wave detectors" *Phys. Rev. D, (Particles and Fields)* 38 8 (1988) 2317-26
- [30] K.A. Strain, B.J. Meers, "Experimental demonstration of dual recycling for interferometric gravitational-wave detectors" *Phys. Rev. Lett.* 66 11 (1991) 1391-4
- [31] G. Heinzel, K.A. Strain, J. Mizuno, K.D. Skeldon, B. Willke, W. Winkler, R. Schilling, K. Danzmann, "An experimental demonstration of dual recycling on a suspended interferometer" *Phys. Rev. Lett.* (1998) In press.
- [32] D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, A. Rüdiger, "Noise behaviour of the Garching 30-meter prototype gravitational-wave detector" *Phys. Rev. D (Particles and Fields)* 38 2 (1988) 423-32
- [33] D. Robertson, E. Morrison, J. Hough, S. Killbourn, B.J. Meers, G.P. Newton, K.A. Strain, H. Ward, "The Glasgow 10m prototype laser interferometric gravitational wave detector" *Rev. Sci. Instr.* 66 9 (1995) 4447-52
- [34] A. Abramovici, W. Althouse, J. Camp, D. Durance, J.A. Giaime, A. Gillespie, S. Kawamura, A. Kuhnert, T. Lyons, F.J. Raab, R.L. Savage, Jr., D. Shoemaker, L. Sievers, R. Spero, R. Vogt, R. Weiss, S. Whitcomb, M. Zucker, "Improved sensitivity in a gravitational wave interferometer and implications for LIGO" *Phys. Lett. A*, 218 3-6 (1996) 157-63
- [35] P. Fritschel, G. Gonzalez, B. Lantz, P. Saha, M. Zucker, "High power interferometric phase measurement" *Phys. Rev. Lett.* 80 15 (1998) 3181-4
- [36] A. Araya, N. Mio, K. Tsubono, K. Suehiro, S. Telada, M. Ohashi, M.-K. Fujimoto, "Optical mode cleaner with suspended mirrors" *Appl. Opt.* 36 7 (1997) 1446-53

- [37] E. Mizuno, N. Kawashima, S. Miyoke, E.G. Heflin, K. Wada, W. Naito, S. Nagano, K. Arakawa, "Effort of stable operation by noise reduction of 100m DL laser interferometer [TENKO-100] for gravitational wave detection" Proc. VIRGO 96, Cascina, World Scientific (1996) 108-110
- [38] V.M. Kaspi, J.H. Taylor, M. Ryba, "High precision timing of millisecond pulsars. III. Long-term monitoring of PSRs B1885+09 and B1937+21" Astr. J. 248 2 pt.1 (1994) 712-28
- [39] B. Bertotti, R. Ambrosini, J.W. Armstrong, S.W. Asmar, G. Comoretto, G. Giampieri, L. Iess, Y. Koyama, A. Messeri, A. Vecchio, H.D. Wahlquist, "Search for gravitational wave trains with the spacecraft ULYSSES" Astron. and Astro. 296 1 (1995) 13-25
- [40] R.W. Hellings, "Gravitational wave detectors in space" Cont. Phys. 37 6 (1996) 457-69

AN INTRODUCTION TO LATTICE QCD

R. D. Kenway

Department of Physics and Astronomy, The University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, Scotland

Abstract

Numerical simulation of QCD is enabling us to calculate strong-interaction effects reliably and thereby to test the Standard Model outside the perturbative regime. I introduce the key features of these simulations, describe the basic first step in which the quark masses are determined and review the status of results for the light hadron spectrum.

1 WHY DO WE NEED NUMERICAL SIMULATIONS OF QCD?

While perturbation theory is reliable for QCD at high energies thanks to asymptotic freedom, the low-energy properties of the theory are intrinsically non-perturbative. It seems unlikely that the complexities of the hadron spectrum and matrix elements will ever be obtained analytically, although we continue to hope for insight into the mechanism by which QCD confines its elementary fields. Fortunately, direct numerical simulation of lattice QCD can proceed from first principles, without any ad hoc assumptions, to compute with arbitrary precision much of the low-energy physics. At present there is a high price attached to doing this, due to the need for very high performance computers.

There are several objectives to this work. In the first instance, we hope to achieve a reliable and systematically-improvable means of calculating non-perturbative aspects of QCD. Perhaps this level of control will lead to an understanding of the mechanism of confinement. Secondly, despite its successes, the Standard Model rests on perturbation theory. Very little of it has been tested non-perturbatively. Many of the Standard Model parameters are obscured experimentally by the need to know hadronic matrix elements. Numerical simulation can provide estimates of these and, perhaps, thereby expose inconsistencies in the Standard Model. Finally, numerical simulation can predict hadronic states, such as glueballs and hybrids, which have yet to be identified experimentally, as well as new phases of QCD, such as the quark-gluon plasma and colour superconductivity, which may exist at non-zero temperature and/or baryon density, and could have important implications for astrophysics.

Thus, there is a lot of physics which we are missing, because it is intrinsically non-perturbative. Numerical simulation is becoming a powerful tool for studying non-perturbative features of quantum field theories and, specifically for QCD, it is likely to play an important role in the search for physics beyond the Standard Model.

In this lecture I will introduce lattice QCD and describe some recent results for the light hadron spectrum (further details can be found in [1]). Christine Davies' lecture [2] will tell you much more about the phenomenological applications, particularly to the physics of heavy quarks. Several excellent reviews of lattice QCD have appeared recently [3, 4, 5] and, if you want to know the current state of the art, the proceedings of the annual lattice conference [6] gives a complete account.

1.1 Confinement from first principles

It is quite straightforward to demonstrate numerically one of the expected features of confinement, namely the linearly-rising potential between a quark and an antiquark. An example is

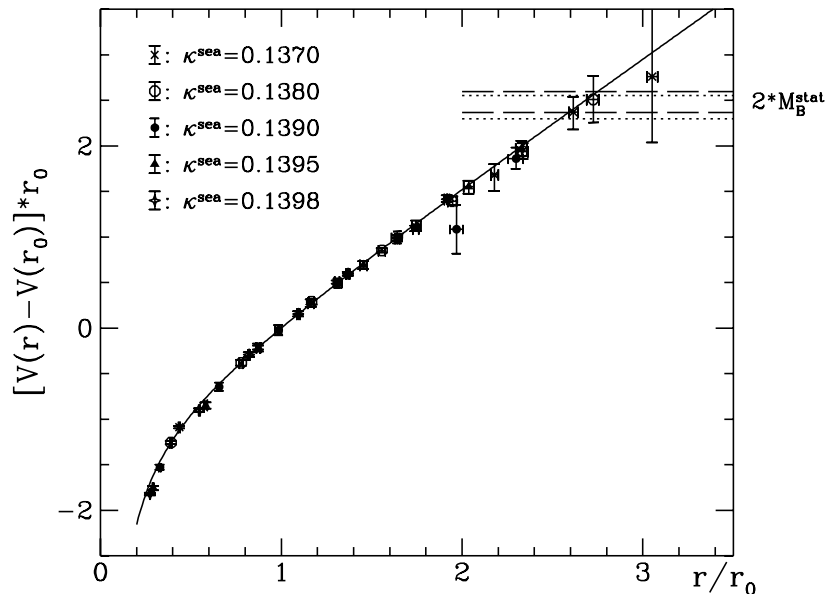


Fig. 1: The potential energy, $V(r)$, between a static quark and antiquark as a function of their separation r (scaled relative to a fixed physical length r_0), computed by the UKQCD Collaboration [7].

shown in Fig 1, where the data are from simulations with different quark masses (parametrised by κ^{sea} in the figure).

Our intuitive picture of confinement is that a flux tube forms between the quark and the antiquark, giving rise to a constant attractive force. The energy in this confining string therefore builds up linearly with separation, until it becomes energetically favourable for the string to break, with creation from the vacuum of a quark-antiquark pair. The energy at which this is expected to happen is roughly the mass of the two mesons which are formed, and this is indicated in Fig 1 by horizontal dashed lines. Once the string breaks, the potential should flatten out. So far, QCD simulations have only just reached large enough distances to be able to observe this, and a clear signal has not yet been obtained. However, string breaking was demonstrated this year in a gauge-Higgs model [8], so this next step in understanding quark confinement is probably close at hand.

1.2 Parameters of the Standard Model

The basic input parameters for QCD are the quark masses. Since quarks are not asymptotic states, due to confinement, the relationship between the quark masses and any physical observable of QCD is complicated. In a quantum field theory, the parameters in the Lagrangian are fixed by a set of renormalisation conditions. The most direct condition to use for quark masses is to require, for each quark flavour, that the mass of a hadron which contains that flavour is the same as observed experimentally.

The resulting quark mass depends on the regularisation scheme used to calculate the hadron mass, ie the lattice, and on the momentum scale, ie the lattice momentum cut-off. It is possible to relate one regularisation scheme to another and to use the renormalisation group to change the momentum scale. By convention, quark masses are quoted in the $\overline{\text{MS}}$ scheme at a scale of 2 GeV.

Other parameters of the Standard Model also require QCD calculations in order to de-

Table 1: The 19 Standard Model parameters (excluding neutrino masses).

11 masses	m_u	m_c	m_t
	m_d	m_s	m_b
	m_e	m_μ	m_τ
	m_Z	m_H	
3 couplings	α	G_F	α_s
3 mixing angles	V_{ud}	V_{us}	V_{ub}
+	V_{cd}	V_{cs}	V_{cb}
1 phase	V_{td}	V_{ts}	V_{tb}
1 vacuum angle	θ		

termine them from experiment. The Standard Model has passed an impressive number of experimental tests, but, even so, we are convinced that it is not a complete theory. Obviously, it excludes the gravitational force. We expect that all four forces have a common origin and that some more complete theory will demonstrate this. Secondly, even excluding the neutrino masses, the Standard Model has 19 free parameters, given in Table 1, which seems too many for them all to be fundamental constants. Surprisingly, only the lepton masses (m_e , m_μ and m_τ), the fine structure constant (α), the Fermi constant (G_F) and two of the CKM matrix elements (V_{ud} and V_{us}) are known to better than 1%. Along with the quark masses, the other elements of the CKM matrix are obscured by hadronic uncertainties.

1.3 The search for new physics

Our interest in knowing better the Standard Model parameters is that the present uncertainties may obscure fundamental inconsistencies. If these are present, they may provide clues to physics beyond the Standard Model. There are several other ways to discover new physics, such as the observation of a supersymmetric particle, or the search for the Higgs boson which must reveal something of the origin of mass, but these may have to wait for the LHC. The most immediate place to look is probably the CKM matrix and the question of why there are three generations.

We know there are exactly three light neutrinos from the observed decay $Z^0 \rightarrow \bar{\nu}\nu$, which provides a measurement of their number:

$$N_\nu = 2.99(2). \quad (1)$$

The consequence of three generations is that the CKM matrix,

$$\begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} = \text{unitary}, \quad (2)$$

and so there are non-trivial relationships between the CKM matrix elements, embodied in the so-called unitarity triangles, such as

$$V_{ub}V_{ud}^* + V_{cb}V_{cd}^* + V_{tb}V_{td}^* = 0. \quad (3)$$

The three complex numbers in Eq (3) form a triangle in the complex plane, whose area is proportional to the amount of CP violation in the Standard Model. If we knew the CKM matrix elements more accurately, we could test the unitarity triangles and, if they close, obtain an indirect measure of CP violation. Of course, we hope to see something going wrong with this picture.

The determination of the CKM matrix elements involves measuring quark decays. The imminent B Factories will provide a wealth of new data on the decays of the b quark, which correspond to the most poorly-known CKM matrix elements. However, b quarks cannot be isolated in the laboratory and always decay inside hadrons. Hence, we have to be able to disentangle the strong-interaction effects in these decays and this usually means being able to compute hadronic matrix elements reliably. Numerical simulation of lattice QCD is essential for getting the most out of B Factories.

2 HOW IS THE COMPUTATION DONE?

Over the 25 years since Wilson invented lattice QCD [9], there have been very considerable technical advances, both in the formulation of QCD on a spacetime lattice and in the numerical algorithms used to simulate it. Progress towards realistic simulations during that time has been as much due to these theoretical developments as to the advances in computer technology. Here, I will try to convey an impression of what, on the one hand, makes QCD a difficult system to simulate and, on the other, gives us confidence that realistic simulations are within reach.

2.1 Lattice discretisation

The starting point is the functional integral for QCD in Euclidean spacetime (ie imaginary time),

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int \mathcal{D}A \mathcal{D}\bar{q} \mathcal{D}q \mathcal{O}[A, \bar{q}, q] e^{-S[A, \bar{q}, q]}. \quad (4)$$

This defines the expectation value of some product, \mathcal{O} , of the gluon, quark and antiquark fields, A , q and \bar{q} , as the average of \mathcal{O} over all possible field configurations, weighted with the exponential of (minus) the Euclidean action, S . Euclidean spacetime is used so that Eq (4) is formally identical to the canonical distribution in statistical mechanics. The resulting probabilistic interpretation of the functional integral then leads naturally to approximating the Euclidean spacetime by a finite $L^3 \times T$ hypercubic lattice, of spacing a , and evaluating the average by Monte Carlo.

It was Wilson [9] who realised that a spacetime lattice could be introduced while preserving exact local gauge invariance at the lattice sites. The non-zero lattice spacing regulates QCD in a very natural way, which emphasises its geometrical aspects and does not require any gauge fixing.

However, the lattice obviously breaks the Euclidean spacetime symmetries, and these must be recovered by tuning the lattice spacing to be much smaller than any other relevant length scale. An engineer's intuition then tells you that the results should become independent of the lattice spacing. Rigorously, this amounts to tuning the gauge coupling to a critical point of the statistical mechanical system, corresponding to asymptotic freedom, ie making contact with the perturbative regime of the theory.

Less obviously, the lattice explicitly breaks chiral symmetry. This symmetry is responsible for ensuring that if the input quark mass is zero, then the quark mass remains zero at all orders in perturbation theory, and the pion is an exact Goldstone boson of the spontaneously-broken chiral symmetry. The fact that the lattice explicitly breaks chiral symmetry means that quark masses are additively renormalised and we have to adjust the up and down quark masses used in the simulations to find a value at which the pion is massless. This is only a minor practical inconvenience, although it complicates the picture of how spontaneous chiral symmetry breaking comes about.

A limitation of today's numerical algorithms is that they only reproduce Eq (4) exactly for degenerate pairs of quark flavours. Approximate algorithms exist for non-degenerate flavours.

As we shall see, this is of little practical relevance today, because the computational cost is such that most simulations have used the quenched approximation corresponding to zero flavours, or, at best, two flavours in an attempt to simulate up and down quarks.

The simulation of lattice QCD proceeds by taking the gauge coupling and the masses of the N_f quark flavours as inputs. The Monte Carlo method is used to evaluate expectation values of products of fields from which, as I will indicate in the next section, it is possible to extract hadron masses and matrix elements. Then the experimental values of $N_f + 1$ hadron masses are used to fix the quark masses and the overall scale, ie the unit of measurement. In practice, this means that the quark masses are adjusted until N_f hadron masses, in units of the $(N_f + 1)$ th hadron mass, match experiment. The gauge coupling is tuned towards zero to approach the continuum limit. The lattice spacing is judged to be small enough when dimensionless ratios of hadron masses become independent of the gauge coupling, which is called ‘scaling’.

2.2 Quantities which can be computed easily

The starting point is usually to calculate hadron energies. These may be obtained from two-point correlation functions, which are expectation values of products of fields localised at two spacetime points, separated by a distance τ in Euclidean time, eg

$$\langle \mathcal{O}^\dagger(\tau)\mathcal{O}(0) \rangle = \langle 0 | T[\hat{\mathcal{O}}^\dagger(\tau)\hat{\mathcal{O}}(0)] | 0 \rangle \quad (5)$$

$$= \langle 0 | \hat{\mathcal{O}}^\dagger e^{-\hat{H}\tau} \hat{\mathcal{O}} | 0 \rangle \quad (6)$$

$$= \sum_n |\langle n | \hat{\mathcal{O}} | 0 \rangle|^2 \frac{e^{-E_n\tau}}{2E_n}. \quad (7)$$

Here, in Eq (5), the basic result which relates path integrals to vacuum expectation values of time-ordered products of Heisenberg fields has been used. Next, in Eq (6) the fields have been transformed to the Schrödinger representation, by introducing the Hamiltonian operator in imaginary time and, finally, in Eq (7), a complete set of states, $\{|n\rangle\}$, with energies E_n , has been inserted. It follows that the two-point function falls exponentially at large Euclidean time, τ , with a rate given by the energy of the lightest state which has a non-zero overlap with $\hat{\mathcal{O}}|0\rangle$. Choosing \mathcal{O} with particular quantum numbers enables us to extract the energy of the lightest hadron with those quantum numbers. At zero spatial momentum, this is just the hadron mass.

Numerically, the masses are obtained by fitting Eq (7) to a sum of exponentials. As a by-product we also obtain the matrix elements in Eq (7). If $\hat{\mathcal{O}}|0\rangle$ has a non-zero overlap with a pseudoscalar meson, eg $\hat{\mathcal{O}} = \bar{q}_1\gamma_0\gamma_5q_2$, then the leading matrix element is proportional to the pseudoscalar decay constant, f_{PS} ,

$$im_{\text{PS}}f_{\text{PS}} = \langle 0 | \bar{q}_1\gamma_0\gamma_5q_2 | \text{PS} \rangle \quad (8)$$

where m_{PS} is the pseudoscalar meson mass. This illustrates the basic technique for extracting masses and matrix elements.

The next level of sophistication is to extract single-particle matrix elements from three-point functions. These are expectation values of products of fields at three different Euclidean times, eg, using a similar argument to that in Eqs (5)–(7),

$$\langle \pi(\vec{p}, \tau_2)\mathcal{O}(\vec{q}, \tau_1)K(0) \rangle = \langle 0 | \hat{\pi}(\vec{p})e^{-\hat{H}(\tau_2-\tau_1)}\hat{\mathcal{O}}(\vec{q})e^{-\hat{H}\tau_1}\hat{K}|0 \rangle \quad (9)$$

$$= \sum_{n,n'} \langle 0 | \hat{\pi}(\vec{p}) | n \rangle \frac{e^{-E_n(\tau_2-\tau_1)}}{2E_n} \langle n | \hat{\mathcal{O}}(\vec{q}) | n' \rangle \frac{e^{-E_{n'}\tau_1}}{2E_{n'}} \langle n' | \hat{K} | 0 \rangle. \quad (10)$$

The single-particle energies and the matrix elements involving the vacuum may all be obtained from two-point functions, as in Eq (7). Hence, at large time separations, $\tau_2 \gg \tau_1 \gg 0$, it is

possible to isolate, for example, the matrix element

$$\langle \pi(\vec{p}) | \bar{s} \gamma_\mu u(\vec{q}) | K(\vec{p} - \vec{q}) \rangle \quad (11)$$

for the semileptonic decay $K \rightarrow \pi e \nu$, or, with other choices of operators, matrix elements for $B \rightarrow K^* \gamma$, $B^0 \bar{B}^0$ mixing, etc.

Unfortunately, multiparticle final states involve complex amplitudes in Minkowski space and are much more difficult to extract from this type of analysis. So the current state of the art is limited to computing matrix elements which involve, or can be related to matrix elements which involve, only single-particle final states. Fortunately, this spans a wide range of useful phenomenology.

2.3 The size of the lattice

The numerical simulation is carried out in terms of dimensionless variables, with a lattice spacing of 1. So a dimensionful quantity must be input to set the scale and, thereby, determine the lattice spacing in physical units. Often the ρ meson mass is used for this purpose. If the simulation is to match experiment, then the ratio of the Compton wavelength of the ρ meson to the lattice spacing must be the same in the simulation and in the laboratory,

$$\frac{\rho \text{ size}}{\text{lattice spacing}} = \frac{m_\rho^{-1}}{1} \Big|_{\text{computer}} = \frac{(770 \text{ MeV})^{-1}}{a} \Big|_{\text{lab}}. \quad (12)$$

We have seen in the previous section how to compute m_ρ , so this equality determines the lattice spacing, a , in inverse MeV.

The key question is how big a lattice is needed, because this equates directly to the cost of the simulation. The box size must be big enough to contain all the hadrons without squashing them significantly. On the other hand, the lattice spacing must be small enough on the scale of the hadrons to resolve all the relevant physics. Thus, we require

$$\text{box size} \gg (\text{masses})^{-1} \gg \text{lattice spacing}. \quad (13)$$

So, if we were trying to simulate mesons built from the four lightest quark flavours, this becomes

$$L = Na \gg (M_\pi, \dots, M_{J/\psi})^{-1} \gg a \quad (14)$$

which would require $N \gg 20$. This type of argument pushes us towards large lattices, which are enormously costly.

Happily, there is an alternative. Since the quark masses may be varied freely in the simulations and we expect the physics to depend smoothly on them, we may simulate QCD using a compressed range of quark masses (and hence smaller lattices) and finally extrapolate the results to the light and/or heavy quark masses which correspond to the real world.

In order to understand how the computational cost varies with N , it is necessary to return to the functional integral in Eq (4), which, on the lattice, becomes a multiple integral over the fields at each site (and link). Computers are particularly inefficient at integrating the anticommuting Grassmann fields representing the quarks and antiquarks. Fortunately, this involves only Gaussian integrals and can be done analytically, with the following result:

$$\int dU d\bar{q} dq e^{-S_G[U] + \bar{q}(\mathcal{D} + m)q} = \int dU [\det(\mathcal{D} + m)]^{N_f} e^{-S_G[U]}. \quad (15)$$

This leaves only the integral over the gluon fields, represented by the link variables U in the above expression, and this can be done using Monte Carlo methods. The evaluation of the

determinant is the most costly part, so that, with today's best algorithms,

$$\text{number of arithmetic operations} \propto \begin{cases} N^{10} & N_f = 2, 4, \dots \\ N^6 & N_f = 0 \text{ (quenched)} \end{cases} . \quad (16)$$

The huge saving from setting $N_f = 0$, which, in Feynman-diagram language, removes internal quark loops, motivates the so-called quenched approximation. It is equivalent to assuming that the quarks and antiquarks which can be excited out of the vacuum by gluons, are very massive. Although there is no physical justification for this assumption, it has been widely used to date, just to enable simulations on large enough lattices to make contact with the real world. As we will see, the quenched approximation turns out to be surprisingly accurate for many quantities. So quenched QCD is rather a good model of strong interactions.

The significance of the computational complexity in Eq (16) is dramatically demonstrated when you consider a simple practical test to see if a particular simulation is close enough to the continuum limit. The obvious thing to do is to halve the lattice spacing and see if the results change significantly. For quenched QCD, the calculation on the lattice with twice the number of sites in each direction costs 64 times more computer time. For full QCD, the factor is over 1000. So, even if the original simulation has been done on a PC, to check it requires a supercomputer!

2.4 Computers

Since the early days, lattice QCD simulations have pushed against the limits of computer speed and, consequently, have played a part in the development of high-performance computing, particularly parallel computing.

Due to the translational invariance and locality of quantum field theories, their lattice versions are intrinsically parallel. Thus, the lattice may be divided up equally amongst an array of microprocessors, each responsible for the simulation on its part of the lattice and communicating with its neighbours to update fields on the boundary between them. The amount of computation which each microprocessor has, grows with the volume of its portion of the lattice, while the amount of communication it has to do, grows only with the surface area. Thus, the communications overhead is relatively small and may even be completely overlapped by simultaneous computation. The bottom line is that the simulation may be implemented so that the speed is directly proportional to the number of microprocessors.

This means that the major limit on QCD simulations is the size of the parallel computer used, which is dictated by the financial budget available. There is no hard technological barrier in view, given the relentless increase in microprocessor speed, which has been doubling every 19 months since the 1950's [10]. Currently, the most highly-parallel machine being used for QCD is the 12,000-processor QCDSF at the RIKEN-Brookhaven Center, which has a peak speed of 600 Gflops and is amongst the fastest computers in the world [11].

2.5 Improved lattice formulations

The N^{10} complexity for full QCD is strong motivation for doing better than just waiting for more powerful computers. Very considerable progress has been made in recent years in improving the lattice formulation to cancel systematically the leading discretisation errors in physical quantities [4]. The obvious consequence of this is that we should see scaling at larger lattice spacings, so that N can be kept small while maintaining a large enough box.

This standard numerical technique is called Symanzik improvement when applied to quantum field theories [12]. For QCD, the $O(a)$ corrections are cancelled by including one additional

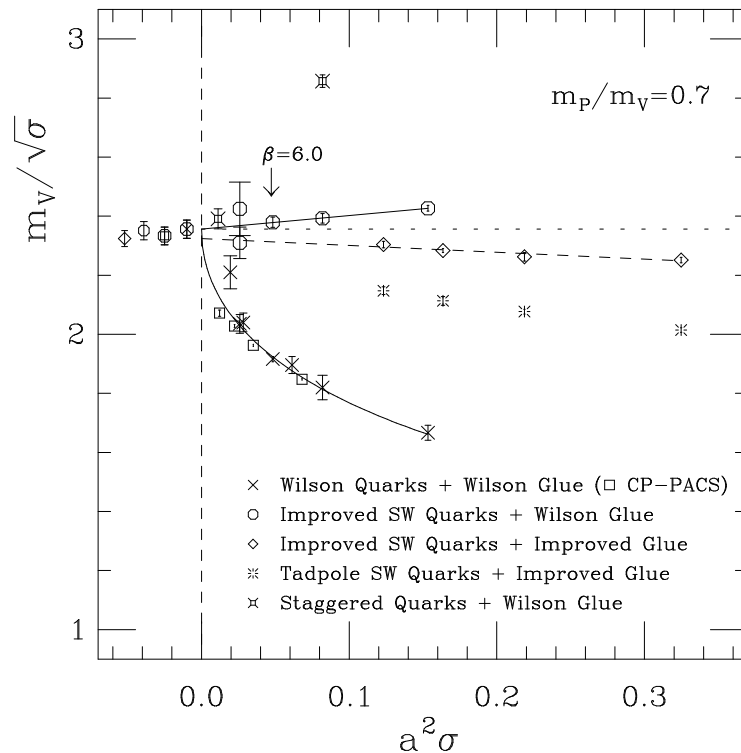


Fig. 2: A comparison of the approach to the continuum limit of the vector meson mass computed using different lattice formulations, including two $O(a)$ -improved actions (denoted by circles and diamonds), for which the leading discretisation errors are proportional to a^2 [13]. Points to the left of the dashed line are the continuum extrapolations of the corresponding data points.

(dimension 5) term in the action,

$$c_{\text{SW}} \frac{a^5}{4} \sum_x \bar{q}(x) i\sigma_{\mu\nu} F_{\mu\nu}(x) q(x), \quad (17)$$

together with explicit linear quark-mass dependence and mixing of the composite fields,

$$\mathcal{O}^{\text{R}} = Z_{\mathcal{O}}(1 + b_{\mathcal{O}} am_q)(\mathcal{O} + \sum_n c_n a\mathcal{O}_n), \quad (18)$$

used in correlation functions, such as Eq (4), (5) and (9). The improvement coefficients (c_{SW} , $b_{\mathcal{O}}$, c_n , etc) may be determined non-perturbatively, eg by imposing symmetries broken by the lattice.

An example of how effective improved actions can be is shown in Fig 2. This plots the vector meson mass, m_V , at fixed quark mass (defined by holding the ratio of the pseudoscalar and vector meson masses constant at 0.7), in units of the string tension, $\sqrt{\sigma}$, as a function of the square of the lattice spacing, in the same units, so that the data from improved simulations should fall on straight lines. The data labelled ‘Wilson Quarks’ is unimproved, showing curvature due to the $O(a)$ errors, and deviating significantly from the continuum limit, $a = 0$, as a increases. The two sets of improved data, labelled ‘Improved SW Quarks’, which employ different choices of lattice gluon action, are close to the continuum limit even at large lattice spacings. The small remaining dependence on a can be confidently extrapolated away. (The other data in this plot refer to different choices of lattice action with various degrees of improvement.) The extra computational cost of using an improved formulation is usually around 20%, so the efficiency gain is very significant.

2.6 The lattice theorist's toolkit

In summary, today's lattice theorist employs the following range of tools.

A supercomputer is needed to generate gauge field configurations and to calculate correlation functions, particularly if the objective is phenomenology, as this often requires large lattices to achieve adequate momentum resolution in the calculation of form factors.

Chiral perturbation theory is needed to guide the extrapolation of simulation results to light quark masses. The high cost of simulating light quarks directly seems to make this extrapolation inevitable for some years yet and it is a major source of uncertainty.

Keeping the lattice spacing as large as possible, means the momentum cutoff is low and some form of extrapolation to large quark masses is needed to study b physics. Heavy-quark effective theory can guide this extrapolation. Alternatively, the heavy quarks can be simulated directly using non-relativistic QCD. For more about this, see [2].

Although I have not discussed this in detail, quantities computed in lattice QCD often need to be translated into a more conventional perturbative regularisation scheme, in order for the results to be useful in phenomenology. This matching involves either a lattice perturbation theory calculation (if the momentum scale is high enough for perturbation theory to be valid), or some non-perturbative renormalisation.

Finally, the use of improved actions to reduce discretisation errors has become almost mandatory, and certainly so for simulations involving dynamical quarks.

3 THE DETERMINATION OF QUARK MASSES

The starting point for any simulation of QCD is to fix the bare quark masses which enter the Lagrangian. A further matching calculation enables these lattice quark masses to be related to quark masses defined in a perturbative regularisation scheme, so that they can be used in phenomenology.

3.1 Fixing the bare quark mass parameters

Conventionally, the bare quark mass enters the lattice QCD action via a so-called hopping parameter, κ , which is essentially its inverse. Simulations are performed for a range of hopping-parameter values in order to find the value, $\kappa = \kappa_{\text{crit}}$, which corresponds to zero quark mass (recall that chiral symmetry is broken explicitly by the lattice and so quark mass is additively renormalised). This is defined as the hopping parameter at which the pseudoscalar meson mass vanishes,

$$m_{\text{PS}}(\kappa = \kappa_{\text{crit}}) = 0, \quad (19)$$

corresponding to it being the Goldstone boson of spontaneously-broken chiral symmetry (although the way this is actually realised in the lattice theory is quite subtle). Chiral perturbation theory tells us how m_{PS} depends on quark mass, when this is small, and so can be used to extrapolate the data to κ_{crit} . Then the bare quark mass for flavour q is

$$m_q a = \frac{1}{2\kappa_q} - \frac{1}{2\kappa_{\text{crit}}}, \quad (20)$$

where κ_q is fixed by matching, for instance, the ratio of the pseudoscalar and vector meson masses to the experimental ratio of the masses of suitably flavoured mesons,

$$\frac{m_{\text{PS}}}{m_{\text{V}}} = \frac{M_{\pi}}{M_{\rho}}, \frac{M_K}{M_{\rho}}, \dots \quad (21)$$

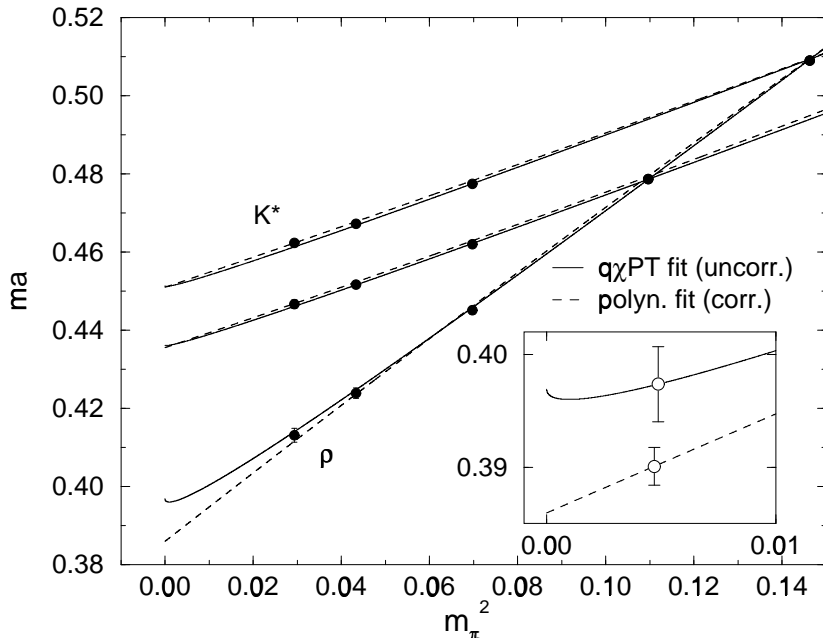


Fig. 3: Quenched QCD data for vector meson masses, obtained by the CP-PACS Collaboration [15], showing a comparison of polynomial and quenched chiral perturbation theory (qχPT) fits to the dependence on quark mass (equivalently the pseudoscalar meson mass squared, m_π^2).

3.2 Chiral behaviour of hadron masses

Simulations close to κ_{crit} are prohibitively costly. Instead, the simulations have to be performed at relatively large quark masses and, consequently, the chiral extrapolation becomes a significant source of error in the resulting estimates for the u and d quark masses, since $\kappa_{ud} \sim \kappa_{\text{crit}}$.

For full QCD, chiral perturbation theory predicts for the Goldstone pion,

$$m_{\text{PS}}^2 = Am_q + O(m_q^2) \quad (22)$$

and for all other hadrons,

$$m_{\text{hadron}} = B + Dm_q + O(m_q^2), \quad (23)$$

where A , B and D are constants.

However, the behaviour of hadron masses is different in quenched QCD. This is because the absence of quark loops eliminates the infinite series of loop diagrams which gives the η' meson its large mass. Instead, the η' meson remains light and, in fact, the singlet two-point function has a double pole, which introduces logarithmic terms in the quark mass [14], namely

$$m_{\text{PS}}^2 = Am_q \{1 - \delta \log m_q\} + O(m_q^2) \quad (24)$$

$$m_{\text{hadron}} = B + C\delta m_{\text{PS}} + Dm_{\text{PS}}^2 + O(m_{\text{PS}}^3), \quad (25)$$

where δ parametrises the strength of these quenched chiral logarithms. The simulation data from the CP-PACS Collaboration [15], which represents the state of the art for the quenched hadron spectrum, are consistent with the presence of quenched chiral logarithms, although, as can be seen in Fig 3, the differences from ordinary chiral perturbation theory only set in at very small quark masses.

3.3 Light quark masses

Since electromagnetic effects are not included in the QCD simulations, the up and down quarks are degenerate in mass. Recent quenched results for their mass, m_{ud} , and the mass of the strange

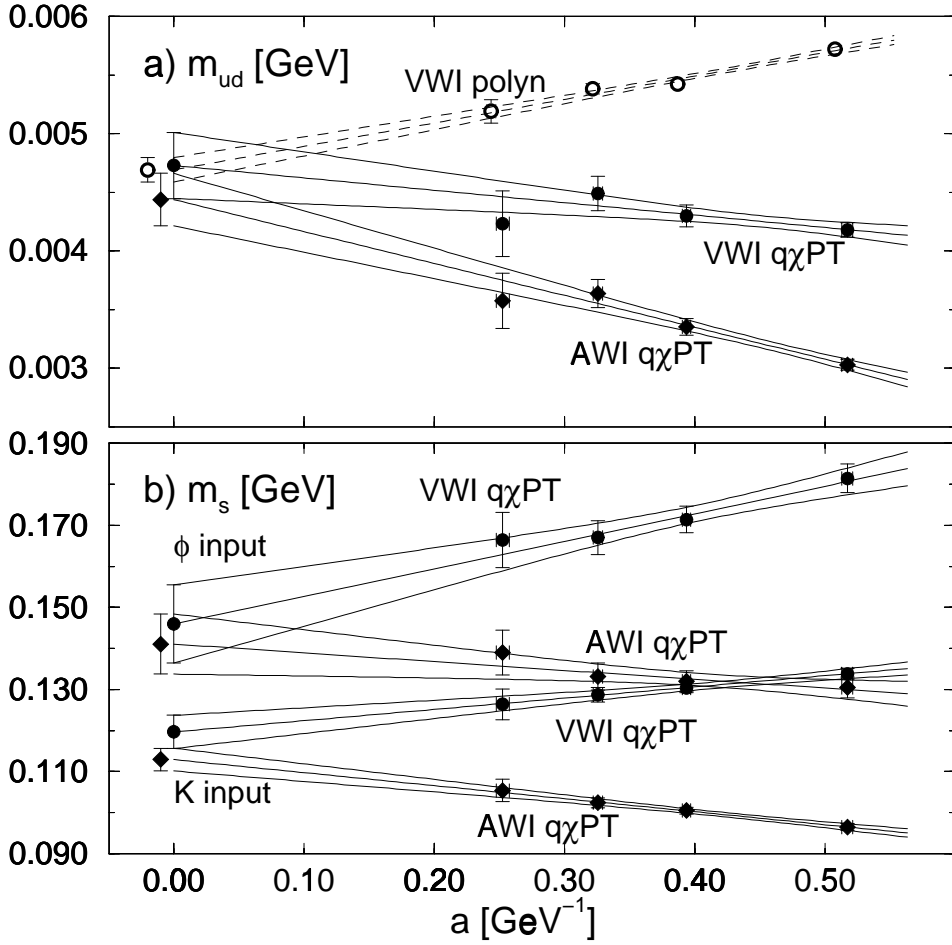


Fig. 4: Quenched QCD data for the light quark masses, obtained by the CP-PACS Collaboration [15] at a range of lattice spacings, and using two different definitions for the renormalised mass: the Vector Ward Identity (VWI) and the Axial Vector Ward Identity (AWI). In all but the top set of data, quenched chiral perturbation theory (qχPT) was used for the chiral extrapolation.

quark, m_s , in the $\overline{\text{MS}}$ scheme at a scale of 2 GeV, are shown as functions of the lattice spacing in Fig 4. There are two different definitions of the renormalised quark mass, depending on which of two Ward Identities is used. Although these give different results at non-zero lattice spacing, both definitions have consistent continuum limits. The result for the quenched QCD up and down quark mass is [15]

$$m_{ud} = 4.6(2) \text{ MeV}. \quad (26)$$

The results for the strange quark mass show the first evidence of a breakdown of the quenched approximation. The two sets of data in Fig 4 have been obtained by using the experimental K and ϕ meson masses to fix the bare strange quark mass, and it is clear that they do not give the same continuum limit. The reason is simply that the quenched strange meson spectrum does not agree with experiment for any choice of the strange quark mass. The continuum results are [15]

$$m_s = \begin{cases} 143(6) \text{ MeV} & (\phi \text{ input}) \\ 115(2) \text{ MeV} & (K \text{ input}) \end{cases}. \quad (27)$$

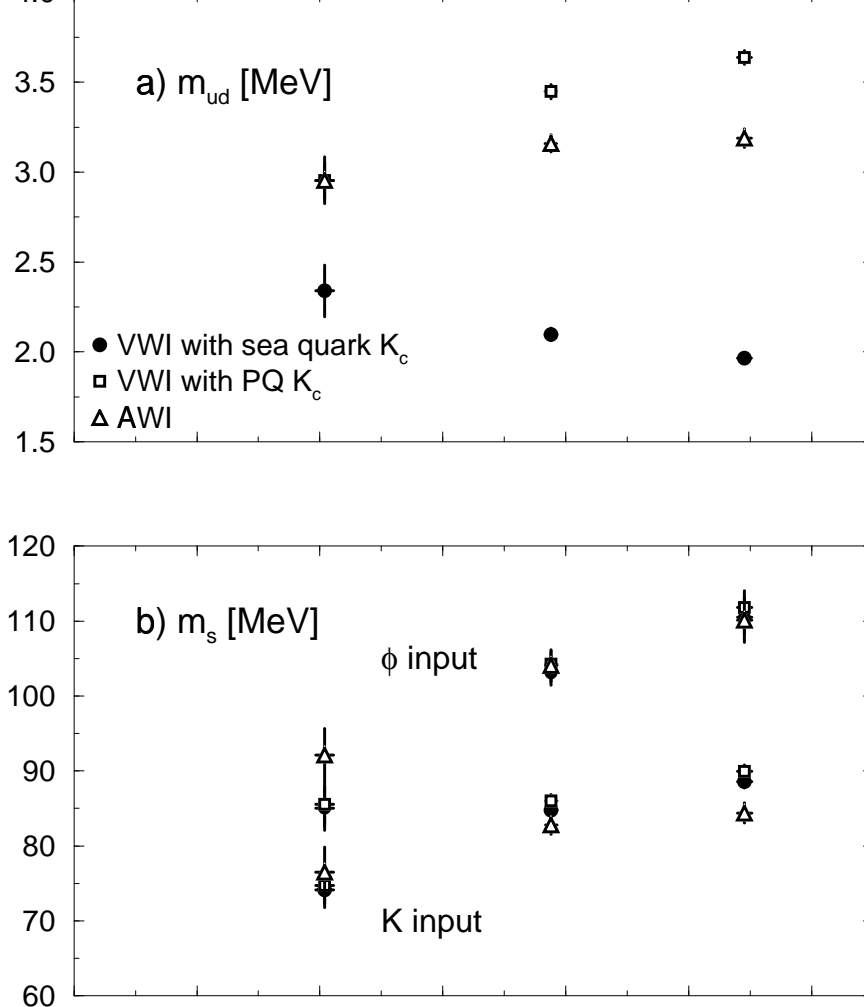


Fig. 5: Quark masses in QCD with two degenerate quark flavours (u and d), obtained by the CP-PACS Collaboration [15] at a range of lattice spacings, using various definitions for the renormalised quark mass. PQ refers to partial quenching, as described in the text, and $K_c \equiv \kappa_{\text{crit}}$.

The next step is to simulate QCD with two degenerate flavours of quark. Since the gluon configurations ‘feel’ the quarks through the determinant term in Eq (15), separate Monte Carlo runs have to be performed for each value of the quark mass used as input. I’ll call this the sea-quark mass. On each set of configurations, we can compute hadron correlation functions built from valence quark fields with a different mass from the sea-quark mass, if we so choose. The two degenerate sea-quark flavours have to be identified with the up and down quarks, at least in the limit of small mass, but the valence quarks can be used to approximate other quenched flavours, such as strange or charm, propagating in gluon fields which are only sensitive to the up and down quarks. The chiral limit may be approached in two ways: by maintaining $\kappa_{\text{valence}} = \kappa_{\text{sea}}$, or by fixing $\kappa_{\text{sea}} = \kappa_{ud}$ and varying κ_{valence} , called ‘partial quenching’. Different values of κ_{crit} result from these two approaches, at non-zero lattice spacing, and this mostly affects the estimate of the up and down quark mass, but the differences appear to go away in the continuum limit.

Some of the best data obtained so far is shown in Fig 5. Even so, the data are not of sufficient quality to justify a continuum extrapolation. The trends are, however, encouraging. Although significant differences between the various definitions are found at non-zero lattice spacing, all these differences decrease as the continuum is approached. Secondly, the strange quark mass estimates obtained from the K and ϕ mesons appear to be converging in the con-

tinuum limit, unlike in quenched QCD. Finally, and most interesting of all, the quark mass estimates are significantly lower than those obtained in the quenched approximation. At a time when it is proving difficult to establish clear effects from including dynamical quarks in our simulations, this is a welcome result!

While the strange quark mass estimates have dropped from around 130 MeV to around 80 MeV, by including dynamical up and down quarks, it should be noted that the latter result still corresponds to a quenched strange quark and so may be an upper bound on the real value.

4 THE QCD SPECTRUM

Attempts to compute the spectrum of QCD began with Wilson's invention of lattice QCD, and represent an essential step in validating QCD. The whole exercise has proved to be much more difficult than originally hoped. This is because it requires demonstrating complete control of all aspects of the simulation. In fact, the quenched QCD light hadron spectrum, which was tackled first for the reasons given above, has turned out to be so close to the real world that, just four years ago, the GF11 Collaboration claimed it agreed with experiment to within 6% [16]. But, we know quenched QCD must be wrong at some level!

4.1 The quenched spectrum and decay constants

It has taken the last four years, and an enormous computational effort, to show that the quenched QCD spectrum does deviate significantly from experiment. This was achieved by the CP-PACS Collaboration [15] and I have already described the inconsistencies they found in the quenched strange meson spectrum. Although it awaits independent confirmation, the CP-PACS tour de force has effectively concluded a 20-year effort to calculate the quenched QCD spectrum.

CP-PACS have obtained high precision results with control of all systematic errors (except quenching itself) on lattices up to $64^3 \times 112$. They perform a continuum extrapolation from data at four lattice spacings, spanning the range $a^{-1} \approx 2 - 4$ GeV, and in a fixed volume of linear size $L \approx 3$ fm. Although not proving the existence of quenched chiral logarithms, their data are consistent with this pathology of quenched QCD and their chiral extrapolations use the quark-mass behaviour predicted by quenched chiral perturbation theory. They are able to conclude that the spectrum is inconsistent with experiment.

The spectrum obtained by CP-PACS is shown in Fig 6, along with the GF11 results for comparison. It is evident that the major achievement has been to reduce the errors to the point where disagreement with experiment, within the uncertainty in the GF11 results, could be exposed. As I have already mentioned, the strange quark mass cannot be fixed unambiguously and, hence, CP-PACS present two spectra in Fig 6, corresponding to two possible choices. Generally, the meson hyperfine splitting is too small in the quenched approximation and it gets worse for charmonium. The octet baryon masses and the decuplet baryon mass splittings are also too small, with M_K as input, although somewhat better with M_ϕ as input.

As we have seen, decay constants are a by-product of spectrum calculations. They are also sensitive to quenched chiral logarithms and CP-PACS's decay-constant data also support the presence of chiral logarithms, with a similar strength, δ in Eq (24), to the spectrum data. The continuum results for f_π and f_K in quenched QCD are significantly smaller than experiment, as shown in Fig 7.

4.2 The QCD spectrum with two degenerate flavours

Present-day simulations with dynamical up and down quarks, and all other flavours quenched, have not reached anything like the same degree of control as has been achieved for fully quenched simulations. The lattices are typically smaller, with $L \sim 2$ fm, although this is not a serious

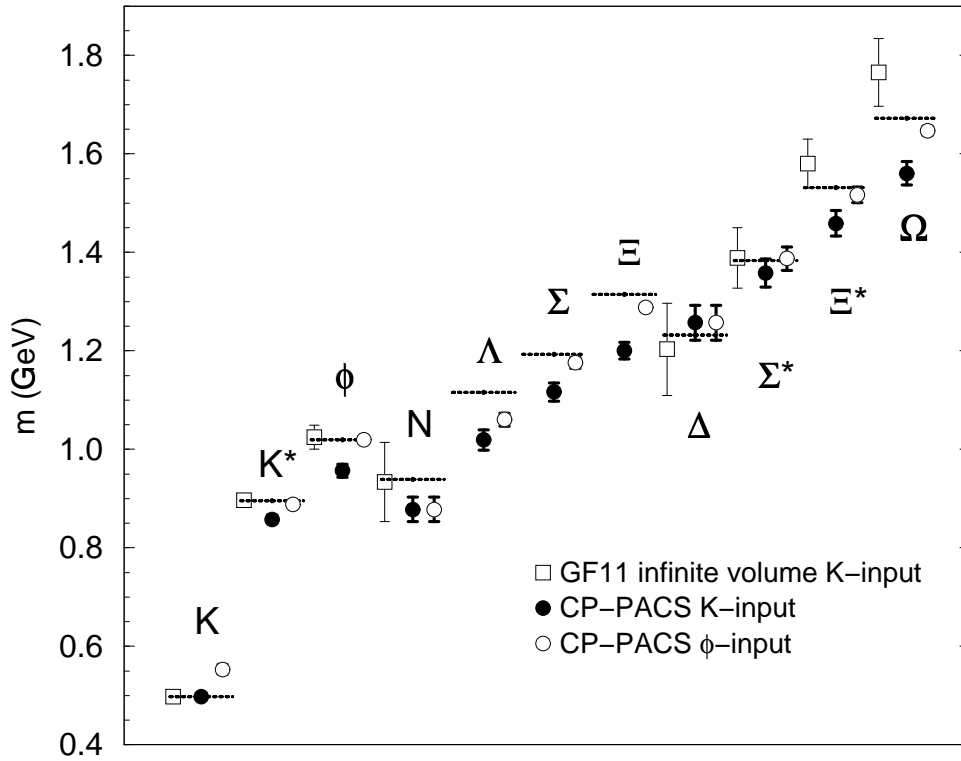


Fig. 6: CP-PACS Collaboration's results for the quenched light hadron spectrum in the continuum limit [15]. Experimental values are denoted by horizontal lines and the results from GF11 [16] are shown for comparison.

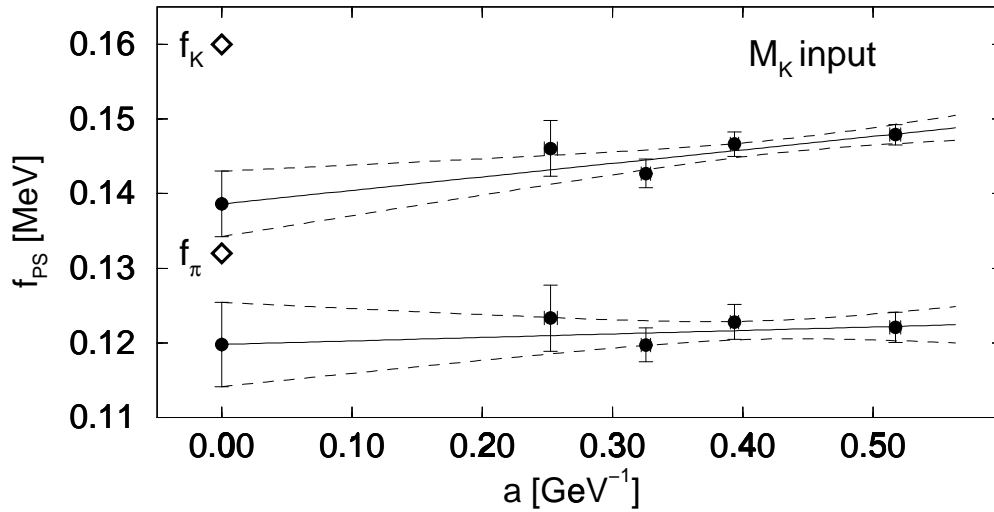


Fig. 7: CP-PACS Collaboration's results for the quenched K meson and pion decay constants in the continuum limit [15]. Experimental values are denoted by diamonds.

problem, because the up and down quark masses actually used are large (so that $m_{PS}/m_V \geq 0.6$) and the hadrons are small. Consequently, a big chiral extrapolation is necessary, from data of relatively poor statistical quality. So it is not particularly surprising that the spectrum obtained agrees with experiment within large errors [15]. It will be a considerable challenge to expose any effects from quenching the heavy quarks, given the effort needed to disprove quenched QCD. The most dramatic result from dynamical-quark simulations so far has been the drop in the quark masses.

4.3 Hybrid mesons

From the point of view of phenomenology, a much more interesting aspect of the hadron spectrum calculations is the ability to predict states which are not found in the quark model and which have yet to be identified experimentally.

Hybrid mesons have gluonic excitations which permit non-quark-model quantum numbers. A while ago, the UKQCD Collaboration showed that, in quenched QCD, the lightest $s\bar{s}$ exotic meson has $J^{PC} = 1^{-+}$ and a mass of 2.0(2) GeV [17].

This year, the calculation has been repeated for $N_f = 2$ sea quarks. Again, the 1^{-+} meson is the lightest exotic and its mass is found to be 1.9(2) GeV in the chiral limit [18], consistent with the UKQCD result. Of course, hybrid mesons can mix with four-quark states, and the effect of this must be understood before the result can be considered relevant to phenomenology. Mixing in the lattice simulation is currently under investigation.

4.4 Glueballs

Glueballs are an important prediction of QCD and the low-lying quenched spectrum has been known for some time [19]. This is already a useful guide for experimental searches [20].

Although there are no results from dynamical-quark simulations yet, the mixing of the lightest glueball with scalar quarkonium states has been studied recently in quenched QCD [21]. The unmixed results are shown in Fig 8. The continuum limit of the $\bar{s}s$ state has a mass well below the lightest glueball, suggesting naively that, of the two experimental candidates, the $f_0(1710)$, rather than the $f_0(1500)$, is predominantly glue.

A crude mixing calculation gives

$$|f_0(1710)\rangle = 0.86(5)|g\rangle + 0.30(5)|s\bar{s}\rangle + 0.41(9)|n\bar{n}\rangle \quad (28)$$

$$|f_0(1500)\rangle = -0.13(5)|g\rangle + 0.91(4)|s\bar{s}\rangle - 0.40(11)|n\bar{n}\rangle \quad (29)$$

$$|f_0(1390)\rangle = -0.50(12)|g\rangle + 0.29(9)|s\bar{s}\rangle + 0.82(9)|n\bar{n}\rangle \quad (30)$$

where $|n\bar{n}\rangle$ is the scalar state built from up and down quarks. The opposite sign of $|s\bar{s}\rangle$ and $|n\bar{n}\rangle$ in $|f_0(1500)\rangle$ is a possible explanation of why its decay to $K\bar{K}$ is suppressed.

5 CONCLUSIONS

I hope I have given you an impression of the challenges and promises of lattice QCD. Ultimately, lattice QCD offers the prospect of a precise, model-independent tool for dealing with the hadronic uncertainties which currently hinder the search for physics beyond the Standard Model at B Factories. To date, quenched QCD has proven to be a suprisingly accurate effective theory of the strong force, but, as a result of arduous and painstaking numerical work, its failure even at the few percent level is now established. The quenched light hadron spectrum has been shown to disagree with experiment and, in particular, the strange quark mass cannot be chosen unambiguously. Simulations with dynamical quarks are underway and sea-quark effects are beginning to show up. Notably, the up, down and strange quark masses are around 40% lower

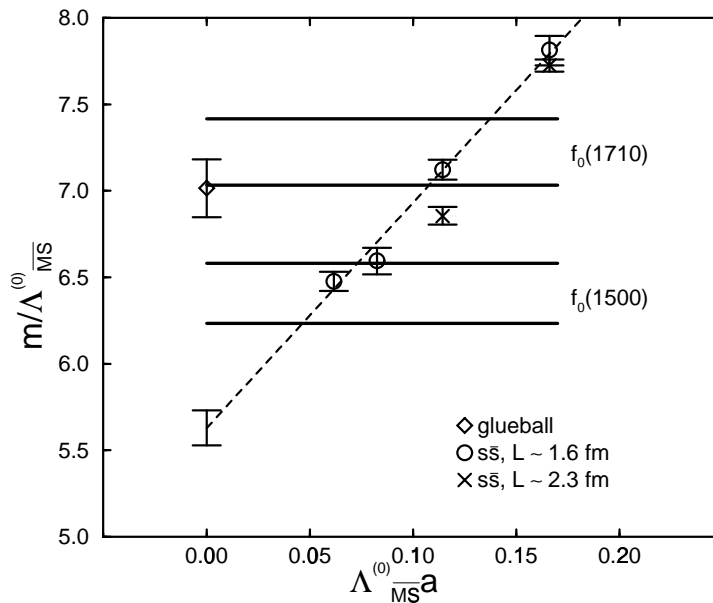


Fig. 8: The continuum quenched QCD glueball mass, and the scalar quarkonium mass as a function of lattice spacing, compared with the experimental values for the masses of the $f_0(1710)$ and $f_0(1500)$ states (the uncertainty in these, represented by the horizontal lines, is largely due to the uncertainty in $\Lambda_{\overline{\text{MS}}}^{(0)}$ used to set the scale) [21].

than in quenched QCD. Continued progress is guaranteed by the relentless advance of computer technology. However, a speedier approach to fully realistic QCD simulations can be anticipated due to further developments in the theoretical formulation, and the range of non-perturbative physics which can be addressed will grow accordingly.

References

- [1] R. D. Kenway, Lattice '98 Proceedings, Nucl. Phys. B (Proc. Suppl.) (1999) in press (hep-lat/9810054).
- [2] C. T. H. Davies, these proceedings.
- [3] R. Gupta, in "Probing the Standard Model of Particle Interactions", LXVIII Les Houches Summer School 1997 (hep-lat/9807028).
- [4] M. Lüscher, in "Probing the Standard Model of Particle Interactions", LXVII Les Houches Summer School 1997 (hep-lat/9802029).
- [5] S. R. Sharpe, ICHEP 98 Proceedings (hep-lat/9811006).
- [6] Lattice '97 Proceedings, eds C. T. H. Davies et al., Nucl. Phys. B (Proc. Suppl.) 63 (1998).
- [7] M. Talevi, Lattice '98 Proceedings, Nucl. Phys. B (Proc. Suppl.) (1999) in press (hep-lat/9809182).
- [8] O. Philipsen and H. Wittig (hep-lat/9807020).
- [9] K. G. Wilson, Phys. Rev. D10 (1974) 2445.

- [10] A. E. Brenner, *Physics Today* 49 (1996) 24.
- [11] D. Chen et al., *Nucl. Phys. B (Proc. Suppl.)* 63A-C (1998) 997.
- [12] K. Symanzik, *Lecture Notes in Physics*, Vol 153 (Springer, New York, 1982); *Nucl. Phys. B* 226 (1983) 187 and 205.
- [13] R.G. Edwards, U. M. Heller and T. R. Klassen, unpublished.
- [14] A. Morel, *J. Physique* 48 (1987) 111; S. R. Sharpe, *Phys. Rev. D* 41 (1990) 3233, *D* 46 (1992) 3146; C. W. Bernard and M. F. L. Golterman, *ibid* *D* 46 (1992) 853.
- [15] R. Burkhalter, *Lattice '98 Proceedings*, *Nucl. Phys. B (Proc. Suppl.)* (1999) in press (hep-lat/9810043); S. Aoki et al., *Lattice '98 Proceedings*, *Nucl. Phys. B (Proc. Suppl.)* (1999) in press (hep-lat/9809120 and hep-lat/9809146).
- [16] F. Butler et al., *Phys. Rev. Lett.* 70 (1993) 2849; *Nucl. Phys. B* 430 (1994) 179.
- [17] P. Lacey et al., *Phys. Lett. B* 401 (1997) 308 (hep-lat/9611011).
- [18] P. Lacey and K. Schilling (hep-lat/9809022).
- [19] M. Teper, in “Confinement, Duality, and Nonperturbative Aspects of QCD”, ed. P. van Baal (Plenum, New York, 1998) 43 (hep-lat/9711011).
- [20] F. E. Close, *Nucl. Phys. B (Proc. Suppl.)* 63A-C (1998) 28 (hep-lat/9802026).
- [21] W. Lee and D. Weingarten (hep-lat/9805029).

HEAVY QUARK PHYSICS FROM LATTICE QCD

C. T. H. Davies

Department of Physics and Astronomy, The University of Glasgow, Glasgow, G12 8QQ,
Scotland

Abstract

I describe lattice QCD results for the spectrum and phenomenology of particles containing bottom and charm quarks. Measurement of the weak decay of these particles, along with theoretical input, will allow the determination of the CKM matrix, key to our understanding of CP violation. To this end there is a big investment in the new experimental B factory programmes and a huge amount of activity among lattice QCD theorists. I review current results and discuss the techniques being used.

1 INTRODUCTION

Bottom and charm quarks are distinguished from up, down and strange by their much heavier mass, larger than a typical QCD scale, Λ_{QCD} , of order a few hundred MeV. They are consequently known generically as heavy quarks, and their bound states as heavy hadrons.

The fact that $m_Q \gg \Lambda_{QCD}$ leads to a simplified dynamics for these quarks which we can take advantage of theoretically. They are, however, still light enough to have a rich spectrum of bound states, unlike the top quark. The top quark is in fact too heavy to be a heavy quark in the sense that we shall use here!

As Richard Kenway described in his lecture [1], the spectrum of hadrons is important as a test of QCD in the non-perturbative regime, and lattice QCD provides an *ab initio* method of calculating the spectrum. Of course the spectrum of heavy hadrons is as much a consequence of QCD as that of light hadrons and the calculation of it on the lattice can be somewhat simpler, as I shall describe. There are two sorts of hadrons containing heavy quarks; those which contain as valence quarks only heavy quarks, and those which contain a mixture of heavy and light quarks. For mesons, the first type are known as ‘heavy-heavy’ mesons and the second as ‘heavy-light’. It is hadrons of the second type, and particularly B mesons, which are of most interest currently and they are the ones that we will concentrate on.

The weak decays of heavy hadrons will be critical in determining the elements of the Cabibbo-Kobayashi-Maskawa matrix and understanding CP violation (see, for example, [2]). Theoretical work on these decays will be necessary alongside the experimental programme and a lot of these theoretical calculations must be done in lattice QCD. Let us take the weak decay of Figure 1 as an example. It is very simple at heart; one quark flavour emits a W particle and changes into another quark flavour; the analog of radioactive β decay of the neutron. It takes place in an hadronic environment, however, because of confinement and thus QCD effects are very important. We must calculate the matrix element for the weak quark current between the hadronic states in lattice QCD. Since this current picks up an element of the CKM matrix, $V_{QQ'}$, from the weak Lagrangian, the experimental result should be $V_{QQ'}$ times the theoretical number, allowing a determination of $V_{QQ'}$.

2 HEAVY QUARKS ON THE LATTICE

At first sight the very nature of heavy quarks seems to make it impossible to include them in lattice QCD.

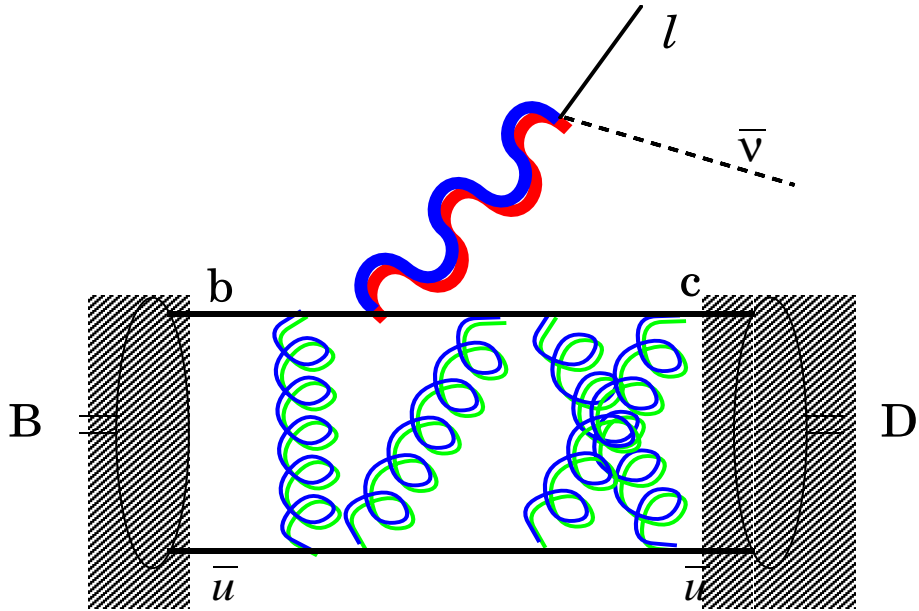


Fig. 1: Semi-leptonic decay of a B meson into a D meson.

An essential feature of lattice QCD calculations is a space-time lattice, introduced to make the calculations tractable. The separation between space-time points is denoted a [1]. In doing the calculations we imagine taking a to zero, but in practice we cannot do this. Instead we must demonstrate that physical results are independent of the unphysical parameter a . This can and will be true if the distance scales important to the calculation, such as $1/\Lambda_{QCD}$, are very much larger than a , so that the discrete nature of our space-time is not important. This condition is $\Lambda_{QCD}a \ll 1$.

For heavy quarks there is a scale associated with their mass and it is not true that this scale in lattice units is very much less than 1, at least for the values of a that we can currently compute with. Typical values of a are 0.1 fm, equivalent to 0.5 GeV^{-1} so that $m_b a \approx 2.5$ and $m_c a \approx 0.75$. What this means (see Figure 2) is that the Compton wavelength of the heavy quark is very much smaller than the lattice spacing and ‘falls through’ the lattice.

All is not lost, however, since the scale associated with the heavy quark mass is actually not an important scale for the physics of heavy quark bound states. This can be seen from the right hand side of Figure 2 - a heavy-light meson on the lattice consists of a heavy quark surrounded by a cloud of light anti-quark (technically known as ‘brown muck’). The size scale of this cloud is set by typical momenta inside the bound state. As we shall see, these typical momenta are much smaller than m_Q , and so the condition $p_Q a \ll 1$ can be satisfied. The physics of the bound states is then well simulated on current lattices.

Inside heavy-light bound states it is clear that any kinetic energy will be associated with the light quark and therefore will be of order Λ_{QCD} . The momentum of the light quark will be of the same order and this will be balanced by the momentum of the heavy quark. Thus $p_Q \approx \Lambda_{QCD}$ and the heavy quark will be non-relativistic. For a b quark inside a B meson we would expect $v/c \approx 0.1$ and for a c quark inside a D meson $v/c \approx 0.3$. The non-relativistic nature of the bound states is also evident if we compare radial and orbital splittings within the heavy-light meson hierarchy to their absolute masses [3]. These splittings (e.g. $B' - B, B^{**} - B$) are around a few hundred MeV (i.e. $\mathcal{O}(\Lambda_{QCD})$) and represent excitations of the light degrees of freedom and therefore typical momenta and energies associated with them. The splittings are

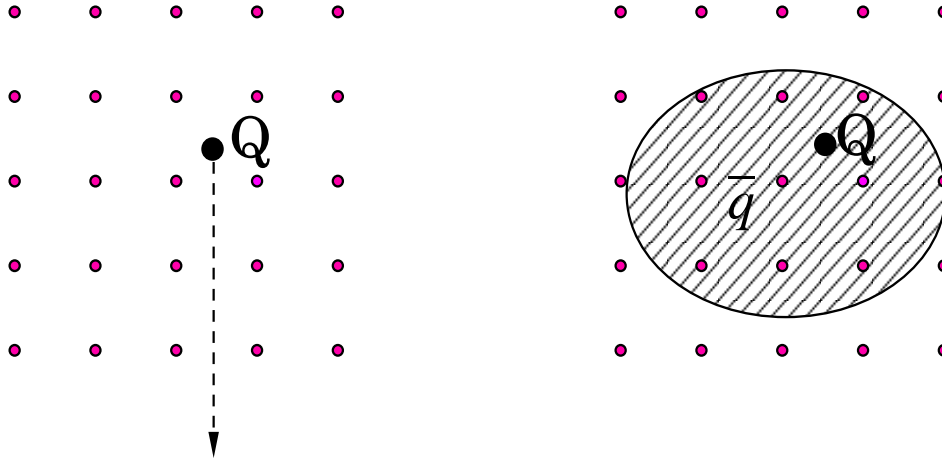


Fig. 2: Heavy quarks and heavy-light mesons on the lattice.

much smaller than the masses of the heavy-light bound states themselves, reinforcing the idea that the heavy quark in these hadrons is non-relativistic. The major contribution to the mass of the hadrons is then just the mass of the heavy quark which simply shifts the overall mass up or down.

Similar arguments hold for heavy-heavy bound states made of both heavy quark and anti-quark. We have two well-documented systems in this category, bottomonium and charmonium, and the mixed bound states of bottom and charm are now being uncovered, beginning with the B_c [4]. If we look again at radial and orbital splittings inside bottomonium and charmonium (e.g. $\Upsilon' - \Upsilon, \chi_b - \Upsilon$), they are of order a few hundred MeV and much smaller than the masses of the mesons themselves. These splittings represent typical kinetic energies of the heavy quarks inside the bound states, and we see again that the heavy quarks are non-relativistic. For b quarks inside an Υ we have $v_Q^2/c^2 \approx 0.1$ and for c quarks inside a ψ we have $v_Q^2/c^2 \approx 0.3$.

It is worth pointing out that the physical picture of heavy-light and heavy-heavy bound states is somewhat different. The heavy-light meson is like the ‘hydrogen atom’ of QCD and the heavy-heavy meson, like positronium. The heavy quarks are non-relativistic in both systems but for different reasons. In heavyonium both heavy quarks are non-relativistic because of the nature of the gluon exchange between them. We cannot neglect the kinetic energy of either quark in calculating the spectrum. For heavy-light mesons the light quark is relativistic and carries all the kinetic energy. We can think of the heavy quark as a static source of colour, in a first approximation for the spectrum. Here we will only discuss results for the spectrum of heavy-light mesons; heavy-heavy meson results are reviewed in, for example, [5].

We can take advantage of the non-relativistic nature of heavy quark bound states in order to perform accurate lattice calculations. The Lagrangian for QCD splits into three (coupled) pieces; the gluonic Lagrangian, the light quark Lagrangian and the heavy quark Lagrangian. The gluonic Lagrangian is taken as the lattice discretisation of $F_{\mu\nu}^2$ and the light quark Lagrangian is based on a discretisation of the Dirac equation [1]. For the heavy quark Lagrangian, however, we should take a non-relativistic form appropriate to the physics, because, as described above, we know that any problems associated with the fact that $m_Q a > 1$ are irrelevant.

There are several ways to write down an effective non-relativistic Lagrangian and they can look quite different from each other even when they have the same philosophy. I shall describe only the simplest Lagrangian here, since we can then see easily what each term is doing. L_Q has separate terms for the heavy quark and anti-quark fields which are now decoupled two-

component spinors. The first few terms for the heavy quark Lagrangian are :

$$L_Q = \psi^\dagger \left(D_t - \frac{\vec{D}^2}{2m_Q} - \frac{\vec{\sigma} \cdot \vec{B}}{2m_Q} \dots \right) \psi. \quad (1)$$

Notice that the mass term has simply been dropped as an overall shift of energy. The first term is just a static colour source moving in the time direction. This is the leading term for heavy-light systems, and gives behaviour independent of heavy quark mass (i.e. flavour) and spin. This is the Heavy Quark Symmetry limit. The second term is the heavy quark kinetic energy and it allows the heavy quark to move in spatial directions on the lattice. For heavy-light systems this term is suppressed by an inverse power of m_Q with respect to the first term. For heavyonium systems, power counting works in a different way and this term is as important as the first one. The third term is the first to refer to the heavy quark spin and it allows a coupling between the heavy quark spin and the chromomagnetic field of its environment. Again, for heavy-light systems, this term is a ‘ $1/m_Q$ ’ term.

3 THE HEAVY-LIGHT SPECTRUM

From this Lagrangian we can deduce a number of features that we would expect in the heavy-light spectrum. The first is that splittings between states which are radial or orbital excitations should be roughly independent of heavy quark mass since they should be dominated by the first term in L_Q . We expect the splittings to behave as constant $\pm \mathcal{O}(\Lambda_{QCD})/m_Q$. There are a number of such splittings that we can study from the Particle Data tables [3] and they are plotted in Figure 3. For the x -axis I use a quantity made of experimentally measurable hadron masses which is related to $1/m_Q$. This is the inverse of the spin-average of the vector and pseudoscalar meson masses, $4/(3m_V + m_{PS})$. We see very flat behaviour as a function of this mass, all the way down to the strange quark mass, although it would be hard to justify this quark as a heavy one! Several of the splittings in Figure 3 are between baryonic states with one heavy quark. These are the so-called heavy-light-light baryons and the dynamics of these bound states is very similar to that of the heavy-light mesons since again there are light degrees of freedom. The two light quarks can take up different internal spin configurations, with $s_{light} = 0$ making the Λ_Q and $s_{light} = 1$ making the Σ_Q and Σ_Q^* . The splitting between the spin-average of the Σ_Q states and the Λ_Q should again be m_Q -independent since it is a spin-excitation of the light degrees of freedom only. This is also true for the splitting between the Λ_Q baryon and the S-wave mesons. In fact these two splittings are incredibly flat as a function of $1/m_Q$, with little sign even of the Λ_{QCD}/m_Q term (see Fig. 3).

Another feature of the spectrum is that splittings between states which differ in their heavy quark spin configuration should be proportional to $1/m_Q$ since they appear for the first term as a result of the third term in L_Q . Again, this is borne out in the Particle Data tables for both mesonic splittings (such as that between the vector and pseudoscalar S-wave states: for the b quark, B^* and B) and baryonic splittings (such as that between the Σ_Q^* and the Σ_Q). These are plotted in Figure 4.

The behaviour of the spectrum can be argued as above by inspection of the Lagrangian and the effect of terms beyond the Heavy Quark Symmetry limit. The absolute size of the splittings, as predicted by QCD, does require a lattice calculation, however, and several of these have been done in the past few years.

Two approaches can be taken to the heavy-light spectrum. One is to use Non-relativistic QCD, which is the lattice discretisation of L_Q described above. There are some additional technicalities to make this rigorously equivalent to QCD for non-relativistic heavy quarks and these are described, for example, in [5]. This is the simplest and probably best approach for the b quark. An alternative is to take a relativistic Lagrangian but simply interpret the results

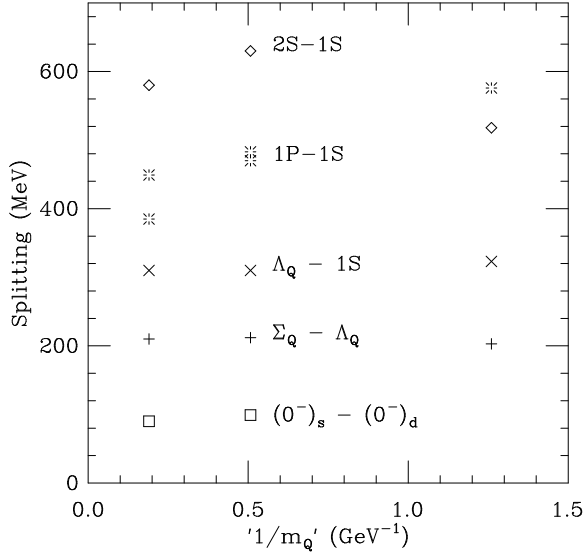


Fig. 3: Experimental results for splittings in the heavy-light spectrum which are independent of the heavy quark spin configuration. These splittings should be independent of the heavy quark mass at leading order in an expansion in $1/m_Q$. Results for three quarks, Q , are given: from left to right, b , c and s . For ‘ $1/m_Q$ ’ on the x -axis I use in fact the inverse of the spin-averaged S-wave meson mass i.e. $4/(3m_V + m_{PS})$. A key to the splittings used is given on the Figure. The states described are spin-averaged where this is necessary and the data exist, e.g. by Σ , I mean the spin-average of Σ and Σ^* .

(such as the energy-momentum relation) in a non-relativistic way [6]. This approach does have the advantage that it matches on to the $m_Q a \ll 1$ region. This may be important for c quarks since it may be possible to reach this region, either by having a fine enough lattice spacing or by compensating for its coarseness by improving the discretisation used [1, 7]. The literature has results from both methods and they agree reasonably well where they overlap.

Figure 5 shows, as an example, lattice results for the B spectrum obtained using NRQCD in the quenched approximation and a comparison with experimental numbers where they exist [8]. Figure 6 shows results for the heavy-light-light baryons. The methodology of the calculations has been described earlier [1]. In brief, we calculate the correlation function of the heavy-light meson between two points separated in time by T lattice spacings. This correlation function is then fit as a function of T to multiple exponentials to extract the smallest exponent. This is given by T times the energy, E , of the lowest energy state that could be created from the vacuum by the operator used in the correlation function. E is then the mass of the ground state meson and its radial excitations can be obtained from the higher exponents. E is in units of the lattice spacing and must be converted to GeV by a determination of a .

To set the scale for the spectrum (i.e. the lattice spacing), the ρ mass was used in Figure 5. The b quark mass was fixed by setting the b mass to its experimental value. Agreement with experiment in Figure 5 is good, except for the (hyperfine) spin splitting between B and B^* (or B_s and B_s^*). This is probably because the calculation was done in the quenched approximation [1]. In this approximation quark/anti-quark pairs popping in and out of the vacuum are ignored and this means that their screening effects on the strong coupling constant are absent. This in turn means that the quenched theory does not ‘run’ properly between different momentum scales. Thus if we fix the lattice spacing from the ρ we expect to get the hyperfine splitting wrong because we believe that it is sensitive to rather higher momenta than those important to

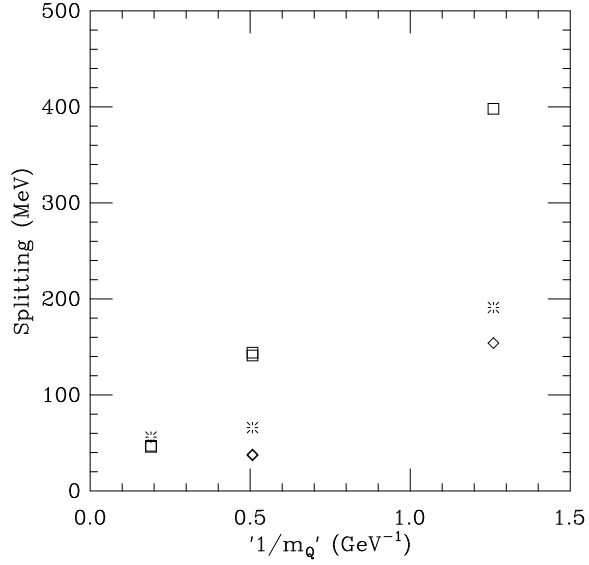


Fig. 4: Experimental results for splittings in the heavy-light spectrum which require a spin-flip of the heavy quark. These splittings should behave as $1/m_Q$ at leading order in an expansion in $1/m_Q$. The splitting in MeV is plotted against the inverse of the spin-averaged S-wave meson mass i.e. $4/(3m_V + m_{PS})$. Squares give the hyperfine splitting between the vector and pseudoscalar S-wave mesons e.g. $B^* - B$ for the b . Bursts give the splitting between the two $s_{light} = 1$ heavy-light-light baryons, e.g. $\Sigma_b^* - \Sigma_b$ for the b . (Note that the experimental value for the b case is preliminary and looks rather strange on this plot in comparison to the other results.) Diamonds give results for spin splittings for P-wave mesons between the spin 2 meson and the heavier of the spin 1 mesons.

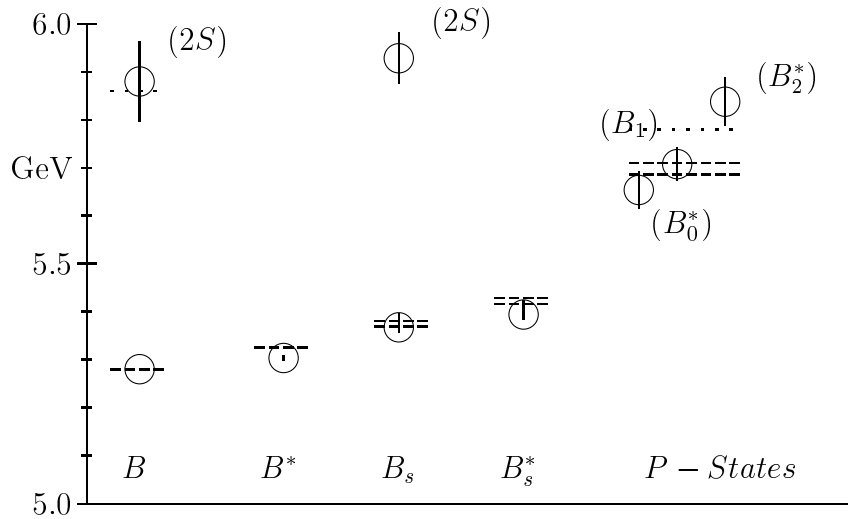


Fig. 5: The B meson spectrum from lattice NRQCD [8]. Errors include statistical errors and systematic errors from fixing the strange quark mass. Dashed lines indicate experimental error bounds from the Particle Data Tables [3]; dotted lines, preliminary experimental results from DELPHI [9, 10].

the ρ . Errors have been seen for other quantities also, most notably when comparing splittings in the Υ spectrum to the ρ mass [5].

The spectrum of D mesons has been obtained using the relativistic Lagrangian with non-relativistic interpretation described above and is given in [11]. The comparison with experiment shows similar features. Results for charm baryons are given in [12]. One advantage of lattice calculations over the real world is that we can also perform calculations for quark masses that don't exist in Nature. This is useful when studying the heavy quark mass dependence, and interpreting it in terms of the Lagrangian, L_Q . See, for example, [8].

Lattice calculations enable us to determine quark masses since the bare mass parameters of the Lagrangian are fixed in a fully non-perturbative way by demanding that a particular hadron mass is correct, for a given value of the lattice spacing. The lattice bare quark mass needs to be converted to an appropriate continuum mass to be useful and this requires a perturbative matching calculation between the lattice and continuum renormalisation schemes. Figure 7 shows results for the \overline{MS} b mass using a lattice bare b mass obtained by fixing the Υ mass to be correct. The running \overline{MS} mass is given at a scale equal to itself by :

$$m_{\overline{MS}}(m_{\overline{MS}}) = Z m_{lattice,bare} \quad (2)$$

with $Z = 1 + \mathcal{O}(\alpha_s)$. Figure 7 gives $m_{\overline{MS}}^b(m_{\overline{MS}}^b) = 4.3(2)$ GeV. Results using the B meson mass are less accurate and not shown. A similar calculation has been done for the charm quark mass using the relativistic Lagrangian approach to the charmonium spectrum [13].

4 DECAY CONSTANTS

B and D mesons can undergo a weak decay to purely leptonic products via the annihilation of their constituent quarks into a W particle. This is illustrated in Figure 8. The rate for the decay depends on the CKM matrix element which connects the constituent quark and anti-quark and a number called the decay constant, f_B or f_D . The decay constant gives, loosely speaking, the probability with which the quark and anti-quark are found inside the meson at the same point and can annihilate. (In potential model language, $f_{meson}^2 M_{meson} = |\psi(0)|^2$).

$$\Gamma(Q\bar{q} \rightarrow l\bar{\nu}) \propto |V_{Qq}|^2 f_{Q\bar{q}}. \quad (3)$$

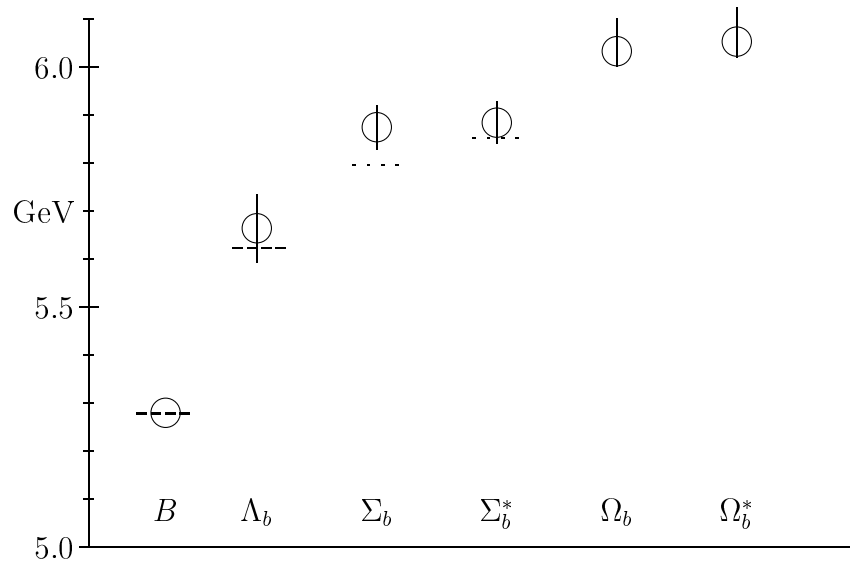


Fig. 6: Baryons with one b and two light quarks from lattice NRQCD [8]. Error bars include statistical errors and systematic errors from fixing the strange quark mass. Dashed lines indicate experimental error bounds from the Particle Data Tables [3]; dotted lines, preliminary experimental results from DELPHI [9].

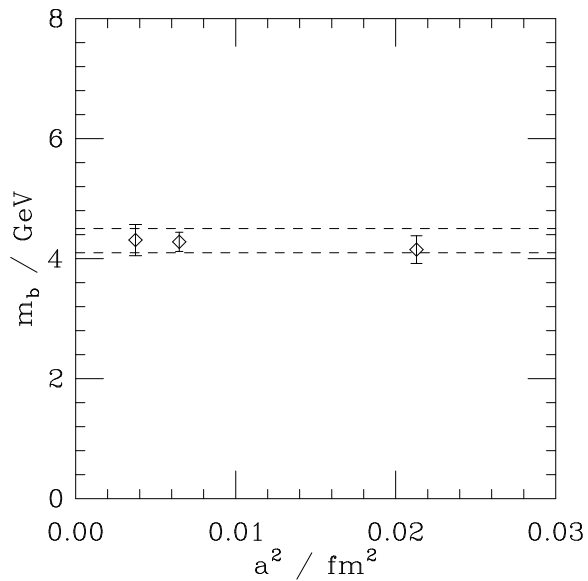


Fig. 7: The b mass in the \overline{MS} scheme at its own scale, as obtained from lattice NRQCD calculations at three different values of the lattice spacing. The bare b mass was fixed by matching to the experimental Υ mass.

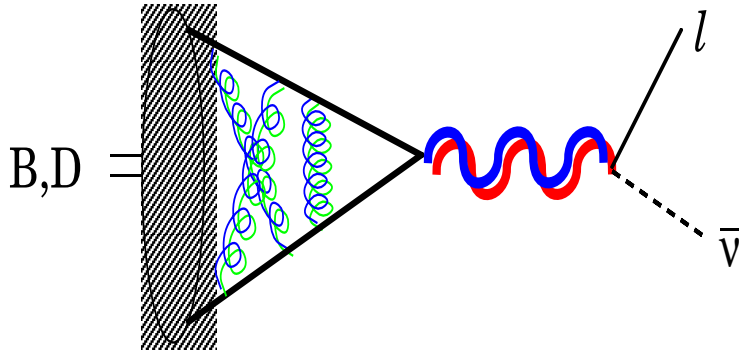


Fig. 8: Leptonic decay of a heavy-light meson.

The decay constant is obtained from the same lattice calculations as those from which we obtain the spectrum. When we fit the correlation function to an exponential in T the exponent gives us the mass, and the amplitude of the exponential gives us the decay constant. In fact the calculation is somewhat more complicated than this for two reasons.

One is that the decay constant is technically the matrix element between the vacuum and the heavy-light meson of the quark current that couples to the W . On the lattice we need a representation of this current which as accurately as possible matches the continuum current. This requires in practice adding terms to the naïve discretisation of the continuum current. The additional terms correct for errors because of the finiteness of the lattice spacing (this is part of the Symanzik improvement programme [1]) and of the heavy quark mass (these are $1/m_Q$ and $1/m_Q^2$ corrections). Because $m_Q a$ is dimensionless of the lattice these two sets of corrections mix together.

The second complication is that, again to match accurately to a continuum matrix element, we must add radiative corrections to the lattice determination of f_B to correct for differences in the renormalisation scheme between the lattice and the continuum. These corrections can be calculated in perturbation theory but require a large ‘mixing and matching’ calculation when the corrections to the current of the paragraph above are included. Such calculations have been done over the past two years for NRQCD and (at a less complete level) for the relativistic heavy quark Lagrangian and this has led to significant improvements in the determinations of f_B and f_D [14].

Figure 9 shows NRQCD lattice results in the quenched approximation for the decay constants of pseudoscalar mesons of different masses as a function of their mass, including these current corrections and renormalisations [15]. Heavy Quark Symmetry predicts that f_{PS} should vary as $1/\sqrt{M_{PS}}$ because the ‘wave function at the origin’ of a heavy-light meson should be independent of the heavy quark mass in the limit in which it acts as a static colour source. From the Figure we can see that the lattice results do see deviations from the HQS limit in the shape of the curve.

The world averages quoted at the recent LAT98 meeting [14] for the B and D in the quenched approximation are :

$$\begin{aligned}
 f_B &= 161(18)\text{MeV} \\
 f_{B_s} &= 181(22)\text{MeV} \\
 f_D &= 198(18)\text{MeV} \\
 f_{D_s} &= 218(20)\text{MeV} \\
 f_{B_s}/f_B &= 1.14(5).
 \end{aligned}
 \tag{4}$$

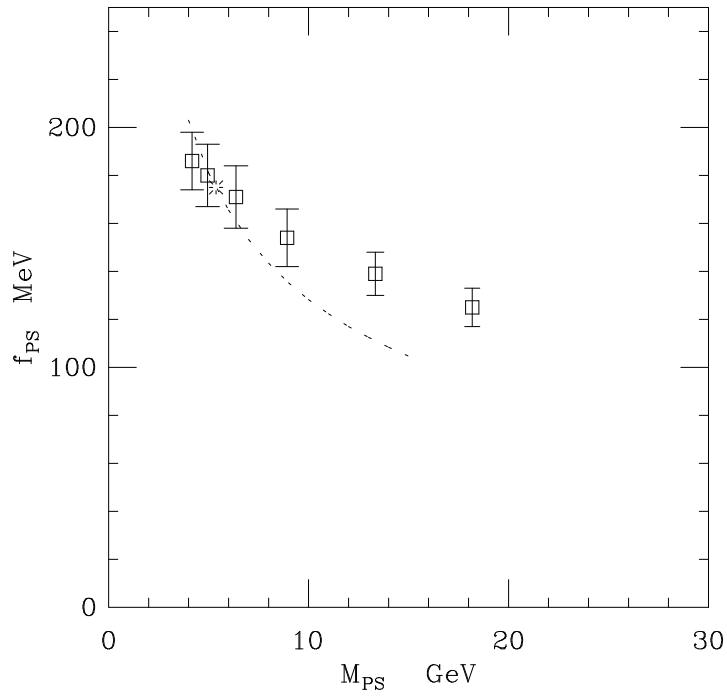


Fig. 9: The decay constant, f , determined for pseudoscalar heavy-light mesons of different masses using lattice NRQCD. The dashed line represents $1/\sqrt{M}$ behaviour expected from Heavy Quark Symmetry, constrained to go through the lattice results for the B_s , marked with a burst.

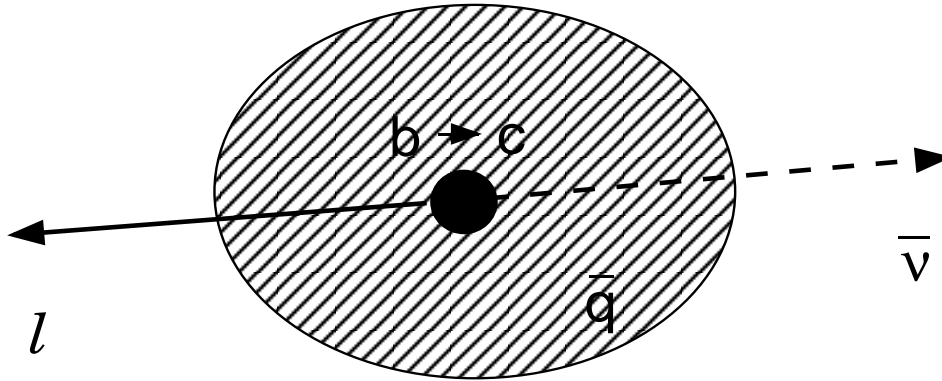


Fig. 10: Semi-leptonic decay of a B meson into a D meson at the zero recoil point.

These numbers are an average over several recent lattice calculations and the error quoted includes statistical and systematic errors within the quenched approximation. These errors are now down to the 10% level and this is a big improvement over the status of, say, two years ago. The central values have fallen too, as various systematic errors described above have been taken care of. The effect of unquenching is still not clear. We expect it to increase the decay constant since, like hyperfine splittings described earlier, it is sensitive to shorter distance scales than the quantities used to fix the lattice spacing. Estimates of the increase expected on unquenching differ from about 20 MeV [16] to 40 MeV [17]. More results on unquenched configurations will certainly follow now that the quenched results are under such good control.

From equation 3 it is clear that an experimental measurement of the leptonic decay rate coupled with a lattice calculation of the decay constant, can yield a determination of certain CKM elements. However there are other kinematic factors not explicit in equation 3, such as helicity suppression of the production of a lepton and anti-lepton from a spin-zero meson, which make the rates from leptonic decay very low and hard to observe [2]. A leptonic decay rate has been measured for the D_s (which gives information on V_{cs}) but measurements for the B are unlikely to be achieved in the near future.

5 SEMI-LEPTONIC MATRIX ELEMENTS

Semi-leptonic decay modes of interest are $B \rightarrow D^{(*)}l\nu$ (illustrated in Figure 1) and $B \rightarrow \pi/\rho l\nu$. These are experimentally more accessible than the purely leptonic modes and, with theoretical calculations of the matrix elements, can yield determinations of the CKM elements V_{cb} and V_{ub} .

Unlike leptonic decays, semi-leptonic decays have 3-body final states and this means a range of momentum transfer, $q^2 = M_W^2$. The hadronic matrix element to be calculated in lattice QCD can then be parameterised in terms a limited set of ‘form factors’ which are functions of q^2 .

For $B \rightarrow D/D^*$ decay a useful kinematic point to study is $q^2 = q_{max}^2 = (m_B - m_{D^{(*)}})^2$. At this point the daughter D/D^* meson has no 3-momentum with respect to the parent B – it is known as the ‘zero-recoil’ point (see Figure 10). Here Heavy Quark Symmetry provides useful insight. The light brown muck in a heavy-light meson is not sensitive to the flavour or spin of the heavy quark when it is acting as a static color source. Thus the replacement of the b quark by a c quark, and even a c quark with spin flipped for a D^* , produces no change. This leads to a normalisation of the form factor at zero recoil. We have

$$\frac{d\Gamma}{d\omega} = \frac{G_F^2}{48\pi^3} (\text{kinematic factors}) |V_{cb}|^2 \mathcal{F}(\omega)^2 \quad (5)$$

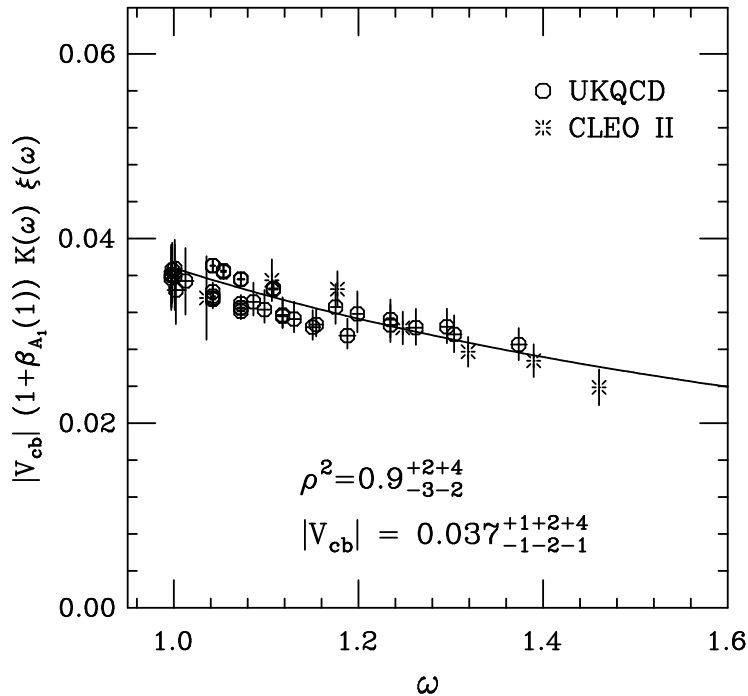


Fig. 11: Lattice results [18] and experimental results for the combination of form factor and CKM element relevant to $B \rightarrow D$ semi-leptonic decays. The comparison yields the slope of the Isgur-Wise function, ρ^2 and V_{cb} given.

with ω equal to $\gamma = 1/\sqrt{(1-v^2/c^2)}$ for the D/D^* meson in the B rest frame. As $\omega \rightarrow 1$, $q^2 \rightarrow q_{max}^2$. In the Heavy Quark Symmetry limit the combination of form factors represented by $\mathcal{F}(\omega)$ becomes equal to a single form factor known as the Isgur-Wise function, $\xi(\omega)$, and takes a universal form independent of the flavour or spin of the heavy quarks involved. In the $\omega \rightarrow 1$ limit $\xi(1) = 1$. For $B \rightarrow D^*$ decay it is apparent that the corrections to Heavy Quark Symmetry in this limit are quite small. The point $\omega = 1$ itself is not experimentally accessible since the rate vanishes at that point [2]. The theoretical challenge is then to map out the shape of the form factors as a function of ω to guide the experimental extrapolation to $\omega = 1$. Figure 11 shows combined results from a 1995 lattice calculation by UKQCD and the CLEO II experiment, in which the form factor away from the zero recoil limit is parameterised by a slope ρ^2 , and the value of V_{cb} determined is also given.

These early lattice calculations generally assumed behaviour based on Heavy Quark Symmetry whereas more recent calculations [14] are testing it more stringently. This requires more technology in terms of mass-dependent renormalisation constants and current mixing and is at a much less complete stage than decay constant calculations. This more general approach will be useful for semileptonic decays such as $B \rightarrow \pi/\rho$ and $D \rightarrow K, K^*$ where Heavy Quark Symmetry has much less to say. Calculations for D to K semi-leptonic decay are in quite good shape, based on a relativistic approach to the c quark [14, 7]. B decays to light mesons are much harder, however, because the typical recoil momenta of the light hadrons are large. The problems of being able to simulate momenta $\vec{p}a \gg 1$ on a lattice with spacing a then reappear, and we expect significant discretisation errors. Current calculations tend to work at small $\vec{p}a$ and extrapolate to the physical region, but it is not necessarily clear how to do this. Much work remains to be done in this area.

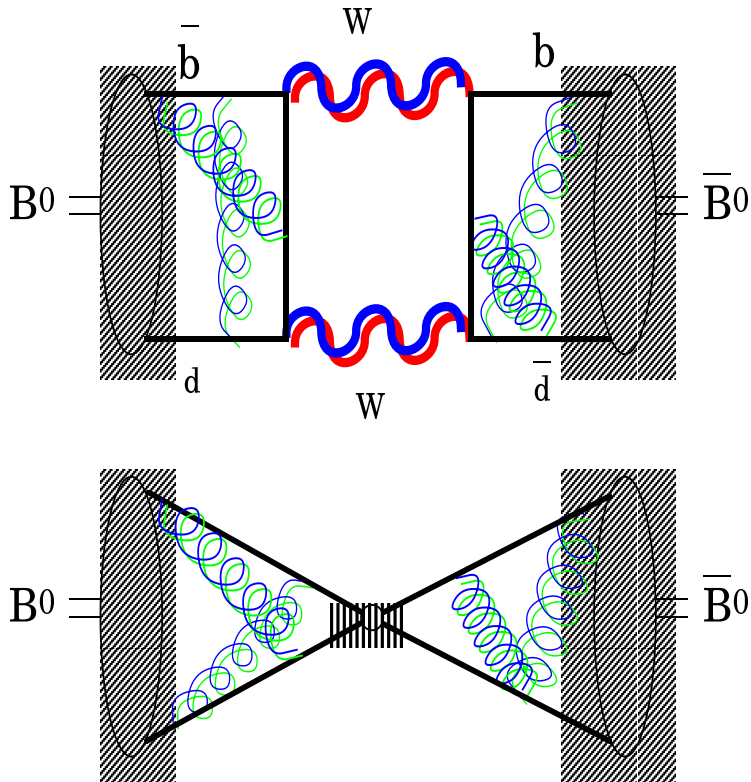


Fig. 12: $B^0 - \bar{B}^0$ mixing. At the top the box diagram and below the equivalent 4-quark operator.

6 NEUTRAL B MIXING

B^0 and \bar{B}^0 can mix via the well-known ‘box’ diagram of Figure 12 in an analogous way to the mixing between K^0 and \bar{K}^0 . This leads to oscillations between these flavour eigenstates with time which depends on the mass difference between the mass eigenstates [2]. The measurement of the oscillations can be combined with a theoretical calculation of the hadronic matrix element of the box diagram to yield a determination of V_{td} since the diagram is dominated by intermediate t quarks.

The mass difference from the box diagram is parameterised as:

$$\Delta m_{B - \bar{B}} = f_B^2 B_B |V_{td}|^2 F(m_t, m_W) \quad (6)$$

where F is a known function. The box diagram matrix element is given as $f_B^2 B_B$ where B_B is the B meson ‘bag constant’. In lattice calculations we imagine the intermediate W bosons and t quarks shrunk to a point since they are too massive to propagate on a lattice with lattice spacing $a \approx 0.1$ fm. Then the box diagram reduces to the matrix element of an effective 4-quark operator, shown below on Figure 12. This then looks rather like the square of the purely leptonic decay diagram 8 and, in the ‘vacuum saturation approximation’ in which the vacuum state is inserted between the two halves of the 4-quark operator, the answer would be f_B^2 . Thus B_B parameterises the difference between the real answer and the vacuum saturation approximation. We expect B_B to be close to 1, which is one reason why a lattice determination of f_B has been so important.

On the lattice we calculate the hadronic matrix element of the 4-quark operator and thus naturally obtain the product $f_B^2 B_B$. By taking ratios with calculations yielding f_B one can obtain B_B separately. Calculations have been done on the lattice using both relativistic and non-relativistic Lagrangians for the b quarks and results are broadly in agreement. The mixing

and renormalisation problem is similar to that for f_B but not all the renormalisation constants with their mass dependence have yet been calculated. A recent review [14] quotes

$$B_{B_d}(m_b) = 0.86(4)(8) \quad (7)$$

$$B_{B_s}/B_{B_d} = 1.00(1)(2) \quad (8)$$

where the errors are statistical and systematic. These numbers are obtained from the quenched approximation but unquenching is not expected to have a big effect on them.

7 CONCLUSIONS

- Good calculations of the spectrum and heavy hadrons are possible on the lattice with some care. In fact, in a lot of cases the results are more accurate than for light hadrons.
- The spectrum and decay constants are fairly ‘mature’ calculations now; other quantities are improving.
- The largest remaining source of systematic uncertainty is that of the quenched approximation. Unquenched calculations will improve over the next few years. Keep your eye on the lattice literature!

References

- [1] R. D. Kenway, these proceedings.
- [2] J. Richman, *Heavy Quark Physics and CP violation*, LXV11 Les Houches Summer School 1997.
- [3] The Particle Data Group, <http://pdg.lbl.gov/> .
- [4] F. Abe *et al*, CDF collaboration, Phys. Rev. Lett. **81** (1998) 2432.
- [5] C. T. H. Davies, lectures given at the 36th Internationale Universitatswochen fuer Kernphysik und Teilchenphysik, 1997, Lecture Notes in Physics, **512** (Springer-Verlag) (1998) 1.
- [6] A. El-Khadra, A. S. Kronfeld and P. Mackenzie, Phys. Rev. **D55** (1997) 3933.
- [7] J. Flynn and C. Sachrajda, hep-lat/9710057, to appear in ‘Heavy Flavours II’, eds A. Buras and M. Lindsa, World Scientific.
- [8] A. Ali Khan, Proceedings of LAT98, Nucl. Phys. B (Proc. Suppl.), in press, hep-lat/9809140.
- [9] DELPHI Collaboration, DELPHI Note 95-107, contribution to EPS ‘95.
- [10] DELPHI Collaboration, DELPHI 96-93 CONF 22, contribution to ICHEP ‘96.
- [11] P. Boyle, Proceedings of LAT97, Nucl. Phys. B (Proc. Suppl.) **63** (1998) 314; P.B. Mackenzie, S. Ryan and J. Simone, *ibid*, 305.
- [12] K. C. Bowler *et al*, UKQCD collaboration, Phys. Rev. **D54** (1996) 3619.
- [13] A. Kronfeld, Proceedings of LAT97, Nucl. Phys. B (Proc. Suppl.) **63** (1998) 311.
- [14] T. Draper, review talk at the LAT98 meeting, Nucl. Phys. B (Proc. Suppl.) in press, hep-lat/9810065.

- [15] A. Ali Khan *et al*, Phys. Lett B**427** (1998) 132.
- [16] C. Bernard *et al* Phys. Rev. Lett. **81** (1998) 4812.
- [17] S. Collins *et al*, hep-lat/9901001.
- [18] K. C. Bowler *et al*, Phys. Rev. D**52** (1995) 5067.

EXPERIMENTAL TECHNIQUES

Tejinder S. VIRDEE

EP Division, CERN, 1211 Geneva 23, Switzerland and

Imperial College of Science Technology and Medicine, London SW7 2BZ, U.K.

Abstract

Experimental techniques used in high energy physics experiments are described. Emphasis has been placed on the techniques to be used in the pp experiments at the Large Hadron Collider. The physics underlying the working of some key detectors is outlined. Factors determining the measurement accuracy attainable with these detectors are discussed. The reasoning behind the designs of the ATLAS and CMS experiments is discussed.

1. INTRODUCTION

The aim of Particle Physics is to answer the two following questions: What are the fundamental constituents of matter? What are the fundamental forces that control their behaviour at the most basic level? Experimentally this involves study of hard particle interactions, determining the identity of the resulting particles and measuring their momenta with as high a precision as possible. Some thirty years ago a single detection device, the bubble chamber, was sufficient to reconstruct the full event information. At the current high centre of mass energies no single detector can accomplish this even though the number of particles whose identity and momenta can be usefully determined is limited [electrons, muons, photons, jets, b-jets, taus and missing transverse energy $E_t(\nu)$]. This leads to a familiar onion-like structure of present day high energy physics experiments. Each layer is specialized to measure and identify different classes of particles.

Starting from the interaction vertex the momenta of charged particles is determined in the inner tracker which is usually immersed in a solenoidal magnetic field. Identification of b-jets can be accomplished by placing high spatial resolution detectors such as silicon pixel or microstrip detectors close to the interaction point. Following the tracking detectors are calorimeters which measure the energies and identify electrons, photons, single hadrons or jets of hadrons. Only muons and neutrinos penetrate through the calorimeters. The muons are identified and measured in the outermost sub-detector, the muon system which is usually immersed in a magnetic field. The presence of neutrinos is deduced from the apparent imbalance of transverse momentum or energy.

These lectures rely heavily on previous literature [1-9]. The emphasis is placed on their use at the future Large Hadron Collider. Hence illustrative examples from ATLAS and CMS are used. These examples are described in detail in Technical Design Reports of various sub-detectors [10,11]. Examples will also be taken from some of the many other experiments, planned or ongoing, using novel techniques.

2. INTERACTION WITH MATTER

2.1 Energy Loss by Charged Particles

Moderately relativistic charged particles, other than electrons, lose energy in matter through the Coulomb interaction with the atomic electrons. The energy transferred to the electrons causes them either to be ejected from the parent atom (*ionisation*) or to be excited to a higher level (*excitation*). The energy loss is given by the Bethe-Bloch equation :

$$-\frac{dE}{dx}\Big|_{ion} = N_A \frac{Z}{A} \frac{4\pi\alpha^2(\hbar c)^2}{m_e c^2} \frac{Z_i^2}{\beta^2} \left[\ln \frac{2m_e c^2 \gamma^2 \beta^2}{I} - \beta^2 - \frac{\delta}{2} \right] \quad (1)$$

where E is the kinetic energy of the incident particle with velocity β and charge Z_i , I ($\approx 10 \times Z$ eV) is the mean ionization potential in a medium with atomic number Z. The notable features of this formula are:

- $1/m_e \Rightarrow$ energy is lost essentially to electrons
- $1/\beta^2 \Rightarrow$ slower particles lose more energy. The slower particles have a longer time in the vicinity of atomic electrons during which to interact.
- at low β the increase in energy loss does not continue down to $\beta = 0$. In a head-on collision the electron can acquire an energy $2m_e v^2$. If this is insufficient to excite electrons to higher states then the particle cannot lose energy. If $I \approx 2m_e v^2$ then the 1st term $\rightarrow 0$.
- energy loss is $\propto Z_i^2$
- the energy loss minimum occurs at $\beta\gamma \approx 4$ (said to be minimum ionizing particles or mips)
- relativistic rise – the relativistic expansion of the electric field means that for relativistic particles interaction with electrons further and further away is possible. The relativistic rise does not continue forever as the polarization of the medium starts screening electrons in the more distant atoms.

A very useful quantity is *areal density* measured in units of g.cm^{-2} . The energy loss of relativistic particles of unit electric charge per unit areal density is found to be roughly the same in all materials with

$$\frac{1}{\rho} \frac{dE}{dx} \approx 1.5 - 2 \frac{\text{MeV}}{\text{g.cm}^{-2}}$$

where ρ is the density of the medium. The energy loss rate in liquid hydrogen, gaseous helium, carbon, aluminium, tin and lead is shown in Fig. 1 [4]. It can be seen that the above approximation is valid for all solids.

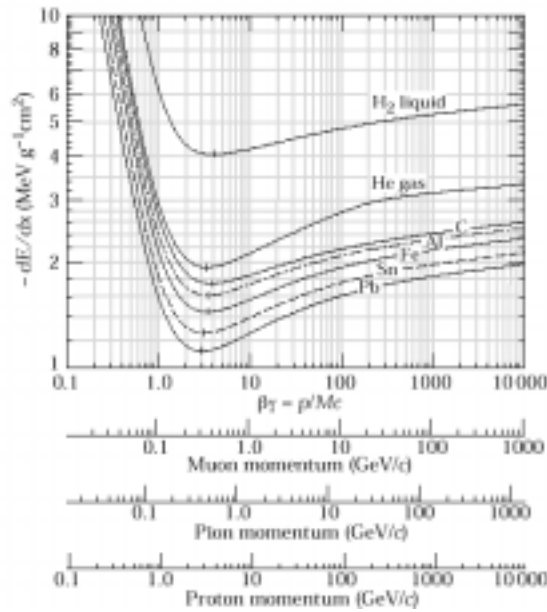


Figure 1: Energy loss rate in liquid hydrogen, gaseous helium, carbon, aluminium, tin and lead.

2.2 Energy Loss by Electrons

Above about 1 GeV radiative processes dominate energy loss by electrons and photons. In the intense electric field of nuclei relativistic electrons radiate photons (*bremstrahlung*), and photons can be converted into electron-positron pairs (*pair creation*).

In dealing with electrons and photons at high energies striking blocks of material (e.g. calorimeters) it is convenient to measure the depth and radial extent of the resulting cascades in terms of *radiation length* (X_0) and *Molière radius* (R_M).

Consider the process of bremsstrahlung. A free electron cannot radiate a photon. However a charged particle emits radiation when it is subjected to acceleration or deceleration. For a given force the acceleration/deceleration is greater the lighter the particle. The Feynman diagram for the bremsstrahlung process is shown in Figure 2. The cross section for the process comprises the coupling constant at the three vertices and the propagator term ($\propto 1/m^2$)

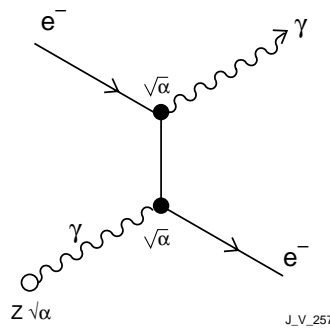


Figure 2: The Feynman diagram for bremsstrahlung

$$\sigma \propto \frac{Z^2 \alpha^3}{m_e^2 c^4}$$

We are interested in $d\sigma/dv$ where v is the energy of the emitted photon. We can make a guess for the expression using dimensional arguments:

$$\frac{d\sigma}{dv} \propto \frac{Z^2 \alpha^3}{m_e^2 c^4} \frac{(\hbar c)^2}{v}$$

Turning this to energy loss per unit distance traversed by the electrons gives

$$-\frac{d\sigma}{dx}|_{rad} = n \int_{v_{min}}^{v_{max}} v \frac{d\sigma}{dv} dv = n \frac{Z^2 \alpha^3 (\hbar c)^2}{m_e^2 c^4} (v_{max} - v_{min})$$

where v_{max} = kinetic energy of electron, $v_{min} \approx 0$ and n is no. of nuclei/unit volume. A numerical factor $[4 \ln(183/Z^{1/3})]$ has to be added describing the effect of the possible range of impact parameters of the electron. At large impact parameters the protons are shielded by atomic electrons. Hence

$$-\frac{dE}{dx}\Big|_{rad} = \left[4n \frac{Z^2 \alpha^3 (\hbar c)^2}{m_e^2 c^4} \ln \frac{183}{Z^{1/3}} \right] E$$

Since $-\frac{dE}{dx} \propto E \Rightarrow \frac{dE}{E} = -B dx \Rightarrow E = E_0 e^{-Bx}$ where A is a constant.

The *radiation length* is defined to be the distance over which the electron loses, on average, all but 1/e of its energy i.e. $X_0 = 1/B$ i.e.

$$X_0 = \left[4n \frac{Z^2 \alpha^3 (\hbar c)^2}{m_e^2 c^4} \ln \frac{183}{Z^{1/3}} \right]^{-1}$$

Infact for Pb, $Z = 82$, $n = 3.3 \cdot 10^{28}$ nuclei/m³, $X_0 \approx 5.3$ mm which is close to the PDG [4] value of 5.6 mm. The radiation length can be approximated as $X_0 \approx \frac{180A}{Z^2} \text{ g.cm}^{-2}$ where A is the mass number

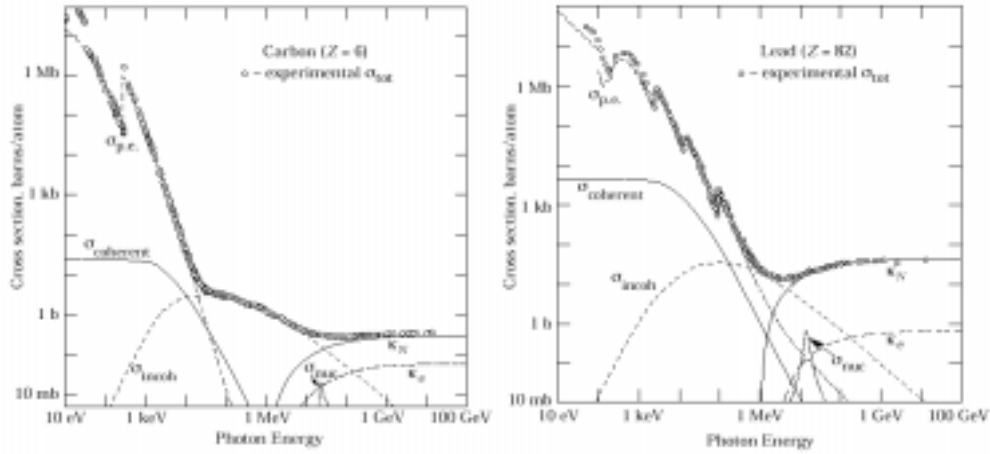


Figure 3: The photon total cross-sections as a function of energy in carbon and lead.

2.3 Energy Loss by Photons

Photons lose energy through photoelectric effect and Compton scattering at low energies and by pair production at relativistic energies. The cross-section for photoelectric effect is given by

$$\sigma_{pe} \approx Z^5 \alpha^4 \left(\frac{m_e c^2}{E_\gamma} \right)^n \quad n = \frac{7}{2} \text{ at } E_\gamma \ll m_e c^2 \text{ and } n \rightarrow 1 \text{ at } E_\gamma \gg m_e c^2$$

with a strong dependence on Z. The cross-section for Compton scattering has been calculated by Klein and Nishina :

$$\sigma_C \approx \frac{\ln E_\gamma}{E_\gamma} \text{ per electron and } \sigma_C^{atom} = Z \sigma_C \text{ per atom}$$

If the energy of the photon is $\gg m_e c^2$ then the dominant energy loss mechanism is pair production and its probability can be deduced, as done in Equation (1) for bremsstrahlung. It is given by:

$$\sigma_{pair} \approx \frac{7}{9} \frac{A}{N_A} \frac{1}{X_0}$$

The probability of a pair conversion in 1 X_0 is $e^{-7/9}$. Since the photon disappears on producing a pair a mean free path length can be defined as

$$L_{pair} = \frac{9}{7} X_0 \text{ independent of energy.}$$

The photon total cross-sections as a function of energy in carbon and lead are shown in Figure 3 [4] which shows the above mentioned dependences.

2.3.1 Critical Energy and Molière Radius

The *critical energy*, ϵ , is defined to be the energy at which the energy loss due to ionisation (at its minimum i.e. at $\beta \approx 0.96$) and radiation are equal (over many trials) i.e.

$$\frac{(dE/dx)_{rad}}{(dE/dx)_{ion}} = \frac{Z\alpha}{\pi m_e c^2} E \beta^2 \frac{\ln 183/Z^{1/3}}{\ln \left[\frac{(2m_e c^2 \beta^2)}{I(1-\beta^2)} \right] - \beta^2} = 1$$

which simplifies to

$$\Rightarrow \epsilon \approx \frac{560}{Z} \text{ (MeV)}$$

The Molière radius gives the average lateral deflection of critical energy electrons after traversal of 1 X_0 and is parameterized as:

$$R_M = \frac{21_{MeV} X_0}{\epsilon} \approx \frac{7A}{Z} \text{ g.cm}^{-2}$$

Table 1: Physical properties of some materials used in calorimeters.

	Z	ρ g.cm ⁻³	I/Z eV	(1/ ρ)dT/dx MeV/g.cm ⁻³	ϵ MeV	X_0 cm	λ_{int} cm
C	6	2.2	12.3	1.85	103	≈ 19	38.1
Al	13	2.7	12.3	1.63	47	8.9	39.4
Fe	26	7.87	10.7	1.49	24	1.76	16.8
Cu	29	8.96		1.40	≈ 20	1.43	15.1
W	74	19.3		1.14	≈ 8.1	0.35	9.6
Pb	82	11.35	10.0	1.14	6.9	0.56	17.1
U	92	18.7	9.56	1.10	6.2	0.32	10.5

2.4 Hadronic Interactions

A high energy hadron striking an absorber interacts with nuclei resulting in multi-particle production consisting of secondary hadrons (e.g. π^\pm , π^0 , K etc.). A simple model treats the nucleus, mass number A, as a black disc with radius R. Then

$$\sigma_{\text{int}} = \pi R^2 \propto A^{2/3} \quad \text{where } R \approx 1.2 \times A^{1/3} \text{ fm}$$

$$\text{infact } \sigma_{\text{inel}} = \sigma_0 A^{0.7} \quad \text{where } \sigma_0 = 35 \text{ mb}$$

In dealing with hadrons it is convenient to measure the depth and radial extent of the resulting cascades in terms of *interaction length* (λ_{int}) which is defined as

$$\lambda_{\text{int}} = \frac{A}{N_A \sigma_{\text{int}}} \propto A^{1/3}$$

The values of the above mentioned parameters for various materials are listed in Table. 1.

3. EXPERIMENTAL MEASUREMENTS: MEASUREMENT OF MOMENTUM

Consider the motion of a charged particle in a uniform solenoidal magnetic field (Fig.4). The radius of curvature, r , is given by:

$$r = \frac{p_T}{0.3B}$$

where r is measured in m, B is the magnetic field strength measured in T and p_T is the momentum perpendicular to B and measured in GeV/c.

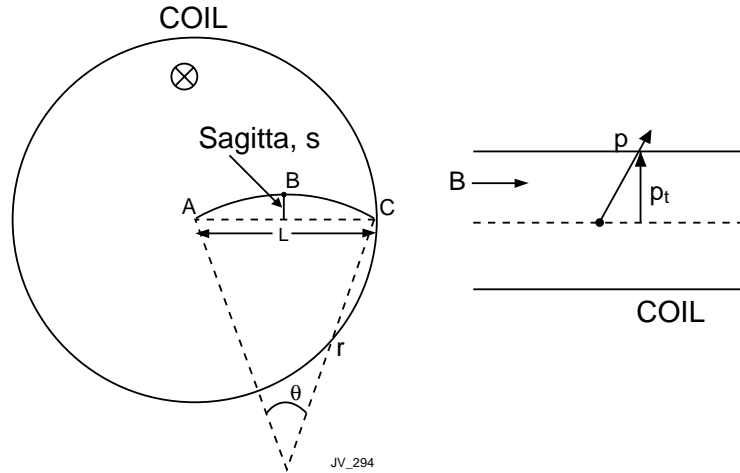


Figure 4: The trajectory of a charged particles in a magnetic field.

The angle θ is given by

$$\sin \frac{\theta}{2} = \frac{L}{2r}$$

If $r \gg L$ then

$$\frac{\theta}{2} \approx \frac{L}{2r} \Rightarrow \theta \approx \frac{0.3BL}{p_T}$$

Therefore the sagitta, s , is given by

$$\begin{aligned}
s &= r - r \cos(\theta/2) \\
&\approx r \left[1 - \left(1 - \frac{1}{2} \frac{\theta^2}{4} \right) \right] \\
&= \frac{r\theta^2}{8} \approx \frac{0.3BL^2}{8p_T}
\end{aligned}$$

As an example $s \approx 3.75$ cm for $p_T=1$ GeV/c, $L=1$ m and $B=1$ T. Suppose the sagitta is measured using points A, B and C. Then

$$s = x_B - \frac{x_A + x_C}{2}$$

$$\therefore ds = dx_B - \frac{dx_A}{2} - \frac{dx_C}{2}$$

assuming $dx_i \approx \sigma(x)$ is the independent single point error

$$(ds)^2 = \sigma^2(x) + \frac{\sigma^2(x)}{4} + \frac{\sigma^2(x)}{4} = \frac{3}{2} \sigma^2(x) \equiv \sigma_s^2$$

where σ_s is the error on the sagitta. The relative momentum resolution can now be estimated as:

$$\begin{aligned}
\frac{dp_T}{p_T} &= \frac{\sigma_s}{s} = \frac{\sqrt{(3/2)} \sigma_x}{s} \\
\frac{dp_T}{p_T} &= \frac{\sqrt{3}}{2} \sigma_x \frac{8p_T}{0.3BL^2} \quad (2)
\end{aligned}$$

Hence the momentum resolution will degrade linearly with increasing p_T but will improve for higher field and larger radial size of the tracking cavity. The latter improvement is quadratic in L ! The next question that can be asked is the arrangement of N measuring points. Uniform spacing is best for minimizing the effect of multiple scattering (see 3.1.2) and the resolution is given by

$$\frac{dp_T}{p_T} = \frac{\sigma_x p_T}{0.3BL^2} \sqrt{\frac{720}{N+4}}$$

For example, $dp_T/p_T \approx 0.5\%$ for $p_T = 1$ GeV/c, $L = 1$ m, $B = 1$ T, $\sigma_x = 200$ μm and $N=10$. For the best momentum resolution $N/2$ points should be grouped at the centre and $N/4$ points at the two ends of the track. Then

$$\frac{dp_T}{p_T} = \frac{\sigma_x p_T}{0.3BL^2} \sqrt{\frac{256}{N}}$$

leading to an improvement in the momentum resolution by a factor of 0.6.

However in a real tracker the errors due to multiple scattering need to be included.

3.1 Multiple Scattering

The electric field close to an atomic nucleus may give a large acceleration to a charged particle. This will result in a change of direction for a heavy charged particle ($m > m_e$). For small particle-

nucleus impact parameters a single large angle scatter is possible. This is described by Rutherford scattering and the angle θ is given by

$$\frac{d\sigma}{d\Omega} \propto \frac{1}{\sin^4 \theta/2}$$

Larger impact parameters are more probable and the scattering angle will be smaller as the nuclear charge is partly screened by the atomic electrons. Hence in a relatively thick material there will be a large number of random and small deflections. This is described by multiple Coulomb scattering. The relative probability of scattering as a function of scattering angle is illustrated in Fig. 5.

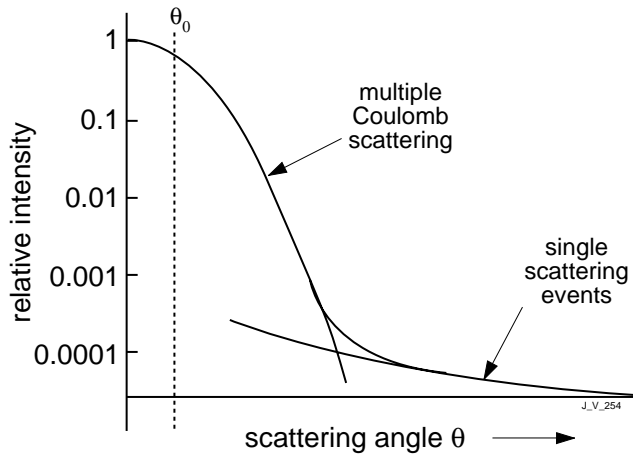


Figure 5: The relative probability of scattering as a function of scattering angle.

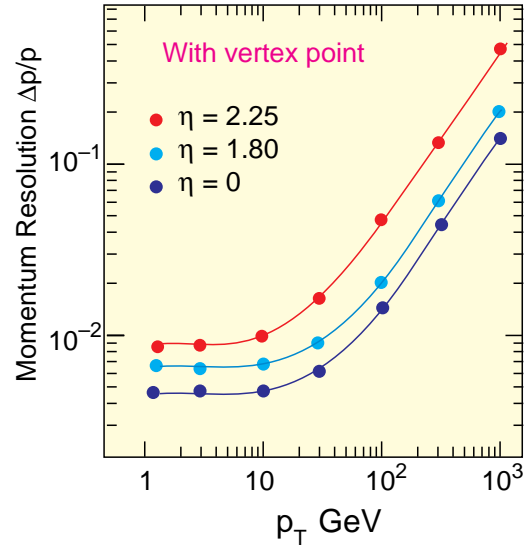


Figure 6: The estimated momentum resolution In CMS as a function of p_T at various η .

The r.m.s. of the scattering angle is given by

$$\theta_0 \approx \frac{13.6 \text{ MeV}}{\beta p c} Z_{inc} \sqrt{\frac{L}{X_0}}$$

where L is the thickness of the material in the tracker and X_0 is the radiation length of the material, Both are measured in m. The apparent sagitta due to multiple scattering is given by

$$s_{ms} = \frac{L\theta_0}{4\sqrt{3}}$$

If the extrapolation error from one measuring plane to the next is larger than the point resolution, i.e. $\theta_0 \Delta r > \sigma_x$, then the momentum resolution will be degraded. The relative momentum resolution due to multiple scattering is then given by

$$\frac{s_{ms}}{s} = \frac{dp}{p} \Big|_{ms} \approx 0.05 \frac{1}{B\sqrt{LX_0}} \quad \text{since } s = \frac{0.3BL^2}{8p} \quad (3)$$

Hence the relative momentum resolution is independent of p and is proportional to $1/B$. For example $dp/p \approx 0.5\%$ for argon gas with $L=1\text{m}$ and $B=1\text{T}$. The estimated momentum resolution

in CMS is illustrated in Fig. 6. The momentum resolution is independent of momentum in the range where the multiple scattering error dominates (up to ≈ 20 GeV/c). The resolution, $\Delta p/p$, above 20 GeV/c is proportional to p .

4 MEASUREMENT OF ENERGY

Neutral and charged particles incident on a block of material deposit their energy through creation and destruction processes. The deposited energy is rendered measurable by ionisation or excitation of the atoms of matter in the active medium. The active medium can be the block itself (*totally active or homogeneous calorimeter*) or a sandwich of dense absorber and light active planes (*sampling calorimeter*). The measurable signal is usually linearly proportional to the incident energy.

An example of the phenomena (bremsstrahlung and pair production) involved in electromagnetic showers is illustrated in Fig. 7 in which a 50 GeV electron is incident on the BEBC Ne/H₂ (70%/30%) bubble chamber in a 3 T magnetic field.

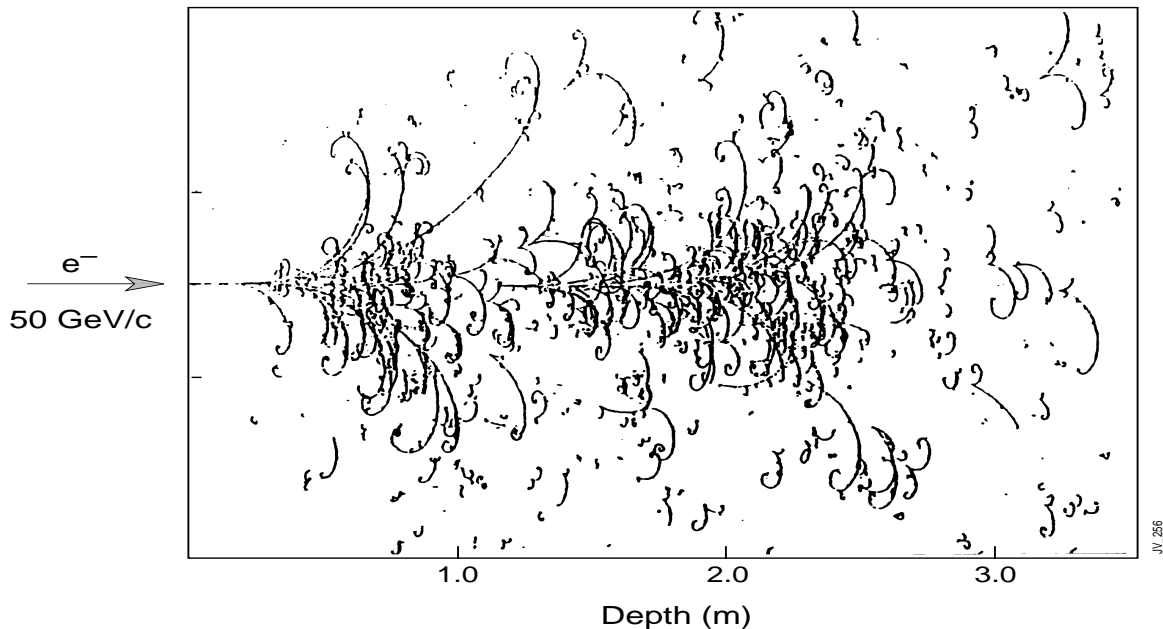
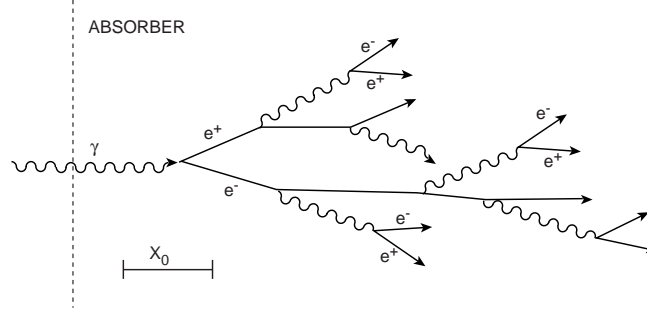


Figure 7: An example of a 50 GeV electron shower in a Ne/H₂ (70%/30) filled BEBC bubble chamber. The radiation length is ≈ 34 cm.

4.1 The Electromagnetic Cascade

4.1.1 Longitudinal Development of the Electromagnetic Cascade

A high energy electron or photon incident on a thick absorber initiates a cascade of secondary electrons and photons via bremsstrahlung and pair production as illustrated in Fig. 8. With increasing depth the number of secondary particles increases while their mean energy decreases. The multiplication continues until the energies fall below the critical energy, ϵ . Ionization and excitation rather than generation of more shower particles dominate further dissipation of energy.



JV217.c

Figure 8: Schematic development of an electromagnetic shower.

Consider a simplified model of development of an electromagnetic shower initiated by an electron or a photon of an energy E . A universal description, independent of material, can be obtained if the development is described in terms of scaled variables

$$t = \frac{x}{X_0} \quad \text{and} \quad y = \frac{E}{\varepsilon}$$

Since in $1 X_0$ an electron loses about $2/3^{\text{rd}}$ of its energy and a high energy photon has a probability of $7/9$ of pair conversion, we can naively take $1 X_0$ as a generation length. In each generation the number of particles increases by a factor of 2. After t generations the energy and number of particles is

$$e(t) = \frac{E}{2^t} \quad \text{and} \quad n(t) = 2^t \quad \text{respectively.}$$

At shower maximum where $e \approx \varepsilon$, the no. of particles is

$$n(t_{\text{max}}) = \frac{E}{\varepsilon} = y \quad \text{and} \quad t_{\text{max}} = \ln \frac{E}{\varepsilon} = \ln y$$

Critical energy electrons do not travel far ($\leq 1X_0$). After the shower maximum the remaining energy of the cascade is carried forward by photons giving the typical exponential falloff of energy deposition caused by the attenuation of photons. Longitudinal development of 10 GeV showers in Al, Fe and Pb is shown in Fig. 9 [12]. It can be noted that the shower maximum is deeper for higher Z materials because multiplication continues down to lower energies. The slower decay beyond the maximum is due to the lower energies at which electrons can still radiate. Both of the above effects are due to lower ε for higher Z materials.

The mean longitudinal profile of energy deposition is given by:

$$\frac{dE}{dt} = Eb \frac{(bt)^{a-1} e^{-bt}}{\Gamma(a)}$$

The maximum of the shower occurs at $t_{\text{max}} = (a-1)/b$. Fits to t_{max} give

$$t_{\text{max}} = \ln y - 0.5 \quad \text{for electron-induced cascades and}$$

$$t_{\text{max}} = \ln y + 0.5 \quad \text{for photon-induced cascades.}$$

The coefficient a can be found using t_{max} and assuming $b \approx 0.5$. The photon induced showers are longer since the energy deposition only starts after the first pair conversion has taken place. The mean free path length for pair conversion of a high energy photon is $X_\gamma = (9/7) X_0$.

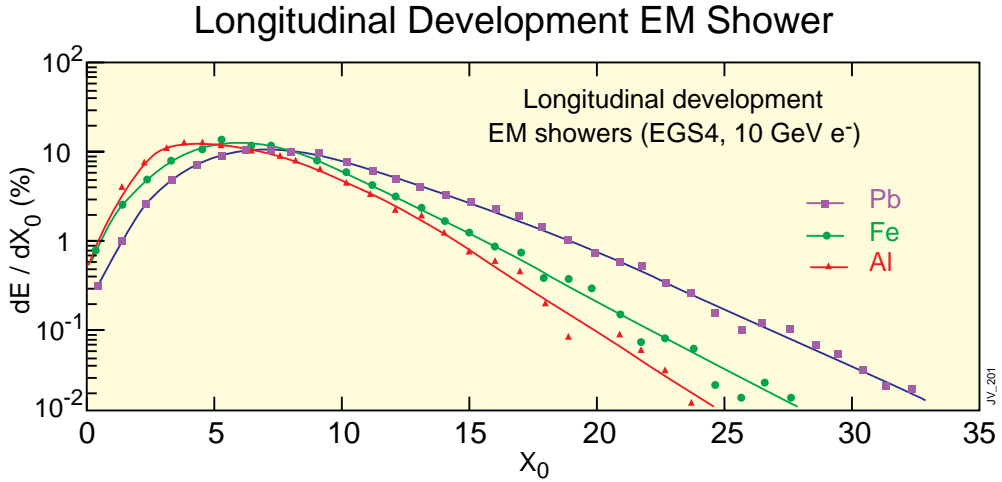


Figure 9: Simulation of longitudinal development of 10 GeV electron showers in Al, Fe and Pb.

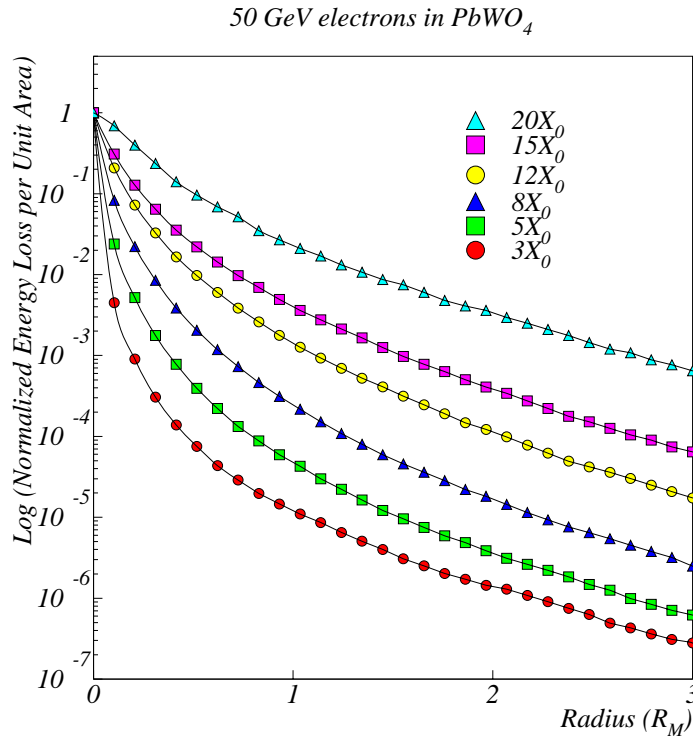


Figure 10: Lateral profile of energy deposition by 50 GeV electrons showers in PbWO_4 at various depths.

4.1.2 Lateral Development of the Electromagnetic Cascade

The lateral spread of an e.m. shower is determined by multiple scattering of electrons away from the shower axis. Also responsible are low energy photons which deposit their energy a long way away from their point of emission, especially when emitted from electrons that already travel at large angles with respect to the shower axis. The e.m. shower begins, and persists, with a narrow core of high energy cascade particles, surrounded by a halo of soft particles which scatter increasingly as the shower depth increases. This is shown in Fig. 10 for 50 GeV electrons incident on lead tungstate [13]. In different materials the lateral extent of e.m. showers scales fairly accurately with the Molière radius. An infinite cylinder with a radius of $\approx 1 R_M$ contains $\approx 90\%$ of

the shower energy. For lead tungstate, and a depth of $26 X_0$, the amount of energy contained in a cylinder of a given radius is shown in Fig. 11. The fact that e.m. showers are very narrow at the start can be used to distinguish single photons from pizeros (see Section 6.6.2).

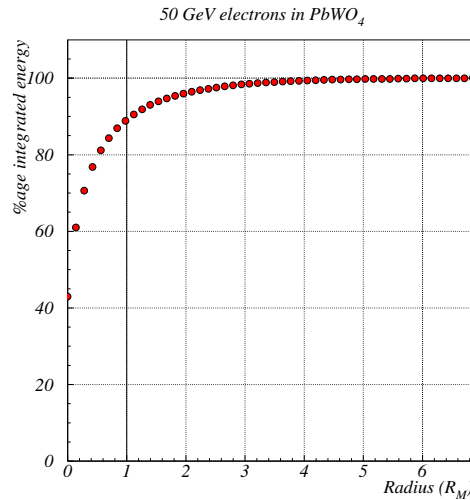


Figure 11: The percentage of energy contained in a cylinder of lead tungstate of different radii.

4.2 The Hadronic Cascade

4.2.1 The Longitudinal Development of the Hadronic Cascade

A situation analagous to that for e.m. showers exists for hadronic showers. The interaction responsible for shower development is the strong interaction rather than electromagnetic. The interaction of the incoming hadron with absorber nuclei leads to multiparticle production. The secondary hadrons in turn interact with further nuclei leading to a growth in the number of particles in the cascade. Nuclei may breakup leading to spallation products. The cascade contains two distinct components namely the electromagnetic one (π^0 's etc.) and the hadronic one (π^\pm , n, etc) one. This is illustrated in Fig. 12.

The multiplication continues until pion production threshold is reached. The average number, n , of secondary hadrons produced in nuclear interactions is given by $n \propto \ln E$ and grows logarithmically. The secondaries are produced with a limited transverse momentum of the order of 300 MeV.

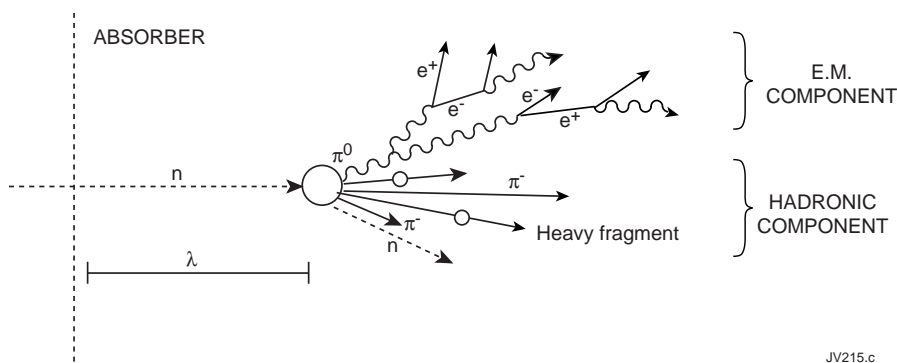


Figure 12: Schematic of development of hadronic showers.

It is convenient to describe the average hadronic shower development using scaled variables

$$v = x/\lambda \quad \text{and} \quad E_{th} \approx 2m_\pi = 0.28 \text{ GeV}$$

where λ is the nuclear interaction length and is the scale appropriate for longitudinal and lateral

development of hadronic showers. The generation length can be taken to be λ . Note $\lambda \approx 35 \text{ A}^{1/3} \text{ g.cm}^{-2}$. Furthermore, if it is assumed that $\langle n \rangle$ secondaries/primary are produced for each generation and that the cascade continues until no more pions can be produced. Then in generation v

$$e(v) = \frac{E}{\langle n \rangle^v}$$

$$e(v_{\max}) = E_{th} \quad \therefore E_{th} = \frac{E}{\langle n \rangle^{v_{\max}}}$$

$$n^{v_{\max}} = \frac{E}{E_{th}} \Rightarrow v_{\max} = \ln(E/E_{th})/\ln\langle n \rangle$$

The number of independent particles in the hadronic cascades compared to electromagnetic ones is smaller by E_{th}/ϵ and hence the intrinsic energy resolution will be worse at least by a factor $\sqrt[3]{(E_{th}/\epsilon)} \approx 6$. The average longitudinal energy deposition profiles are characterised by a sharp peak near the first interaction point (from π^0 s) followed by an exponential fall-off with scale λ . This is illustrated in Fig. 13. The maximum occurs at $t_{\max} \approx 0.2 \ln E + 0.7$ (E in GeV).

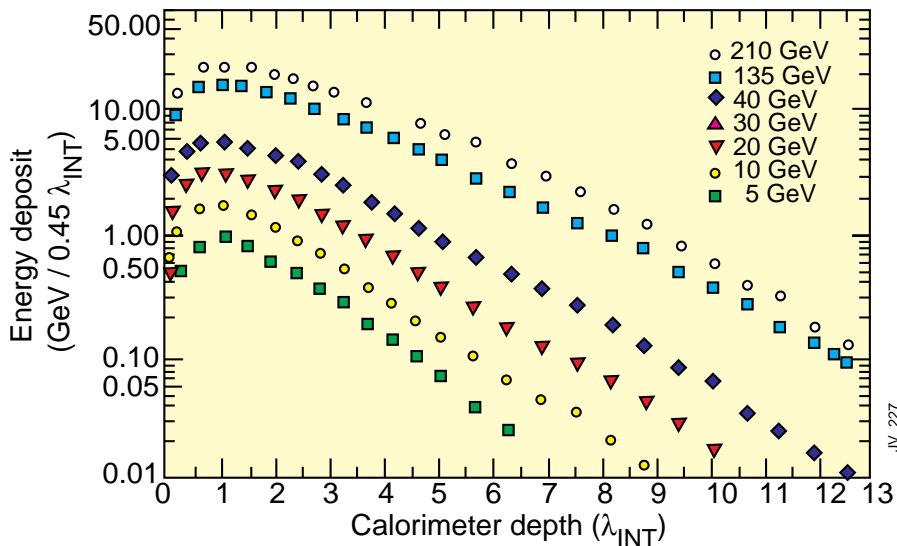


Figure 13: longitudinal profile of energy deposition for pion showers of different energies.

A parameterisation for the depth required for almost full containment (95%) is given by $L_{0.95}(\lambda) \approx t_{\max} + 2\lambda_{att}$ where $\lambda_{att} \approx \lambda E^{0.13}$. Fig. 13 shows that over 9λ are required to contain almost all the energy of high energy hadrons. However there is a considerable variation from one hadronic shower to another as illustrated in Fig. 14 [13]. The peaks arise from energy deposited locally by π^0 s produced in the interactions of charged hadrons. These interactions take place at differing depths from shower to shower. The energy carried by π^0 s also varies considerably from shower to shower.

4.2.2 The Lateral Development of the Hadronic Cascade

The secondary hadrons are produced typically with $\langle p_t \rangle \approx 300 \text{ MeV}$. This is comparable to the energy lost in 1λ in most materials. At shower maximum, where the mean energy of the particles is $E_{th} \approx 280 \text{ MeV}$, the radial extent will have a characteristic scale of $R_\pi \approx \lambda$. High energy hadronic showers show a pronounced core, caused by the π^0 component with a characteristic transverse scale of R_M , surrounded by an exponentially decreasing halo with scale λ . This is illustrated in Figure 12 for a lead/scintillating fibre calorimeter [14].

270 GeV Incident Pions in Copper

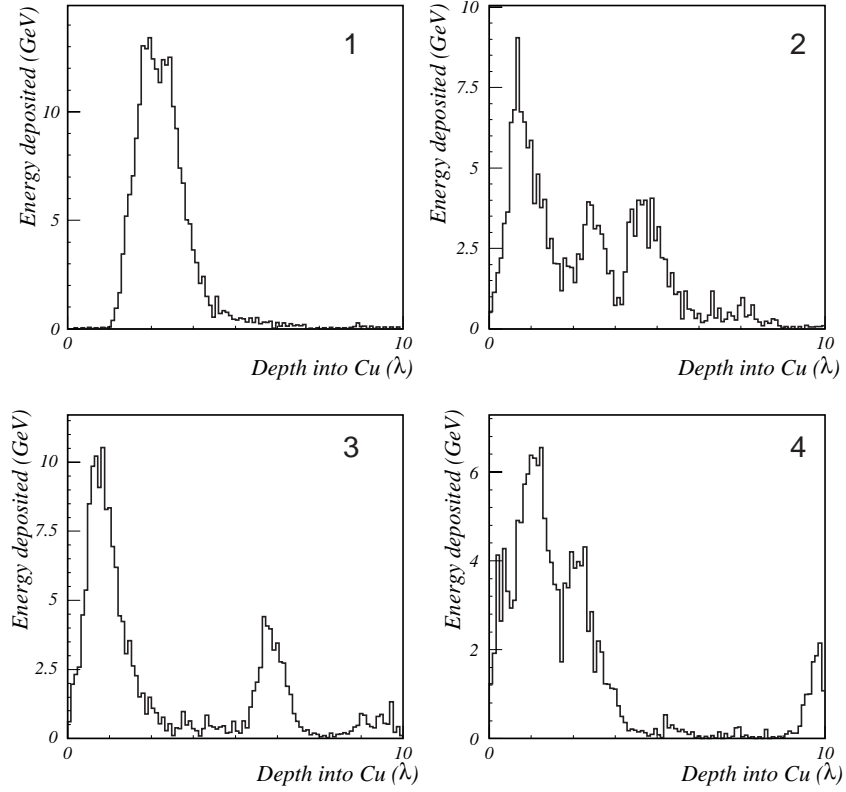


Figure 14: A simulation of the development of four representative pion showers in a block of copper.

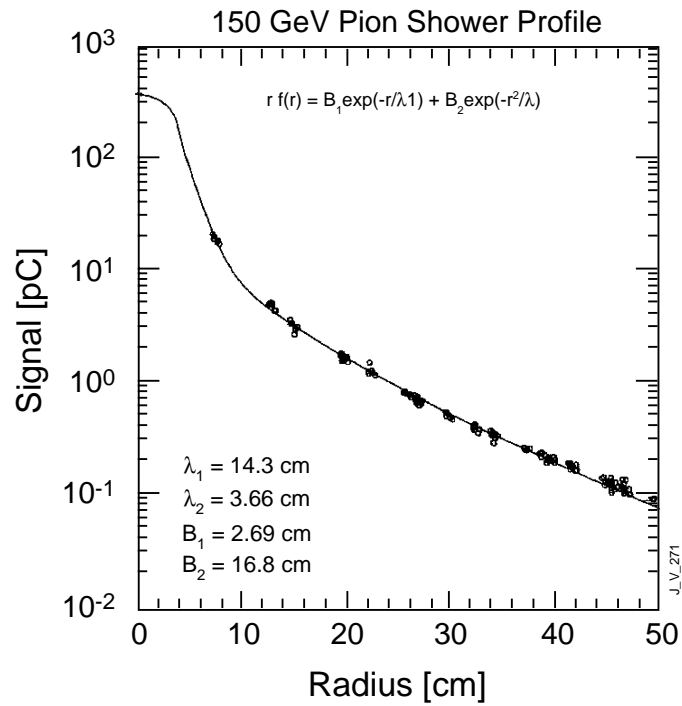


Figure 15: The lateral profile of energy deposition of pion showers.

5. ENERGY RESOLUTION

The energy resolution of calorimeters is usually parameterised as :

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \otimes \frac{b}{E} \otimes c$$

where the r.h.s. is the square root of the quadratic sum of the three terms.

The first term, with coefficient a , is the *stochastic or sampling* term and accounts for the statistical fluctuation in the number of primary and independent signal generating processes, or any further process that limits this number. An example of the latter is the conversion of light into photo-electrons by a photo-device.

The second term, with coefficient b , is the *noise* term and includes:

- the energy equivalent of the electronics noise and
- the fluctuation in energy carried by particles, other than the one(s) of interest, entering the measurement area. This is usually labeled pileup.

The last term, with coefficient c , is the *constant* term and accounts for:

- imperfect quality of construction of the calorimeter
- non-uniformity of signal generation and/or collection
- cell-to-cell inter-calibration error
- the fluctuation in the amount of energy leakage from the front or the rear (though somewhat increasing with energy) the volume used for the measurement of energy,
- the contribution from the fluctuation in the e.m. component in hadronic showers.

The tolerable size of the three terms depends on the energy range involved in the experiment. The above parametrisation allows the identification of the causes of resolution degradation. The quadratic summation implies that the three types of contributions are independent which may not always be the case.

5.1 Intrinsic Electromagnetic Energy Resolution

It is instructive to look at homogeneous calorimeters in which all the energy is deposited in the active medium. If the shower is fully contained then the intrinsic energy resolution is determined by the fluctuation in the number, n , of ions or photons produced. If W is the mean energy required to produce an electron-ion pair (or a photon) then $n = E/W$, and

$$\frac{\sigma}{E} = \frac{\sqrt{n}}{n} = \sqrt{\frac{W}{E}}$$

However the fluctuation is smaller as the total energy deposited (= incident energy) does not fluctuate. The improvement in resolution is characterised by the Fano factor, F , as

$$\frac{\sigma}{E} = \sqrt{F} \times \sqrt{\frac{W}{E}} = \sqrt{\frac{FW}{E}}$$

F is dependent on the nature of processes that lead to energy transfer in the detector including ones that do not lead to ionisation e.g. phonon excitations.

Consider calorimeters used for the spectroscopy of low energy (\approx MeV) gamma rays. The two commonly used detectors are inorganic scintillators (e.g. NaI) and semiconductor detectors (e.g. Ge). The energy resolution of the Ge detector is superior and is measured to be $\sigma \approx 180$ eV for photons carrying 100 keV. The above formula gives $\sigma = \sqrt{(FEW)} \approx 195$ eV where $F_{\text{Ge}}=0.13$ and $W=2.96$ eV. It should be noted that without the Fano factor $\sigma \approx 540$ eV!

Other phenomena may limit the number of signal generating events. Lead glass shower detectors are based on the detection of Cerenkov light, produced by the electrons and positrons with kinetic energies greater than ~ 0.7 MeV. This means that at most $1000 / 0.7 \sim 1400$ independent particles, per GeV of deposited energy, produce Cerenkov light. The resolution is then dominated by the fluctuation in this number and thus cannot be better than $(\sigma/E)_n \geq 3\% / \sqrt{E}$. This is further limited by photo-electron statistics as only about 1000 photo-electrons are generated when using photomultipliers to detect the scintillation light. This leads to an additional loss of resolution given by $(\sigma/E)_{pe} \approx 3\% / \sqrt{E}$.

5.2 Energy Resolution of Sampling Electromagnetic Calorimeters

When the very best energy resolution is not required, sampling calorimeters are employed. The shower energy is measured in active layers, often of low Z , sandwiched in between passive absorber layers of high Z materials. Only a fraction of the shower energy is dissipated in the active medium and the energy resolution is dominated by the fluctuation in this fraction. If the energy loss in an active layer is much smaller than that in the absorber layer then the number of independent charged particles crossing an active layer can be approximated by $n=E/\Delta E_{abs}$ where ΔE_{abs} is the energy lost by a minimum ionizing particle (m.i.p.) in the absorber layer.

Now $\Delta E_{abs} = t_{abs} \times (dE/dx)$ where t_{abs} is measured in units of X_0 . Hence

$$\frac{\sigma}{E} = \frac{\sqrt{n}}{n} \propto \frac{\sqrt{t_{abs}}}{\sqrt{E}}$$

For a fixed thickness of an active layer the energy resolution improves with decreasing absorber thickness. The above formula is not valid if the crossings between consecutive active layers are correlated, i.e. when t_{abs} is small. A generally valid formula is:

$$\frac{\sigma_s}{E} = \frac{5\%}{\sqrt{E}} (1 - f_{samp}) \Delta E_{cell}^{0.5(1-f_{samp})}$$

where ΔE_{cell} is the energy deposited in a unit sampling cell i.e. 1 active and 1 absorber layer. f_{samp} is labeled the *sampling fraction* and is the fraction of the total energy that is deposited in the active medium. As $f_{samp} \rightarrow 1$, $\sigma_s = 'a' \rightarrow 0$ (usually $a \neq 0$ due to imperfections in calorimeter systems) and as $f_{samp} \rightarrow 0$, $\sigma_s \propto \sqrt{\Delta E_{cell}} \propto \sqrt{\Delta E_{abs}}$.

The sampling fraction can be calculated as follows. If d is the thickness of active layer then

$$f_{samp} = 0.6 f_{mip} = 0.6 \frac{d \left(\frac{dE}{dx} \right)_{act}}{\left[d \left(\frac{dE}{dx} \right)_{act} + t_{abs} \left(\frac{dE}{dx} \right)_{abs} \right]}$$

$f_{mip} \approx 2/(12.75+2) \approx 13.5\%$ for a sampling calorimeter with 1cm Pb and 1cm scintillator plates

The fractional energy resolution as a function of $\sqrt{(d/f_{samp})}$ is shown in Figure 16 [5]. Clearly the energy resolution of gas calorimeters will be poor as the sampling fraction tends to be very low.

5.3 Energy Resolution of Hadronic Calorimeters

Hadronic calorimeters, because of the large depth required ($\approx 10\lambda$), are by necessity sampling calorimeters. The response of a sampling electromagnetic calorimeter can be expressed as

$$E_{vis} = e E$$

where E , E_{vis} are incident and visible energies respectively and $e = f_{samp}$, the electromagnetic

sampling fraction. Similarly the response of a hadronic sampling calorimeter is

$$E_{vis} = e E_{em} + \pi E_{ch} + n E_n + N E_{nucl}$$

where E_{em} , E_{ch} , E_n , E_{nucl} are respectively the energy deposited by electromagnetic component, charged hadrons, low energy neutrons and energy lost in breaking up nuclei. Each component has its own sampling fraction. N is normally very small but E_{nucl} can be large e.g. it is $\approx 40\%$ in Pb calorimeters. Hence the ratio of the response to electromagnetic and hadronic showers i.e. e/h is usually > 1 and the hadronic calorimeter is said to be *non-compensating*.

In hadronic calorimeters the fluctuation in the visible energy has two sources :

- sampling fluctuations as in the e.m. case which can be reduced by finer sampling and
- intrinsic fluctuation in the shower components (δE_{em} , δE_{ch} etc.) from shower to shower as seen in Fig. 14.

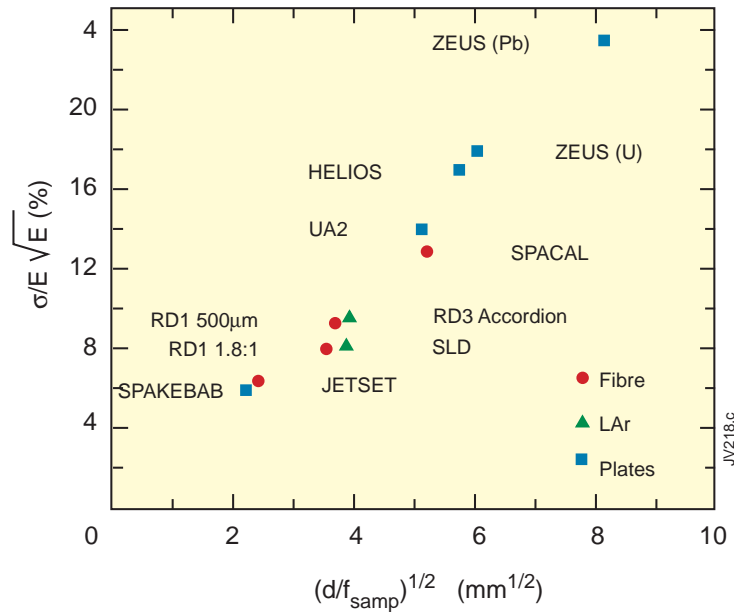


Figure 16: The fractional energy resolution of some calorimeters as a function of $\sqrt{(df_{samp})}$

Therefore the stochastic term is given by

$$a_h = a_{samp} \oplus a_{intr}$$

$$\left(\frac{\sigma}{E}\right)_{samp} = \frac{a_{samp}}{\sqrt{E}} \quad \text{where } a_{samp} \approx 10\% \sqrt{\Delta E_{cell}}$$

$$\left(\frac{\sigma}{E}\right)_{intr} = \frac{a_{intr}}{\sqrt{E}} + c$$

where c is the constant term which depends on e/h and vanishes for a compensating calorimeter.

5.3.1 The Intrinsic Hadronic Energy Resolution

There is a considerable event-to-event fluctuation in the e.m. component (F_0) of hadronic showers. A large event-to-event fluctuation in the neutral fraction is evident. Although at relatively low energy most of the e.m. component is produced in the first interaction, there is a rise in the fraction as the energy of the incident hadron increases, and hadrons further down the cascade

have enough energy to produce neutral pions. This can be seen in Fig. 17 [15] which shows the result of a simulation of pions of 20 GeV and 200 GeV incident on lead.

It usually turns out that the response to electrons and photons i.e. the e.m. component (labeled e) differs from that due to charged hadrons i.e. the non-e.m. component (labeled h). If E is the incident energy the response to electrons (E_e) and charged pions (E_π) can be written as :

$$E_e = e E, \quad E_\pi = [e F_0 + h (1 - F_0)] E \quad \text{leading to}$$

$$\frac{e}{\pi} = \frac{(e/h)}{[(e/h)F_0 + (1-F_0)]}$$

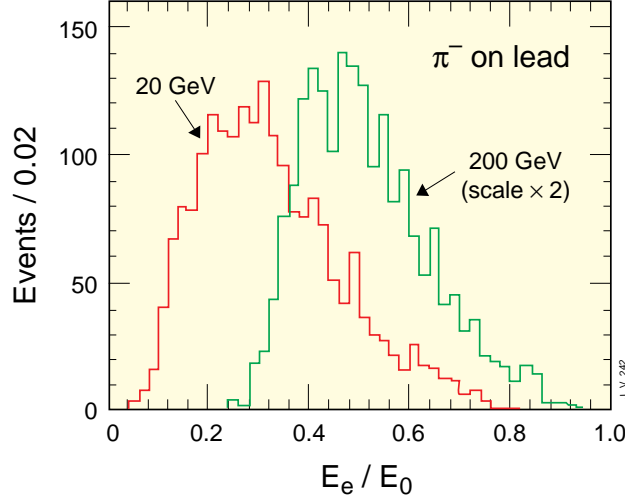


Figure 17: Distribution of e.m. energy fraction for charged pions incident on lead.

If $e/h = 1$ the calorimeter is said to be compensating.

Consider $dE_\pi = [(e - h) dF_0] E$. Then

$$\frac{dE}{E} = \frac{dF_0 [(e/h) - 1]}{[(e/h)F_0 + (1-F_0)]}$$

Hence the fractional error depends on e/h , F_0 and dF_0 . If $e/h=1$ then there is no contribution due to the fluctuation dF_0 . For example:

$$\frac{dF_0}{F_0} = \frac{df_0}{f_0} \sim \frac{1}{\sqrt{f_0 \langle n \rangle}}$$

i.e. for a 200 GeV hadron, $\langle n \rangle \approx 9$, $dF_0 \approx 0.6 \Rightarrow (dE/E)_{\text{comp}} \approx 3.5\%$.

$$\left. \frac{dE}{E} \right|_{\text{comp}} \sim \frac{1}{\sqrt{\ln E}} \quad \text{and} \quad \rightarrow 0 \quad \text{as} \quad E \rightarrow \infty \quad \text{since} \quad \langle n \rangle \propto \ln E$$

This aspect is illustrated by calorimeters using quartz fibres as active media. Charged particles traversing the fibres generate Cerenkov light which is guided to photomultipliers by the fibres themselves. Such a technique is employed by CMS for calorimetry in the very forward region ($3 < |\eta| < 5$) [11: Hadron Calorimeter TDR]. The aim is to measure the energies of, and tag, high energy jets from the WW fusion process. The signal in the calorimeter arises predominantly from the electromagnetic component as charged hadrons have a very high Cerenkov threshold when compared to that of electrons. Hence e/h is very large and the energy resolution at high energies will be dominated by the fluctuation in F_0 . The resolution should improve as $1/\ln E$ rather than as

$1/\sqrt{E}$ as illustrated in Fig. 18 which shows the measured energy resolution of the CMS copper/quartz fibre calorimeter. Also shown is the resolution after subtraction of the contribution from photostatistics. It should be noted that the photostatistics contribution is sizeable as only about 1 photoelectron per GeV is generated.

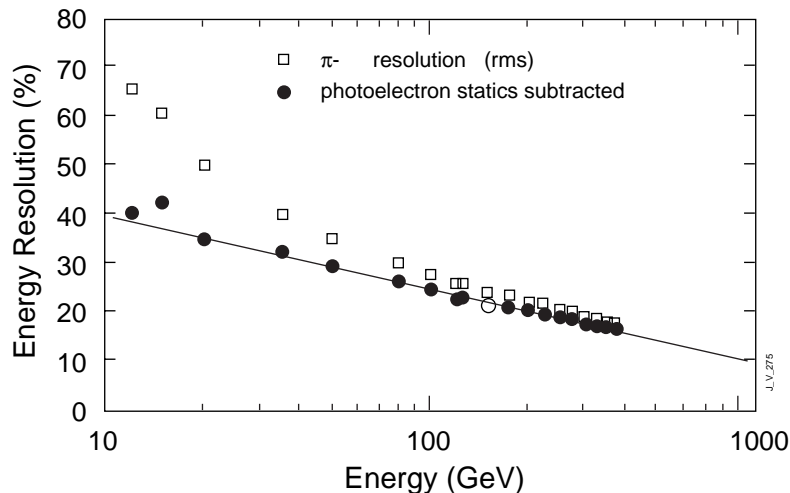


Figure 18: The measured pion energy resolution of a copper/quartz fibres calorimeter.

If $|e/h| \geq 10\%$ the performance of the calorimeter is compromised because of the fluctuation in the π^0 content of the cascades. This leads to:

- a non-Gaussian measured energy distribution for mono-energetic hadrons,
- an e/π ratio that is different from unity and that varies with energy,
- a non-linear response in energy to hadrons,
- an additional contribution to the relative energy resolution (σ/E),
- a σ/e that does not improve as $1/\sqrt{E}$.

These effects have been observed and are detailed in reference [12].

5.3.2 Compensation

The degree of compensation is expressed by the energy independent ratio e/h . The e/h ratio cannot be measured directly but can be inferred from the energy dependent e/π signal ratios. Two relations between the signal ratio $e/\pi(E)$ and e/h given by Groom [15] and Wigmans [16] are:

$$\frac{e}{\pi} = \frac{e/h}{1 + (e/h - 1)F_0}$$

$$F_0 = 1 - (E/0.76)^{-0.13} \quad D. \text{ Groom}$$

$$\text{or } F_0 = 0.11 \ln E \quad R. \text{ Wigmans}$$

It is instructive to see how the energy is dissipated by a hadron in a Pb absorber. The breakdown of the dissipated energy is as follows:

- 42% in breaking up nuclei and not rendered measurable (invisible)
- 43% by charged particles
- 12% by neutrons with kinetic energy ~ 1 MeV
- 3% by photons with an energy ~ 1 MeV.

The sizeable amount of invisible energy loss means that hadronic calorimeters tend to be under-compensating ($e/h > 1$).

Compensation can be achieved in three ways;

- boost the non-e.m. response using depleted uranium,
- suppress e.m. response
- boost the detectable response to low energy neutrons.

The ZEUS Collaboration [17] have found that achieving compensation for U/scintillator and Pb/scintillator calorimeters requires absorber/scintillator plate thickness ratios given by 1:1 and 4:1 respectively. They also used a technique of interleaved calorimeters to determine the intrinsic energy resolution of U and Pb calorimeters. This is accomplished by reading out odd and even scintillator layers separately. The results are as follows:

hadrons	Pb	$\sigma_{\text{samp}} = 41.2 \pm 0.9\% / \sqrt{E}$	$\sigma_{\text{intr}} = 13.4 \pm 4.7\% / \sqrt{E}$
	U	$\sigma_{\text{samp}} = 31.1 \pm 0.9\% / \sqrt{E}$	$\sigma_{\text{intr}} = 20.4 \pm 2.4\% / \sqrt{E}$
electrons	Pb	$\sigma_{\text{samp}} = 23.5 \pm 0.5\% / \sqrt{E}$	$\sigma_{\text{intr}} = 0.3 \pm 5.1\% / \sqrt{E}$
	U	$\sigma_{\text{samp}} = 16.5 \pm 0.5\% / \sqrt{E}$	$\sigma_{\text{intr}} = 2.2 \pm 4.8\% / \sqrt{E}$

The intrinsic fluctuations in a compensating Pb calorimeter are smaller than those for a U one. However the sampling has to be much coarser for Pb calorimeter leading to a much poorer e.m. energy resolution. ZEUS therefore chose U as the absorber material. It can also be seen that for compensating Pb and U calorimeters the energy resolution is dominated by sampling fluctuations and is given by

$$\sigma_{\text{samp}} = \frac{11.5\% \sqrt{\Delta E_{\text{cell}}(\text{MeV})}}{\sqrt{E(\text{GeV})}}$$

The sampling fluctuations for hadrons are larger than those for e.m. showers by a factor of 2. From the above it is evident that very good e.m. energy resolution is incompatible with $e/h=1$.

5.4 Jet Energy Resolution

Hadronic calorimeters are primarily used to measure the energies of jets and hence the most important quantities that characterize them are:

- jet energy resolution and energy linearity,
- missing transverse energy resolution.

In hadron-collider experiments the energy of jets is often estimated by adding the energy contained in a cone, with half angle ΔR , where $\Delta R = \sqrt{(\Delta\eta^2 + \Delta\phi^2)}$ in pseudorapidity (η) and ϕ space, and whose axis is centred on a seed cell with an energy above a pre-defined threshold. The jet energy resolution is limited by effects from:

- algorithms used to define jets (energy is dependent on cone radius, lateral segmentation of cells etc.),
- the fluctuation in the particle content of jets due to differing fragmentation from one jet to another,
- the fluctuation in the underlying event,
- the fluctuation in energy pileup in high luminosity hadron colliders
- magnetic field.

A figure of merit of a hadron calorimeter is di-jet mass resolution. For the purposes of measuring the jet energy resolution low p_t di-jets ($50 < p_t < 60$ GeV), high p_t di-jets ($500 < p_t < 600$ GeV) and

high mass di-jets ($3 < m_Z < 4$ TeV) at the LHC can be used [18]. The mass resolution for the three categories v/s cone size, ΔR , is shown in Fig. 19a for a perfect calorimeter with no underlying event. The mass resolution improves with increasing cone size. However when running at high luminosity there are ≈ 30 minimum bias events which accompany the event of interest. The fractional mass resolution as a function of cone size is plotted in Fig. 19b. Also plotted is the case when energy is estimated using only the towers above a certain energy threshold (low p_t events – $E_t > 0.3$ GeV, others $E_t > 1$ GeV).

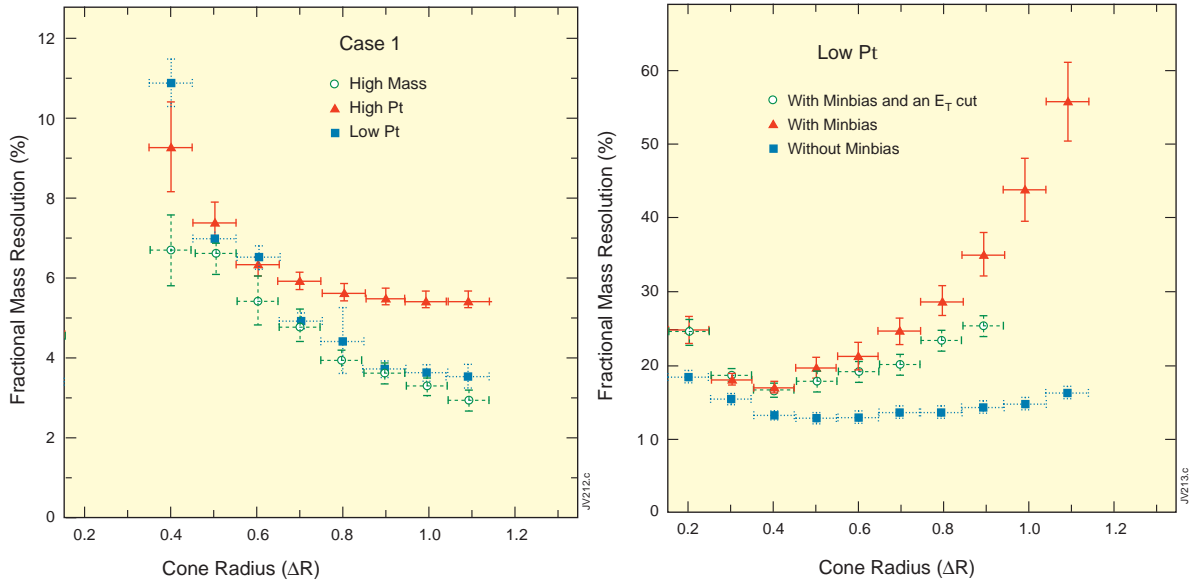


Figure 19: The fractional jet-jet mass resolution as function of cone radius a) perfect calorimeter, b) with 30 minimum bias events overlapped with the event of interest.

From the above it can be seen that in hadronic colliders the uncertainties caused by jet fragmentation (fluctuation of energy inside a pre-defined cone size) and underlying event are very significant in comparison with instrumental effects such as energy resolution, magnetic field, threshold E_T etc.). Hence the mass resolution finally depends on the physics itself. At high luminosities the resolution is degraded if the cone-size is too small (some signal energy is excluded) or if the cone size is too large (significant pileup energy is included). In order to obtain the best mass resolution the cone size has to be optimised for each process and instantaneous luminosity.

The mass resolution due to the angular error, $d\theta$, in defining the jet axis is given by:

$$\frac{dM}{M} = \frac{p_T}{M} d\theta$$

Only highly boosted and low mass di-jets (e.g. boosted Zs from $H \rightarrow ZZ$) will have a significant contribution from the angular error. This is illustrated in Fig. 20 [18].

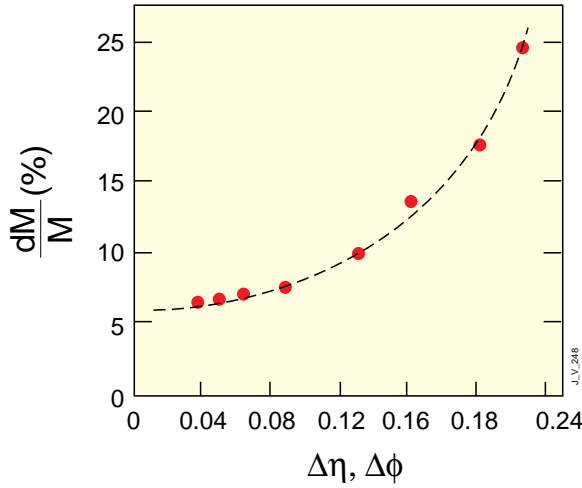


Figure 20: The fractional jet-jet mass resolution as a function of the tower size.

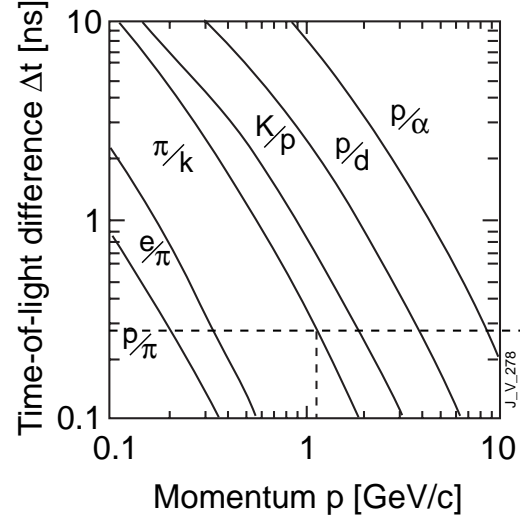


Figure 21: The difference in the time of flight as a function of time for different pairs of particles.

In experiments at e^+e^- machines the jet energy resolution can be improved by using the centre of mass energy to constrain the energies of jets if the jet directions are measured relatively precisely.

6 IDENTIFICATION OF PARTICLES

6.1 Identification of Particles using Time of Flight

At low momenta the difference in the time taken by different particles to traverse a certain distance can be used to distinguish between them. For a particle, with momentum p and mass m_1 , the time taken to traverse a distance L is

$$t_1 = \frac{L}{\beta_1 c}$$

Hence the difference in the time of flight for two particles with masses m_1 and m_2 is

$$\Delta t = \frac{L}{c} \left(\frac{1}{\beta_1} - \frac{1}{\beta_2} \right) = \frac{L}{c} \left(\sqrt{1 + m_1^2 c^2 / p^2} - \sqrt{1 + m_2^2 c^2 / p^2} \right) \approx \frac{Lc}{2p^2} (m_1^2 - m_2^2)$$

The difference Δt is illustrated in Fig. 21 for different pairs of particles as a function of momentum and for a path length of 1m. Plastic scintillators have usually been used to measure the time to a precision of about 300 ps. This enables π/K separation up to ≈ 1 GeV. ALICE [19], the heavy ion experiment on the LHC, will use parallel plate chambers which give a timing precision of ≈ 100 ps enabling π/K separation up to ≈ 2 GeV/c.

6.2 Identification of Particles using Specific Energy Loss

The energy loss of charged particles traversing a medium is given by Equation 1. The ‘truncated’ mean energy loss, measured in the OPAL jet-chamber [20], for different species of particles is shown in Fig. 22. The gas used is 80%/20% Ar/CH₄ at NTP. A large number of samples is used with each sample corresponding to a path-length of ≈ 1 cm. The energy loss measured in each sample has a considerable fluctuation as can be seen from the distribution for minimum ionising pions ($400 \text{ MeV}/c < p < 800 \text{ MeV}/c$) shown in Fig. 23. The distribution is known after Landau

who first calculated it. There is a long tail to the distribution and OPAL reject 30% highest charge samples when determining the mean energy loss per track.

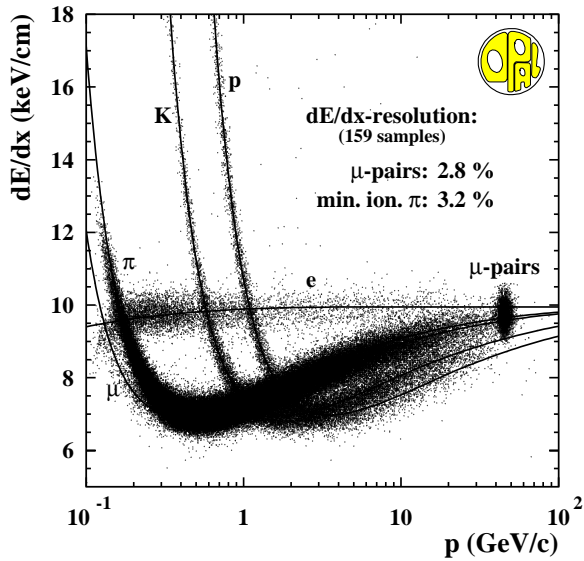


Figure 22: The truncated mean energy loss as a function of momentum for different species of particles.

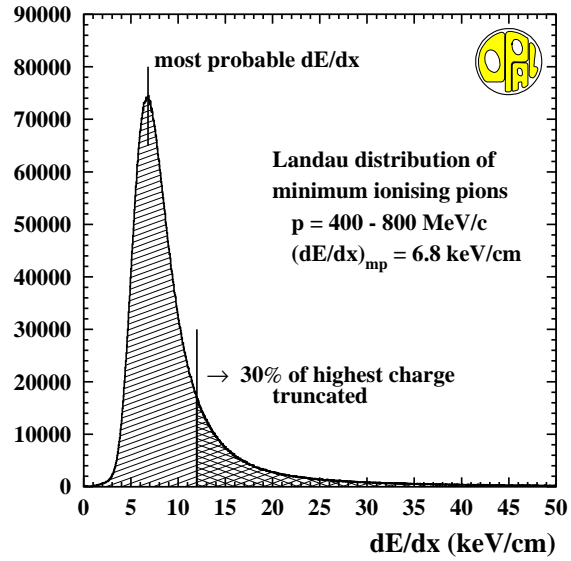


Figure 23: The energy loss distribution for minimum ionising pions.

The difference in the energy loss at a given momentum by various particles can be used to distinguish between them. Since

$$p = m\beta\gamma \quad \text{and} \quad \frac{dE}{dx} \propto \frac{1}{\beta^2} \ln(\beta^2\gamma^2)$$

a simultaneous measurement of p and dE/dx will yield the mass of the particle. The average energy loss for an electron, muon, pion, kaon and proton in 80%/20% Ar/CH₄ gas mixture at NTP is shown in Fig. 22. It can be seen that separation of 10 GeV/c pions and kaons, at 2σ level, requires a $\sigma(dE/dx)$ of $\approx 3\%$. The separation power in OPAL is summarized in Fig. 24.

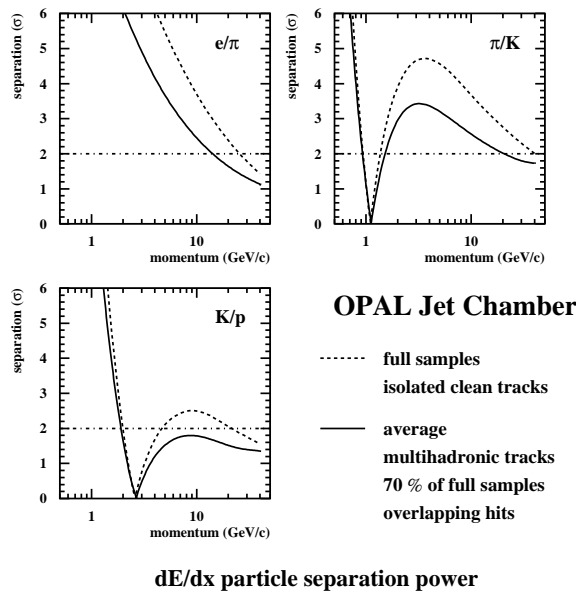


Figure 24: The separation power in OPAL

6.3 Identification of Particles using Cherenkov Radiation

Cherenkov radiation is emitted when a charged particle traverses a dielectric medium, with a refractive index n , at a velocity which is higher than the velocity of light in that medium i.e. if

$$v_{particle} > \frac{c}{n} \quad \text{or} \quad \beta > \beta_{thr} = \frac{1}{n}$$

At each point of emission 'Huygen's wavelets' are generated which add constructively along a cone with half-angle given by the Cherenkov angle (see Fig. 25). The outer surface of the cone constitutes a wavefront. The process is similar to the generation of a sonic shock wave by supersonic aircraft.

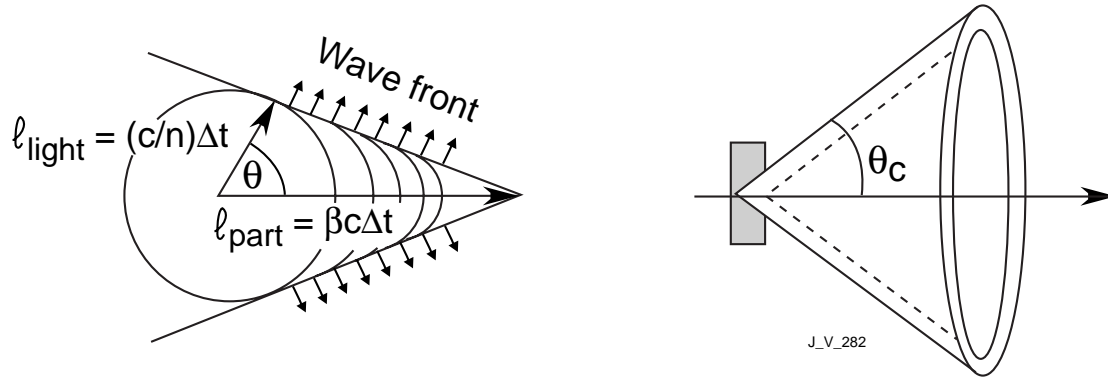


Figure 25: The construction of the Cherenkov light wave-front and the formation of a ring image.

The Cherenkov angle is given by

$$\cos \theta_c = \frac{1}{\beta n} \quad \text{with} \quad n = n(\lambda) \geq 1 \quad \text{and} \quad \theta_{max} = \cos^{-1}\left(\frac{1}{n}\right)$$

The number of photons emitted per unit length of radiator and unit wavelength interval are

$$\frac{d^2 N}{dx d\lambda} = 2\pi z^2 \alpha \frac{1}{\lambda^2} \sin^2 \theta_c$$

$$\text{or} \quad \frac{d^2 N}{dx dE} = \frac{\alpha}{\hbar c} \sin^2 \theta_c \approx 365 \sin^2 \theta_c \quad eV^{-1} cm^{-1}$$

with the characteristic $1/\lambda^2$ dependence. The energy loss through Cherenkov radiation is small ($\approx 1\%$) when compared with that due to ionization. Values of the above parameters for a few radiators are given in Table. 2.

The number of photoelectrons (pe) detected in a photo-device is given by

$$N_{pe} = 370 L \int \epsilon_{col}(E) \epsilon_{det}(E) \sin^2 \theta_c(E) dE$$

where L is the path length in the radiator, ϵ_{col} is the efficiency for collecting the Cv light, ϵ_{det} is the quantum efficiency of photo-conversion of the photodevice. Typically for a photomultiplier, with sensitivity in the range 350-550 nm, $N_\gamma \approx 450 \sin^2 \theta_c \text{ cm}^{-1}$.

Table 2: Parameters for some selected Cherenkov radiators.

Medium	n-1	θ_{\max}	$\pi_{\text{thr}}(p)$ GeV/c	N_{γ} (eV ⁻¹ cm ⁻¹)
Air	1.000283	1.36°	5.9	0.21
Isobutane	1.00217	3.77°	2.12	0.94
Aerogel	1.0065	6,51°	1.23	4.7
Aerogel	1.055	18.6°	0.42	37.1
Water	1.33	41.2°	0.16	160.8
Quartz	1.46	46.7°	0.13	196.4

Two types of Cherenkov detectors are used for particle identification. Threshold Cherenkov detectors use the existence of a threshold for radiation to make a simple yes/no decision based on whether a particle is above/below a threshold velocity ($\beta=1/n$). Ring-Imaging Cherenkov (RICH) detectors can use the dependence of the Cherenkov cone half-angle on the velocity to test a given hypothesis for the mass of the particle with known momentum.

6.3.1 Threshold Cherenkov Detectors

The number of photons emitted depends on the velocity of the particle and is given by

$$N_{\gamma} \propto \sin^2 \theta_C = 1 - \frac{1}{\beta^2 n^2} = 1 - \frac{1}{n^2} \left(1 + \frac{m^2}{p^2} \right)$$

An example of the use of threshold Cherenkov counters comes from BaBar experiment at SLAC [21]. Two aerogel radiators are used: A1 with $n=1.055$ and A2 with $n=1.0065$ leading to the following conditions (Fig. 26) :

- $p > 0.4$ GeV/c, π in A1 give light,
- $p > 1.2$ GeV/c, π in A1 and A2 give light,
- $p > 1.4$ GeV/c, K in A1 give light,
- $p > 4.2$ GeV/c, K in A1 and A2 give light.

Hence π/k separation can be obtained in the range below 4.2 GeV/c which is adequate for the study of CP violation in BaBar.

6.3.2 Ring Imaging Cherenkov (RICH) Detectors

RICH detectors determine the identity of particles by measuring the Cherenkov angle, θ_C , once the momentum has been measured precisely. The principle of operation is illustrated in Fig. 27 [22]. All photons, emitted at the same angle, are focused by a spherical mirror, placed at radius R_M from the point of origin of the particle, to a spherical detecting plane at a radius $0.5 R_M$. The detecting plane will see a ring of photon impacts whose radius can be measured once the centre is known from the tracking system. In realistic detectors this simple concept is modified by magnetic field effects and the requirement of placing the detectors outside the path of the particles.

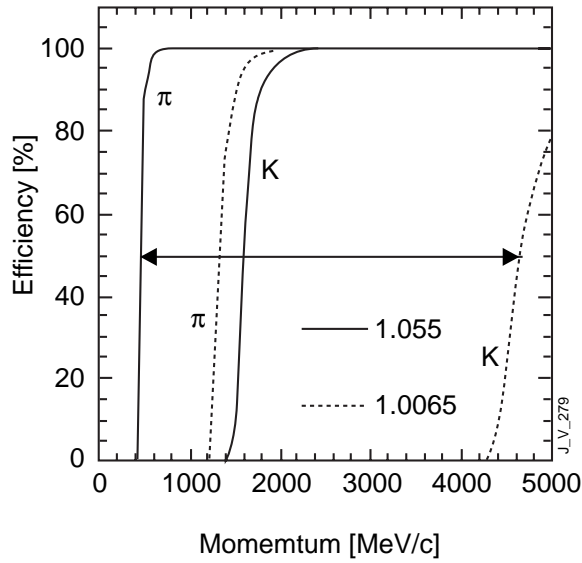


Figure 26: The pion and kaon thresholds for the two aerogel radiators used in BaBar.

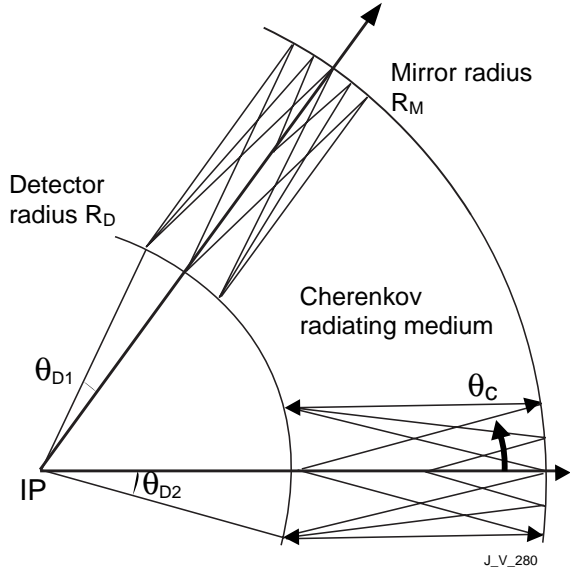


Figure 27: The principle of operation of ring-imaging Cherenkov detectors.

The angle θ_c is given by:

$$\theta_c = \cos^{-1}\left(\frac{1}{\beta n}\right) = \cos^{-1}\left(\frac{E}{pc} \frac{1}{n}\right) = \cos^{-1}\left(\frac{\sqrt{p^2 + m^2}}{pc} \frac{1}{n}\right)$$

The error in the measurement of the angle is minimized by minimizing the error on the localization of a photo-conversion, σ_θ , and maximizing the number of photo-electrons, N_{pe} . Two particles with masses m_1 and m_2 can be distinguished by n_σ up to a momentum p given by

$$p = \frac{1}{\sqrt{n_\sigma}} \sqrt{\frac{(m_2^2 - m_1^2) \sqrt{N_{pe}}}{2 \tan \theta \times \sigma_\theta^{pe}}}$$

As an example π/K can be separated up to 75 GeV/c at a 3σ level for $N=20$ pe, $\sigma_\theta=1$ mrad and $\theta=31$ mrad (CF_4).

The LHCb experiment [23] will use two RICH detectors to provide π/K separation in the momentum range from 1 to ≈ 100 GeV/c. The first RICH detector (Fig. 28) is a combined gas (C_4F_{10})-aerogel device and the second one is a gas device. The Cherenkov light will be detected by recently developed hybrid photo-detectors (HPD), sensitive to visible and near-UV light using small silicon-pad pixels to give unambiguous 2-D space points. The response for many triggers is shown in Fig. 29: there are ≈ 4 photoelectrons/event on the air ring (HPD1) and ≈ 1 photoelectron/event on the aerogel ring (HPD2-7).

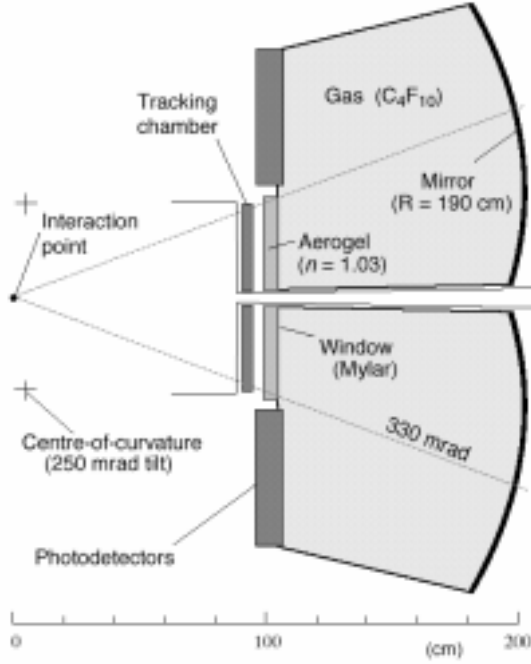


Figure 28: A combined gas(C_4F_{10})/aerogel RICH HPDs.device for LHCb.

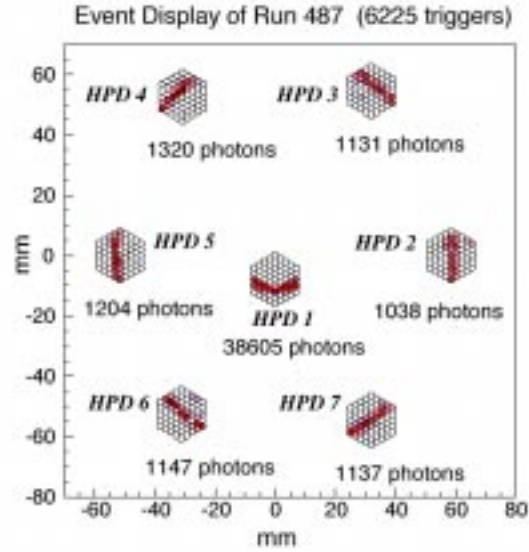


Figure 29: The Cherenkov light is detected by The response for many triggers is shown.

6.4 Identification of Electrons using Transition Radiation Detectors

The existence of transition radiation was predicted by Ginzburg and Franck in 1946. Transition radiation is emitted when a charged particle moves from a medium of refractive index n_1 to a medium of a different index n_2 . This may be thought of as an apparent acceleration. The changes in the medium are achieved by using thin foils of material like polyethelene and air.

Consider

$$p = \gamma m v \Rightarrow m = \frac{1}{\beta c \gamma} p$$

$$\therefore \left(\frac{\Delta m}{m} \right)^2 = \frac{1}{\beta^2 c^2} \left(\frac{\Delta \gamma}{\gamma} \right)^2 + \left(\frac{\Delta p}{p} \right)^2$$

If the momentum is precisely measured (i.e. $\Delta p/p$ is small) then the mass resolution at high momenta is $\propto \gamma$. The radiated energy/boundary to vacuum is given by:

$$W = \frac{1}{3} \alpha \hbar \omega_p \text{ i.e. } W \propto \gamma$$

where $\hbar \omega_p$ (≈ 20 eV for polyethelene) is the plasma frequency

The X-rays of transition radiation are emitted at a small angle w.r.t. the charged track ($\theta=1/\gamma$). The energy of the photons and the number of photons per boundary are given by

$$E_{ph} = \frac{\gamma}{3} \hbar \omega_p \text{ and } N_{ph} \approx \frac{W}{\hbar \omega_p} \propto \alpha \approx \frac{1}{137}!$$

Hence many transitions are needed. The detector therefore consists of a stack of many thin foils with a high Z detecting gas for effective photo-conversion of X-rays. The particle must traverse a minimum thickness to efficiently emit transition radiation, This is $\approx 20 \mu\text{m}$ for polyethelene.

6.4.1 ATLAS Transition Radiation Tracker

The ATLAS TRT [10. Inner Tracking TDR] comprises straw tube proportional chambers embedded in polyethylene fibres (see Fig. 30). Standard radiators for TR are made out of 15-20 μm thick polypropylene foils with a regular 200-300 μm spacing. Such regular radiators provide the highest radiation yield because the thickness and spacing can be optimised for each detector concerned. The ATLAS geometry however does not allow use of foil radiators. Foam radiators made out of polyethylene have been found to be the best but not as good as foil radiators as the variation of wall thickness and spacing are large. Furthermore it was found that properly oriented polyethylene fibres are almost as performant.

The gas in the straw tubes has to be an efficient X-ray absorber and hence dense gases such as Xe are employed. In order to attain high rate capability the gas has to be fast and ATLAS use CF_4 . In order to ensure stable operation at high gains a quenching gas such as CO_2 is needed. The gas chosen by ATLAS has the following composition: 70%:20%:10%/ Xe: CF_4 : CO_2 .

The probability to observe in a single straw an energy deposit from 200 GeV electrons above a given threshold is shown in Fig. 31 as a function of the this threshold for a variety of radiators (and no radiator). The best radiators are those yielding the highest probability/straw for an energy threshold of $\approx 6-7$ keV. The probability to exceed such a threshold is a factor 3-4 higher with than without a radiator. The difference between the probability for electrons and pions to deposit large amounts of energy in straws is shown in Fig. 32. In ATLAS the average number of TR hits (signaled by large energy deposits) for 30 GeV electrons is ≈ 6 per track compared with ≈ 1 for pions of the same energy.

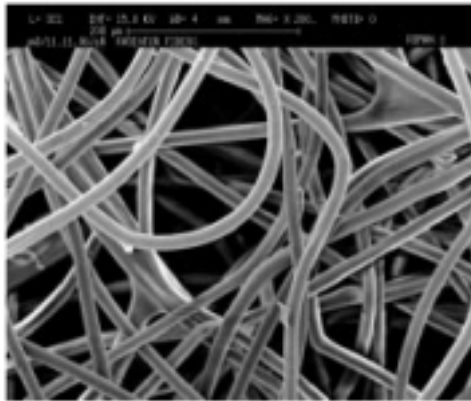


Figure 30: The detail of the polyethelene fibre radiator.

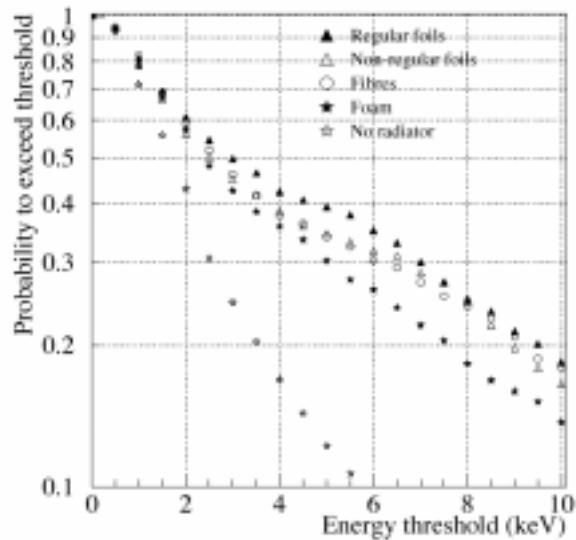


Figure 31: The probability of 200 GeV electrons to exceed a given threshold with and without radiators.

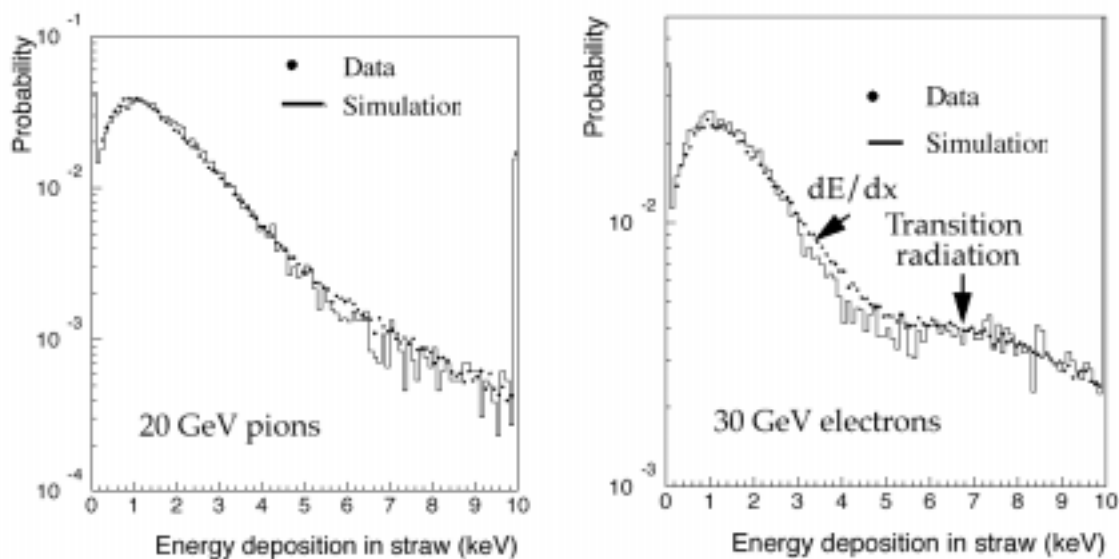


Figure 32: The difference between the distribution of energy deposits in the ATLAS TRT for 20 GeV pions and 30 GeV electrons showing clearly the onset of transition radiation for electrons.

6.5 Identification of b-jets

There are two methods for tagging b-jets. Charged leptons with relatively high momentum and a large momentum transverse to the jet axis arise mainly from semi-leptonic decays of b-hadrons. Hadronic decays of b-quarks can be enriched by taking advantage of the relatively long (≈ 1.5 ps) lifetimes of bottom hadrons. The long lifetime leads to secondary vertices which are separated from the primary vertex. Hence one looks for:

- relatively high transverse momentum electrons or muons within a jet arising from the semi-leptonic decay of a b-quark.
- one or more charged tracks within a jet with a significant impact parameter (defined to be the distance of closest approach of the track from the primary vertex)
- a secondary vertex consistent with the flight path of a B-meson (see Fig. 33).

The latter two methods require measuring layers close to the interaction vertex. Both CMS and ATLAS have several layers of pixel detectors. The precision with which the impact parameter can be measured is determined by:

- the closeness of the first measuring layer from the interaction vertex,
- the number of measurements close to the interaction vertex,
- the spatial resolution of the measured points and
- the amount of material in these layers leading the degradation in the significance of impact parameter due to multiple scattering.

Some of these points are illustrated in Fig. 34 that presents the results of a simulation of the CMS inner tracker. The points plotted are for two different radii of the first pixel layer (4 cm or 7.7 cm), differing number of tracks with different significance of the impact parameter. The estimated impact parameter resolution for 10 GeV tracks in CMS is around 15 (>20) μm for the first pixel layer at 4 (7.7) cm.

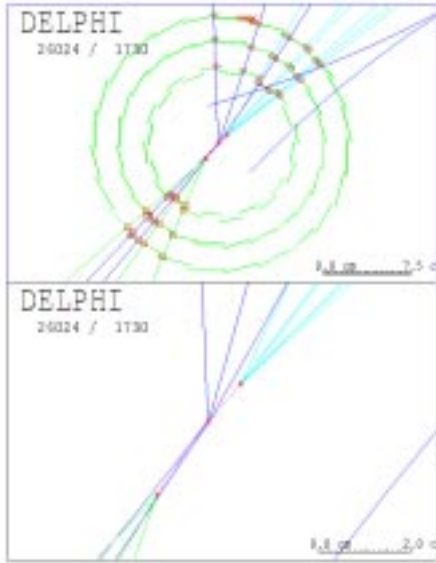


Figure 33: A DELPHI (LEP) event with two reconstructed B-hadron vertices.

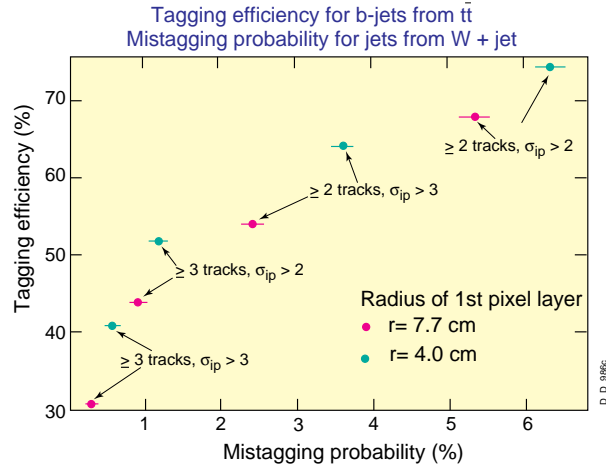


Figure 34: Tagging efficiency for b-jets and mistagging probability in CMS for various conditions.

6.6 Particle Identification Using Calorimeters

Below we consider some ways in which calorimeters can be used to identify isolated electrons and photons from hadrons and jets.

6.6.1 Isolated electromagnetic shower-jet separation

The largest source of electromagnetic showers is from the fragments of jets, especially π^0 s. A leading π^0 taking most of the jet energy can fake an isolated photon. There are large uncertainties in jet production and fragmentation. Furthermore the ratio of production of di-jets to irreducible di-photon background is $\approx 2 \cdot 10^6$ and γ -jet/irreducible $\gamma\gamma$ is ≈ 800 . Hence a rejection of ≈ 5000 against jets is needed. A certain rejection factor (≈ 20) can be obtained by simply asking for e.m. showers with a transverse energy greater than some threshold with the energy measured in a small region.

Jets can be distinguished from single electromagnetic showers by

- demanding an energy smaller than some threshold in the hadronic compartment behind the electromagnetic one
- using isolation cuts
- demanding a lateral profile of energy deposition in the ECAL consistent with that from an electromagnetic shower.

Using these criteria ATLAS [10: LAr TDR] estimates that the rejection factor against jets can be ≈ 1500 for a photon efficiency of 90%. This is illustrated in Fig. 35 where the effect of various cuts is shown: a) the energy (E_T^{had}) in the hadron calorimeter compartment behind the e.m. one of size $\Delta\eta \times \Delta\phi = 0.2 \times 0.2$ should be less than 0.5 GeV, b) e.m. isolation (R_{isol})— more than 90% of the energy is contained in the central 3×5 e.m. cells compared with that in central 7×7 e.m. cells, c) lateral shower profile (R_{lateral})— look for an e.m. core such that the central 4 towers contain more than 65% of the shower energy, d) shower width in η (σ_η). The distribution for jets is shown as dashed histogram whereas the full histograms depict single photons.

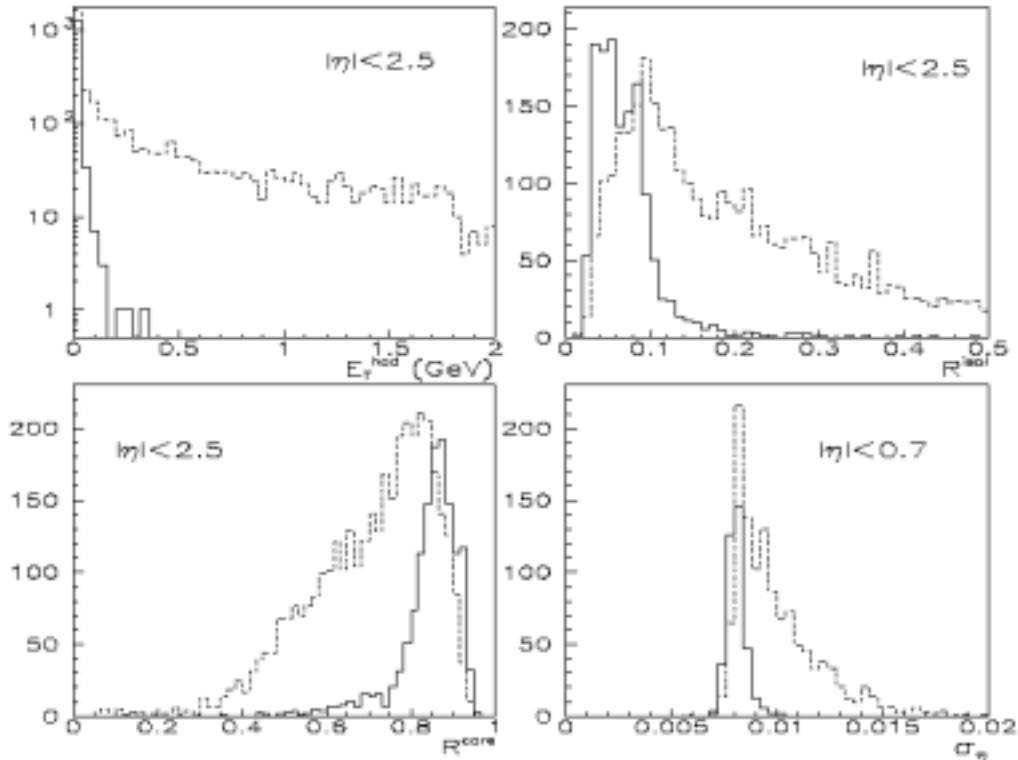


Figure 35: The distributions used to cut against jets. Solid histogram is for photons and the dashed one for jets. See text for explanation.

6.6.2 Photon – π^0 separation

After the application of the above criteria only jets resulting in leading π^0 s can fake genuine single photons. Further rejection can only be achieved by the recognition of two e.m. showers close to each other. CMS [11: ECAL TDR] uses the fine lateral granularity ($\approx 2.2\text{cm} \times 2.2\text{cm}$) of their crystals and a neural network algorithm that compares the energy deposited in each of the 9 crystals in a 3×3 crystal array with that expected from a single photon. Variables are constructed from the 9 energies, x and y position of impact and a pair measuring the shower width. The fraction of π^0 s rejected is shown in Figure 36.

The narrowness of the e.m. shower in the early part can be used to reject events consisting of two close-by e.m. showers. Planes of fine pitch orthogonal strips after a pre-shower, placed at a depth of $\approx 2.5 X_0$, can also be used to distinguish π^0 s from single photons. Results using 2mm pitch strips are shown in Fig. 36.

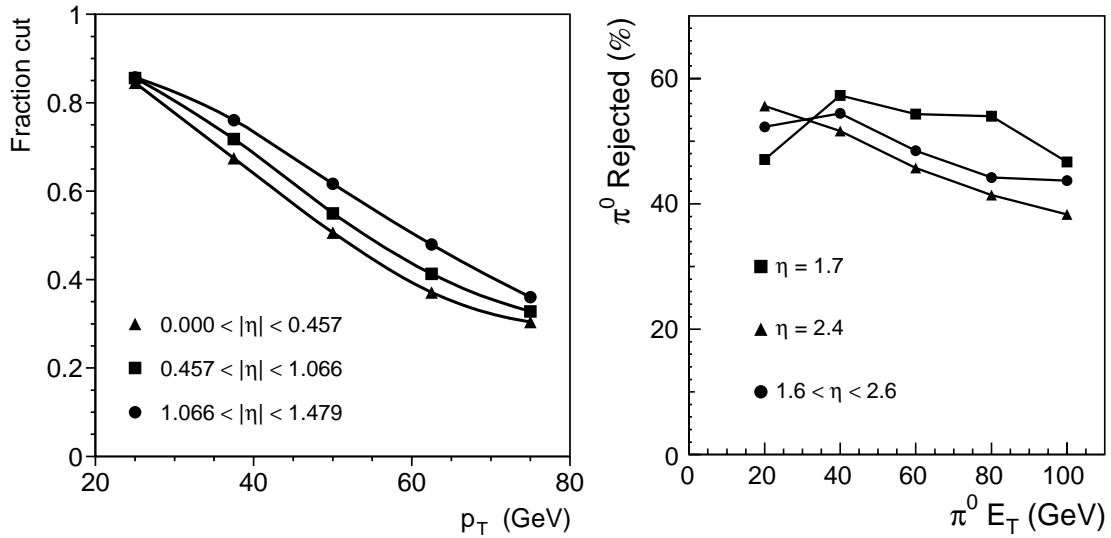


Figure 36: lhs) Fraction of pizeros rejected using lateral shape of energy deposit as a function of p_T rhs) Variation of pizero rejection as a function of η using two planes of orthogonally oriented 2mm pitch Si strips after 2 and 3 X_0 .

6.6.3 Electron-hadron Separation

A high energy pion faking an electron leads to the contamination of signals using prompt electrons. At LHC in order to bring down the rate of fake electrons from this source to a factor ≈ 10 below that from the genuine sources (e.g. $b \rightarrow e X$, $W \rightarrow e \nu$ etc.) an $e-\pi$ separation of ≥ 1000 is required for $p_T \geq 10$ GeV/c.

The electron-hadron separation is usually based on the difference in the longitudinal and lateral development of showers initiated by electrons and charged hadrons. One or more of the following can be used to achieve the desired pion rejection power when detecting electrons :

- a preshower detector between $\approx 1.5 - 4 X_0$
- lateral segmentation
- longitudinal segmentation including a hadron calorimeter
- energy - momentum matching

The ultimate rejection power is limited by the charge exchange process or the first hadronic interaction, which results in one or several π^0 's taking most of the energy of the incoming hadron. The shower from such hadrons then looks like an e.m. shower. Therefore sampling of showers early in their longitudinal development is important.

The separation power for single particle, using (i - iii) is shown in Fig. 37 [24]. The structure of the calorimeter consisted of :

- towers of a lateral size of $\sim 11 \times 11$ cm (effective $X_0 \approx 8$ mm),
- 8-fold longitudinal segmentation, the first four samplings (2mm U / 2.5 mm TMP) with thickness of 3, 6, 10, 7 X_0 leading to a total of 1λ , the next two (5mm U/ 2.5mm TMP) each with thickness of 0.7λ and the last two (5cm Fe/ 1cm scintillator) each with thickness of 2.5λ .
- a position detector placed at a depth of 3 X_0 . The rejection power, as a function of energy, using (ii), (iii) and (iv) individually and then all combined is shown.

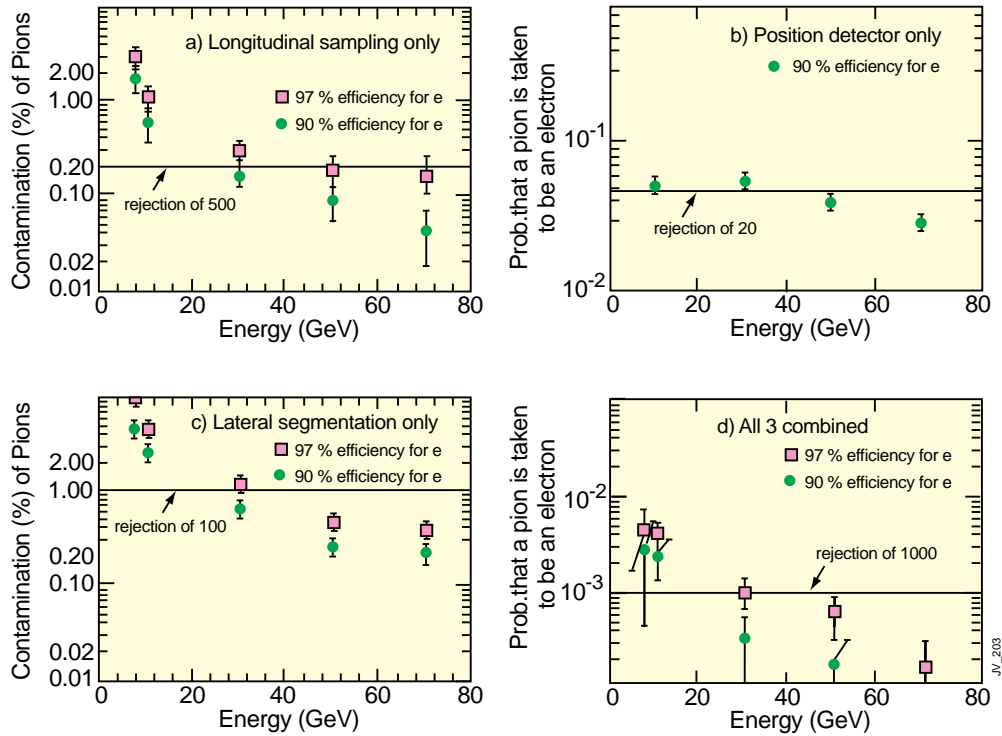


Figure 37: The probability that a single pion is taken to be an electron using a) longitudinal profile only, b) preshower detector only, c) lateral profile only and d) all three combined.

6.7 Identification of Muons

Muons are identified by their penetration power through the material of calorimeters which absorb the electrons, photons and hadrons. The depth required to absorb the hadrons is shown in Fig. 38 for pions of various energies. It can be seen that about two metres equivalent of iron is needed to absorb most of the energy of the hadrons. This corresponds to over 10λ of material. Insufficient depth of material can allow debris from hadronic showers to emerge and hence cause false identification of a hadron as a muon. The added confusion can also lead to difficulty in matching muon-tracks in jets and increase in trigger rate. Two scenarios are illustrated in Fig. 39. The study was carried out in H2 beamline at CERN using a 3T magnet followed by slabs of magnetized iron. A calorimeter is installed inside the magnet. One shows a hadron starting a shower late and contaminating the first muon station and the second with a secondary muon penetrating the muon system.

The identification and the measurement of the momentum of muons is accomplished by tracking in magnetic field. Two configurations are possible, tracking in air-filled magnetic field or in magnetized iron.

Extra material is usually required when muons are tracked through an air-filled field in order to decrease hadronic punchthrough from the calorimeters. However the magnetic field can help by sweeping the soft debris that may 'punch-through'.

Muons of energies above a few hundred GeV generate their own background when traversing magnetized iron. The critical energy of muons in iron is 350 GeV and hard bremsstrahlung (sometimes labeled catastrophic energy loss) can spoil muon tracking. A simulation of a 1 TeV muon traversing iron is shown in Fig 40. When tracking in iron several muon stations are required separated by a sufficient thickness of iron so as to kill the e.m. shower before the following station is reached.

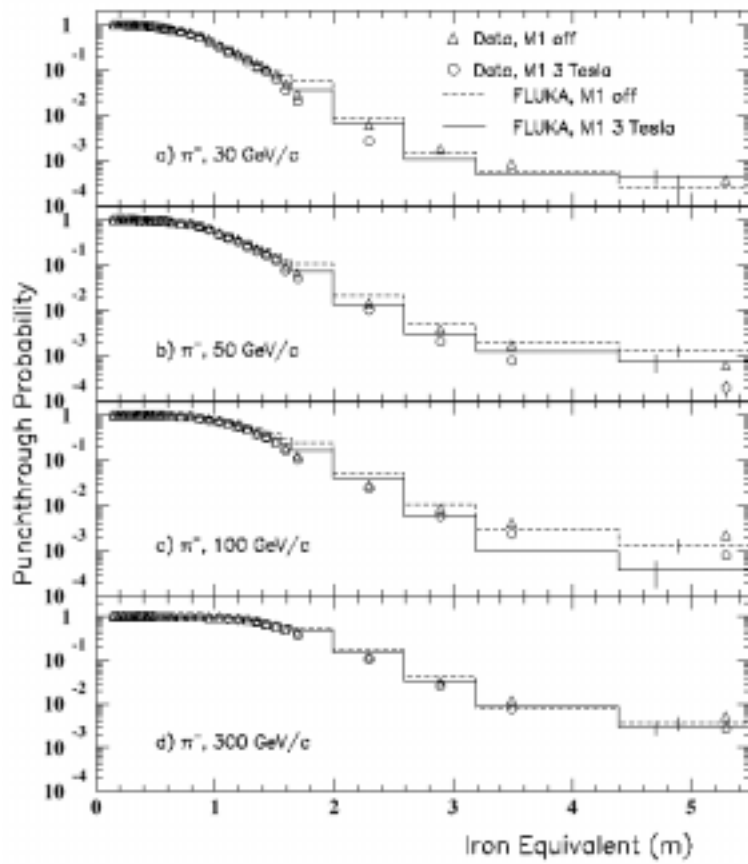


Figure 38: The punchthrough probability as a function of thickness of iron.

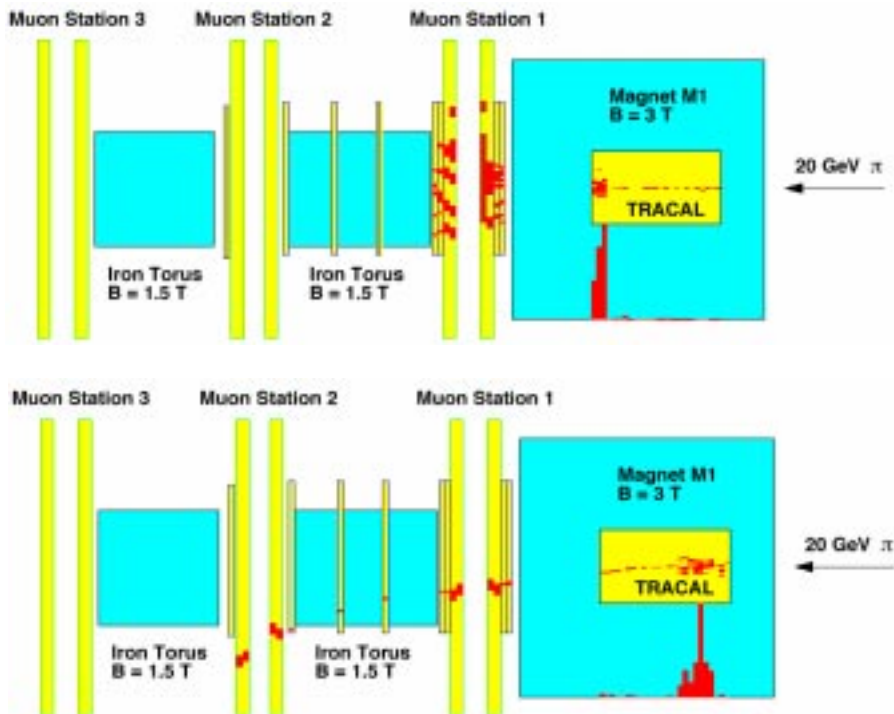


Figure 39: Two examples of punchthrough—one leading to confusion in the muon station and the other leading to a secondary muon.

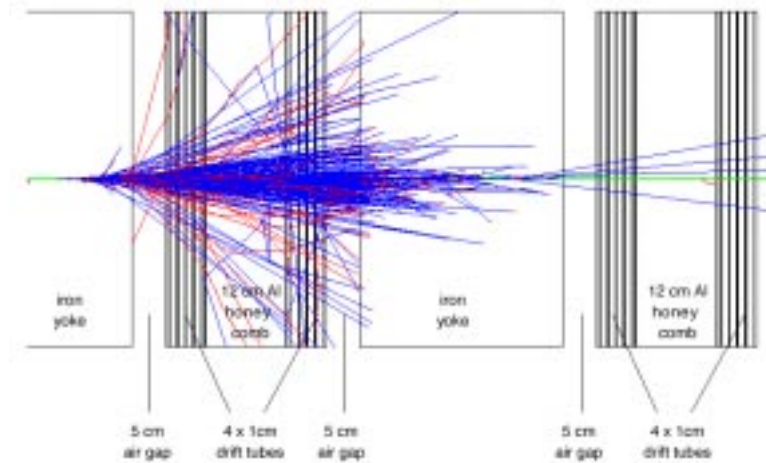


Figure 40: The simulation of a 22 GeV electromagnetic shower generated by a photon radiated by a 1 TeV muon.

7 THE EXPERIMENTAL CHALLENGE AT THE LHC

In the search for high-mass objects and rare signatures, high \sqrt{s} and high luminosity are required. The main LHC machine parameters (proton-proton mode) are a centre of mass energy of $\sqrt{s}=14$ TeV, design luminosity of $L=10^{34} \text{ cm}^{-2} \text{ s}^{-1}$ and a bunch crossing interval of 25 ns. The large proton-proton inelastic cross-section ($\approx 70 \text{ mb}$) leads to some 10^9 interactions/sec. These parameters lead to formidable experimental challenge [6].

The event selection (trigger) must reduce the billion interactions/s to ≈ 100 events/s for storage. The short bunch-crossing period has implications for the design of the readout and trigger systems. It is clearly not feasible to make a trigger decision in the time between bunch crossings, yet new events occur every crossing and a trigger decision has to be made for every crossing. This requires relatively complicated *pipelined* trigger processing and readout, where many bunch crossings are processed concurrently by a chain of processing elements. The 1st Level trigger decision takes $\approx 3 \mu\text{s}$ so the data must be stored in pipelines for $\approx 3 \mu\text{s}$.

At design luminosity a mean of ≈ 20 minimum bias events will be superposed on the event of interest. Around 1000 charged tracks emerge from the interaction region every 25 ns. Thus, the products of an interaction under study may be confused with those from other interactions in the same bunch crossing. This problem, known as *pileup*, clearly becomes more severe if detectors with a response time longer than 25 ns are used. The affect of pileup can be reduced by using highly granular detectors with good time resolution, giving low *occupancy* (fraction of detector elements that contain information) at the expense of having large numbers of detector channels.

The high particle fluxes emanating from the interaction region lead to high radiation levels requiring radiation hard detectors and front-end electronics.

LHC detectors are therefore not just larger versions of the previous generation of HEP detectors. Many years of R&D have been needed to develop detectors and electronics that could survive the harsh environment of LHC.

8 THE PROTON-PROTON EXPERIMENTS AT THE LHC

The p-p General Purpose Detectors (GPDs) at the LHC follow closely the onion-like structure discussed in the introduction. The single most important aspect of the overall detector design is the magnetic field configuration for the measurement of muon momenta. The choice strongly

influences the rest of the detector design. The two basic configurations are solenoidal and toroidal. The closed configuration of a toroid does not provide magnetic field for inner tracking. Since a detector without magnetic inner tracking cannot adequately study a number of important physics topics an additional inner solenoid is required to supplement a toroid. Large bending power is needed to measure precisely high momentum muons or other charged tracks. This forces a choice of superconducting technology for solenoids whereas both superconducting (air or iron core) and warm (iron core) are possible for toroids. Below we briefly discuss the advantages and drawbacks of each of the configurations.

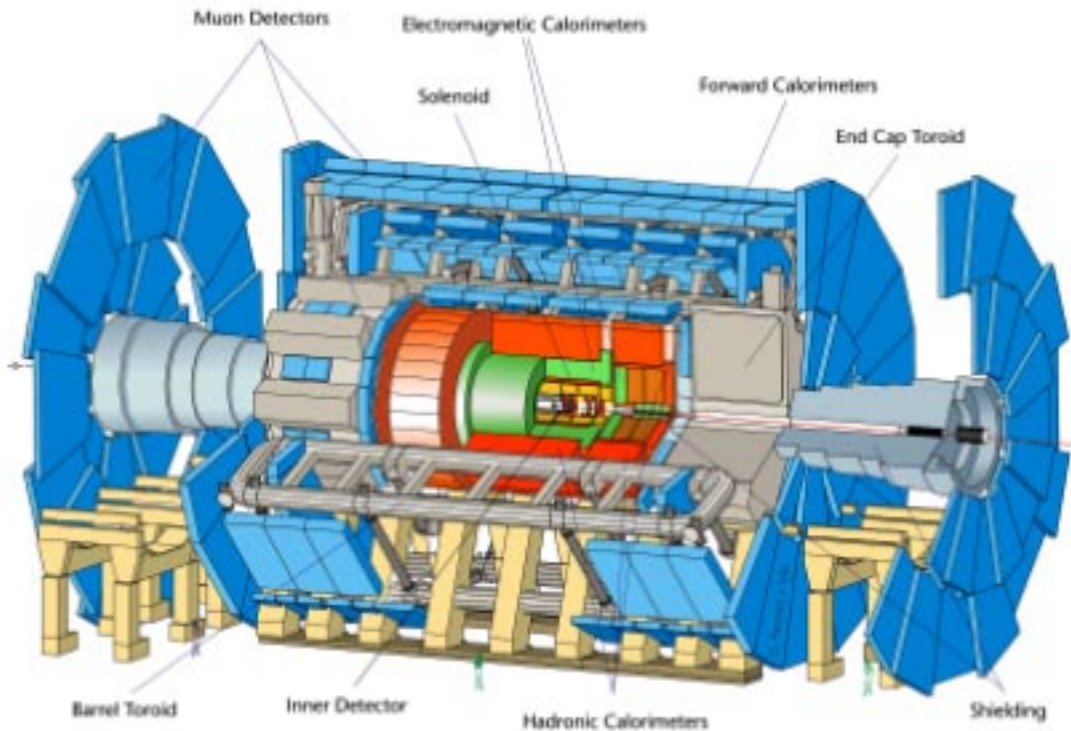


Figure 41: The 3-D view of the ATLAS detector displaying the various sub-detectors.

8.1 ATLAS

The overall detector layout is shown in Fig 41. The magnet configuration uses large superconducting air-core toroids consisting of independent coils arranged with an eight-fold symmetry outside the calorimetry. The magnetic field for the inner tracking is provided by an inner thin superconducting solenoid generating a field of 2T.

The inner detector is contained a cylinder of length 6.8 m and radius 1.15 m. It consists of a combination of ‘discreet’ high-resolution Si pixel and microstrip detectors in the inner part and ‘continuous’ straw-tube tracking detectors with transition radiation capability in the outer part of the tracking volume. Highly granular liquid-argon (LAr) e.m. sampling calorimetry covers the pseudorapidity range $|\eta| < 3.2$. In the endcaps the LAr technology is used for the hadronic calorimeter. The forward LAr calorimeters, extending the coverage to $|\eta| = 4.9$ are also housed in the same cryostat. The barrel part of the hadronic calorimetry is provided by Fe/ scintillator-tile sampling calorimeter using WLS fibres. The calorimetry is surrounded by the muon spectrometer. The air-core toroid system encloses a large field volume. The muon chambers, grouped into three stations, are placed in the open and light structure to minimize effects from

multiple scattering. The muon spectrometer defines the overall dimensions of the ATLAS detector with a diameter of 22m and a length of 46 m. The weight of the detector is about 7000 tons.

8.2 The Compact Muon Solenoid (CMS)

The overall layout is shown in Fig. 42. At the heart of CMS sits a superconducting solenoid. In order to achieve a good momentum resolution within a compact spectrometer without making stringent demands on muon-chamber resolution and alignment a high magnetic field is required. CMS has a long (13 m), large bore ($\phi=5.9$ m) and high field (4T) solenoid. The field is large enough to saturate 1.5 m of iron which is thick enough to accommodate four muons stations to ensure robustness and full geometric coverage. Each muon station consists of many measuring planes. These consist of aluminium drift tubes in the barrel region and Cathode Strip Chambers (CSCs) in the endcap region.

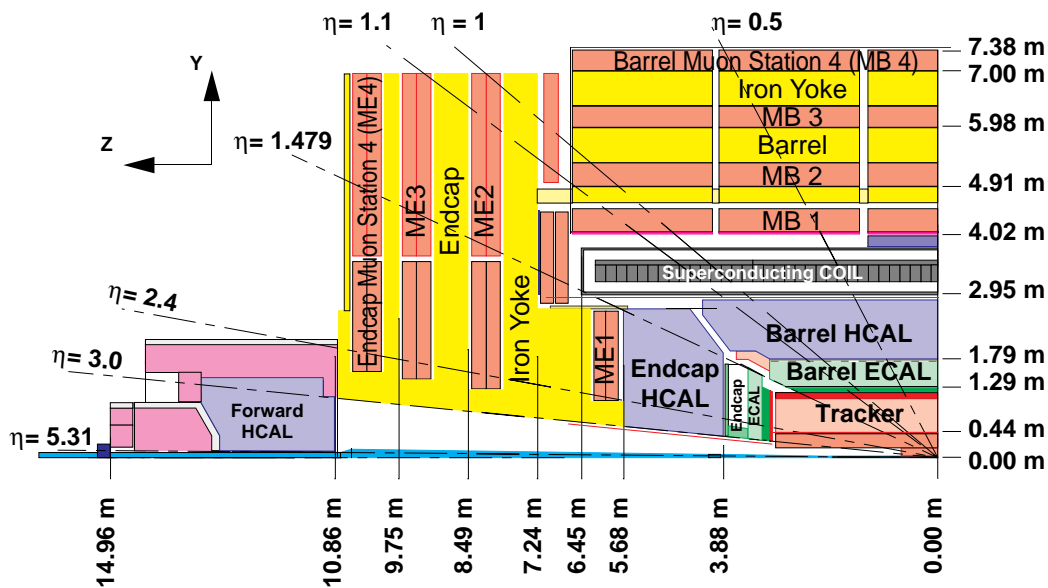


Figure 42: The transverse cut of the CMS detector.

The bore of the magnet is large enough to accommodate the inner tracker and the calorimetry inside the coil. The tracking volume is given by a cylinder of length 6 m and a diameter of 2.6 m. In order to deal with high track multiplicities tracking detectors with small cell sizes are used. Solid-state and gas microstrip detectors provide the required granularity and precision. Pixel detectors placed close to the interaction region improve the measurement of the track impact parameter and secondary vertices. The electromagnetic calorimeter (ECAL) uses lead tungstate (PbWO_4) crystals. A preshower system is installed in front of the endcap ECAL for π^0 rejection. The ECAL is surrounded by a copper/scintillator sampling hadronic calorimeter. The light is channeled by clear fibres fused to wave-length shifting fibres embedded in scintillator plates. The light is detected by photodetectors that can provide gain and operate in high axial magnetic fields (proximity focussed hybrid photodiodes). Coverage up to rapidities of 4.7 is provided by a Cu/quartz fibre calorimeter. The Cerenkov light emitted in the quartz fibres is detected by photomultipliers. The forward calorimeters ensure full geometric coverage for transverse energy measurement. The overall dimensions of the CMS detector are: a length of 21.6 m, a diameter of 14.6 m and a total weight of 14500 tons.

8.3 Muon Systems

8.3.1 ATLAS

The number of field lines crossed by a muon track in toroids is constant. In the endcap region the magnetic field increases as $1/R$. Hence toroids have the property that the transverse momentum resolution is constant over a wide range of pseudo-rapidity. The integral of $B \cdot dl$ ($\propto 1/\sin\theta$) compensates for the Lorentz boost in the forward direction. In an air-core toroid a good stand alone momentum resolution can be reached as long as the quantity BL^2 is large enough (cf. Equation 2). Two drawbacks of the toroidal configuration are:

- the bending does not take place in the transverse plane and hence benefit cannot be drawn from the precise knowledge of the beam-beam crossing point ($20 \mu\text{m}$ at LHC), and
- a solenoid is needed to provide field for the inner tracker, opening the debate of whether the coil should be place before or after the electromagnetic calorimetry.

The design criteria for the muon system can be obtained by requiring that an unambiguous determination is made of the sign for muons of 1 TeV. This implies that $\Delta p/p \approx 10\%$. The sagitta, s , for a track of momentum p in a uniform magnetic field is given by $s=0.3BL^2/8p$. In the case of ATLAS where $B \approx 0.6\text{T}$, $L \approx 4.5 \text{ m}$ $s \approx 0.5 \text{ mm}$ for $p_\mu = 1 \text{ TeV}$. This implies that the sagitta has to be measured with a precision of $\approx 50 \mu\text{m}$. For muon system as large as in ATLAS precision of this nature presents special challenges of spatial and alignment precision. From the term BL^2 (see Equation 2) it is clear that a large magnet is required. However it is not easy to generate a high field over a large volume. By considering Ampere's theorem we can estimate the current required, Now

$$2\pi RB = \mu_0 nI \Rightarrow nI = 20 \times 10^6 \text{ At}$$

$$\text{i.e. } 2.5 \times 10^6 \text{ At for 8 coils}$$

where n is the number of turns and I is the current. ATLAS employ $2 \times 2 \times 30$ turns leading to a current of $I=20 \text{ kA}$. Such currents can only be considered in the context of superconducting magnets.

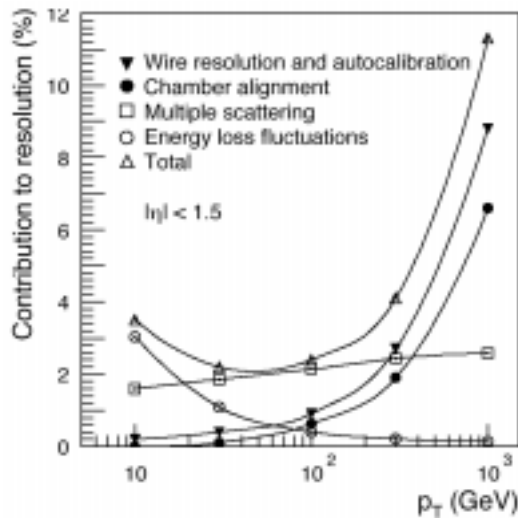


Figure 43: The various contribution to the momentum resolution for muon in the ATLAS detector.

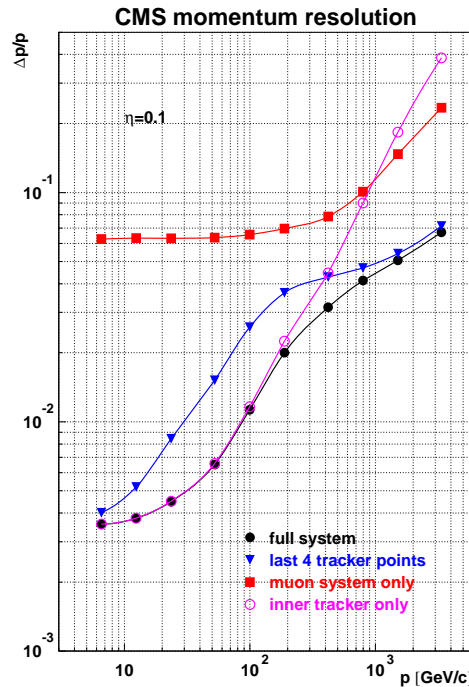


Figure 44: The muon momentum resolution using the the inner tracker and the muon system.

The basis of measurement in the barrel part of the ATLAS spectrometer is to measure a position on the muon trajectory before and after the magnet, and a third point between the other two.

$$\frac{\Delta p}{p} = 26.7 \sigma \sqrt{\left(\frac{1}{2N_1} + \frac{1}{N_2}\right)} \frac{p}{BL^2} \quad (\%)$$

where p is in GeV, B in T and σ, L is in m. In ATLAS $\sigma \approx 70 \mu\text{m}$, $N_1 \approx 6$, $B \approx 0.6 \text{T}$, $L \approx 4.5 \text{m}$ implying $\Delta p/p \approx 0.8 \%$ at 100 GeV which is very close to the value found in simulations.

The momentum resolution is limited by energy loss fluctuation in the calorimeters at small momenta and by detector resolution at high momenta, whereas multiple scattering effect is approximately momentum independent as can be seen in Fig. 43. The momentum resolution is typically 2-3% over most of the kinematic range apart from very high momenta, where it increases to $\approx 10\%$ at $p=1 \text{ TeV}$.

8.3.2 CMS

A large $\int B \cdot dL$ can be obtained for a modest size using high field solenoids. The bending, which takes place in the transverse plane, starts at the primary vertex. For tracks which pass through the end of the solenoid the momentum resolution worsens as $L_c \cdot \tan\theta / r_c$ where L_c and r_c is the length and radius of the solenoid. The effect can be attenuated by choosing a favourable dimensional ratio (length/radius). For CMS coil, where a high magnetic field is chosen, the challenge lies in the production of a reinforced superconducting cable that can take an outwards pressure of about 60 atm.

The field generated in a solenoid is given by $B = \mu_0 n I$. For CMS with $B=4 \text{T}$, $n=2168$ implying that $I \approx 20 \text{ kA}$ again requiring superconducting technology.

Centrally produced muons are measured three times: in the inner tracker, after the coil and in the return flux. If multiple scattering and energy loss are neglected then the muon trajectory beyond the return yoke extrapolates back to the beam-line due to the compensation of the bending before and after the coil. This fact can be used to improve the momentum resolution at high momenta. The sagitta is given by perpendicular distance between the outermost inner tracking points and the line joining the beam to the muon beyond the return yoke. The muon momentum resolution is illustrated in Fig. 44.

6.3.3 Muon Detectors

Two kinds of muon detectors are used at LHC serving complementary purposes. These are gaseous drift chambers that provide accurate position measurement for momentum determination and ‘trigger’ chambers, such as resistive plate chambers, that have a short response ($\leq 25 \text{ ns}$) for precise bunch crossing identification but less accurate position measurement. The former category of detectors can also provide a first level-trigger on muons. In LHC GPDs the rate in the barrel region ($\approx 10 \text{ Hz/cm}^2$) is two orders of magnitude smaller than in the endcaps. This rate is due mainly to hits induced by photons from neutron capture. The neutrons are evaporation neutrons produced by breakup of nuclei in hadronic showers. Hence drift chambers are replaced by faster chambers such as cathode strip chambers. Since the dominant background is neutron induced which usually affecting two detecting layers, each of the muon stations comprises several (≈ 6) layers of detectors.

The operation of the gas chambers is discussed later.

9. INNER TRACKING

The most powerful way to ‘see’ the event topology is by using the inner tracker. The role of the inner tracker is to measure the momentum and impact parameter of charged tracks with minimal disturbance. The figures of merit are the track finding efficiency, the momentum resolution and the secondary vertex resolution. As described earlier the inner tracker plays a crucial role in the identification of electrons, taus and b-jets.

During the 60’s the bubble chamber was the detector of choice for tracking. However it was superseded by electronic detectors as HEP moved to the study of lower cross-section phenomena. The bubble chambers had a low repetition rate and lacked sufficient triggering capability. Recently large detectors such as ALEPH and DELPHI at LEP have used ‘electronic bubble chambers’ in the form of Time Projection Chambers (TPCs). These give 3-D spatial information with high granularity and some particle identification capability is in-built using dE/dx measurements. However these are not used in the LHC GPDs as the electron drift time is long (25-45 μ s). They are suitable for LEP as the event rate is low and the bunch crossing interval is large. The tracking detectors at the LHC have to deal with very high particle rates ($\approx 4 \cdot 10^{10}$ particles/s emerging from the interaction point) and very short bunch crossing time (25 ns). Furthermore the target momentum resolution for 100 GeV tracks is almost an order of magnitude better at the LHC than at LEP. Hence Si pixel and microstrip detectors, and short drift-time gaseous detectors (straw or MSGCs) are used.

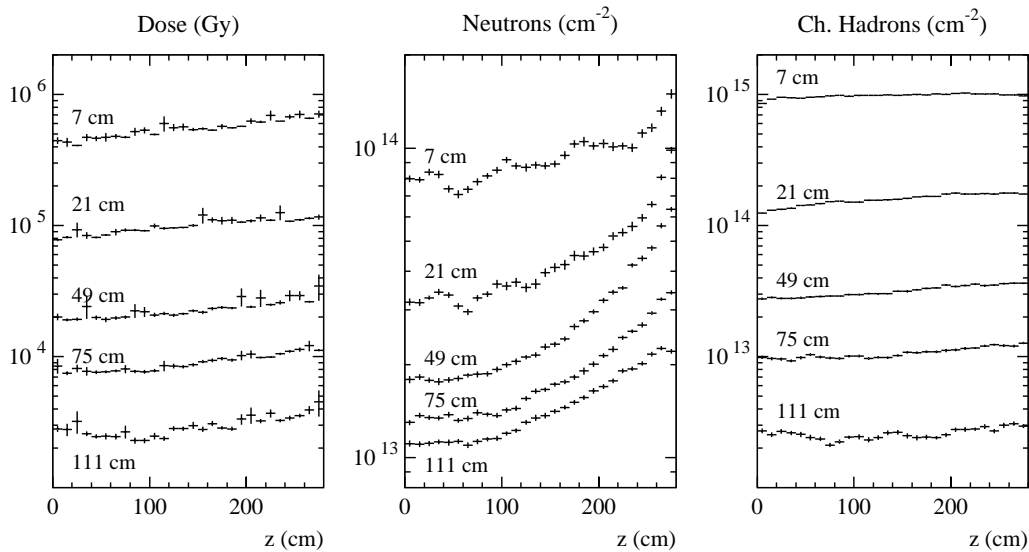


Figure 45: The integrated dose, neutron and charged hadron fluence in the inner tracking cavity for and integrated luminosity of $5 \cdot 10^5 \text{ pb}^{-1}$ corresponding to the first 10 years of LHC operation.

The radiation levels in the CMS tracking system are illustrated in Fig. 45 for an integrated running time of 10 years corresponding to $5 \cdot 10^5 \text{ pb}^{-1}$. Three regions can be delineated. Pixel detector are placed closest to the interaction vertex where the particle flux is the highest. The typical size of a pixel is about $150 \mu\text{m} \times 150 \mu\text{m}$ leading to an occupancy of about 10^{-4} per LHC crossing. In the intermediate regions the particle flux is low enough to enable use of Si microstrip detectors with typical cell size of $10 \text{cm} \times 75 \mu\text{m}$ leading to an occupancy of $\approx 1\%$ /LHC crossing. In the outermost regions of the inner tracker the particle flux has dropped sufficiently to allow

use of gaseous detectors. Typical cell size in CMS is $25\text{cm} \times 250 \mu\text{m}$ giving an occupancy of a few percent.

We shall now look at the operation of gaseous and semi-conductor detectors,

9.1 Gaseous Tracking Detectors

Fast charged particles ionise atoms of a gas. If W is the energy required to create an electron-ion pair then the total total number of electron-ion pairs is

$$n_{total} = \frac{\Delta E}{W} = \frac{dE}{dx} \frac{\Delta x}{W}$$

In fact $n_{total} \approx 3-4 n_{primary}$. Fig. 46 shows the number of primary electrons for various gases. For a gap of 1 cm of Argon about 100 electron-ion pairs are created. A signal consisting of only 100 electrons is not easy to detect as the noise of fast amplifiers tends to be $\approx 1000 e^-$'s. Hence one needs amplification in the gas. Consider a cylindrical cell, with grounded walls and a very thin anode wire placed at the axis. The electric field at a distance r from the wire can be calculated using Gauss' theorem and is given by

$$E(r) = \frac{CV_0}{2\pi\epsilon_0} \frac{1}{r}$$

Consider a charged track that traverses the cell (Fig. 47). The electrons from the electron-ion pairs drift towards the anode wire. Close to the wire the electric field is sufficiently high for the electrons to gain enough energy to ionise further atoms. This leads to an exponential increase in the number of e-ion pairs.

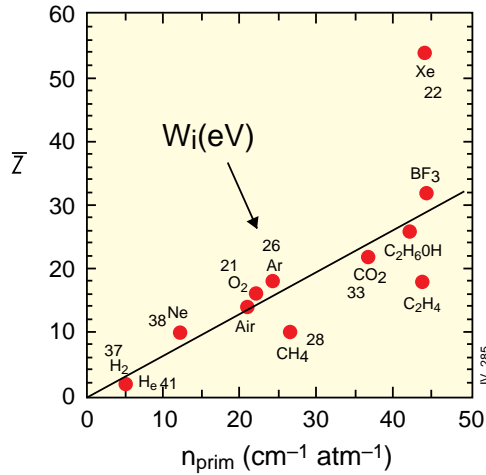


Figure 46: The number of primary clusters per cm in various gases.

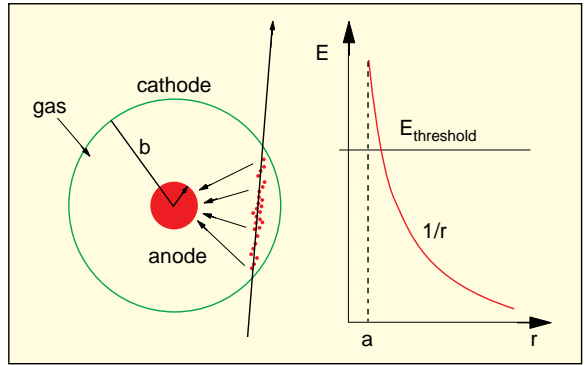


Figure 47: Schematic of a charged track traversing a cylindrical gas chamber (e.g. a straw tube).

The probability that an electron will produce an ionising collision with an atom in a distance dr is

$$N_a \sigma_i dr$$

where N_a is the no. of atoms/unit volume. The increase in the number of electrons after dr then is

$$dn = n N_a \sigma_i dr$$

Let $N_a \sigma_i = \alpha$. α is known as the 1st Townsend coefficient. The larger is the value of α the larger is the number of collisions per unit distance. The mean free path length (λ) between collisions can

be defined as $\alpha = 1/\lambda$. α is a function of r as it varies with electric field which usually varies with r . Hence

$$dn = n \alpha(r) dr$$

$$\therefore n = n_0 \exp \int_a^{r_c} \alpha(r) dr$$

where n_0 is the no. of electrons initially present. Therefore the gain, M , is given by

$$M = \frac{n}{n_0} = \exp \int_a^{r_c} \alpha(r) dr$$

In fact $M \propto e^{CV_0}$

It is interesting to calculate what happens near an anode wire. From Table 3 it can be deduced that 50% (90-%) of the electrons are produced ≈ 2.5 (10) μm from the anode wire.

Table 3: Values of various parameters that determine gain in gas wire-detectors.

r (μm)	E (kV/cm)	α (ip/cm)	$\lambda=1/\alpha$ (μm)
10	200	4000	2.5
20	100	2000	5
100	20	80	125
200	10	1	1 cm

9.1.1 Detector Gas Mixtures

Avalanche multiplication occurs in all gases. However gases should have the following properties: a low working voltage, stable operation at high gain, high rate capability, long lifetime and fast recovery. The principal component of a desirable gas is usually a noble gas such as argon. These allow multiplication at relatively low electric fields. They do not have molecules and hence the electrons only suffer elastic collisions with little loss of energy. Electrons can easily be absorbed by complex molecules. Argon is usually preferred as it gives more primary ionisation than He or Ne and is significantly cheaper than Kr or Xe. However a counter full of a noble gas (e.g. argon) does not allow stable operation. During the avalanche process many Ar atoms are excited and decay emitting UV photons (e.g. with an energy of 11.6 eV). These UV photons can strike the cathode (usually clad with copper which has an ionisation threshold of 7.7 eV) and eject photoelectrons which give rise to another avalanche. There is therefore a positive feedback and a continuous discharge sets in. A chamber filled with pure Ar suffers such breakdown at relatively low gain. Gases are added which 'quench' the secondary avalanches. Polyatomic gases have many non-radiative vibrational and rotational excited states over a wide energy range. If a chamber contains a fraction of such a gas, its molecules will absorb energy from excited argon atoms by colliding with them or by dissociating into smaller molecules. Since $\tau_{\text{emission}} \gg \tau_{\text{collision}}$ the UV photon emission is eliminated or quenched. The presence of a quenching gas can allow an enormous increase in stable gain obtainable. Isobutane (C_4H_{10}), methane (CH_4) and many hydrocarbons and alcohols are such gases.

9.1.2 Operation Modes of Chambers

We can look at various operation modes of gas chambers (Fig. 48) as the potential difference is increased [25]. At very low voltages electrons begin to be collected but *recombination* of electrons and ions is the dominant process. At higher voltages all the electrons and ions are collected and the chamber is said to operate in the *ionisation mode*. At a certain higher voltage called the threshold voltage (V_T) the electric field close to the surface of the anode is large enough to begin the process of *multiplication*. Increasing the voltage (V_0) beyond V_T results in gains $\geq 10^4$ with the detected charge being proportional to the deposited energy. The chamber is said to operate in the *proportional mode*. At even higher voltages the proportionality is gradually lost due to the distortion of the electric field caused by space charge around the anode (*limited proportionality mode*). The region of limited proportionality eventually ends in a region of saturated gain i.e. the size of the signal is the always the same i.e. independent of the initial deposition of energy. The chamber is said to operate in the *Geiger mode*.

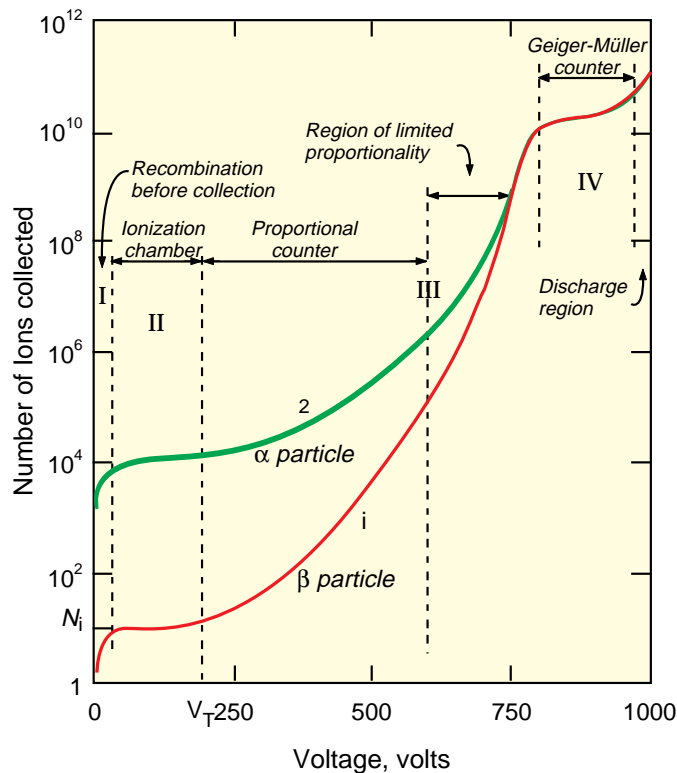


Figure 48: The various modes of operation of gas wire detectors.

9.1.3 Time development of the Signal

Consider a single primary electron drifting towards the anode (at ≈ 5 cm/ μ s) and into the region of increasingly high electric field (Fig. 49). At a radius of typically a few wire radii the electric field becomes large enough for the primary electron to gain enough energy to cause ionisation. Due to lateral diffusion a drop-like avalanche surrounding the wire develops. The whole process of exponential avalanche multiplication lasts ≈ 1 ns. The electrons are collected very fast (in ≈ 1 ns) as the drift distance is only a few microns. The positive ions drift slowly towards the cathode. However *the signal on the electrodes is induced by the movement of charges*. Since the electrons move a very short distance they induce a very small signal. The size and time development of the induced signal is determined by the ion drift. It is given by:

$$I(t) = \frac{Q}{2t_0 \ln(b/a)} \left(\frac{1}{1+t/t_0} \right)$$

where t_0 is the characteristic time. For example, the total drift time for ions in Ar at NTP is 550 μs for $a = 10\mu\text{m}$, $b = 8\text{mm}$, $C = 8 \text{ pF/cm}$, $\mu^+ = 1.7 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$, $V_0 = 3\text{kV}$. The growth time is very fast (1/1000 of the total drift time). The normal practice is to terminate the counter with a resistor R such that the signal is differentiated with a time constant RC allowing very short pulses. A high rate capability is therefore possible.

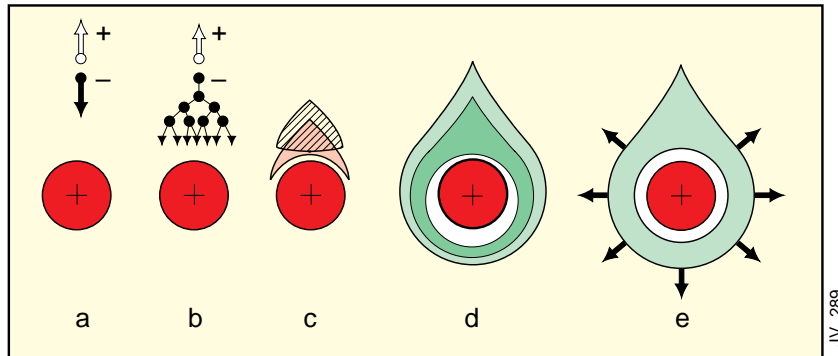


Figure 49: The schematic drawing of the development of an avalanche in a wire chamber.

9.1.4 Drift Chambers

The working of a drift chamber is illustrated in Fig. 50. The electrons liberated along the path of the charged track drift towards the anode wire. The spatial information is obtained by measuring the time of drift of electrons. The traversal of the particle is signaled by a scintillator or by the bunch crossing time in collider experiments. The stop on a TDC is given by the arrival of the electrons. The cell size in drift chambers is much larger than in multi-wire proportional chambers thus relatively fewer number of wires and electronics channels are needed.

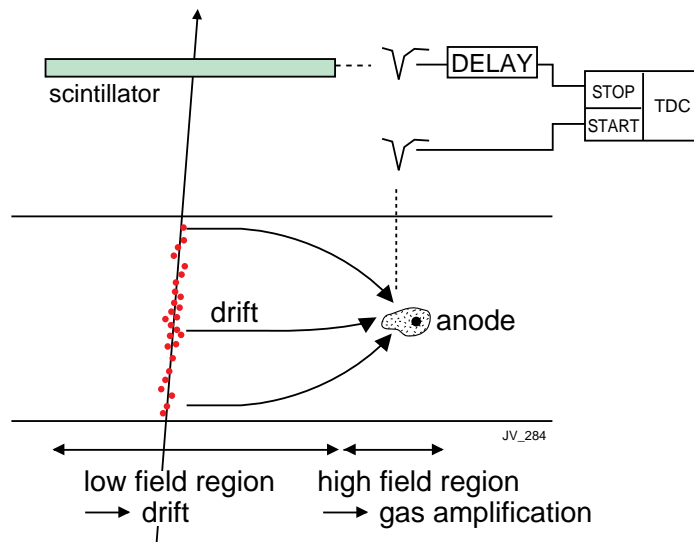


Figure 50: The principle of operation of a drift chamber

The desired properties of the gases used in drift chambers are listed below:

- the gas should have a high purity - electrons can be captured if electro-negative impurities

are present. The longer the drift path length the higher the required level of purity.

- the gas should exhibit saturation of drift velocity i.e. drift velocity that is constant over small voltage shifts away from the working voltage. The precision then becomes insensitive to field inhomogenities, changes in voltage and temperature etc.

- the gas should be fast for high counting rate. The maximum counting rate is limited by the total drift time.

Below we consider two drift chambers; the monitored drift tubes (MDTs) used in the ATLAS and drift tubes (DTs) used in the CMS muon system.

The MDTs are cylindrical aluminium tubes with a length of about 5 m, a diameter of 3 cm and a wall thickness of $400\mu\text{m}$ with a $50\mu\text{m}$ diameter central W-Re wire. The tubes are operated with a non-inflammable $\text{Ar-CH}_4\text{-N}_2$ (91%-5%-4%) gas mixture at a pressure of 3 bars. The wire is set at a potential of $\approx 3300\text{V}$ and the electric field at the wire is $\approx 200\text{ kV/cm}$ yielding a gain of about 20,000. The maximal drift time is $\approx 500\text{ns}$. The distance-time relation is shown in Fig. 51a and good linearity is achieved over almost the full drift path. The measured resolution is $\approx 80\mu\text{m}$ (Fig. 51b).

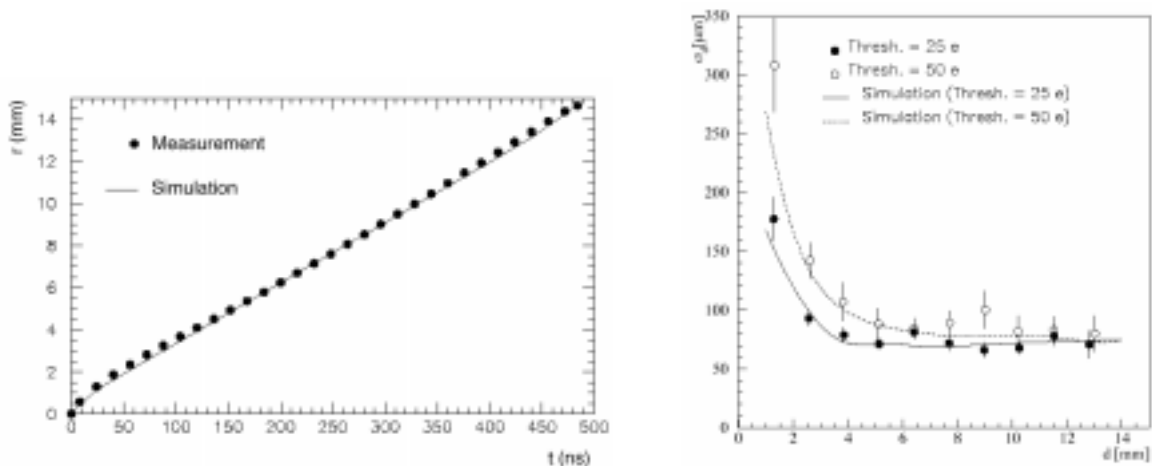


Figure 51: a) The distance-time relation and b) the measured position resolution for the ATLAS MDTs.

The layout of the CMS muon DT cell is illustrated in Fig. 52. Three field-shaping electrodes are employed to assure a linear distance-time relationship. The side cathode is set at $\approx 1800\text{V}$, the top and bottom strips at $\approx 1800\text{V}$ and the wire is run at $\approx 3600\text{V}$. The gain is about 90,000. The gas mixture is Ar-CO_2 (85%-15%). The drift velocity as a function of the electric field is shown in Fig. 53. The measured resolution is found to be $\approx 200\mu\text{m}$ (Fig. 54)

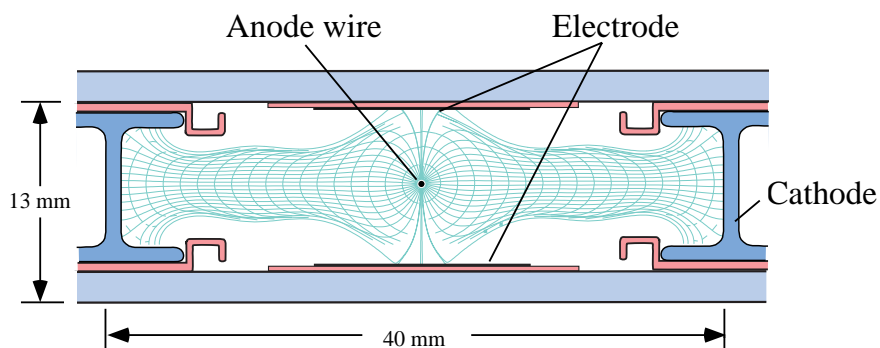


Figure 52: The layout of the drift cell of the CMS drift tubes

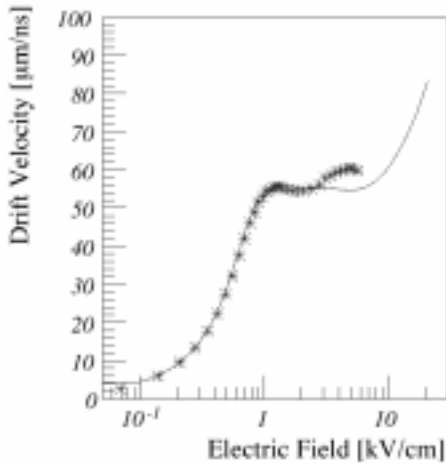


Figure 53: The drift velocity as a function of the electric field in Ar-CO₂. The calculated one is shown as a line.

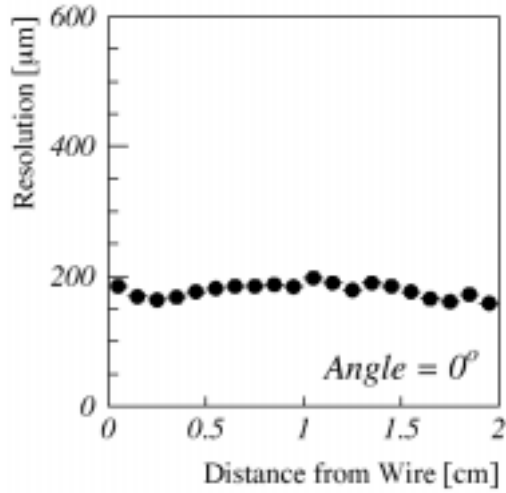


Figure 54: The measured position resolution in the CMS drift tubes.

9.1.5 The Time Projection chamber

The time projection chamber (TPC) is a 3-D imaging drift chamber. An example is the ALEPH TPC illustrated in Fig. 55 [26]. It consists of a large gas filled cylinder with a thin HV electrode in the middle plane. The magnetic and electric fields are parallel, uniform and along the axis. The ends of the cylinder are covered by sector arrays of proportional anode wires. Parallel to each wire are rectangular cathode pads.

The ALEPH TPC has a diameter of 3.6 m and a length of 4.4m. It is filled with Ar-CH₄ (91%-9%) at atmospheric pressure. The electron drift time is 45 μs. The r-φ coordinate is obtained by interpolating signals induced on precisely located cathode pads (6.2mm×30mm) and the z coordinate from the drift time. Particle identification for low momenta particles can be achieved by using the dE/dx information. Diffusion is significantly reduced by the axial magnetic field and the performance is improved by laser calibration. Various systematics are removed by making straight line fits to trails of ionisation caused by laser beams shot into the gas volume. The performance is given by:

$$\sigma_{R\phi} \approx 170 \mu m, \sigma_z \approx 750 \mu m \text{ and } \frac{\sigma_{p_t}}{p_t} \approx 0.1 p_t \oplus 0.3 (\%)$$

where p_t is in GeV.

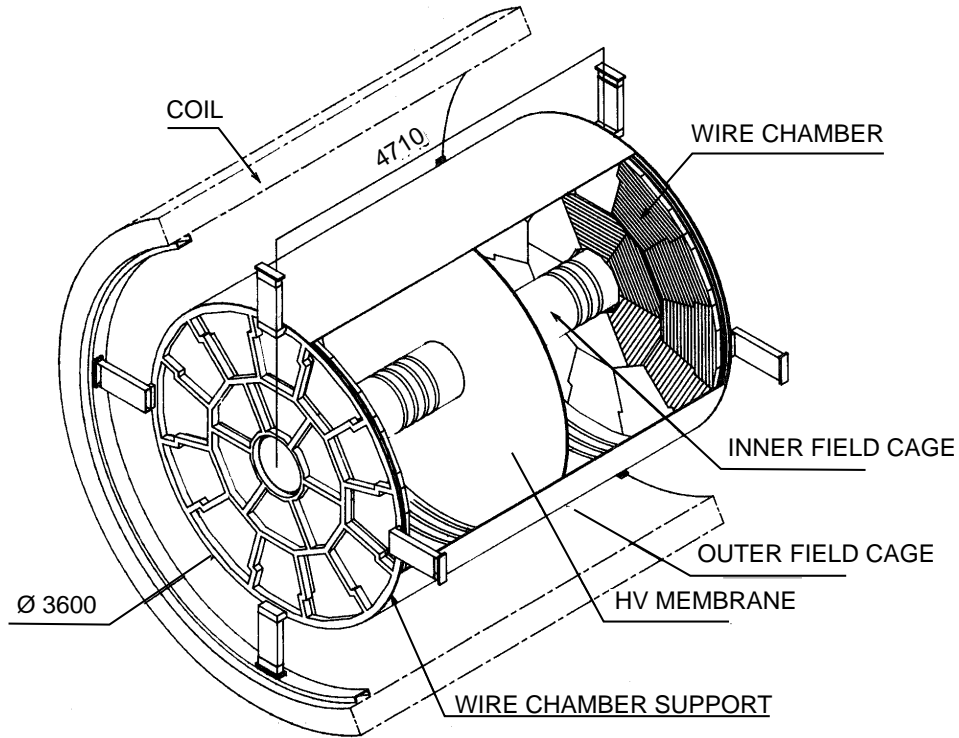


Figure 55: The illustration of the ALEPH time projection chamber.

9.1.6 Microstrip Gas Detectors

Microstrip gas chambers (MSGCs) are made using micro-electronics technology where the precision of photo-lithography is $\approx 0.1\text{-}0.2 \mu\text{m}$ [27]. This allows the overcoming of two major limitations of multi-wire proportional chambers (MWPCs). The spatial resolution in MWPCs, orthogonal to the wire is limited by wire spacing. The limit is around 1mm due to mechanical and electrostatics considerations. The rate capability of MWPCs is limited by the long ion collection time that is typically several tens of μs .

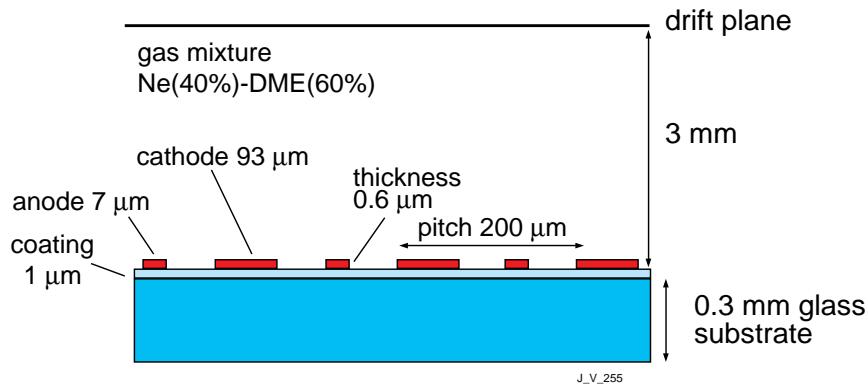


Figure 56: The principle of operation of a microstrip gas chamber.

The MSGC is a miniaturized version of a MWPC. The principle of operation of an MSGC is illustrated in Fig. 56. The wires are replaced by thin strips imprinted on an insulating support e.g. glass to prevent the electrostatic forces from distorting or breaking them. Hence their spacing and width can be reduced. The cathode and anode strips are laid on the substrate over which is placed a planar electrode generating the drift field. The gas gap between the substrate and the drift

electrode is usually about 3mm. The MSGCs have been operated up to rates of 10^6 particles/mm²/s.

For the chambers to be used in CMS the drift plane is set at $-3500V$, the cathode at around $-520V$ and the anode is grounded. The primary electrons drift along the uniform electric field (Fig.57) until they get close to the anode where multiplication takes place. A major advantage of the MSGC is the much faster signal development since the ions have only to drift the short distance of about $60 \mu m$ from the anode to the cathode as shown in Fig. 58.

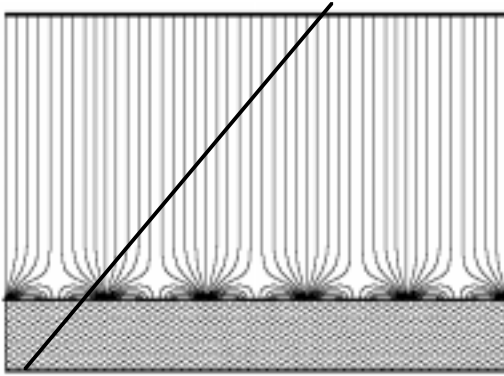


Figure 57: Illustration of the drift field lines.

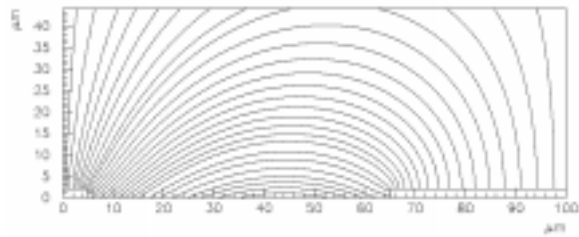


Figure 58: The drift path of ions.

Many parameters can significantly affect the performance. Coating the surface of glass with resistive coating (e.g. Pestov glass with $r \approx 10^{16} \Omega/cm^2$) stops charging-up of the substrate at high rates and renders the detector independent of the bulk electrical characteristics of the substrate. Strips have to be made out of gold to slow aging and decrease attenuation of the collected charge along the strips if they are long. A reduction in the maximum operating voltage for MSGCs exposed to heavily ionising tracks and at high rates has been observed. A proposed solution is to passivate the cathode edges and has been adopted by CMS.

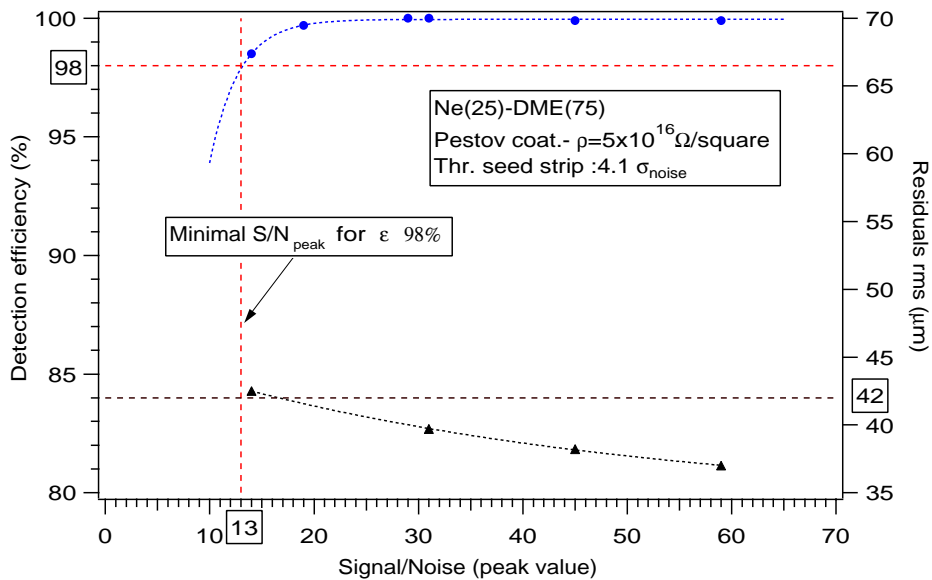


Figure 59: The measured performance of the CMS MSGCs.

The gas mixture used by CMS is Neon-DiMethylEther (40%-60%). DME generate a large number of primary clusters (60 clusters/cm and 180 electrons/cm for a mip). From many points of view 100% DME would be the best gas but it would require a cathode voltage too close to the breakdown voltage. The performance of CMS chambers with an anode-cathode pitch of 200 μm is shown in Fig. 59.

The average cluster size is over 2 and a resolution of about 45 μm can be achieved by using the centroid of the charge deposition.

In a more recent development, the Gas Electron Multiplier (GEM [28]) the gain is achieved in two stages. The first stage of amplification is carried out in a thin polymer foil (kapton), metal-clad on both sides and perforated with a high density of holes with photo-lithographic processing. When a suitable potential difference is applied between the two sides, the field in the hole is large enough to allow multiplication. The second stage can be a conventional MSGC. This is illustrated in Fig. 60. The MSGC can then be operated at a lower voltage promising a wider margin for safe operation as illustrated in Fig. 61.

Much research is still being carried out on the miniaturized gas chambers.

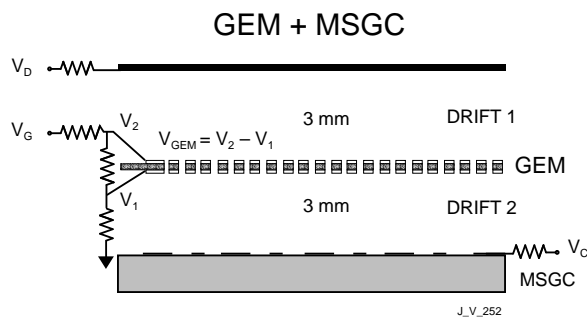


Figure 60: Schematic drawing of GEM+MSGC 2-stage gas microstrip detector

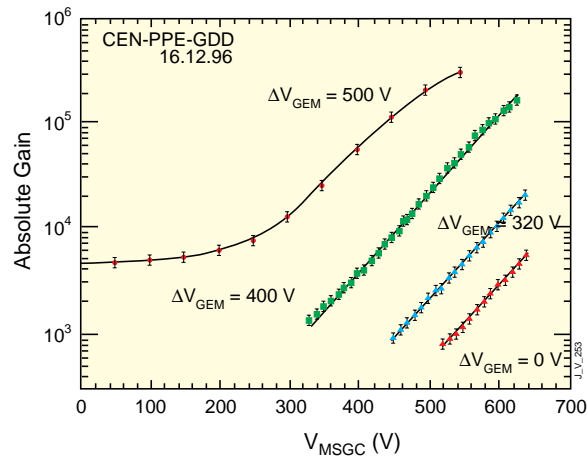


Figure 61: The various combinations of GEM and MSGC potential difference for a given gain.

9.2 Silicon Detectors

Solid state detectors have long been used for the measurement of low energy photons. When used for tracking they have several advantages :

- a large number of carriers are produced. Charged particles traversing Si create electron-hole (e-h) pairs. It takes about 3.6 eV of deposited energy to create an e-h pair compared with about 30 eV to create e-ion pair in gas detectors. The typical thickness of Si detectors is about 300 μm leading to the creation of about $3 \cdot 10^4$ e-h pairs on average. This is large enough not to require multiplication.

- the mobility of electrons and holes is 1450 and 450 cm^2/Vs respectively. The typical operating voltage is $\approx 100\text{V}$ and both the electron and hole charges drift to the respective electrodes within 10 ns. Hence silicon detectors have a fast response and both charge carriers equally induce the detected pulse.

- microelectronic techniques are used for the production of silicon detectors. Silicon detectors therefore have small pitch ($\geq 50\mu\text{m}$) but strip lengths are usually limited in length ($\leq 12\text{cm}$).

9.1.7 Signal generation in Silicon Detectors

The energy levels of atoms become energy bands in a regular assembly of atoms such as crystals.

Valence and conduction bands of energy levels are formed in crystalline materials due to the periodic lattice structure. In the valence band the electrons are bound to specific lattice sites whereas in the conduction band they are free to move through the crystal. At non-zero temperatures it is possible for valence electrons to get enough energy to get into the conduction band (the band gap energy is 1.12 eV). In a pure intrinsic (undoped) material the electron density, n , and the hole density, p , are equal i.e. $n_i = p_i$. For silicon $n_i = 1.45 \times 10^{10} \text{ cm}^{-3}$. Hence in a silicon detector with an area of 1 cm^2 and a thickness of $300 \mu\text{m}$ there are 4.5×10^8 free charge carriers but only $\approx 3.10^4$ e-h pairs are created by the passage of a mip. For the signal to be significant the number of free charge carriers has to be considerably reduced. This is done by 'depleting' the detector using reverse biased p-n junctions.

The p-n junction is made by doping to create *p-type* and *n-type* silicon.

Silicon sits in Group IV of the periodic table i.e. it has 4 outer electrons and can form 4 covalent bonds. If a small concentration (few ppm) of a pentavalent impurity (e.g. P or As) is added one electron is left over after all the covalent bonds are formed. It is very lightly bound and can easily be promoted to the conduction band without creating a corresponding hole. P or As are known as *donor* impurities. Donor electrons are not part of the regular lattice and can occupy a position in the normally forbidden gap (near the top of the gap). Thermal excitation is sufficient to ionize a large fraction of the donors (N_D) and $n = N_D$. The added concentration of electrons increase the rate of e-h recombination shifting the equilibrium between electrons and holes. The concentration of holes decreases but

$$n p = n_i p_i$$

e.g. at room temperature, $n_i = p_i = 10^{10} \text{ cm}^{-3}$ and if donor impurities are $N_D \approx 10^{17} \text{ atoms/cm}^3$, $n = 10^{17} \text{ cm}^{-3}$ and $p = 10^3 \text{ cm}^{-3}$. Charge neutrality is maintained by the presence of ionized donor impurities which cannot migrate as they are fixed to the lattice sites.

If a small concentration (few ppm) of trivalent impurity (e.g. B) is added there is one fewer electron and one covalent bond is left unsaturated. The vacancy represents a hole. Other electrons can be captured to fill this vacancy but are on the whole less firmly bound to the specific sites. are formed. Such impurities are labeled *acceptor* impurities. These lie near the bottom of the forbidden gap. Thermal excitation always assures that electrons are available to fill the vacancies. Hence a large fraction of the acceptor sites are filled i.e.

$$p \approx N_A \quad \text{with} \quad n p = n_i p_i \quad \text{and} \quad p \gg n$$

A measure of impurity level is the electrical *conductivity* or its inverse *resistivity* e.g. an impurity concentration of $10^{13} \text{ atoms/cm}^3$ leads to a resistivity of $500 \Omega \cdot \text{cm}$.

Heavily doped materials with an unusually high impurity concentration are labeled n^+ or p^+ and have a high conductivity. Such doping is often used for electrical contacts.

Consider now the formation of a p-n junction. Start with a wafer of *p-type* crystal with an original acceptor concentration N_A . Assume that one side is left exposed to the vapour of *n-type* impurity so that the left side becomes *n-type*. The density of electrons in *n-type* material is much higher than in the *p-type* one. There is a net diffusion of conduction electrons into *p-type* material where they quickly recombine with holes. The electrons moving out of *n-type* material leave immobile +ve charges and net -ve charge on the p-side is established. The accumulated space charge creates an electric field that diminishes the tendency for further diffusion. The region over which the imbalance exists is called the *depletion region*. The concentration of electrons and holes in this region is $\approx 100/\text{cm}^3$! In our case the region will extend deeper into p-side than the n-side. For electron-hole pairs created in the depletion region electrons will be swept towards the *n-type* material and the holes towards the *p-type*. The application of reverse bias (n-side made +ve) extends the depletion region and the depletion depth is given by

$$d \cong \sqrt{\frac{2\varepsilon V}{eN}} \quad \text{and in PDG } d \approx 0.5(0.3)\mu\text{m} \times \sqrt{\rho V} \quad \text{for } n\text{-type (p-type)}$$

where N is the lower dopant concentration, ρ is the resistivity (typically between 1-10 k Ω .cm). Typically for $d \approx 300\mu\text{m}$ and $\rho = 5 \text{ k}\Omega\text{.cm}$ the bias voltage is $\approx 70\text{V}$.

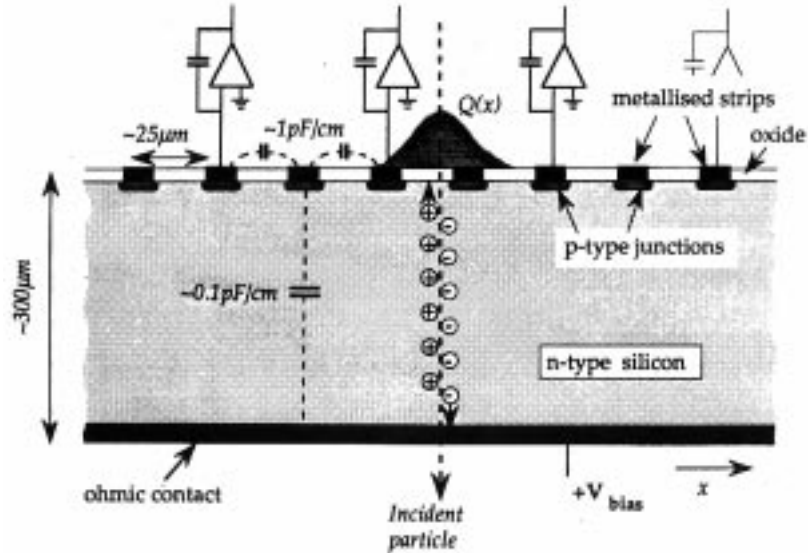


Figure 62: Schematic of a silicon microstrip detector.

1.1.3 Radiation Damage in Silicon Detectors

There are two types of effects that take place in silicon after high levels of irradiation: surface and bulk. The bulk damage is more important than surface damage. High energy hadrons interacting with nuclei, requiring a transfer of only $\approx 15\text{eV}$, can displace Si atoms from their lattice positions. The simplest defects are *vacancies* where a Si atom is absent from its site and *interstitials* where a Si atom occupies a position intermediate between other atomic sites. Disruption of the crystal symmetry leads to the formation of unwanted energy levels in the forbidden gap. The presence of these energy levels considerably increases the leakage current in radiation damaged detectors. The volume current density is observed to increase linearly with particle fluence with

$$\frac{\Delta I}{V} = \alpha \phi$$

where V is the volume (in cm^3), ϕ is the fluence in particles/ cm^2 and $\alpha = 2.10^{-17} \text{ A/cm}$ for minimum ionizing protons and pions after long term annealing.

The ultimate limit to detector lifetimes is given by significant changes in the substrate dopant density during and after irradiation. This effect is poorly understood. The substrate eventually becomes *p-type* irrespective of the initial type. This is shown in Fig. 63. The depletion voltage initially decreases and then increases without limit with increasing particle fluence. Partially depleted detectors can still operate but the charge collection in the undepleted volume is slow and occurs with reduced efficiency due to trapping of electrons. The dopant changes continue after irradiation has stopped BUT can be arrested if the detectors are kept below 0°C . The manufacture of Si detectors has improved substantially and high voltage operation (bias voltage $\geq 500 \text{ V}$) is nowadays possible. It is expected that Si microstrip detectors can be operated to fluences beyond a few times 10^{14} cm^{-2} .

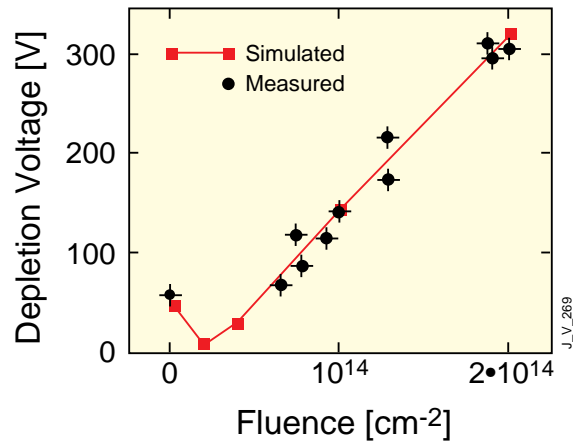
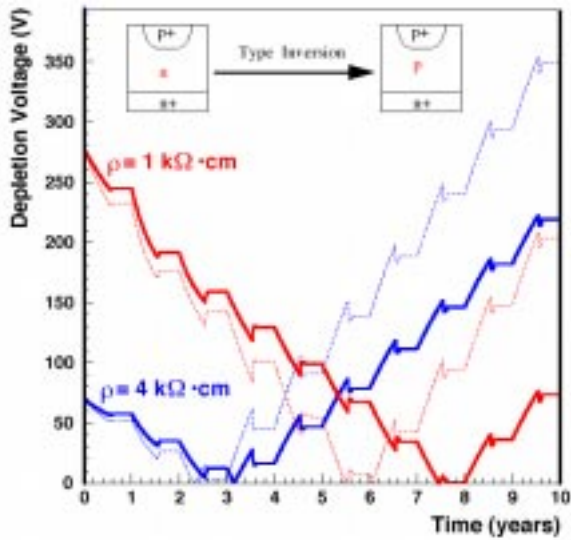


Figure 63: a) The bias voltage to achieve full depletion as a function of running time at the LHC at high luminosity

b) the bias voltage to achieve full depletion as a function of the neutron fluence.

With increasing particle fluence the depletion voltage increases without limit. To obtain a detailed picture of the signal loss after irradiation we consider Si pixel detectors irradiated with pions of 300 MeV/c at extremely small angles. The thickness of the Si was 300 μm and the size of the pixels was $125 \mu\text{m} \times 125 \mu\text{m}$, Effects due to charge trapping can be separated from charge lost due to reduced depletion depth. The method and results are illustrated in Fig. 64.

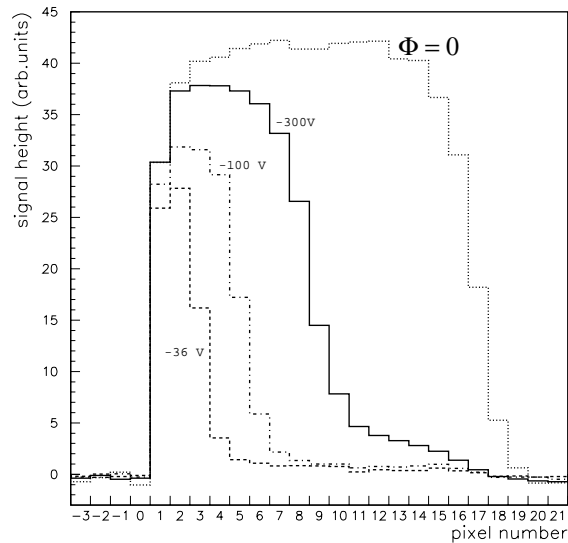
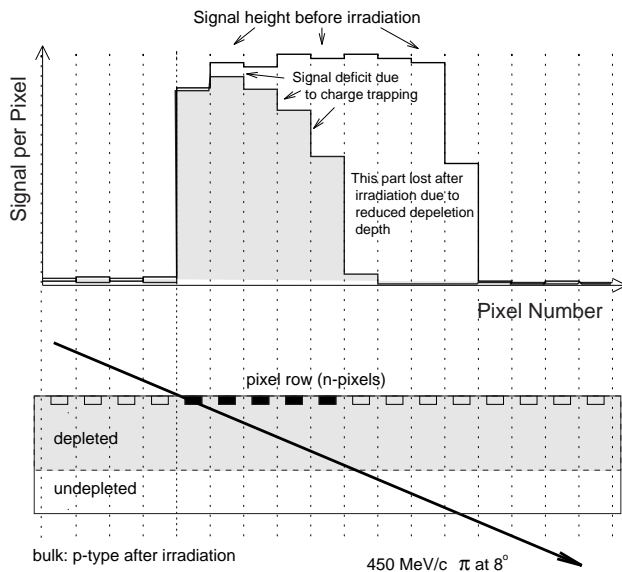


Figure 64: Minimum ionising pions (450 MeV/c) graze the pixel array at an angle of 8° (bottom). The signals from the hit pixel row exhibit the response of pixels at various depths (top). Each pixel sense a depth of $17.5 \mu\text{m}$.

b) Depth profiles of charge collected from the array irradiated with $6 \cdot 10^{14}$ pions/cm². The dotted histogram is for the un-irradiated array.

The detector was kept at 3°C and was irradiated to $6 \cdot 10^{14}$ pions/cm². Twenty days after irradiation the signals from various pixels were measured. The reduced depth of the depletion can clearly be discerned from the pixels that yield little or no charge. The decrease in signal with increasing depth of traversal is due to trapping of charge in the depleted region. The trapping of charge can be kept below $\approx 10\%$ by applying a high bias voltage.

10. ELECTRONICS NOISE

Noise is any unwanted signal that obscures the desired signal. Therefore noise degrades the accuracy of the measurement. There are two types of noise: *intrinsic* and *extrinsic* noise. The intrinsic noise is generated in the detector or electronics and cannot be eliminated though possibly reduced. The extrinsic noise is due to pickup from external sources or unwanted feedback (e.g. ground loops, power supply fluctuations etc.) and is usually eliminated by proper design.

Intrinsic noise has two principal components namely :

- thermal noise (Johnson or Nyquist noise) – *series noise*

Any resistor, R , will develop a voltage across its ends whose average value is zero but r.m.s. is

$$\langle v^2 \rangle = 4ktR\Delta f$$

- shot noise – *parallel noise*

This source arises from fluctuation in the charge carriers and is given by

$$ENC^2 = \frac{4ktR_s(C_d + C_{in})^2}{\tau} I_s + I_n \tau I_p$$

where C_d is the detector capacitance, C_{in} is the input capacitance of the amplifier, I_n is the leakage current, τ is the shaping time and I_s , I_p are series and parallel noise integrals (≈ 1 for $(RC)^2$ shaping). For example, for $\tau = 50$ ns, and a leakage current of $1 \mu A$, $ENC \approx 800$ electrons. Further examples are considered in Sections 9.3 and 10.2.

10.1 Electronics for LHC Experiments

The main components of electronics systems are:

- front-end, signal processing,
- data transmission,
- power supplies, services, ...

The features that differentiate the electronics of the LHC experiments from e.g. LEP experiments are:

- high speed signal processing
- signal pileup
- high radiation levels
- larger number of channels (large data volume),
- new technologies

For example the challenges for the inner tracker electronics are:

- signals are small and fast response must be preserved. Hence long leads cannot be used and the preamplifiers must be mounted on the detectors themselves.

- the data must be held in pipeline memories awaiting Level-1 decision. It is not feasible to transfer data off of the detector at a rate of 40 million events/s for millions of channels. Hence the pipeline memories must be located on the detectors. Consideration has to be given to how the signals are taken out.

- the several millions of channels will dissipate a considerable amount of heat (power dissipation has to be kept as low as possible; the goal is \leq few mW/channel). This leads to the question of how the electronics are cooled.

The above leads to difficult engineering and systems challenges. All this has to be accomplished whilst keeping the amount material in the tracker to the minimum to minimize multiple scattering

and conversion or bremsstrahlung.

10.1.1 Electronics of Sub-detectors

The characteristics and requirements for the electronics for the various sub-detectors can be summarized as follows:

tracking: large number of channels ($\approx 10^7$'s of millions), limited energy precision and limited dynamic range (< 8 -bits). The power dissipation/channel has to be low ($\approx \text{mW/ch}$) and the electronics have to withstand very high radiation levels (neutron fluence of 10^{15} n/cm^2 , integrated doses of 10^7 's of Mrads).

calorimetry: : medium number of channels ($\approx 10^5$), high measurement precision (12-bits), large dynamic range (16 to 17-bits), very good linearity and very good stability in time. The power constraints and the radiation levels ((neutron fluence of 10^{13} n/cm^2 , integrated doses of 100's of krad) are not as stringent as for the tracker.

muon system: the large surface area that needs to be instrumented means that the electronics are distributed over large area and the radiation levels are low.

The generic LHC readout system is illustrated in Fig. 66.

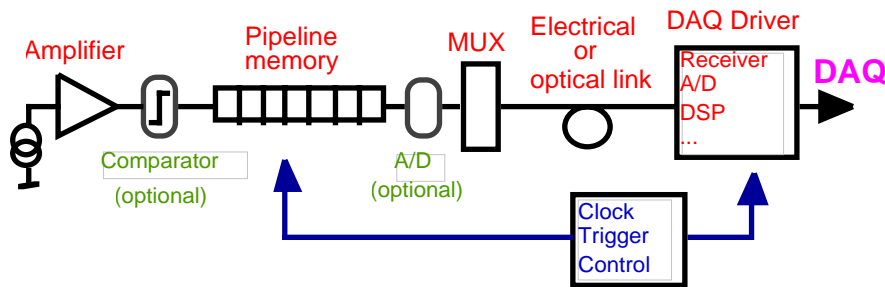


Figure 66a: A generic readout system for a p-p experiment at the LHC.



Figure 66b: The deconvolution of the signal of a silicon detector. The amplifier has a time constant of 50ns.

The functions that are common to all systems are amplification, analogue to digital conversion, association to beam crossing, storage prior to trigger, deadtime-free readout, zero suppression and formatted storage prior to access by the data acquisition, calibration control and monitoring. Most of these features can be illustrated by using as an example the microstrip tracking electronics chain of CMS. The scheme is illustrated in Fig. 67.

Each microstrip is read-out by a charge sensitive amplifier with $\tau=50 \text{ ns}$. The output voltage is sampled at the beam crossing rate of 40 MHz. Samples are stored in an analogue pipeline for up to the Level-1 latency of $\approx 3.2 \mu\text{s}$. Following a trigger a weighted sum of 3 samples is formed in an analog circuit. This confines the signal to a single bunch crossing and gives the pulse height (Fig. 66b). The buffered pulse height data are multiplexed out on optical fibres. The output of the transmitting laser is modulated by the pulse height for each strip. The light signals are transformed into electrical pulses by a Si photodiode and digitized. After some digital processing (zero suppression etc.) the data are formatted and placed into dual port memories for access by the data acquisition. The electronics noise/channel of the tracking system is about 1000 to 1500

electrons before and after irradiation respectively.

The calorimeter and muon systems have also to generate the primitives (energy or momentum values) for the first-level trigger.

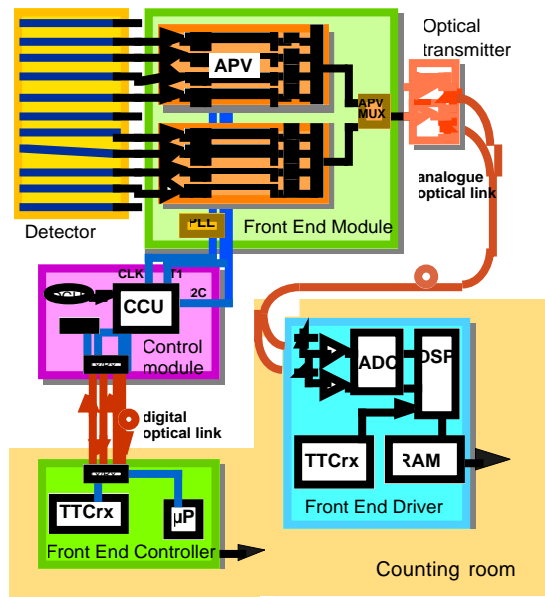


Figure 67: Schematic of the CMS Tracker readout and control system

11. INORGANIC SCINTILLATORS

The desirable properties of a scintillator include:

- a high efficiency of conversion of deposited energy into scintillation light,
- a conversion to light that is proportional to the energy deposited,
- a high light output,
- a medium that is transparent to its emitted light,
- a short luminescence decay time,
- a refractive index $n \approx 1.5$ for efficient coupling to photosensors
- radiation hardness for LHC operation.

No material simultaneously meets all these criteria. Inorganic scintillators (e.g. sodium iodide) have the best light output and linearity whilst organic scintillators (e.g. plastic scintillator) have faster light output but smaller light yield and display saturation of output for radiation with high linear energy transfer. Two types of light emission are possible: *fluorescence* resulting in prompt emission of light in the visible wavelength range and *phosphorescence* resulting in slower emission of light at longer wavelengths. In particle physics inorganic scintillators are only used for electromagnetic calorimetry.

The most demanding physics channel for an electromagnetic calorimeter at the LHC is the two-photon decay of an intermediate-mass Higgs boson. The background is large and the signal width is determined by the calorimeter performance. The best possible performance in terms of energy resolution is only possible using fully active calorimeters such as inorganic scintillating crystals.

Inorganic scintillators have crystalline structure. The valence band contains electrons that are bound at the lattice sites whereas electrons in the conduction band are free to move throughout the crystal. Usually in a pure crystal the efficiency of scintillation is not sufficiently large. A small amount of impurity, called an activator, is added to increase the probability of emission of visible

light. Energy states within the forbidden gap are created through which an electron, excited to the conduction band, can de-excite. Passage of a charged particle through the scintillator creates a large number of electron-hole pairs. The electrons are elevated to the conduction band whereas the +ve holes quickly drift to an activator and ionize it. The electrons migrate freely in the crystal until they encounter ionised activators. The electrons drop into the impurity sites creating activator excited energy levels which de-excite typically with $T_{1/2} \approx 100$ ns. In a wide category of materials the energy required to create an electron-hole pair is $W \approx 3E_g$ e.g. in sodium iodide (NaI), $W \approx 20$ eV. In thallium doped sodium iodide [NaI(Tl)] the number of emitted photons $N_\gamma \approx 40000/\text{MeV}$ with an energy of ≈ 3 eV.

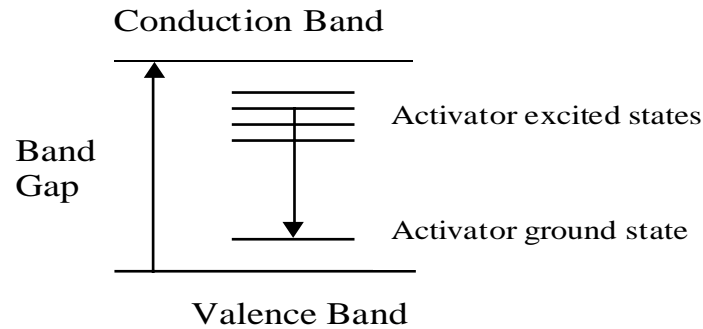


Figure 26: The energy level diagram for a scintillating crystal containing an activator

The consequence of luminescence through activator sites is that the crystal is transparent to its own scintillation light because the emission and absorption bands do not overlap and self-absorption is small. The shift towards longer wavelengths is known as *Stokes' shift*.

The scintillation mechanism in crystals without activators is more complex. For example, in lead tungstate the intrinsic emission in the blue is through excitons localized on the Pb site whereas the green emission is due to defects in the crystalline structure linked to oxygen vacancies [29].

The properties of various crystals used in high energy experiments are given in Table 3. The parameters of some of the recently designed crystal calorimeters are given in Table 4 [30].

Table 3: Properties of various scintillating crystals.

Crystal		NaI(Tl)	CsI(Tl)	CsI	BaF ₂	BGO	CeF ₃	PbWO ₄
Density	g.cm ⁻²	3.67	4.51	4.51	4.89	7.13	6.16	8.28
Rad. length	cm	2.59	1.85	1.85	2.06	1.12	1.68	0.89
Molière radius	cm	4.5	3.8	3.8	3.4	2.4	2.6	2.2
Int. length	cm	41.4	36.5	36.5	29.9	22.0	25.9	22.4
Decay Time	ns	250	1000	35	630	300	10-30	<20>
				6	0.9			
Peak emission	nm	410	565	420	300	480	310-340	425
				310	220			
Rel. Light Yield	%	100	45	5.6	21	9	10	0.7
				2.3	2.7			
d(LY)/dT	%/°C	≈ 0	0.3	- 0.6	- 2.0	- 1.6	0.15	-1.9
Refractive Index		1.85	1.80	1.80	1.56	2.20	1.68	2.16

Table 4: Parameters of various experiments using scintillating crystals.

Experiment		KTeV	BaBar	BELLE	CMS
Laboratory		FNAL	SLAC	KEK	CERN
Crystal Type		CsI	CsI(Tl)	CsI(Tl)	PbWO ₄
B-Field	T	-	1.5	1.0	4.0
Inner Radius	m	-	1.0	1.25	1.3
No. of crystals		3,300	6,580	8,800	76,150
Crystal Depth	X ₀	27	16-17.5	16.2	26
Crystal Volume	m ³	2	5.9	9.5	11
Light Output	p.e./MeV	40	5,000	5,000	4
Photosensor		PMT	Si PD	Si PD	APD*
Gain of photosensor		4,000	1	1	50
Noise / channel	MeV	Small	0.15	0.2	30
Dynamic Range		10 ⁴	10 ⁴	10 ⁴	10 ⁵

* APD: Si avalanche photodiode

11.1 Radiation Damage in Crystals

All crystals suffer from radiation damage at some level. It is rare that irradiation affects the scintillation mechanism itself. However formation of colour centres takes place leading to absorption bands. A colour centre is a crystal defect that absorbs visible light. A high concentration of blue light colour centres makes crystals yellowish. The simplest colour centre is an F-centre where an electron is captured in an anion vacancy. The consequence of colour centre production is a decrease in the light attenuation length leading to a decrease in the amount of light incident on the photosensor. This is illustrated in Figure 27 for various samples of PbWO₄ crystals grown under differing conditions. The crystals were irradiated using γ s, incident at the front of the crystal, from a ⁶⁰Co source.

Extensive R&D has been carried out over the last 5 years by CMS in order to improve the radiation hardness of PbWO₄ crystals [29]. Generally the strategy has been to decrease the concentration of defects that lead to colour centre production by optimizing the stoichiometry (the concentration of PbO and WO₃ in the melt) and annealing after the growth of the crystal. The remaining defects are compensated by specific doping, e.g. by pentavalent elements on the W site and trivalent on the Pb site, and by improving the purity of the raw materials. The levels of improvement can be seen from Fig 69. The most recent crystals of lead tungstate have shown very good resistance to irradiation. This is illustrated in Fig 70. The loss of collected light, for crystals doped with both niobium and yttrium, show a decrease in the collected light of less than 2% at saturation. The effect of irradiation can also be dose-rate dependent

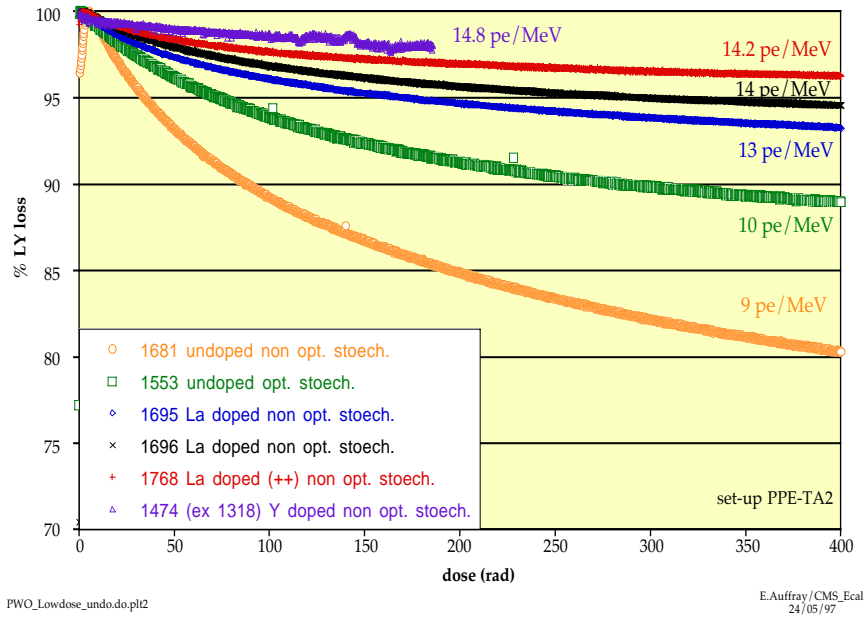


Figure 69: The loss in the collected light as a function of dose (delivered at ≈ 0.15 Gy/hr) for crystals grown under various conditions.

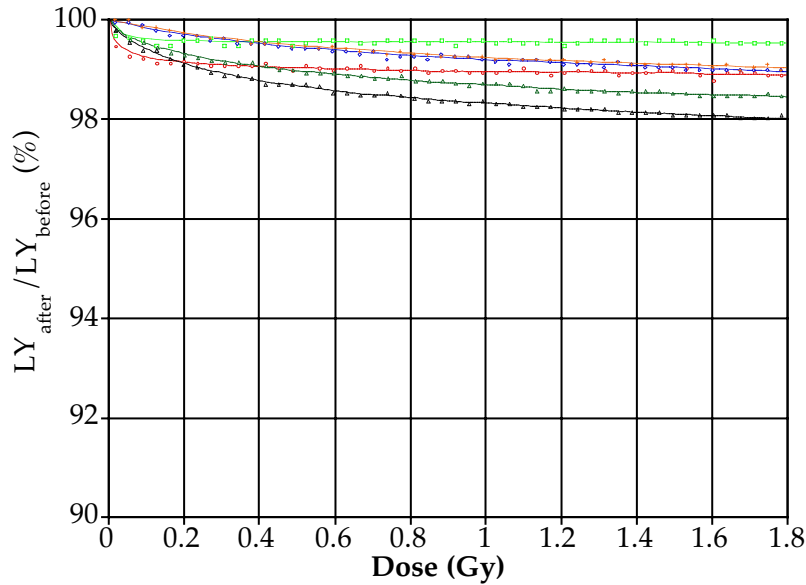


Figure 70: The loss in the collected light as a function of dose (delivered at ≈ 0.15 Gy/hr) for recent PbWO_4 crystals doped with Niobium and Yttrium.

The radiation dose expected at the shower maximum for the lead tungstate calorimeter of CMS, over the first ten years of LHC operation, is below 4,000 Gy in the barrel region ($|\eta| < 1.5$), $\approx 70,000$ Gy at $|\eta| \approx 2.5$ rising to 200,000 Gy at $|\eta| \approx 3.0$. Furthermore the expected dose rate at design luminosity, and shower maximum, is below 0.3 Gy/h in the barrel region, ≈ 6 Gy/h at $|\eta| \approx 2.5$ rising to 15 Gy/h at $|\eta| \approx 3.0$.

11.2 Performance of CMS Lead Tungstate Crystals

Several matrices of improving quality have been tested in electron beams over the last few years. Radiation damage leads to a decrease in the attenuation length and hence in the collected light. As the efficiency of the scintillation mechanism is not affected by irradiation the energy resolution will not be affected as long as the attenuation length does not fall below ≈ 2 -3 times the length of the crystal. The small loss of light can be corrected by regularly measuring the response to a known amount of light injected into crystals. This has been demonstrated in beam tests [31].

Results from a recently tested prototype are shown in Fig 71. The distribution of the sum of energy in 9 crystals for electron of an energy of 280 GeV is shown. An excellent energy resolution is measured without significant tails. The measured energy resolution is also shown. The stochastic term is expected to be $< 3\%$ in the final calorimeter since the surface area of the photosensor will be doubled.

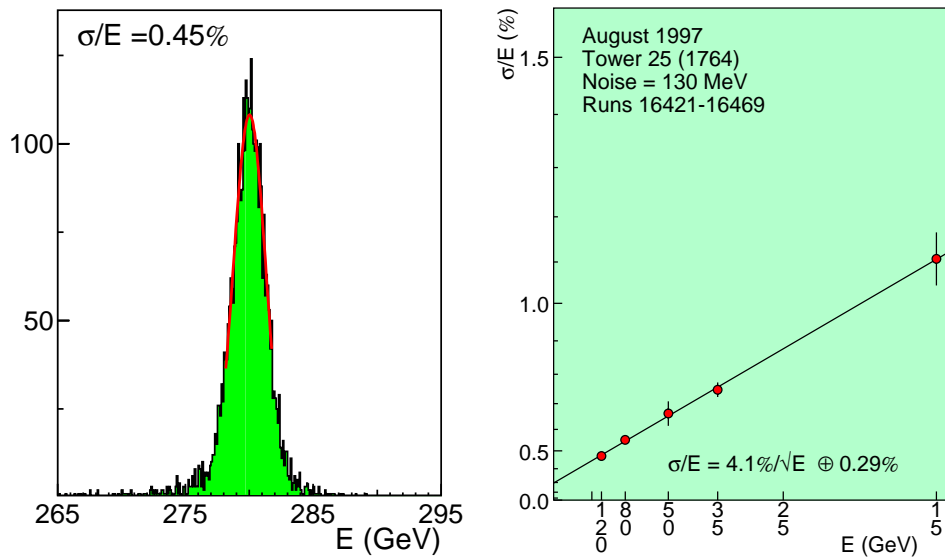


Figure 71: a) The distribution of the sum of energy in 9 crystals for an electron of an energy of 280 GeV, b) the measured energy resolution

11.3 Photosensors

11.3.1 Photomultipliers

The contribution to the energy resolution from the process of conversion of light to photoelectrons can be significant. For example, in a lead glass calorimeter about 10,000 Cerenkov photons/GeV impinge on the photomultiplier. The conversion leads to about 1000 photoelectrons/GeV and hence the contribution to the stochastic term will be

$$\left(\frac{\sigma}{E}\right)_{pe} = \frac{\sqrt{1000}}{1000} \approx 3.2\%$$

The maximum number of independent e^\pm particles, given that the Cerenkov threshold is 0.7 MeV, is $1000/0.7$ per GeV i.e. $n = 1400 e^\pm$. This leads to an additional contribution to the energy resolution i.e. $(\sigma/E)_n = (\sqrt{1400})/1400 \approx 2.7\%$. The observed resolution then becomes

$$\frac{\sigma}{E} = \sqrt{\left(\frac{\sigma}{E}\right)_n^2 + \left(\frac{\sigma}{E}\right)_{pe}^2} \approx 4.5\%$$

An energy resolution of $\sigma/E \sim 5\% / \sqrt{E}$ for e.m. showers has been measured in a large lead glass array [32].

11.3.2 Silicon Avalanche Photodiodes

The light output from PbWO_4 crystals is low. These crystals are deployed by CMS in a 4T transverse magnetic field and the use of photomultipliers is excluded. Unity gain Si photodiodes cannot be used since even the small rear shower leakage from $25 X_0$ deep crystals considerably degrades the energy resolution. This is due to the fact that the photodiode response to ionising radiation is significant compared with the signal due to scintillation light. Hence CMS use Si avalanche photodiodes (APDs) with a gain of about 50. The particularity of these novel devices, over and above photomultipliers, is the noisy amplification process. The working principle of these devices is shown in Fig 72.

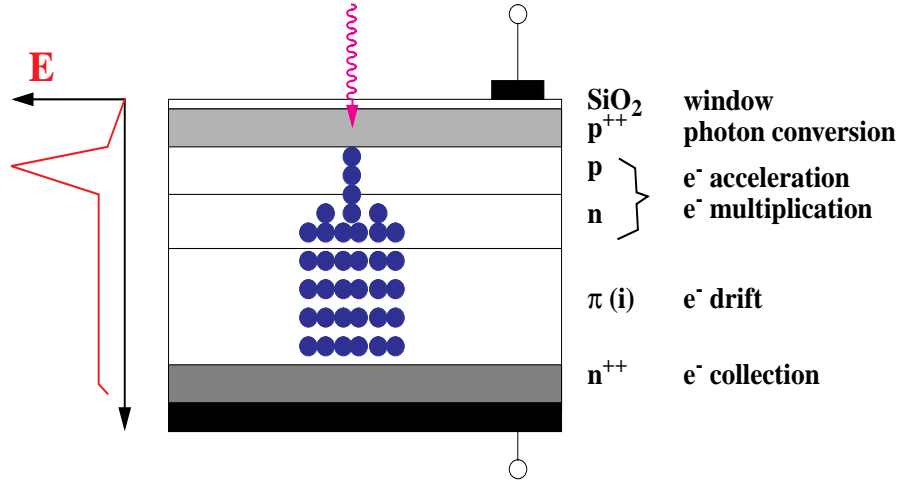


Figure 72: The working principle of a Si avalanche photodiode.

Consider a crystal with a light yield of N_γ photons/MeV. $N_\gamma E$ photons hit the APD for an energy deposit E . Assuming a quantum efficiency Q (which can easily be $\approx 85\%$ for APDs),

No. of photoelectrons is $N_{pe} = N_\gamma E Q$

Then the photostatistics fluctuation is $\pm \sqrt{N_{pe}}$

If there is no fluctuation in the gain process then the no. electrons transferred to the amplifier is (M =gain) $M N_{pe} \pm M \sqrt{N_{pe}}$

BUT if the multiplication process is noisy and the gain itself has a fluctuation, σ_M , then the no. of electrons is $M N_{pe} \pm \sqrt{M^2 + \sigma_M^2} \sqrt{N_{pe}}$

Hence the photostatistics contribution to the energy resolution becomes

$$\frac{\sigma_{pe}(E)}{E} = \frac{1}{\sqrt{N_\gamma E Q}} \sqrt{\frac{M^2 + \sigma_M^2}{M^2}} = \frac{1}{\sqrt{N_\gamma E Q}} \sqrt{F}$$

where F is called the 'excess noise factor' and quantifies the induced degradation in the energy

resolution due to fluctuations in the amplification process. Typically for APDs $F \approx 2$ and for photomultipliers $F \approx 1.2$. Some properties of APDs, from two manufactures, are listed in Table 5.

Table 5: Some properties of APDs

Parameter	Hamamatsu	EG&G
Active Area	25 mm ²	25 mm ²
Quantum Efficiency at 450nm	80%	75%
Capacitance	100 pF	25 pF
Excess Noise Factor, F	2.0	2.3
Operating Bias Voltage	400-420 V	350-450 V
dM/dV x 1/M at M=50	5%	0.6%
dM/dT x 1/M at M=50	-2.3%	-2.7%

11.4 System Aspects

A real calorimeter is a system comprising active media, electronics chain and mechanical structure, all enclosed in an environment that must be kept stable. Hence many factors have to be considered in order to maintain the resolution achieved in beam tests. For example, in the case of the CMS ECAL, the temperature of the crystals has to be maintained to within 0.1% since both the crystal and the photosensor each have a temperature dependence of the output signal of $d(\text{Signal})/dT \approx -2\%/^{\circ}\text{C}$. This requires a powerful cooling system and a hermetic environmental shield. To maintain uniformity of response across crystals the mechanical structure has to be thin and preferably made of low-Z material. No load from one crystal should be transferred to its neighbours. A 300 μm glass fibre alveolar structure has been chosen by CMS. The electronics system has to provide a stable response, deliver high resolution digitization (12-bits) and a large dynamic range (≈ 16 -bits) whilst preserving a low electronics noise per channel (< 40 MeV/channel). Furthermore, the on-detector electronics must be radiation hard and have as low a power consumption as possible.

More information on the systems aspects of calorimeters can be found in the ATLAS [10] and CMS [11] Technical Design Reports.

12. CALORIMETRY USING NOBLE LIQUIDS

Calorimeters using liquid filled ionization chambers as detection elements have several important advantages. The absence of internal amplification of charge results in a stable calibration over long periods of time provided that the purity of the liquid is sufficient. The number of ion pairs created is large and hence the energy resolution is not limited by primary signal generating processes. The considerable flexibility in the size and the shape of the charge collecting electrodes allows high granularity both longitudinally and laterally.

The desirable properties of liquids used in ionization chambers include:

- a high free electron or ion yield leading to a large collected charge,
- a high drift velocity and hence a rapid charge collection,
- a high degree of purity. The presence of electron scavenging impurities leads to the reduction of electron lifetime and consequently a reduction in the collected charge.

The properties of noble liquids are given in Table 6.

12.1 Charge Collection in Ionisation Chambers

Ionisation chambers are essentially a pair of parallel conducting plates separated by a few mm and with a potential difference in an insulating liquid (e.g. liquid argon).

Consider what happens when a single ion-pair is created at a distance $(d-x)$ from the +ve electrode (Fig. 73). The electron drifts towards the +ve electrode and induces a charge

$$Q = - e \frac{(d-x)}{d}$$

where d is the width of the gap. Assuming that the electron drifts with a velocity v , and the time to cross the full gap is v_d , then the induced current is

$$i(t) = \frac{dQ}{dt} = - e \frac{v}{d} = - \frac{e}{t_d}$$

The contribution from the drifting ions can be neglected as their drift velocity is about three orders of magnitude smaller than that of electrons.

Table 6: Properties of noble liquids.

		LAr	LKr	LXe
Density	g/cm ³	1.39	2.45	3.06
Radiation Length	cm	14.3	4.76	2.77
Moliere Radius	cm	7.3	4.7	4.1
Fano Factor		0.11	0.06	0.05
Scintillation Properties				
Photons/MeV		-	1.9 10 ⁴	2.6.10 ⁴
Decay Const. Fast	ns	6.5	2	2
Slow	ns	1100	85	22
% light in fast component		8	1	77
λ peak nm		130	150	175
Refractive Index @ 170nm		1.29	1.41	1.60
Ionization Properties				
W value	eV	23.3	20.5	15.6
Drift vel (10kV/cm)	cm/ μ s	0.5	0.5	0.3
Dielectric Constant		1.51	1.66	1.95
Temperature at triple point	K	84	116	161

Now consider the case where a charged particle traverses the gap (Figure 31b). Suppose N ion-pairs are produced and are uniformly distributed across the gap. The fraction of electrons still moving at a time t after traversal is $(t_d-t)/t_d$ for $t_d < t$. Therefore

$$i(t) = - Q_0 \frac{v}{d} \left(1 - \frac{t}{t_d} \right)$$

where $Q_0 = Ne$ and the current is at its maximum at time $t = 0$ and disappears once all the charges have crossed the drift gap. This time is about 400 ns for a 2 mm LAr gap. Hence

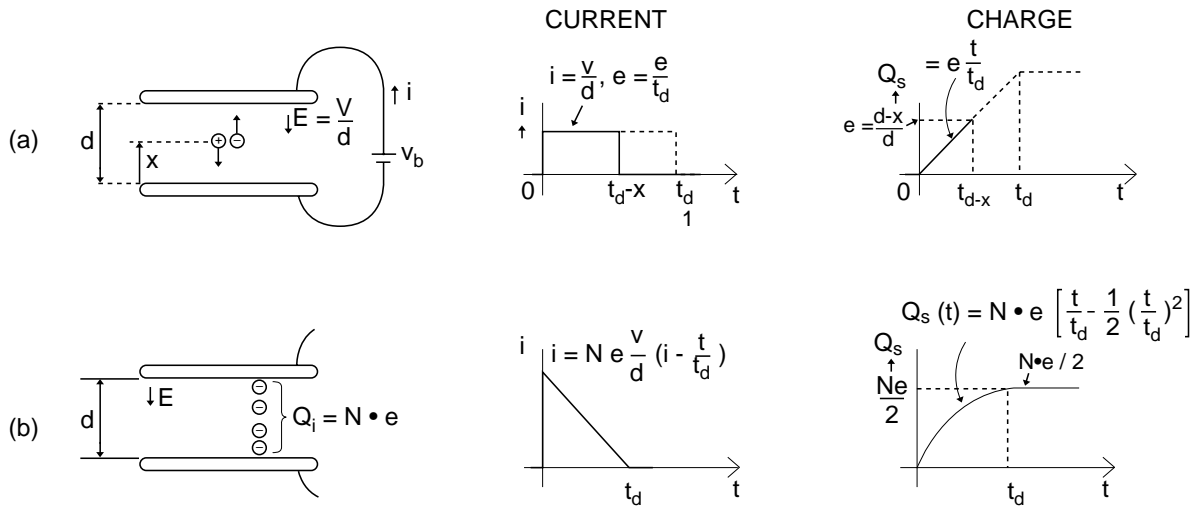
$$q(t) = \int_0^t i(t) dt = -Q_0 \left(\frac{t}{t_d} - \frac{t^2}{2t_d^2} \right) \text{ for } t < t_d$$

The total collected charge (for $t > t_d$) is

$$Q_c = \frac{Q_0}{2}$$

The factor two is due to uniform distribution of ionisation. During drift the electrons can be trapped by impurities. Then the induced current will be reduced. In fact if the electron lifetime is τ

$$i(t) = \frac{Q_0}{t_d} \left(1 - \frac{t}{t_d} \right) e^{-t/\tau} \text{ for } t < t_d$$

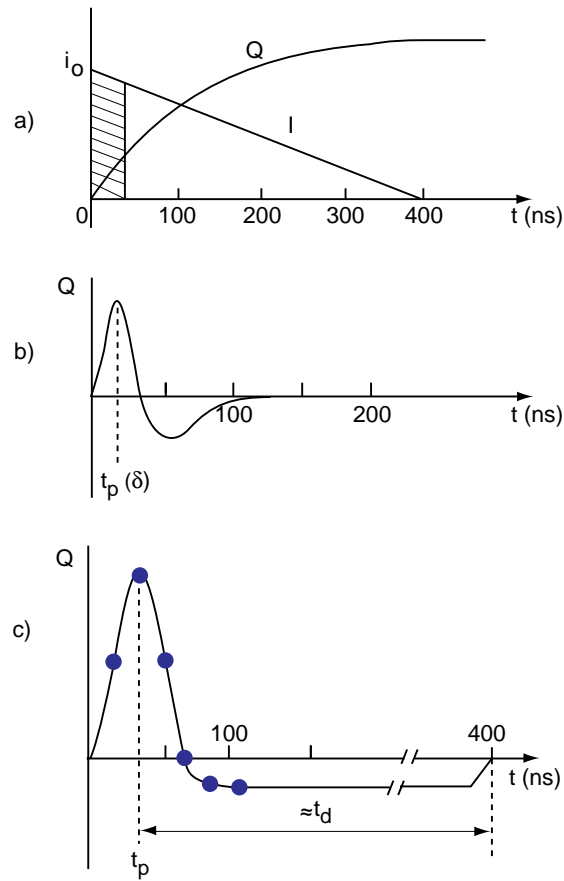


J.V.274

Figure 73: The current and charge for a) a single e-ion pair, b) uniformly distributed e-ion pairs.

12.2 Signal Shapes

As discussed above, for a long electron ‘lifetime’ the induced current has a triangular shape with a duration equal to the electron drift time t_d (Fig. 74a). The total collected charge is also shown. It is clear that a device based on full charge collection will be slow, and hence not suitable for use at the LHC. However the energy information is contained in the initial current i_0 . The information can be extracted and high rate operation made possible by clipping the signal with fast bipolar shaping (Fig. 74b). If the system impulse response has zero integrated area then pileup does not produce a baseline shift. For a peaking time, t_p that is much faster than drift time i.e. $t_p \ll t_d$, the output response becomes the first derivative of the current pulse (Fig. 74c). The height of the output pulse is proportional to the initial current. However, with respect to full charge collection, the energy equivalent of the electronics noise will increase as this scales with $1/\sqrt{\tau}$, where $\tau (=RC)$ is the shaper time constant. At high luminosities, pileup also influences the choice of the value of τ . Pileup scales as $\sqrt{\tau}$. As an example, the optimized value for τ gives $t_p \approx 40$ ns for the ATLAS “accordion” e.m. calorimeter..



JV_204

Figure 74: a) Induced current and integrated charge, b) bipolar shaping function and c) the shape of the output pulse, all as a function of time.

12.3 Examples of Noble Liquid Calorimeters

Conventionally ionization chambers are oriented perpendicularly to the incident particles. However in such a geometry it is difficult to

- realize fine lateral segmentation with small size towers, which in addition need to be projective in collider experiments,
- implement longitudinal sampling,

without introducing insensitive regions, a large number of penetrating interconnections, and long cables which necessarily introduce electronics noise and lead to significant charge transfer time. To overcome these shortcomings a novel absorber-electrode configuration, known as the 'accordion' (Fig. 75, [33]), has been introduced, in which the particles traverse the chambers at angles around 45° .

In a variant, the NA48 [34] experiment has chosen an arrangement of electrodes that is almost parallel to the incident particles. With such structures the electrodes can easily be grouped into towers at the front or at the rear of the calorimeters. In ATLAS the absorber is made of lead plates, clad with thin stainless steel sheets for structural stiffness and corrugated to the shape shown in Fig. 75. Details of the sampling structure are also shown. The read-out electrodes are made out of copper clad kapton flexible foil and kept apart from the lead plates by a honeycomb structure.

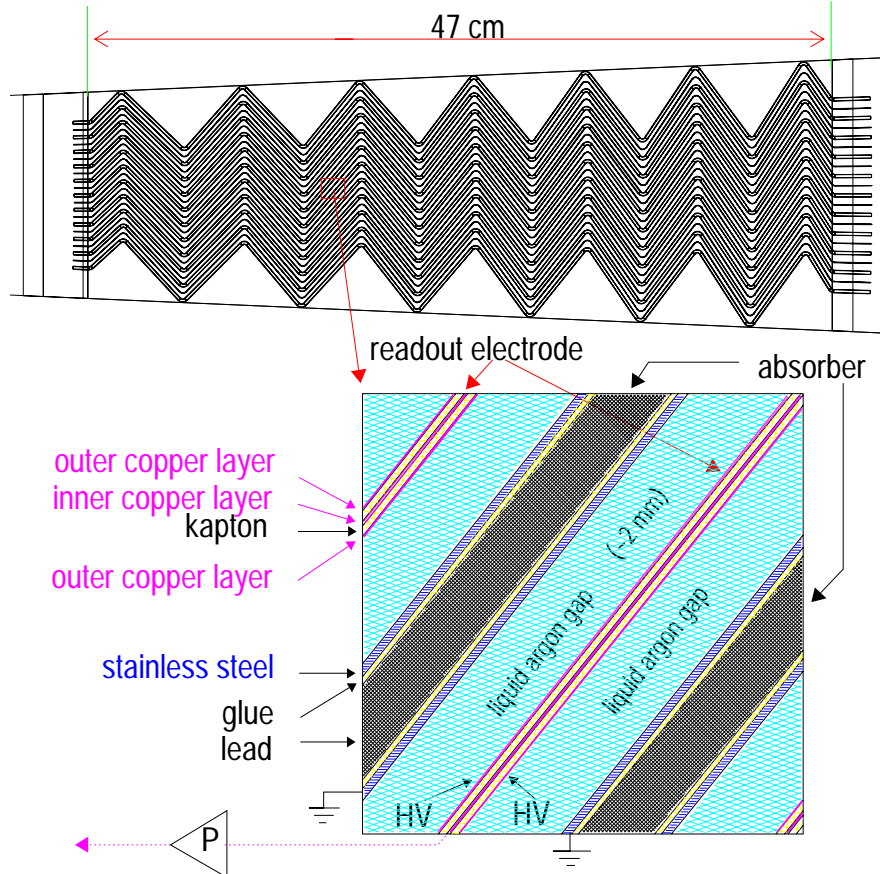


Figure 75: Top) the “accordion” structure of absorber plates of the ATLAS ECAL, below) details of the electrode structure.

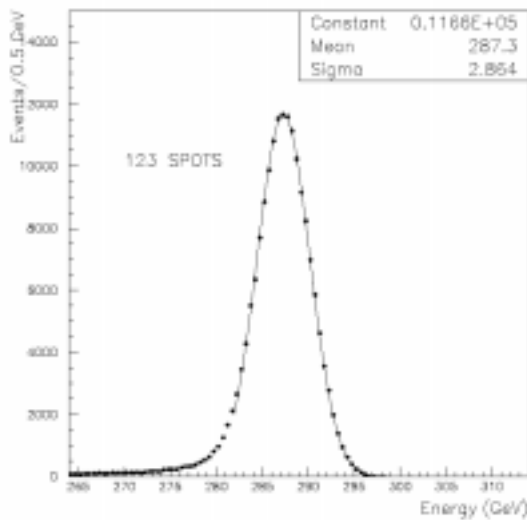


Figure 76: The distribution of the reconstructed energy for 300 GeV electrons in the ATLAS ECAL.

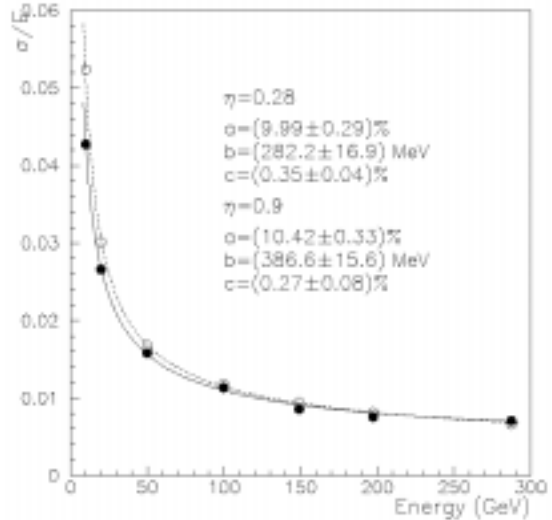


Figure 77: The fractional energy resolution for the ATLAS barrel prototype ECAL.

The results from a beam test of a large ATLAS prototype are shown in Figs. 76 and 77. The electron shower is reconstructed using a region of 3x3 cells each of a size of $\approx 3.7 \text{ cm} \times 3.7 \text{ cm}$. The distribution of reconstructed energy for 300 GeV electrons, over a large area, is shown in

Fig. 76. The fractional energy resolution is shown in Fig. 77 and can be parameterised as

$$\frac{\Delta E}{E} \approx \frac{10\%}{\sqrt{E}} \oplus \frac{0.28_{\text{GeV}}}{E} \oplus 0.35\%$$

where E is in GeV. The response of more than 150 cells over a large area has also been measured. The cell-to-cell non-uniformity is measured to be $\approx 0.58\%$. The major contributions come from mechanics (residual ϕ -modulation, gap non-uniformity, variation of absorber thickness) and calibration (amplitude accuracy). The large flux of isolated electrons from W or Z decays will be used to establish cell-to-cell intercalibration.

13. COMBINED E.M. AND HADRONIC CALORIMETRY

The LHC pp-experiments have put more emphasis on high precision e.m. calorimetry. This is not compatible with perfect compensation. For example the electromagnetic energy resolution of the compensating ZEUS U-calorimeter is modest. The e.m. (σ_E) and hadronic (σ_h) resolutions are given by

$$\frac{\sigma_E}{E} = \frac{17\%}{E} \quad \text{and} \quad \frac{\sigma_h}{E} = \frac{35\%}{E}$$

Nevertheless it is very important to ensure:

- a Gaussian hadronic energy response function (a moderate energy resolution is acceptable),
- hermiticity
- linearity of response, especially for jets.

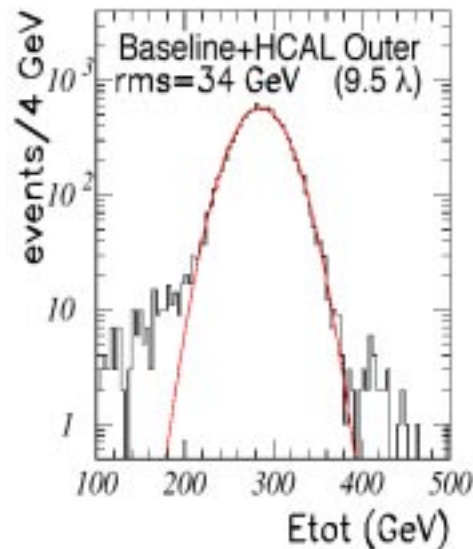


Figure 78: Distribution of reconstructed energy of 300 GeV pions [8, HCAL TDR].

As an example in CMS this is done by introducing multiple longitudinal samplings. Reading out separately the first scintillator plate, placed behind the e.m. calorimeter, allows a distinction to be made between the cases where an e.m. shower has developed in the crystals (little signal from the first scintillator) and the ones where a hadronic shower has started (signal from the first scintillator). The energy observed in the first scintillator therefore allows a correction to be made. In fact the correction can be somewhat 'hard-wired' by choosing an appropriate thickness for the scintillator. The longitudinal leakage can also be up-weighted by increasing the thickness of the

last scintillator. The measured energy distribution for 300 GeV pions in the CMS baseline is shown in Fig. 78. The tails are kept below a few percent.

The test beam results of the combined calorimetry of ATLAS (LAr ECAL and Fe/Scintillator HCAL) are shown in Fig. 79. The data are compared with results from two simulation codes namely FLUKA and GCALOR. Use is made of three energy-independent corrections for the:

- intercalibration between the e.m. and hadronic calorimeter
- energy lost in the cryostat wall separating the two calorimeters
- non-compensating behaviour of the e.m. calorimeter. A quadratic correction is made.

The above procedure minimizes the fractional energy resolution but results in a systematic underestimation of the reconstructed energy: by 20% at 30 GeV and decreasing to $\approx 10\%$ at 300 GeV. Other weighting methods, which have the effect of simultaneously minimizing the non-linearity and the energy resolution can also be employed.

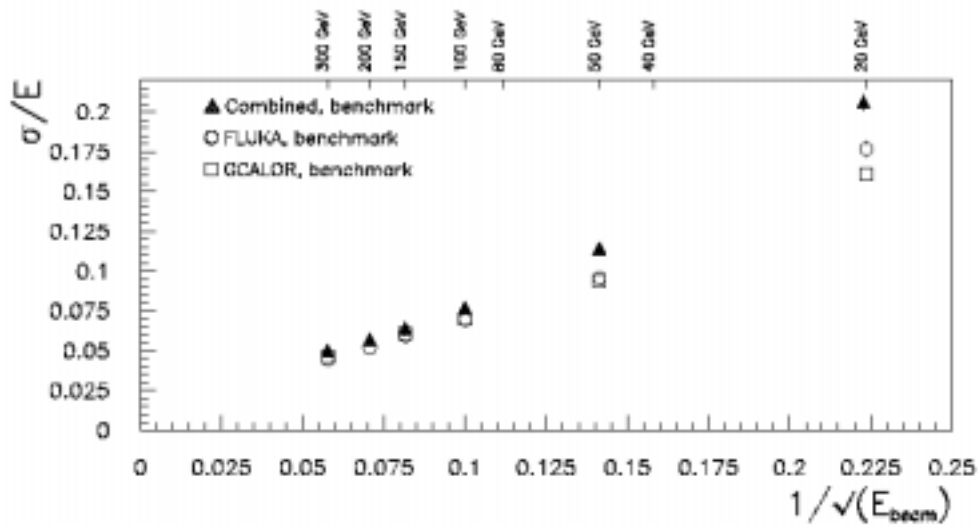


Figure 79: The energy resolution for pions compared with FLUKA and GCALOR simulation codes.

The hadronic cascade simulation codes such as FLUKA, GHEISHA and GCALOR have improved substantially and can now be used with some confidence in the design of hadron calorimeters.

14. TRIGGER AND DATA ACQUISITION SYSTEM

The bunch crossing rate at the LHC is 40 MHz. At high luminosity about 1 billion pairs of protons are interacting every second. The role of the trigger and data acquisition system is to look at (almost) all the bunch crossings, select about one hundred of these containing the most interesting events, collect all the detector information corresponding to these and record on permanent storage for offline analysis. This is a daunting task because the selection process:

- must be highly efficient. Since only data from about 100 crossings can be recorded the vast majority of the events have to be rejected. However none of the few expected rare events should be missed.
- should not introduce any bias.
- should cause as little deadtime as possible.
- must use data from the same crossing for all sub-detectors. This requires synchronisation of millions of channels.

- needs an information super-highway as the 20 or so interactions every 25 ns lead to the generation of 40,000 Gbits/s. The data flow has to be reduced as quickly as possible by high selectivity.

- is carried out in real time i.e. one cannot go back and recover lost events. It is essential to monitor the selection process.

A typical trigger and data acquisition system consists of four parts : the detector electronics, the calorimeter and muon first level trigger processors, the readout network and an on-line event filter system. As an example the functional view of the CMS system is shown in Fig.80

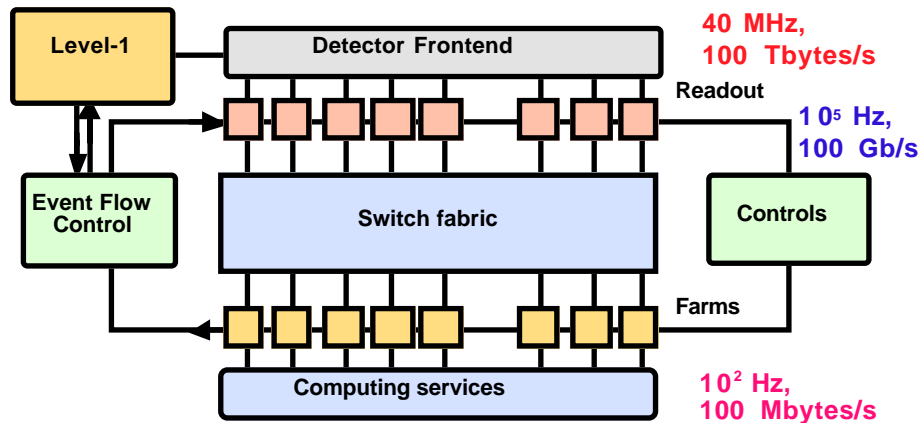


Figure 80: The functional view of the CMS trigger and data acquisition system.

The Level-1 Trigger System is required to reduce the bunch-crossing rate of 40 MHz to an event rate of 100 kHz. Upon receipt of a Level-1 trigger the data from the pipelines are first transferred to ‘derandomizing’ memories that can accept the very high instantaneous input rate (the Level-1 can accept several events within the space of \approx ten crossings even though the average rate is much lower). These memories are emptied into readout buffers: usually many individual channels are multiplexed over a single readout link. After further signal processing (e.g. digitisation, deconvolution) zero suppression and/or data compression takes place before the reduced amount of data are placed in dual-port memories for access by the DAQ system. Each physics event (\approx 1 Mbytes large) is contained in about 500 front-end *Readout* buffers. To further analyse the event it is necessary to transfer the data from the 500 Readout units to a single processor running the appropriate physics selection algorithm. The input rate of 100 kHz is thus reduced to the 100 Hz of sustainable physics. The ‘event building’ is performed using a data switch, somewhat similar to a telephone exchange; in fact use will certainly be made of switching technologies from the telecommunications industry. The most important elements, and also the most difficult ones to develop, are the front-end buffers, the switch that will connect these memories to the processor farm and the physics selection selection algorithms.

14.1 Level-1 Trigger

Since the detector data are not all promptly available and the selection process is highly complex, it is carried out by successive approximations called trigger levels.

The search for new physics involves the study of hard interactions. Hence the first trigger step (Level-1) looks for any one or a combination of the following entities: high transverse momentum muons, high transverse energy photons, electrons, or jets, significant missing E_t (to find neutrinos). Track stubs in the muon system or energy deposits in the calorimeters are used to create the so-called trigger objects: isolated e.m. clusters, muons, jets, E_t^{miss} etc. The selection is

based on e.m. and/or hadronic clusters and/or muons with transverse energy and momentum above certain pre-loaded thresholds in trigger processors.

The time required to make the Level-1 decision is between 2 to 3 μs . Most of the time is taken by propagation delays on cables between the detector and the underground counting room where the trigger logic is housed. There is not much time to combine information from sub-detectors and only elementary operations with elementary conditions are possible. Reduced granularity (e.g. information from groups of 25 crystals are combined to form one trigger-tower, of typical size $\Delta\eta \times \Delta\phi \approx 0.1 \times 0.1$, in the case of CMS ECAL) and reduced resolution (e.g. 8-bits instead of full 12-bit information is used for energy in the CMS ECAL trigger towers) data are used to form trigger objects.

An example from ATLAS of how an isolated e.m. calorimeter is selected is shown in Fig. 81. There are some 4000 ECAL and geometrically matching 4000 HCAL trigger towers each giving an 8-bit value every 25 ns.

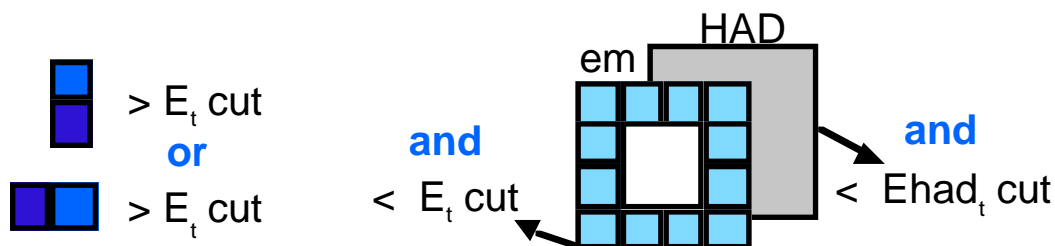


Figure 81: Selection of isolated e.m. clusters at Level-1 Trigger in ATLAS

The Level-1 selection proceeds as follows: the energy in a ‘hot’ trigger tower is combined with that from the hottest one in either η or ϕ (Fig. 81) the transverse energy of the sum should be greater than some E_t^{cut} , an isolation condition is applied. For an isolated cluster the transverse energy in the 12 towers surrounding the central 2x2 has to be smaller than E_t^{em} . Furthermore the transverse energy leaking into the HCAL (sum energy in the 16 HCAL towers behind) for an e.m. shower has to be smaller than E_t^{had} .

The above algorithm is applied for each of the 4000 window positions representing a massive computing task. Pipelined and parallel processing is employed. Pipelined processing means that the logic is organised in a chain of operations to be performed one after the other for each crossing. Each processing element in the chain performs its function in 25 ns and passes its result to the next element in the chain. Data corresponding to successive bunch crossings follow each other down the processing ‘pipe’. Parallel processing means that many processing elements act in parallel, for example performing the same operations on different data.

The above algorithms can be extended to triggers on taus or jets. For a τ -trigger, the vertical and horizontal sum transverse energy of the two (2x1) or four (2x2) trigger towers in both ECAL and HCAL should be greater than some cut E_t^{cut} . The isolation is applied only in the ECAL as above. For jet triggers the transverse energy is summed in 4x4 trigger towers in both the ECAL and the HCAL. A sliding window can be employed centred on blocks of 2x2 towers.

For muon triggering ‘roads’ are defined from one station to the next. The width of the road depends on the desired p_t threshold. The calculation allows for magnetic deflection and multiple scattering. An example, from CMS, is shown in Fig.82.

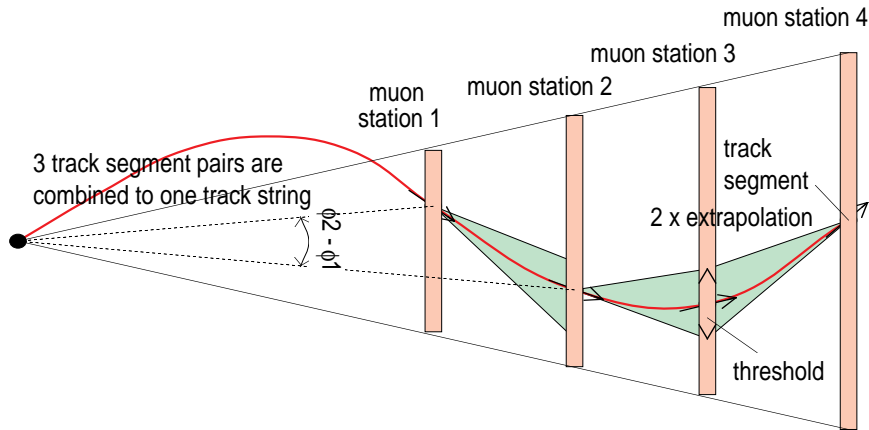


Figure 82: Schematic of Muon Level-1 trigger in CMS using 'roads'.

Each of the barrel stations consist of two sets of 4 planes of drift tubes measuring the r - ϕ coordinate. The two sets are separated by about 20 cm and hence a 'primitive' giving the track direction can be created. This is used to project to the next station and if a consistent primitive is found there the process is continued. This allows muon finding and a yes or no answer for the trigger.

The trigger rates for e.m clusters for ATLAS and muons for CMS are illustrated in Figs. 83 and 84. The efficiency curves are also shown. Clearly for the lowest possible rate the turn-on in the efficiency curve should be as steep as possible (ideally a step function).

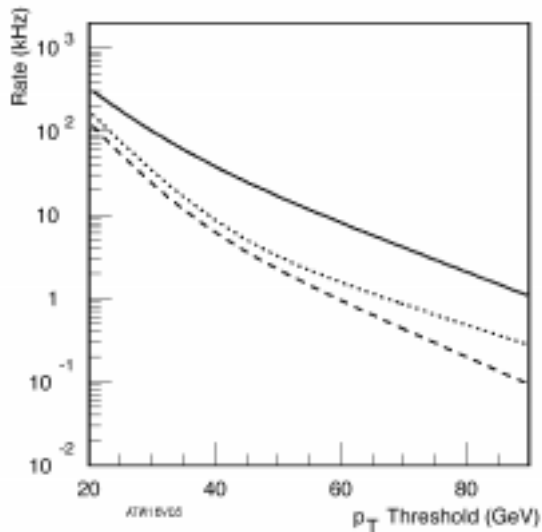


Figure 83: Inclusive electron trigger rate at high luminosity ($10^{34} \text{ cm}^{-2} \text{ s}^{-1}$), without isolation (solid), requiring only hadronic isolation (dotted) and requiring both electromagnetic and hadronic isolation (dashed).

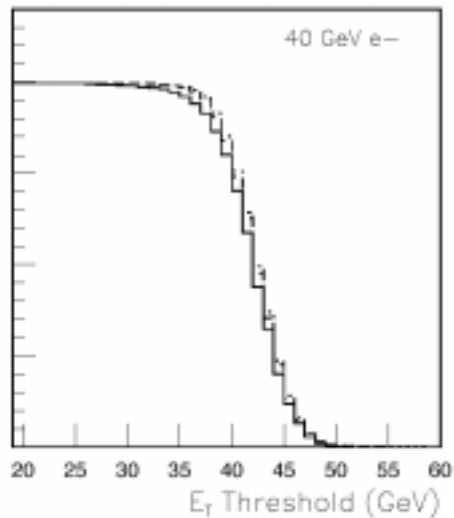


Figure 84: Trigger efficiency curve for 40 GeV E_T electrons at high luminosity ($10^{34} \text{ cm}^{-2} \text{ s}^{-1}$).

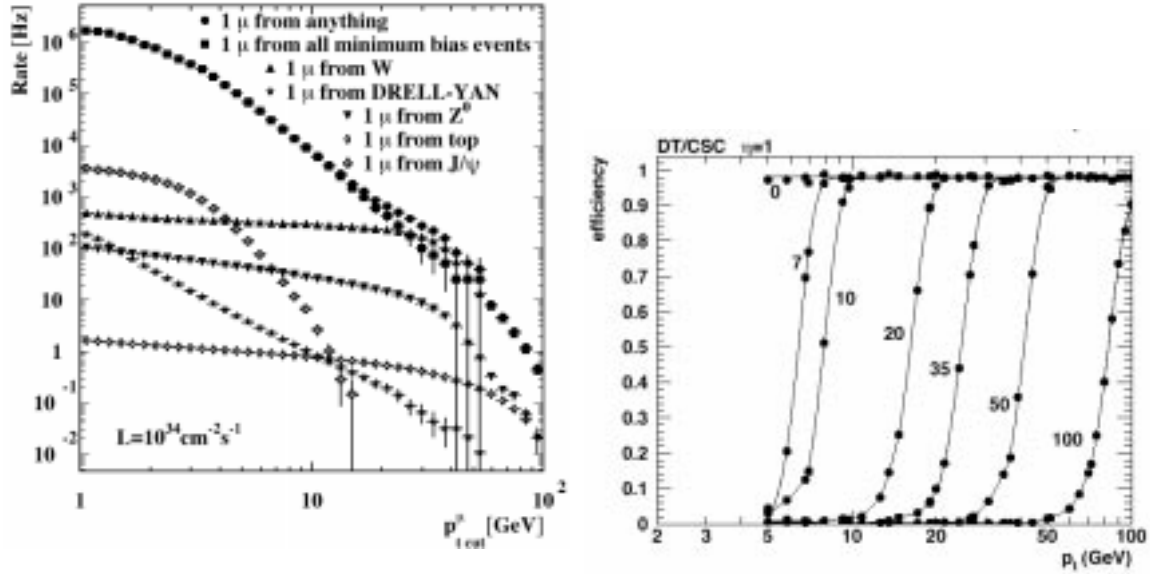


Figure 84:

For events within the geometric acceptance of the detector the trigger ‘cocktail’ and the rates at high luminosity ($10^{34} \text{ cm}^{-2} \text{ s}^{-1}$) for CMS are detailed in Table. 7. The physics efficiency (for events within the geometric acceptance of the detector) using these cuts for some example channels is as follows: $H(80 \text{ GeV}) \rightarrow \gamma\gamma - \approx 99\%$, $H(150 \text{ GeV}) \rightarrow 4l - \approx 100\%$, $pp \rightarrow t\bar{t}b\bar{b} \rightarrow eX - \approx 90\%$, $\Sigma \text{SUSY} - \approx 85\%$. The cumulative rate of 30 kHz is predicted. However a margin for error is taken and both ATLAS and CMS are designing their systems to handle at least 100 kHz of input rate.

Table 7: Trigger rates in CMS running at high luminosity for 90% efficiency of selection

Type Calo.	E_t^{cut} GeV	Individ. kHz	Increm. kHz	Type Muon	E_t^{cut} GeV	Individ. kHz	Increm. kHz
Sum E_t	400	0.3	0.3	1μ	20	7.8	7.8
E_t^{miss}	200		0.9	2μ	4	1.6	9.2
1 e.m.	33		5.3	$1\mu+1e.m$	4/8	5.5	14.4
2 e.m.	20		1.3				
1 jet	140		1.0	$1\mu+1 \text{ jet}$	4/40?	0.3	14.4
Multijets	various		3.0	$1m+ E_t^{\text{miss}}$	4/60?	1.0	15.3
1 em+1jet	14/50		0.3				
Cumul.			12.1	Cumul.		≈ 15	

Numbers need checking

14.2 Higher Level Triggers and Data Acquisition

The output rate from the Level-1 trigger has to be reduced to 100 Hz by using more complex algorithms. In CMS all this will be carried out in a processing farm. The farm will consist of about 500 computers with the capability of 1000Gips (Giga instructions per second). The aim is to initially bring to the farm only the full granularity and full-precision calorimeter and/or muon system data. The Level-1 trigger will point to the region of interest (ROI) for further analysis. A

factor of 5-10 reduction in rate is to be achieved by refining the energy or momentum measurement and applying a sharper cut. The isolation condition is also refined. The tracker data corresponding to the ROI may then be pulled into the processor through the switch. For example matching e.m. clusters or track segments in the muon system to tracks in the ROI, both in space and in energy/momentum, will enable the rate to be further reduced. At each level of refinement more data are brought into the processor but for fewer and fewer events. The final step will use the full event data and almost full event reconstruction and physics analysis will be carried out. In this way the bandwidth needed in the switch can be minimized. Nevertheless a switch network of about 500 Gbits/s is required. It is estimated that the data rate that will be handled by the LHC experiment event builders is equivalent to the data rate exchanged by World Telecom today (1998).

15. CONCLUSION

Much R&D has been carried out during the last 8 years to develop detectors that could cope with the harsh conditions anticipated in the pp LHC experiments. These detectors are not just bigger versions of the current detectors but are substantially different, innovative and at the frontier of technology. Improvements in techniques used in particle detection have always been essential to explore uncharted territory. The ATLAS and CMS detectors should be capable of discovering whatever Nature has in store at the TeV energy scale.

ACKNOWLEDGEMENTS

I would like to thank Nick Ellis for the invitation to give these lectures. I would also like to sincerely acknowledge the unknowing help from many colleagues (S. Cittolin, N. Ellis, D. Green, G. Hall, C. Fabjan, D. Fournier, C. Joram, C. Seez, P. Sphicas, to name a few). I have liberally borrowed from their presentations and write-ups in various schools and conferences. I would like to thank Guy Martin for drawing many plots in the lectures and this write-up.

REFERENCES

- [1] Techniques for Nuclear and Particle Physics Experiments, W. R. Leo, Springer, 1994.
- [2] Single Particle Detection and Measurement, R. S. Gilmore, Taylor and Francis, 1992.
- [3] Physics of Particle Detectors, D. Green, Cambridge Univ. Press, to be published.
- [4] Rev. of Particle Physics, C. Caso et al., Euro Phys. J. C3(1998)1, <http://pdg.lbl.gov/>
- [5] D. Fournier and L. Serin, Experimental Techniques, 1995 European School of HEP, CERN 96-04, 1996.
- [6] N. Ellis and T. S. Virdee, Experimental Challenges in High Luminosity Collider Physics, Ann. Rev. Nucl. Part. Sci. 44(1994)609.
- [7] C. Joram, Particle Detectors, CERN Academic Training Programme, 1998.
- [8] G. Hall, Electronics at LHC, CERN Academic Training Programme, 1998.
- [9] P. Sphicas, Trigger/DAQ at LHC, CERN Academic Training Programme, 1998.
- [10] ATLAS Technical Design Reports, <http://atlasinfo.cern.ch/Atlas/GROUPS/Notes.html>
- [11] CMS Technical Design Reports, <http://cmsdoc.cern.ch/LHCC.html>
- [12] R. Wigmans, Ann. Rev. Nucl. Sci. and Part. Sci., **41**(1991)133.
- [13] D. Barney, Private communication.

- [14] D. Acosta et al., SPACAL, Nucl. Instr. and Meth., **A294**(1990)193.
- [15] D. Groom, To appear in Proc. of Intl. Conf. on Calorimetry in High Energy Physics, Tucson, 1998.
- [16] R. Wigmans, Nucl. Instr. and Meth., **A259**(1987)389.
- [17] G. Drews et al., Nucl. Instr. and Meth., **A290**(1990)335, and H. Tiecke (for ZEUS Calorimeter Group), Nucl. Instr. and Meth., **A277**(1989)42.
- [18] A. Beretvas et al., CMS TN/94-326 (1994).
- [19] R. Apsimon et al., Nucl. Instr. and Meth., **A305**(1991)331.
- [20] ALICE Technical Design Reports, <http://www1.cern.ch/ALICE/documents.html>
- [21] M. Hauschild et al., Nuc. Instr. and Met., **A314**(1992)74.
- [22] BaBar
- [23] J. Seguinot and T. Ypsilantis, Nucl. Instr. and Meths., **142** (1977)377.
- [25] F. Sauli, Techniques and Concepts of High Energy Physics, ed T. Ferbel, Vol. 11(1983)301, New York: Plenum.
- [26] D. Decamp et al., Nucl. Instr. and Meths. **A294**(1990)121.
- [27] A. Oed, Nucl. Instr. and Meths. **A263**(1988)351.
- [28] F. Sauli, Nucl. Instr. and Meths. **A386**(1997)531.
- [29] A. Annenkov et al., CMS NOTE 1998/041 and references therein.
- [30] G. Gratta et al., Ann. Rev. Nucl. Part. Sci. **44**(1994)453.
- [31] E. Auffray et al., Nucl. Instr. and Meth., **A412**(1998)223.
- [32] M. A. Akrawy et al., Nucl. Instr. and Meth., **A290**(1990)76.
- [33] D. Fournier, Nucl. Instr. and Meth., **A367**(1995)5.
- [34] D. Schinzel, Proc. of Intl. Wire Chamber Conference, Vienna, 1998.

JINR PROGRAMME IN HIGH ENERGY PHYSICS

A. Sissakian

Joint Institute for Nuclear Research, Dubna, Russia

Abstract

Information about the current status of Joint Institute for Nuclear Research and its High Energy Physics programme is presented.

1. GENERAL INFORMATION ABOUT JINR.

1.1. Historical background.

The Joint Institute for Nuclear Research (JINR) is situated in Dubna, a small town located about 120 km north of Moscow. This town appeared in the late 1940s in hard times after the end of World War II as part of the USSR Nuclear Defense Programme on the initiative of Igor Kurchatov, an outstanding Soviet scientist who was responsible for the Programme at that time and who also had a perfect understanding of the importance of fundamental research.

JINR — as the international centre — was founded in March of 1956 (in Khrushchev times) on mutual agreement of the governments of Albania, Bulgaria, China, Czechoslovakia, Hungary, the Socialist Republic of Vietnam, the German Democratic Republic, the Democratic People's Republic of Korea, Mongolia, Poland, Romania, and the Soviet Union. The USSR handed over to the new institution two laboratories located in Dubna. After the establishment of JINR — an international organization for fundamental research in nuclear science — Dubna became an open town.

It is necessary also to mention that the formation of various scientific fields of research at JINR was initiated by a number of some outstanding scientists of the Institute's Member States including the first directors Dmitri Blokhintsev and Nikolai Bogoliubov.

1.2. JINR's Charter; membership and internal organization.

The Joint Institute was created in order to unify intellectual and material potential of Member States to study fundamental properties of matter.

The Charter of JINR was adopted in 1956, later on it was revised and newly adopted in 1992.

In accordance with the Charter the activity of the Institute is realized on the basis of its openness, mutual and equal cooperation for all interested parties to participate in research.

The main governing body of our Institute is the Committee of the Plenipotentiaries of the Member-State Governments. Each State contributes to the JINR budget.

The JINR structure is illustrated on Figure 1.

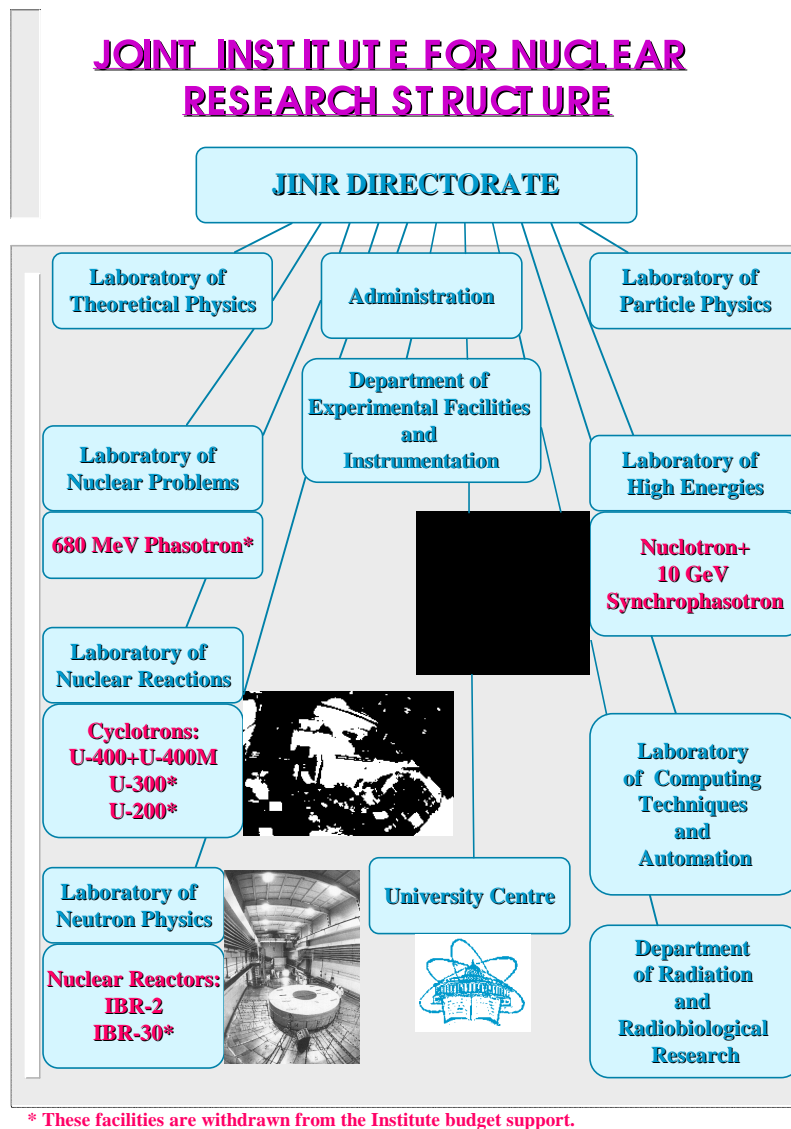


Figure 1. Joint Institute for Nuclear Research Structure.

In accordance with new Charter more than 1/3 of Scientific Council now consists of non-member state scientists (from CERN, France, Germany, Italy, USA). H. Shopper, P. Spillantini, G. Piragino, F. Legar, B. Peyod, F. Dydak, C. Detraz, G. Trilling, M. Della Negra, L. Masperi, et al. are among the members of SC.

The aim of the Institute is:

- to carry out theoretical and experimental investigations on adopted scientific topics;
- to organize the exchange of scientists in carrying out research, of ideas and information by publishing scientific papers, organizing conferences, symposia etc.;
- to promote the development of intellectual and professional capabilities of scientific personnel;

- to maintain contacts with other national and international scientific organizations and institutions to ensure the stable and mutual cooperation;
- to use the results of investigations of applied character to provide supplementary financial sources for fundamental research by implementing them into industrial, medical and technological developments.

The results of investigations carried out at the Joint Institute for Nuclear Research can be used solely for peaceful purposes to the benefit of mankind.

So until the late 80's Dubna was a centre which unified the efforts of leading research groups of nuclear sciences of the so-called "socialist countries" and the Soviet Union.

After the disintegration of the USSR the membership of JINR underwent the following changes: The majority of East European countries, such as Poland, the Czech and Slovak Republics, Bulgaria, Romania continue to be Member States of our Institute and contribute in the budget. Germany stays as an observer and makes a substantial financial contribution. Most of the former Soviet Union republics which became independent states of the CIS (Common-wealth of Independent States) entered JINR as new members.

I would like here to remind you the words of great Russian writer A. Chekhov who said: "...there is no national science as no national multiplication table. If the science is a national one — this is not a science anymore...". JINR is a perfect illustration of this idea...

The participation in the Institute can be realized in different forms: on the basis of membership, bilateral and multilateral agreements to perform particular scientific programmes. JINR Member States contribute financially to the Institute's activity and have equal rights in its management.

JINR has at present 18 Member States (Table 1).

Table 1
JINR Member States

Armenia	Moldova
Azerbaijan	Mongolia
Belarus	Poland
Bulgaria	Romania
Cuba	Russian Federation
Czech Republic	Slovak Republic
Georgia	Ukraine
Kazakhstan	Uzbekistan
D.P. Republic of Korea	Vietnam

JINR has special cooperation agreements concluded on 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);

Recently the Agreements were signed with UNESCO and CLAF (Latin America Centre on Physics).

Among the major partners with whom JINR has long-term cooperation 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 USA.

JINR is also an associated member of EPS, has observer and/or members in ICFA, ECFA, IUPAP and other international unions. The latest political changes in Eastern Europe and especially in Russia have been making JINR more and more open. New collaborating countries are welcomed to join JINR.

Today JINR is a large centre with a total staff close to 6000 including services and workshop. Approximately 1100 scientists work in it. Among them there are about 40% from member-states (but Russia).

The internal JINR organization is determined by scientific specialization. There are 7 Laboratories in the Institute:

- Bogoliubov Laboratory of Theoretical Physics (BLTP);
- Laboratory of High Energies (LHE);
- Laboratory of Particle Physics (LPP);
- Laboratory of Nuclear Problems (LNP);
- Flerov Laboratory of Nuclear Reactions (FLNR);
- Frank Laboratory of Neutron Physics (FLNP);
- Laboratory of Computing Techniques and Automation (LCTA).

Each Laboratory (but BLTP) has its own design and construction divisions which develop and manufacture non-standard equipment for particle accelerators, detectors and other experimental facilities. The staff of these divisions totals about 370 engineers, technicians and workers.

A number of associate Experimental Physics Workshops are also part of the Institute.

The personnel of the JINR Experimental Physics Facilities Division totals about 400. It is equipped with everything necessary to manufacture large-sized non-standard facilities, electronics, and has technological lines for constructing detectors for high energy physics. It was there that the main units of JINR's heavy ion cyclotrons U-400 and U-400M were constructed in recent years, as well as the Nuclotron — a new superconducting accelerator for relativistic nuclear physics. It is an excellent result especially taking into account a difficult economic situation in Russia of the last years.

As certainly known well, nuclear science and in particular its frontier Particle Physics or High Energy Physics is rather an expensive field of research. Of course it is very well understood now that deep fundamental studies have always resulted in huge technological benefits. And the great discoveries of the past such as electricity, magnetism, etc. have never been paid off, and all the investments to fundamental science in the whole world is still a negligible part compared to the benefit that mankind got from it. It would be fare to say that mankind is eternally indebted to fundamental science.

Now few words about JINR's international cooperation.

Despite the present hard economic and financial situation in most of JINR's Member States, which of course has greatly affected the ongoing research programme of the Institute, many scientific groups from Dubna continue to participate in largest projects of world's major centres.

The intensity of JINR international cooperation events can be demonstrated by the following:

- approximately 1200 our specialists participated in 1997 in joint experiments and international conferences in outside JINR;
- more than 1000 scientists from collaborating laboratories and centres visited Dubna annually;

- each year JINR organizes about 50 conferences, workshops and other scientific meetings (among them the International Conference for HEP accelerators — HEACC-98 — was held in Dubna on September, 1998);
- Together with CERN JINR participated in the organization of the European School on HEP annually (formerly JINR-CERN School of Physics since 1970);
- JINR scientists participated in more than 150 international conferences held world-wide annually.

1.3. On forthcoming JINR reforming.

Since the beginning of 1998 JINR started a programme of new reforms to be accomplished within 3-4 years. The urgency of reforming JINR is dictated by a number of factors, and first of all by the difficult economic and financial situation, that will most probably continue in the future. Unfortunately, despite the decisions of the Committee of Plenipotentiaries, many of the JINR Member States fail to fulfill their financial obligations toward the Institute. During the past several years, the Directorate had to make extraordinary efforts to secure adequate budgetary funding, but the actual implementation of budget was about 60-65% of the approved annual one (37.5 M\$ per year). It is quite clear that such low level of financing urges the reforms in order to concentrate the financial and human resources only on the most important directions of JINR's activity. One of the timely and important steps toward this should be a reasonable solution to enhance centralization of the Institute management and reduce the excessive number of administrative functions and services in all the Laboratories.

The reforms are proposed to be done in two stages. The main tasks to be solved at the first stage are as follows:

- Centralized operation and management of the JINR basic facilities;
- Optimization and reduction of infrastructure;
- New staff policy.

The Directorate believes that a guaranteed stable operation of the JINR facilities should be achieved through a centralized management of the facilities and by a policy of tight economy of resources. Now we can mention with satisfaction that the reforms have already yielded the first positive results in this area.

The reform in the field of the overgrown infrastructure of JINR is directed to a maximum reduction of doubling services and transfer of a part of the buildings for rent, including by JINR self-supported divisions.

The proposed reforms envisage a general reduction of the Institute's personnel with a tempo not less than 25% during 3 years. The reduction is aimed at:

- a higher inflow of young scientists to JINR,
- an increase in salaries.

The reduction is regarded by the Directorate as a delicate question requiring measures for social protection of the dismissed personnel, and will be done in cooperation with the municipal authorities of the town Dubna.

The second stage of reforms will be connected with the field of scientific research.

2. JINR IS A MAJOR PARTNER OF WORLD'S HEP LABORATORIES.

2.1. JINR's international cooperation in high energy physics.

Broad international cooperation is one of the most important principles of the JINR activity. Almost all investigations are carried out in a close collaboration with JINR member-state scientific centres as well as international and national institutions and laboratories of the world. The most effective cooperation is realized with such institutes of Russia as IHEP (Protvino), Kurchatov Institute in Moscow, Institute of Nuclear Physics in Gatchina near St. Petersburg, ITEP (Moscow), INR (Troitsk), Lebedev Institute of Physics (Moscow), Moscow State University et al.

A fruitful scientific cooperation is being held with CERN, especially in the last years, as well as with many physics laboratories in USA, France, Germany, Italy, and other countries. Cooperation with Nuclear Scientific Centre of Peking University (China) is being developed, a Protocol on collaboration has been signed between JINR and the Institute of Modern Physics of Academia Sinica. The JINR Directorate is ready to maintain constant and long-term contacts with laboratories of other countries as well.

The Figure 2 shows a current status of JINR International Cooperation in HEP.

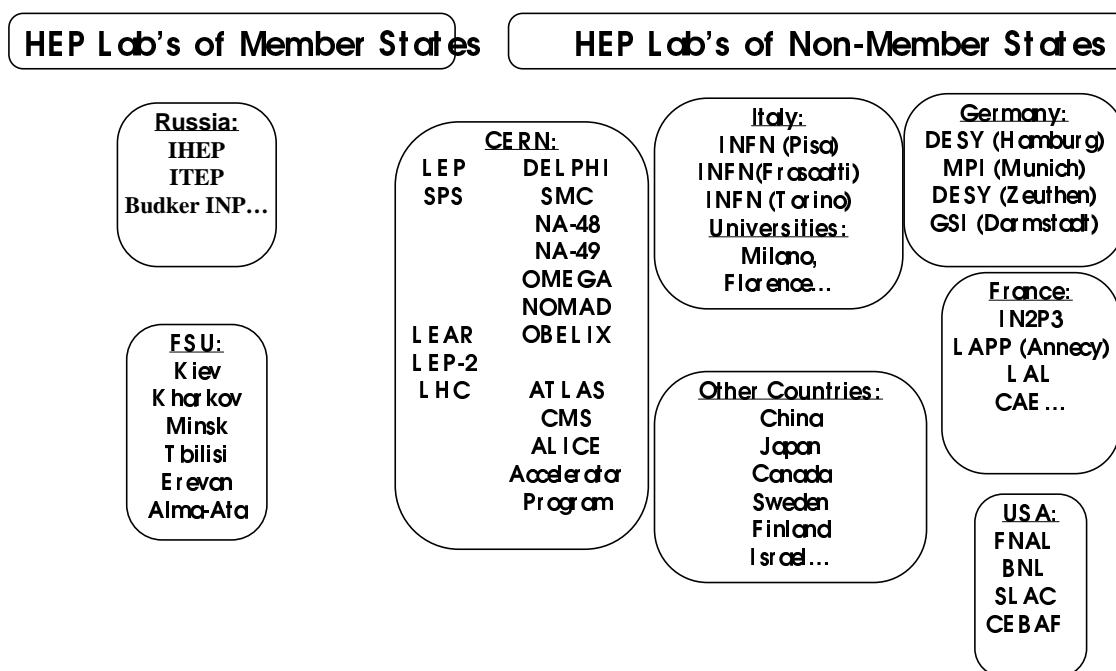


Figure 2. JINR International Cooperation in HEP.

2.2. Cooperation with CERN.

Dubna physicists are involved in a big part of the CERN experimental programme. The general Agreement between JINR and CERN was signed in 1992, but cooperation between two international organizations has a very long history.

The Table 2 shows the experimental projects in which Dubna research groups are involved at CERN.

Table 2
CERN Experiments where Dubna is involved

Project	Location	a) main goals b) JINR contribution
NA48	SPS CERN	a) Highest precision direct CP-violation searching in neutral kaons decays. b) Subsystems design & construction; data taking runs. Data analysis.
NA49	SPS CERN	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 , \bar{p} , d and \bar{d} . Data analysis.
NOMAD	SPS CERN	a) Search for $\nu_\mu \rightarrow \nu_\tau$ and $\nu_\mu \rightarrow \nu_e$ oscillations. b) Data taking & analysis; new proposal preparation.
COMPASS	SPS CERN	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 Calorimeter, Muon Detector, large area track chambers.
DIRAC	PS CERN	a) 5% accuracy test of low en. QCD by 10% precision ($\pi^+\pi^-$) atom life time measurement. b) JINR proposed experiment; Drift Chambers; secondary particles channel; trigger development; MC simulation & software. Data taking RUNs, data processing & analysis.
DELPHI	LEP CERN	a) Precision measurement of $m(W)$, search for new particles, etc. b) Continue to maintain Hadron Calorimeter and Surround Muon Chambers; physics analysis.
ATLAS	LHC CERN	a) General purpose $p\bar{p}$ -experiment. b) Subsystems: calorimeters; muon, transition radiate detection; radiate hardness tests; physics software & simulation; trigger and data acquisition.
CMS	LHC CERN	a) General purpose $p\bar{p}$ -experiment. b) Subsystems: forward mu-station, hadron END CAP Calorimeter; e/m cal preshower; simulation. Heavy ion physics.
ALICE	LHC CERN	a) Heavy ions 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 Accelerator Complex Elements	LHC CERN	a) Development & construction of LHC beams formation & control system elements. b) Design & construction of transverse oscillation damping system. Simulation & prototypes study.

The Figure 3 gives a picture on the scale of Dubna contribution to some CERN Projects.

The very first MODULE 0 for the Atlas Hadron Barrel Tile Calorimeter was successfully assembled in Dubna on the 6 meters in diameter rotating table of the milling-boring shop-machine.

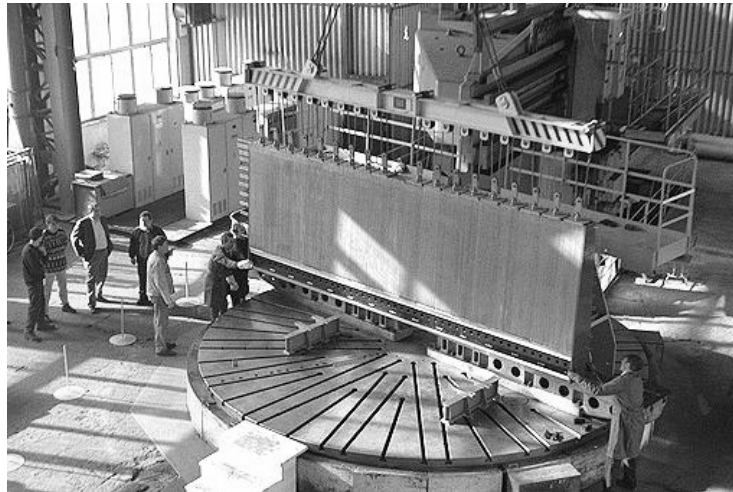


Figure 3. MODULE 0 in Dubna.

Russian, Slovakian, Belarus industries were involved in manufacturing of the module. JINR, IHEP, ANL, PISA, Barcelona, Prague contributed a lot to this activity.

2.3. Cooperation with IHEP (Protvino).

JINR scientists are carrying out experiments at IHEP's U-70 proton synchrotron with the help of such set-ups as Tagged Neutrinos, EXCHARM, HYPERON, Neutrino Detector, and others. The most essential features of our scientific programme on 70 GeV accelerator of Protvino are summarized in this Table.

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.
TARGGED NEUTRINO COMPLEX	Verification of the universal features of weak interactions; search for rare decays in neutrino interactions; search for CP-violation in K-decays.
PROZA-DIBARION	Measurements of polarization parameters of πN and NN interactions.

3. JINR'S SCIENTIFIC POTENTIAL.

3.1. Technical possibilities of JINR Laboratories for HEP experiments.

Technical possibilities of JINR Laboratories for HEP experiments are as follows:

track and semiconductor detectors (Laboratory of Particle Physics), superconducting magnetic systems, polarized targets; cryogenic systems in Laboratory of High Energies; wire proportional chambers, pressurized drift tubes, electromagnetic and hadron calorimeters, radiation-proof big-sized scintillation counters in Lab. of Nuclear Problems, radiation tests by fast neutrons in Laboratory of Neutron Physics, etc.

3.2. Dubna accelerators and reactors.

The main fields of the Institute's investigations 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.

The major facilities of the Institute for experimental investigations are the nuclotron, synchrophasotron, phasotron, U-400 and U-400M cyclotrons, IBR-2 and IBR-30 neutron reactors.

The new superconducting accelerator Nuclotron was put into operation three years ago. It will enable to perform a wide programme of research in relativistic nuclear physics. The injection complex is being developed consisting of a buster, linac and ion sources. This complex will allow one to accelerate nuclei from hydrogen to uranium with the intensity from 10^{13} to 10^8 particles per pulse respectively in the energy range of 6-7 GeV per nucleon. Polarized deuteron beams are foreseen.

Research programme on NUCLOTRON is executing and I give one recent new result on the determination by studying of cumulative protons production with $\theta_{pp}=109^\circ$. Experiment was performed on internal deuteron 2 GeV/nucleon beam of NUCLOTRON with the carbon target on orbit. It was found that in the cumulative protons production the transversal dimension (r_0) of the interaction region for incoming deuterons is noticeably larger than for incoming protons (Figure 4).

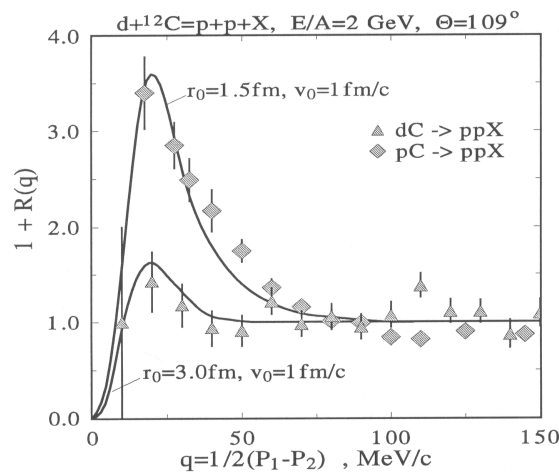
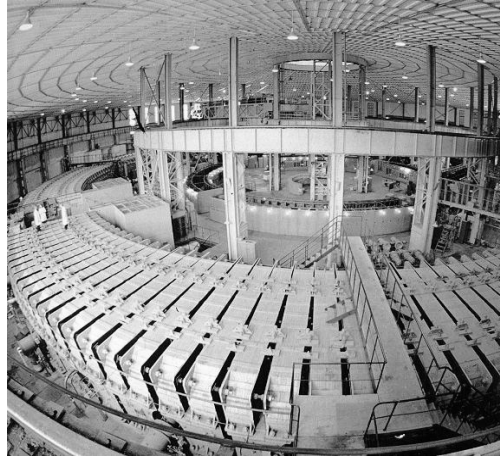


Figure 4. The Studying of Cumulative Protons Production on Nuclotron.

Synchrotron is an accelerator of 10 *GeV* protons put into operation in 1957. In the 70's the acceleration of nuclei heavier than hydrogen, that is deuterium, lithium, carbon, fluorine and magnesium, was accomplished in the broad energy spectrum from a few hundred *MeV* to 4.5 *GeV* per nucleon. Average densities of beams range from 10^4 to 10^{11} *ion/cm²s* depending on the atomic number of accelerated nuclei and experimental requirements (Figure 5).



Particles obtained	Number of particles per cycle	Energy of particles
protons	4×10^{12}	8-10 GeV
deutrons	1×10^{12}	3.6 GeV/nucleon
${}^3\text{He}$	2×10^{10}	
${}^4\text{He}$	5×10^{10}	
${}^7\text{Li}$	2×10^9	
${}^{12}\text{C}$	1×10^9	
${}^{16}\text{O}$	5×10^7	
${}^{20}\text{Ne}$	1×10^4	
${}^{24}\text{Mg}$	5×10^6	
${}^{28}\text{Si}$	3×10^4	
${}^{32}\text{S}$	3×10^3	

Figure 5. Synchrophasotron — accelerator of polarized protons and deuterons.

Synchrophasotron beams attracts physicists from all the world. Wide international collaboration SPHERE has recently performed the study of the nuclear matter at short distances in experiments with polarized deuteron beams to separate contribution from S and D components in the deuteron wave function.

In order to clarify the reactions mechanism and the structure of non-nucleon degrees of freedom the new experiment with the polarized deuteron fragmentation into cumulative hadrons has been performed. The observed difference of MODEL — to — experiment data is especially large for $K=0.2 \text{ GeV}/c$ where it would be natural to expect the manifestation of non-nucleon degrees of freedom in the deuteron wave function (Figure 6):

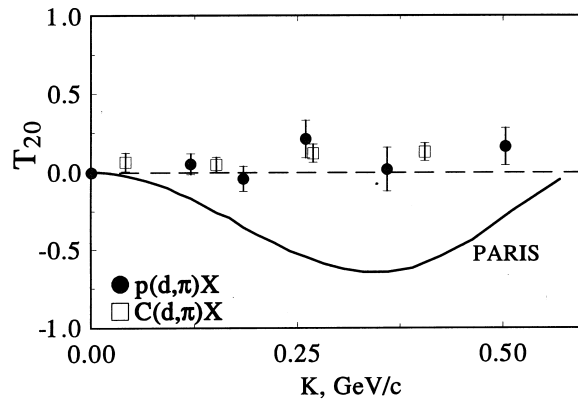


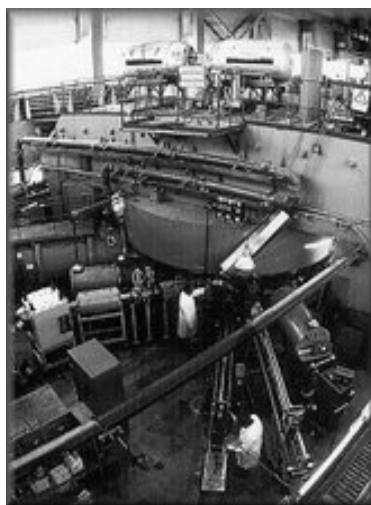
Figure 6. The Measurement of the Tensor Analyzing Power T_{20} in Inclusive Polarized Deuteron Fragmentation into Pion at Zero Angle.

Phasotron is an accelerator of 680 MeV protons. It was put into operation in 1949, reconstructed in 1984 and represent the oldest basic facility of JINR. 10 beam channels are available at this machine which are used to carry out experiments with pions, muons, neutrons and protons. 5 secondary beams are designed to carry out medical investigations. The intensity of the extracted proton beam is 2 mA . The research programme at the phasotron includes low energy proton-nuclei interactions, mu-catalysis physics. Radiochemistry and applied physics (including cancer therapy, proton, pion, meson beam, radioisotopes etc.) are currently the noticeable part of phasotron programme.

U-400 is a heavy ion isochronous cyclotron constructed in 1978. The range of accelerated nuclei is $(A/Z)=4 - 20$, energy is $650 Z^2/AMeV$, beam intensity is $10^{12} - 10^{14}$ ion/s .

U-400M is an isochronous cyclotron put into operation in 1991-92 to accelerate heavy ions. It produces ion beams of atomic masses from 4 to 100 and maximum energy up to 25 $MeV/nucleon$ and beam intensity from 10^{12} up to 10^{14} ion/s . It is designed to operate in the cyclotron U-400+U-400M complex and allows to accelerate ions from hydrogen up to uranium in the range of energy 120 — 20 MeV per nucleon respectively with the average beam intensity of 4 — $10^{13} - 10^{11}$ ion/s . The main areas of research include synthesis of new elements, investigation in the chemistry of new transferium elements, and studies of the radioactive decay of heavy nuclei far from beta-stability.

The Figure 7 shows the cyclotrons U-400 and U-400M.

U-400**U-400M**

Cyclotron	Particles obtained	Energy of Particles	Beam Intensity ions/sec
U-400	ions B÷Zr	650 Z ² /AMeV	10 ¹² —10 ¹⁴
U-400M	ions B÷Zr	120—20 MeV per nucleon	4×10 ¹³ —10 ¹¹

Figure 7. Accelerators U-400 and U-400M.

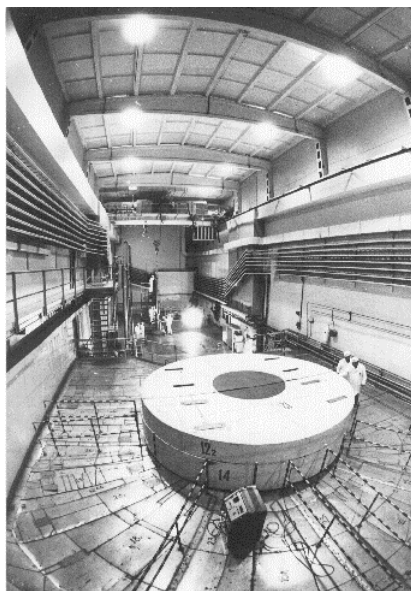
On the U-400M accelerator there was recently obtained quite a new result. In the study of the reaction of 2 neutrons transfer from ⁶He to ⁴He we have detected — for the first time — the di-neutron bounded state in the nuclei field. Secondary ⁶He was registered in CMS angular interval of 10 — 160°. The observed effect is the increase of ⁶He yield at the angles $\Theta > 100^\circ$.

The experiment is expected to be continued with another configuration which provides a detection of low-energy ⁶He in angular range more than 160 degrees.

Another very interesting achievement is that during 1998 experiment there was synthesized the most heavy — of known — nuclei ²⁸³Z₁₁₂ in the reaction chain $U + ^{48}\text{Ca} \rightarrow (\text{compound}) \rightarrow 3n + ^{283}\text{Z}_{112}$. The searches of 114th element of the Mendeleev periodic table are in progress. This element expectedly has a long time of life. Discovery of such an element would manifest the existence of the “island of stability”.

IBR-2 is a pulsed reactor with average thermal power 2 MW and peak power in pulse 1500 MW. The wide programme of the condensed matter studies was performed on this reactor. Power pulses with a frequency of 5 Hz are generated by reactivity modulators (Fig. 8).

IBR-30+LUE-40 is a pulsed neutron source consisting of old pulsed reactor IBR-30 and electron 40 MeV linac LUE-40. The average heat power of the reactor is 10 kW, instant pulse power is 150 MW. It generates neutron pulses with frequency about or less than 100 Hz and duration of 4.5 μs. The total neutron yield is 5×10^{14} n/s, the flux of fast neutron at surface of active core is about 10^{12} n/cm²s.



Reactor	Particles obtained	Thermal Power	Pulse Power	Frequency of Power Pulses
IBR-2	neutrons	2 MW	1500 MW	5 Hz
IBR-30+	neutrons			
LUE-40	γ-quanta	10 kW	150 MW	≤100 Hz

Figure 8. IBR-2 — pulsed reactor:

3.3. Dubna High Performance Computing Centre.

The very important element of the Dubna infrastructure is the Satellite Space Station. Through this station and Laboratory of Computing Techniques and Automation we develop of interfaces with international computer networks to provide a prompt contact between JINR and other research centres of the world community of scientists.

Regarding the computing facilities, at JINR there is a whole diversity of computational problems in various fields of physics, which need powerful computing resources. They involve problems of theoretical and mathematical physics, solid state physics problems, experimental data processing problems, especially in HEP. Following the world tendency in the field of computing for science and higher schools and the progressive requirements of users, the LCTA

developed a conception of establishing a High Performance Computing Centre (HPCC) at JINR. It should be read as a balanced development of four main components of HPCC, namely:

- Telecommunication systems:
 - External communication channels (INTERNET);
 - High-speed JINR ATM Backbone;
- Systems for powerful computations and mass data processing:
 - General High-performance server;
 - Clusters of workstations of JINR laboratories and experiments;
 - Computing farms (PC-farms);
- Data storage system:
 - File servers system based on AFS;
 - Mass storage system;
 - Information servers and database servers;
- Software support systems:
 - Systems for application creation and maintenance;
 - Visualization systems.

4. DUBNA AS AN EDUCATIONAL CENTRE.

4.1. The University Centre of JINR.

JINR gradually changes from a purely scientific research institution to an international centre in which fundamental science, applied research, and engineering are closely connected with the university education process. In 1991, the University Centre (UC) of Moscow State University, Moscow Engineering Physics Institute, and Moscow Institute of Physics and Technology was established at JINR.

The UC students come from many institutes and universities of Russia, the former Soviet Union, and JINR Member States. Students of 4th and 5th years and graduates are invited to study at the UC for two years.

The students complete here their university education. Classes include not only ordinary courses in physics, but also intensive courses on subjects defined on the basis of JINR research.

Structurally, it takes the form of a new satellite “students” laboratory. Its prototype is the currently working University Centre. This new training function of JINR is supposed to be oriented to international demand.

The Centre offers the following full-time graduate programmes:
Nuclear Physics, Particle physics, Condensed Matter Physics.

In the above three fields, the UC also offers the full-time theoretical physics programmes on the basis of the Bogoliubov Laboratory of Theoretical Physics.

The UC offers also Technical Physics, and Radiobiology.

The full-time educational programme of the University Centre is two years long, though it has also become a practice to accept students for shorter periods, such as one or two-month intense courses on some selected topic. The working language for foreign students is English.

Post-graduate students are also admitted to attend lectures on selected topics and take part in scientific research at the JINR Laboratories.

Students have wide access to the Laboratories of JINR and can work with scientists and staff of the Institute, as well as to study under professors who are eminent in their fields.

Graduate and post-graduate studies at the UC are based immediately on JINR's research conducted at a wide variety of world-renown facilities, for example, heavy ion accelerator U-400, ion beam from the U-400M cyclotron, the nuclotron — a superconducting accelerator of relativistic nuclei, the IBR-30 neutron booster, and the IBR-2 pulsed reactor (which is especially fruitful in condensed matter research).

Special importance is attached to the language education. Russian and English are taught here as a second language.

The UC has a post-graduate training license from the State Committee of Higher Education of Russia. The post-graduate students are trained in the large scale of specialties: Physics of nuclei and elementary particles, Theoretical physics, Charged particle beam physics and accelerator technique, Computational mathematics, Solid state physics, Physical experiment technique, High energy physics...

The International University of Nature, Society, Man (the University President is JINR Director V.G. Kadyshevsky, University Rector is O.L. Kuznetsov, President of the Russian Academy of the Natural Sciences) was opened in 1994 in Dubna. University works in the very tight cooperation with JINR University Centre.

4.2. International contacts of the UC.

International scientific educational contacts have become a regular and well-established UC's activity.

The UC has always kept high profile in the organization and conduction of international scientific schools and training courses. Here are some typical examples.

In 1995, the UC actively participated in the organization of two schools — on theoretical physics (jointly with the Laboratory of Theoretical Physics) and on neutron physics (jointly with the Laboratory of Neutron Physics). In September-October 1995, the International Nuclear Information System (INIS) courses of IAEA were conducted in Dubna, which was largely assisted by the UC. In 1996, the UC and Laboratory of Particle Physics organized jointly the Young Scientist School on Problems of the Charged Particle Acceleration. Within the frames of the cooperation between IAEA and JINR, the 9-week International Regional Post-Graduate Educational Course on Radiation Protection was held in 1996 on the basis of the UC. In 1998, the UC has conducted the International Summer School in memory of Bruno Pontecorvo.

The UC is a participant of the European Mobility Scheme for Physics Students (EMSPS). The European Physical Society has appointed the UC one of the Russian Federation coordinators in the EMSPS.

By the Partnership Agreement between JINR and European Physics Education Network (EUPEN), the UC is included in EUPEN's "Thematic Network" in Physics.

UC also maintains ongoing contacts with CERN in the training of students and young scientists.

A number of graduate students from Western Europe had their specialized practice at JINR's laboratories, which was coordinated and assisted by the UC.

UC also receives student groups from Europe coming here with visits of acquaintance.

5. PLANS FOR FUTURE.

JINR has the following projects for the development of new basic facilities:

IREN (Intense Resonance Neutron Source) is a project aimed at constructing a high-flux pulsed neutron source to carry out investigations with resonance neutrons.

The facility will comprise a modern 200 MeV electron linac and subcritical plutonium booster having neutron multiplication coefficient 30. The pulse rate is 150 Hz, duration 0.4 μ s, total neutron yield $\sim 10^{15}$ n/s. The research programme for the IREN includes investigation of P- and CP-violations in slow neutron interaction with nuclei and other fundamental nuclear physics topics. This project started of realization from 1994. It will be put into operation in 1999.

τ -Factory is a project of a new accelerator complex to comprise an electron-positron collider with an energy of 1.7 — 3 GeV in each beam and an injection system.

The research programme for this complex will include the study of Tau-lepton and τ -neutrino physics, τ -charmonium spectroscopy, CP violation, charmed baryon physics and meson spectroscopy. The design of specialized synchrotron radiation source (8 — 10 GeV, NK-10) is under consideration.

Particularly we investigate the possibility of the creation the Synchrotron Radiation Source at JINR on the base of the NIKHEF (Amsterdam) AmPS machine.

Accumulated scientific experience, available high qualification “human resources” and our realistic estimates of our financial possibilities give the reasonably balanced vision of our nearest perspectives.

JINR’s plans for the nearest future:

- Development of methodical and computing possibilities for participation in experimental programmes of the world’s largest HEP laboratories (CERN, FNAL, IHEP and others).
- IREN construction.
- Development of the injector complex of the Nuclotron.
- Further development of the JINR University Centre.
- The use of JINR’s advanced infrastructure for holding international conferences, meetings and schools.

In conclusion of this part of my talk I would like to express the following opinion:

The combination of JINR’s attractive scientific programme, development of new technologies and our recent initiatives in the educational field make our Institute an interesting and promising partner.

6. THE STATUS OF HEP IN RUSSIA AND FSU-COUNTRIES.

Speaking of the programme in high energy physics in the territory of the former Soviet Union, one has to note the existence of serious economic difficulties. Yet we believe that they are temporary. This region has a rich background of applied and fundamental sciences. The world’s largest proton accelerator in operation in the late 60’s was in Protvino, Russia. The prestige of our HEP physicists in theoretical investigations, in accelerator and detector technologies was always very high.

Among the participants of European and world collaborations one can often see such research centres as

Institute for High Energy Physics (Protvino),
Institute of Theoretical and Experimental Physics (Moscow),
St. Petersburg Institute of Nuclear Physics (Gatchina),
Budker Institute for Nuclear Physics (Novosibirsk),
Institute for Nuclear Research (Troitsk, Moscow),
Moscow State University,
Lebedev Institute of Physics (Moscow),
Kurchatov Institute (Moscow),

Moscow Engineering Physics Institute,
Yerevan Institute of Physics and Yerevan University (Armenia),
Institute of Physics (Azerbaijan),
Belarus State University,
Institute for High Energy Physics (Alma-Ata, Kazakhstan),
Kharkov and Kiev institutes (Ukraine),
Institute of Nuclear Physics (Tashkent, Uzbekistan)

and others.

These Institutes actively participate in HEP research programmes at CERN, DESY, FNAL and some other world scientific centres.

The main accelerating facilities used for research in high energy physics in Russia are:

- the 76 GeV proton synchrotron (IHEP, Protvino, near Serpukhov),
- the 7×7 GeV positron-electron storage rings VEPP-4 (Institute for Nuclear Physics of the Siberian branch of the Russian Academy of Sciences, Novosibirsk),
- the synchrotron for acceleration of protons (10 GeV) and atomic nuclei, and Nuclotron (Joint Institute for Nuclear Research, Dubna),
- the proton synchrotron of the Institute for Theoretical and Experimental Physics (Moscow) accelerating protons up to 9.3 GeV,
and others.

Besides, a number of proton accelerators with energies up to hundreds of MeV, phasotron is also available operating in Russia at the Joint Institute for Nuclear Research (Dubna) and isochronous cyclotron — at St. Petersburg Institute of Nuclear Physics.

An intensive linear proton accelerator is also constructed at the Institute for Nuclear Research (Troitsk). The first stage of the accelerator has been completed.

In IHEP (Protvino, Director — A.A. Logunov) the works to build new accelerator U-600 are in progress.

Another direction of HEP research is connected with plans of the Institute for Nuclear Physics (Novosibirsk, Director — A.N. Skrinsky) to create colliding electron-positron beams on the basis of linear electron accelerators (VLEPP).

The international experience in dealing with heavy-current accelerators shows that a meson factory is a generator of high-intensity secondary beams of pions, muons, neutrons, neutrinos, polarized nucleons, hydrogen neutral atoms and is a unique tool for investigations in nuclear and elementary particle physics. That is why the construction of Moscow meson factory by the Institute for Nuclear Research (scientific leader — A.N.Tavkhelidze, director — V.A. Matveev) is another important direction for fundamental studies in physics in Russia.

In terms of its parameters, every large accelerator is a unique physical installation. The high cost and sophistication of experimental apparatus urges the necessity of a wide international scientific and technical cooperation. Such cooperation in operating huge accelerating machines makes it possible to develop faster and more efficiently complicated experimental facilities comprising numerous detectors, electronic equipment and control systems.

CONCLUSION.

This short review presents only some general information about our research centre in Dubna. However, I'd like to emphasize that over its 42 years of existence JINR has become a well-known international scientific centre, which incorporates the fundamental research of the structure of matter, development and application of high technologies, and university education

in the relevant fields of knowledge. The scientific policy pursued by the JINR Directorate was developing in the context of the world scientific trends. At the same time last years were marked with a struggle for survival and preservation of the Institute as a unified scientific centre in the time of radical political changes and serious economic difficulties in Russia and some of the Member States. This struggle was very often waged under conditions close to the extreme ones. Nevertheless, due to the joint efforts, the Institute has survived and continues to contribute significantly to world science in the field of particle physics, nuclear physics, and condensed matter physics.

Successful implementation of new reforms should become an impulse for the dynamic development of this international scientific centre in the 21st century.

Acknowledgments.

I would like to express my appreciation to Prof. J. Budagov, Dr. O. Pukhov and Mrs. M. Studenova for their help in this report preparation.

MALT WHISKY - AN INTRODUCTION TO THE SPIRIT OF SCOTLAND

Peter M. Dryburgh

University of Edinburgh, The King's Buildings, Edinburgh EH9 3JL, Scotland, U.K.

Abstract

The historical connection between malt, grain and blended whisky is used to introduce the making of malt whisky and outline the preparation of malt from barley, the processes of fermentation and distillation and the maturation of the strong spirit in oak casks. The geographical basis for the classification of malt whiskies is mentioned. Ten whiskies, ranging from the lightly flavoured Bladnoch to the heavily peated and phenolic Ardbeg, are discussed in terms of their individual characteristics and the distilleries which produced them.

INTRODUCTION

An article on malt whisky might seem out of place in a publication devoted to a summer school on high energy physics but the malt whisky tasting held on 25th August at St. Andrews, although essentially a social event, was the occasion for some serious study of this fascinating subject and seemed to be enjoyed by all participants. The purpose of this article is to expand on the brief notes which were issued at the event.

Blended whisky has satisfied the world's demand for Scotch whisky since the turn of the century. Despite the overwhelming emphasis given to blended whisky, and its consequent impact upon world markets, there have always been people to whom the only true whisky was the original single malt, produced in the traditional way and, nowadays, increasing numbers of discriminating drinkers are exploring these single malt whiskies in much the same way that more people are learning about wine.

Blended whisky was introduced late in the 19th century, when industrial stills started to produce large quantities of cheap alcohol from fermented cereal mashes other than malt. The spirit produced had little character but was legally 'grain whisky'. Skilled blenders added malt whiskies to it to produce blended whisky, which was so successful that, not only is the word 'Scotch' used in most languages, but individual blends like Johnnie Walker, Haig and Chivas Regal are familiar throughout the world.

The traditional malt whisky of Scotland was prepared originally in farmhouses, crofts and castles from the simplest of local ingredients, barley and water. If a moist, sprouting barley seed is heated and dried over a fire, the tiny plant is killed but the sugars and other substances produced to feed it are left in the seed. The product of this cereal infanticide is malt, and its importance lies in the fact that the sugars can be fermented by the action of yeast and water, just as in the brewing of beer. A most important source of flavour in malt whiskies is the peat smoke from the fire used to dry the malted barley, the way in which the aromatic constituents of the 'peat reek' are incorporated into the drying malt being very reminiscent of the process of kippering. The amount of peat smoke used in the malting depends upon local tradition.

By the time that the weakly alcoholic solution produced by the fermentation has been distilled twice in simple copper stills - pot stills - it has turned into a colourless, pungent spirit containing about 70% alcohol. The slow magic which turns this fierce, raw spirit into mature, malt whisky takes

place in oak casks. (Old sherry casks are highly prized because of the special finish which they give to the mature whisky.) The law requires Scotch whisky to have been matured in the cask for a minimum period of three years, but malt whiskies have to be kept for much longer to develop their full character. The practice of allowing the whisky to mature is the only important feature which distinguishes the process used in distilleries from that of the ancient cottage industry, but it is of crucial importance in the development of flavour, aroma and body.

Until recently, there were more than 100 malt distilleries distributed all over Scotland from Wigtownshire to Wick and on the islands of Orkney, Jura, Mull, Skye and Islay. Malt whiskies are usually classified as Highland, Lowland and Islay malts, the Highland and Lowland areas lying north and south of a boundary joining Greenock and Dundee.

LOWLAND MALTS

Lowland malts are of a light character as no peat smoke is used in the malting process and they can be pleasant and flowery. At their best, they are excellent aperitifs. Bladnoch was made in the most southerly distillery in Scotland until the distillery was closed in 1993. Like a number of the earlier lowland whiskies, Bladnoch was triply distilled and it has a fresh, simple, slightly lemony taste and an aroma like that of cut grass. Rosebank, generally regarded as the best of all lowland malts, was made in Falkirk but, sadly, the distillery has been shut and the buildings converted to other purposes. At its best - probably 8 years old, the original standard age for bottling it - Rosebank demonstrates just how distinguished a lowland malt can be, while still retaining the essential features of lightness and floweriness.

HIGHLAND MALTS

The highland malts are so numerous and various that any small selection can give only a cursory introduction to their potential.

Balvenie is made in Dufftown, one of the important distillery towns north of the Grampians but, despite the fact that the distillery was built in 1892, the whisky has been available in bottles only since 1971. The distillery is operated owned by the Grant family and lies close to their much better known distillery, Glenfiddich. Although the two distilleries are so close, even sharing the same water supply, Balvenie is surprisingly different from its well known neighbour. Balvenie has a spicy and smoky taste but the smoky element is not dominant and the final taste strikes a splendid balance between smoke, honey, oak and fruit. Balvenie shows in a most enjoyable way all the important features of a good malt without any excessive dominance by any of them.

A much lamented victim of distillery closures is Glen Mhor, which was situated in Inverness but was wantonly demolished more than ten years ago. Glen Mhor has been described as the greatest after-dinner drink in existence. It is hard to discuss this whisky without straining the vocabulary: it is smooth, mellow, big and complex. Many brandies are unable to match the depth and subtlety of Glen Mhor.

Whiskies from any particular area often have certain generic characteristics but there are some whiskies which stand out as individuals having little in common with any others. One such individual is Old Pulteney, made at the most northerly distillery on the mainland at Wick. Old Pulteney seems to derive its considerable strength of taste from something other than peat. It matures rapidly for such a heavy style of whisky and is excellent as normally bottled at 8 years old. This is a whisky to enjoy before or after food. The taste is long and appetising with a faintly salty character. It is a plain, robust and satisfying whisky and deserves to be more widely known.

It is taking something of a liberty to include two Orcadian whiskies with the northern highlands, particularly as Orcadians have at least as much affinity with Scandinavia as with Scotland, but all systems of whisky classification are general and approximate and no one can dispute that Orkney lies well north of the Highland Boundary. The two malt whiskies produced in Orkney are

Highland Park and Scapa and although Highland Park is much the better known, Scapa is a whisky of real quality.

Highland Park occupies a site associated with whisky distilling since before 1790 and is one of the few remaining distilleries to carry on the traditional process of floor malting. Highland Park whisky brings to the nose to a decided suggestion of peat smoke but there is nothing heavy-handed about it and the overall impression of the taste is of substantial quality, good balance and depth.

Scapa distillery lies on the shore of Scapa Flow, that great stretch of sheltered water which provided the base for the British fleet in the two world wars. The distillery is in sight of Highland Park but is very different in character. Scapa has a rather flowery aroma, a slightly sweet initial taste and a long aftertaste. Scapa is a brisk sort of whisky but it is certainly not unsophisticated and is worthy of considerable study.

2 ISLAY MALTS

Islay Malts are sometimes considered as being all heavily peated and massive in taste. This is not the case and the eight whiskies of Islay are easily separated into three distinct groups, one of which contains Lagavulin, Laphroaig and Ardbeg, usually considered the archetypal heavy whiskies, a second group consisting of Bowmore, Caol Ila and Port Ellen, and a third group of two, Bruichaddich and Bunnahabhain, which have much more in common with whiskies from the northern highlands than with their close neighbours in Islay. The three different types were represented at the tasting by Bunnahabhain, Caol Ila and Ardbeg.

Bunnahabhain is made using malt which has no peat smoke used and the water comes from a deep borehole so it contains no trace of the extracts of peat which often colour and flavour water which has been obtained from surface supplies. To the nose, Bunnahabhain presents a fresh and attractive aroma, while the taste is nutty and spirituous but beautifully balanced. It has an almost oily texture which gives it a general sense of richness in the mouth. Like all good whisky it has a long, lingering aftertaste.

Caol Ila is produced in a distillery first built in 1846 but rebuilt in 1974. The distillery overlooks the Sound of Islay from where it takes its Gaelic name and has a magnificent view of the island of Jura. Caol Ila represents perfectly what is often regarded as a typical Islay whisky; smoky, peaty and spicy to the nose and with a slightly medicinal taste. All malt whisky is best drunk unchilled but whisky of this weight and character suffers particularly from being served cold.

Ardbeg Distillery has been closed for a number of years but was recently bought and reopened. This was very good news for the many enthusiasts who enjoy this powerful and unique drink. Ardbeg contains the highest concentration of polychlorinated phenolic compounds found in any whisky and this accounts for its decidedly medicinal taste and powerful assault on the taste buds. The distillery still uses the old floor malting process and is careful to prevent much air circulating in the kiln so that the malt is subjected to a huge concentration of peat smoke. Ardbeg is not to everybody's taste but, to those who enjoy it, it is an experience rather than just a drink and demands to be treated with respect.

Whether you enjoy the delicate and ethereal, the substantial and satisfying or sheer Wagnerian splendour, you should remember that malt whisky is a complex creature of spirits and volatiles and to chill it is to paralyse it. Keep it warm and it will sing for you.

DOLLY: THE SCIENCE BEHIND THE WORLD'S MOST FAMOUS SHEEP

H.D. Griffin

Roslin Institute (Edinburgh), Roslin, UK

Abstract

'The Roslin Institute near Edinburgh is one of the world's leading centres for research on farm animals. Its main expertise is in animal genetics, with the results benefiting the livestock breeding and production industries in particular. Its work was, however, little known to the general public. A single lamb changed all that. **Dr Harry Griffin**, Assistant Director at the Institute, explains.....'

Dolly was the first mammal cloned from a cell from an adult animal. She was derived from cells that had been taken from the udder of a 6-year old Finn Dorset ewe and cultured for several weeks in the laboratory. Individual cells were then fused with unfertilised eggs from which the genetic material had been removed. Two hundred and seventy seven of these 'reconstructed eggs' - each now with a diploid nucleus from the adult animal - were cultured for 6 days in temporary recipients. Twenty nine of the eggs that appeared to have developed normally to the blastocyst stage were implanted into surrogate Scottish Blackface ewes. One gave rise to live lamb, Dolly, some 148 days later.

DOLLYMANIA

Dolly was born on 5 July 1996. It took Ian Wilmut and his colleagues a few weeks to complete the experimental work and several more months before they had written a paper and had it accepted by *Nature*. Publication was confirmed for 27 February, 1997 and by then plans were well advanced to handle the expected media interest. We knew that *Nature* was featuring Dolly in the weekly press release that it distributes each Friday. The information in these releases is 'embargoed' until 7 o'clock on the Wednesday evening before publication and this is intended to provide journalists time to research their stories more carefully before going to press. However, ideas of a relaxing weekend were shattered by two phone calls late on Saturday warning us that the *Observer* would be running the story the next day.

A quick visit to the newsagents first thing on Sunday confirmed our story was on the front page. The *Observer's* science writer Robin McKie had resisted taking a sensational line but he had raised fears about human cloning. Our main aim then was to ensure that the other papers have some solid facts to report. Two of the team at Roslin and four at our PR agency *De facto's* office in Basingstoke worked throughout the day to answer calls from around the world and set up schedules for reporters, TV crews and photographers desperate to visit Roslin.

On Monday, Dolly provided a major story in most of the papers and Roslin was besieged by reporters and TV crews from all over the world. ABC, NBC, CNN and CBS all beamed live interviews across the Atlantic via satellite vans in the car park. Phones rang incessantly and extra spokesmen were drafted in to help. Dolly quickly became the most photographed sheep of all time and was invited to appear on a chat show in the US. Astrologers asked for her date of birth and the shares of our collaborators PPL rose sharply. President Bill Clinton called on his bioethics commission to report on the ethical implications within 90 days and Ian Wilmut was invited to testify to both the UK House of Commons and the US Congress. Dolly Parton said she was 'honoured' that we had called our sheep after her.

In the first week after the story broke, we estimated that together the Roslin Institute, PPL and *De Facto* answered more than 2000 telephone calls, talked at length to close on 100 reporters and provided access to Dolly to 16 film crews and more than 50 photographers from all over the world. After all this, the actual publication of the paper in *Nature* on Thursday 27 February seemed an anticlimax.

‘Science Breakthrough of the Year’

For developmental biologists, Dolly’s existence challenges one of the fundamental tenets of developmental biology. Most scientists had thought that differentiation - the gradual process of specialisation that allows the fertilised egg to develop into the hundreds of cell types that make up the whole animal- was irreversible. After all, even over a 90 year lifespan a liver remains a liver, a nerve cell a nerve cell. The production of a live lamb from a cell taken from the udder of a 6-year old ewe demonstrated that differentiated cells are not immutable: it was possible to take a differentiated cell and turn its biological clock back to zero. To start its life all over again. And for this, Dolly was voted ‘Science Breakthrough of 1997’ by the editors of the prestigious US journal *Science*, ahead of NASA’s Pathfinder mission to Mars and advances in cyclotron design.

At first Dolly was a clone alone, but in recent months clones of adult mice have been reported from Hawaii and cloned calves from both New Zealand and Japan. These additional reports demonstrate that Dolly is not an ‘anecdote’ and that it should be possible to clone from a variety of different cell types and species.

Cloning of humans

The intense media interest was not with sheep. The immediate assumption was that cloning of humans was just around the corner and that seemed to trigger an explosion (at least in the media) of fears about the future. However, in contrast to the barrage of calls from the press, the relatively small size of our postbag suggested that the general public was much more sanguine. They seemed to know that the scenarios that the media were imagining were unlikely to be realised and could, as a consequence, be enjoyed in safety. Just like ‘*the X files*’.

Much of the media speculation was based on science fiction rather than good sense, with the *Times* and a Dr Patrick Dixon taking our award for the most outrageous list of ‘reasons’ for cloning humans. A common misunderstanding was that clones would be somehow less than human. However, most of us have already met a human clone - an identical twin - and no one would seriously suggest that we should produce twins to order as sources of spare parts. Others seemed to believe that cloning would produce an identical photocopy, rather than a child that would grow up with a personality and behaviour all of its own.

Those speculating about potential uses in humans need to recognise that the technology is very much in its infancy. Our experiments in sheep produced several lambs with developmental abnormalities that died in late pregnancy or soon after being born. A clone will inherit somatic mutations from the donor and this in turn may lead to premature ageing or a higher incidence of cancer later in life. These risks alone mean that it would be grossly unethical to perform similar experiments in women now, however desperate they might be to have a child.

Experience with cloning in farm animals may identify ways of reducing risk but this is likely to take many years. In the meantime, society as a whole has time to contemplate which uses of the technology might be acceptable and which would not. Views will inevitable change over time: the first test tube baby, Louise Brown, was born in 1979 and *in vitro* fertilisation is now commonplace. Attitudes will also vary between countries: many feared cloning would be exploited by megalomaniacs but in our view it is far more likely that the ethical boundaries will be tested by the *in vitro* fertilisation clinics in the US or the Far East.

Nuclear transfer

Cloning by nuclear transfer is not itself novel. The technique was first reported in frogs in 1953 and has been used widely since in amphibians to study early development. This work showed that the first few cell divisions after fertilisation produce cells that are totipotent (i.e. they can develop into all of the cell types that make up the whole animal). As the embryo develops further, the cells lose this property and the success of nuclear transfer rapidly declines. Some nuclear transfer experiments using cells from adult frogs produced viable embryos, but these never developed beyond the tadpole stage.

Nuclear transfer in mammals proved to be more difficult. The cloning of mice using nuclei from very early embryos was reported in 1977 but this work was not repeatable and interest among developmental biologists waned. Research on nuclear transfer in cattle continued, stimulated by the potentially large commercial benefits of multiplying elite embryos. Artificial insemination allows each bull to have thousands of offspring but each cow can only produce 5 or 6 calves in a lifetime. Multiple ovulation embryo transfer (MOET) and cloning by embryo splitting have been used to partially redress this imbalance, but these techniques have limited potential for further development. By contrast, nuclear transfer has, at least in principle, the ability to produce an unlimited number of identical animals.

By the middle of the 1980's several research groups from around the world had produced cloned sheep and cattle by transferring nuclei directly from early embryos. In 1985 Steen Willesden in the US had produced live calves by nuclear transfer from embryos that had progressed to the 64- and 128-cell stage and this was the first suggestion that nuclear transfer in mammals was possible from at least partially differentiated cells.

In the early 1980's the then Animal Breeding Research Organisation initiated a programme designed to produce transgenic sheep and cattle that would secrete human proteins in their milk. Using the beta- lactoglobulin promoter, John Clark and colleagues were able to direct expression of the transgene to the mammary gland. This success led to the setting up in 1987 of PPL Therapeutics and in 1989 the production of Tracy, a transgenic sheep that secreted 35 g of a human protein- alpha-1-antitrypsin- in her milk. Over the same period other groups had developed other transgenic livestock, including genetically modified pigs for use as sources of organs for transplantation to human patients.

At this time the only way of producing transgenic livestock was by pro-nuclear injection. This procedure involves the introduction of 200-300 copies of the transgene into a recently fertilised egg which is then implanted in a surrogate mother. Only 2-3% of eggs give rise to transgenic offspring and only a small proportion of these express the added gene at sufficiently high levels to be of commercial interest. It was also only possible to add genes.

If animals can be derived from cells in culture, then it is possible to carry out much more specific genetic modifications, including the removal or substitution of specific genes. This has been achieved in mice using embryo stem (ES) cells but to date no one has yet succeeded in obtaining ES cells from cattle, sheep or pigs. After learning of Willesden's success, Ian Wilmut thought that nuclear transfer might provide an alternative.

The major breakthrough came in 1995 when Keith Campbell, Ian Wilmut and colleagues produced live lambs - Megan and Morag - by nuclear transfer from cells from early embryos that had been cultured for several months in the laboratory. The key element in this success was the induction of quiescence in the donor cells. However, at this stage they did not know if they had also stumbled on a particular amenable cell type simply by chance.

Additional experiments were performed to test if successful nuclear transfer technique was restricted to embryo-derived cells or could be carried out with a wider range of cell types. Nuclear transfer was carried out

from embryo-derived cells, foetal fibroblasts and - in collaboration with PPL Therapeutics - cells from an adult ewe. Four lambs were born from embryo cells, two from the foetal cells and one, subsequently named Dolly, from an adult cell. Their identity was confirmed by DNA testing and the results published in *Nature* on 27 February 1997.

Practical applications of cloning

The ability (*via* nuclear transfer) to derive live animals from cultured cells provides an alternative way of producing transgenic farm animals. Moreover, the ability to manipulate many millions of cells at once opens up the possibility of much more specific genetic modifications, including the deletion or substitution of specific genes or the introduction of the single letter changes in the whole of the genetic code that are characteristic of many human genetic diseases.

human therapeutic proteins

Human proteins are in great demand for the treatment of a variety of diseases. Whereas some can be purified from blood, this is expensive and runs the risk of contamination by AIDS or hepatitis C. Proteins can be produced in human cell culture but costs are very high and output small. Much larger quantities can be produced in bacteria or yeast but the proteins produced can be difficult to purify and they lack the appropriate post-translational modifications that are needed for efficacy *in vivo*.

By contrast, human proteins that have appropriate post-translational modifications can be produced in the milk of transgenic sheep, goats and cattle. Output can be as high as 40 g per litre of milk and costs are relatively low. PPL Therapeutics, one of the leaders in this field, recently announced that alpha-1-antitrypsin produced from a transgenic flock is now being used to treat cystic fibrosis patients in phase 2 clinical trials. In the US, Genzyme Transgenics have focussed on goats and their lead product, tissue plasminogen activator, is also in clinical trials.

An immediate advantage of producing transgenic animals by nuclear transfer is that it uses less than half of the experimental animals than does pronuclear injection. It is also possible to specify the sex of the offspring and thereby substantially reduce the time taken to generate a production flock. PPL and the Roslin Institute have already used this approach to produce Polly and Molly, two transgenic sheep that secrete the human blood clotting factor IX in their milk.

At present such transgenic animals produce a human protein in addition to the normal complement of milk proteins. This restricts yield and with some proteins there would be a major advantage in removing one or more of the endogenous milk proteins. This is the case with serum albumin where the total demand for treatment of burns and other trauma is estimated at over 600 tonnes each year. In this case production would be in cattle and the aim would be to replace the bovine albumin gene with its human equivalent.

nutraceuticals

There are a number of potential opportunities for altering the nutritional content of milk. For example, cow's milk is ideal for calves but not for premature infants. Gene targeting using nuclear transfer will allow milk to be produced in which one or more of the normal cow's proteins have been replaced by human proteins, thereby improving its nutritional quality for these special 'consumers'. Other people have an immune response to specific proteins in milk or are intolerant of lactose and gene targeting would allow the creation of herds of cows that produced milk lacking the problem components.

xenotransplantation

Over the past 20 years transplantation of hearts and kidneys has become almost routine. There is an shortage of suitable organs for transplant (about 5000 each year in the UK) and many patients have to suffer prolonged dialysis and/or die as a result. Transgenic pigs are being developed as sources of organs to meet the shortfall.

The pigs produced so far have an added human protein such as compliment inhibitory factor that coats the pig tissues and is intended to prevent immediate rejection of the transplanted heart or kidney.

The ability to remove genes could have a major impact on the success of such xenotransplants. Surprisingly, most of the antibodies in our blood that would the react against a pig organs recognise a single carbohydrate linkage, galactose alpha (1,3) galactose. This sugar residue does not appear to be functionally important because it is not present in humans and monkeys. Elimination of the glycosyltransferase responsible for attaching this sugar residue by targeting the relevant gene in pigs is therefore expected to greatly reduce hyperacute rejection of transplanted organs.

animal models of disease

Mice in which specific mutations have been deliberately introduced have often proved very useful as models for studying human genetic diseases. In some cases, differences between mice and humans means that the effects of the introduced mutation are not the same as in the human genetic disease. This is the case for the cystic fibrosis mutation, where differences in chloride channel mean that the knock out experiment in mice has little effect on lung physiology.

Nuclear transfer will extend the range of species in which gene targeting will be possible and thereby provide better models to test treatments for human diseases. Lung physiology in sheep and humans is very similar and the gene targeting to introduce the cystic fibrosis mutation in sheep would be expected to produce a similar phenotype in homozygous animals to that in man. The deliberate creation of a genetic disease in a large animal raise public concerns and the ethical justification of creation of such a model would need to be well argued.

cell therapy

Intact cells are already used to treat patients suffering from a number of diseases, including leukemia and Parkinson's disease. In most cases these cells have to be obtained from close relatives to avoid problems of immune rejection.

The fact that Dolly was cloned from a cell taken from an adult ewe shows that even differentiated cells can be 'reprogrammed' into all the cell types that make up an intact animal. The only way that we can perform this dedifferentiation step now is by 'incubating' cells in the cytoplasm of an unfertilised egg but when we know more about the mechanisms involved, then it may be possible that human cells can be reprogrammed without the use of a human egg. This would allow the patient's own cells to be used in cell therapies, thereby avoiding the time, expensive and uncertainty of tissue matching. Cells would removed from the patient, converted into the desired cell type in the laboratory and then reintroduced into same patient for treatment.

cloning in farm animal production

In addition to providing a route to gene targeting in livestock, nuclear transfer could be used to deliver what is the popular image of cloning: that is the production of, at least in principle, unlimited numbers of genetically identical animals.

There are major practical hurdles to overcome before this could be a routine procedure in livestock

production. We first need to show the techniques can be used in cattle and pigs since it is probably only in these species that the benefits are likely to justify the cost. Non-surgical means would be needed for embryo transfer and success rates would have to be dramatically improved. Previous experience with new technologies such as artificial insemination and multiple ovulation embryo transfer suggest that it may be 10-20 years before this could be possible

Nevertheless, the main advantage of cloning would not be in selection programmes but in the more rapid dissemination of genetic progress from elite herds to the commercial farmer. At present this is achieved through artificial insemination - which supplies only half the genes- and by limited use of embryo transfer. This process is not that efficient and in dairy cattle, the performance of the average cow is probably some 10 years behind the best. With cloning, it would be possible to remove this difference. Farmers who could afford it would receive embryos that would be clones of the most productive cows of elite herds and thereby lift the performance of their herds to that of the very best within one generation. This would be a one-off gain, since from then on the rate of genetic progress would return to that of the elite herds.

In this scenario, breeding companies would sell cloned embryos in much the same way as they now sell semen. Farmers would choose cloned embryos for high merit beef bulls or dairy cows from catalogues that described the genetic merit for a series of economically important traits, including fertility, health and longevity. The cloned embryo would be delivered to the farm much in the same way as semen straws are today, perhaps from breeders overseas.

A major risk would be loss of genetic diversity but this could be avoided by systems that ensured that breeding companies produced a limited number of clones of each genotype and restricted the number of each of the clones that could be sold to any one producer. Although the herds of some producers might consist entirely of cloned animals, the fact that they were clones of different elite animals may actually increase genetic diversity on some farms.

genetic conservation

Although cloning is associated in most people's minds with a loss of genetic diversity, the techniques that were used to produce Dolly will also provide new approaches to genetic conservation. With increasing commercial pressures, many indigenous breeds adapted to local conditions are under threat from imported breeds that are being reared in intensive farming systems. The local breeds may contain valuable genes that confer heat tolerance or disease resistance and there is an urgent need to prevent their extinction. Current methods of conservation involve storage of frozen semen or embryos but are time consuming and costly. As a consequence, the future of only a small proportion of endangered breeds is being addressed.

The new techniques developed at Roslin may provide much simpler and more effective means of conserving breeds. Blood samples, skin biopsies or even hair follicles might be suitable sources of cells that could be grown briefly in the laboratory and then frozen in liquid nitrogen for long term storage or transfer to other endangered populations.

To some these applications seem much less glamorous than the prospect of the cloning of human beings. In reality, they have the potential to provide radical new treatments for a wide variety of serious diseases that affect many millions of people world-wide. By contrast, the cloning of human beings- if and when it happens- is likely to remain a marginal activity with little impact on society beyond the immediate participants.

Further information on cloning, including references to work cited, is available on the Institute's web site on

www.ri.bbsrc.ac.uk

|

ORGANIZING COMMITTEE

Tatyana Donskova	donskova@cv.jinr.ru
Nick Ellis	nick.ellis@cern.ch
Sinead Farrington	s.m.farrington@sms.ed.ac.uk
Ian Knowles	egil.lillestol@cern.ch
Egil Lillestøl	i.knowles@ed.ac.uk
Sasha Olchevski	alexander.olchevski@cern.ch
Steve Playfer	s.m.playfer@ed.ac.uk
Susannah Tracy	susannah.tracy@cern.ch
Alan Walker	a.walker@ed.ac.uk

LECTURERS

Christine Davies	cdavies@itp.ucsb.edu
Peter Dryburgh	peter.dryburgh@ed.ac.uk
John Ellis	john.ellis@cern.ch
Mike Green	m.b.green@damtp.cam.ac.uk
Harry Griffin	harry.griffin@bbsrc.ac.uk
Jim Hough	j.hough@physics.gla.ac.uk
Richard Kenway	rdk@epcc.ed.ac.uk
Peter Litchfield	p.j.litchfield@rl.ac.uk
Chris Llewellyn-Smith	john.ellis@cern.ch
Michelangelo Mangano	michelangelo.mangano@cern.ch
Yossi Nir	ftnir@weizmann.weizmann.ac.il
Victor Novikov	novikov@heron.itcp.ru
John Peacock	jap@roe.ac.uk
Susan Rowan	s.rowan@physics.gla.ac.uk
Alexey Sisakyan	sisakian@jinr.dubna.su
Johanna Stachel	stachel@ceres1.physi.uni-heidelberg.de
Jim Virdee	tejinder.virdee@cern.ch

DISCUSSION LEADERS

Patricia Ball	patricia.ball@cern.ch
Gerhard Buchalla	gerhard.buchalla@cern.ch
Vittorio Del Duca	vdd@ph.ed.ac.uk
Georgi Dvali	dvali@ictp.trieste.it
Belén Gavela	gavela@delta.ft.uam.es
Christoph Greub	greub@butp.unibe.ch
Valery Khoze	valery.khoze@Inf.infn.it
Oleg Teryaev	teryayev@thsun1.jinr.dubna.su

LIST OF STUDENTS

Markus Adams	markus@physik.uni-dortmund.de
Martin Aleksa	martin.aleksa@cern.ch
Reyes Alemany	reyes.alemany@cern.ch
Fabio Ambrosino	fabio.ambrosino@na.infn.it
John Andress	john.andress@bristol.ac.uk
Ignacio Arino	arinyo@ecm.ub.es
Joao Bastos	bastos@piranha.fis.uc.pt
Marc Beckmann	marc.beckmann@desy.de
Andrea Biguzzi	biguzzi@hep.phy.cam.ac.uk
Catherine Biscarat	biscarat@clermont.in2p3.fr
Vladimir Borisov	borisov@sunhe.jinr.ru
Edouard Boudinov	i76@nikhef.nl
Othmane Bouhali	bouhali@hep.iihe.ac.be
Giacomo Bruno	giacomo.bruno@cern.ch
Daniel Buirra Clark	d.buirra-clark1@physics.ox.ac.uk
Ernst-Jan Buis	ernst-jan.buis@cern.ch
Giorgio Cabibbo	giorgio.cabibbo@roma1.infn.it
Susana Cabrera Urban	susana.cabrera.urban@cern.ch
Anselmo Cervera	acervera@axnd02.cern.ch
Dmitry Chernyak	d.v.chernyak@inp.nsk.su
Nicola Coppola	coppola@bo.infn.it
Robin Coxe	robin.coxe@cern.ch
Mikolaj Cwiok	cwiok@fuw.edu.pl
Agnieszka Czermak	czermak@chall.ifj.edu.pl
Tatiana da Silva	tati@ifjf.br
Olivier Deschamps	odescham@in2p3.fr

Lec Drage	drage@hep.phy.cam.ac.uk
Joerg Dubbert	joerg.dubbert@physik.uni-muenchen.de
Andrey Dudarev	dudarev@nusun.jinr.ru
Malcolm Ellis	ellis@physics.usyd.edu.au
Peter Fagerstroem	peterf@physics.utoronto.ca
Harald Fox	harald.fox@cern.ch
Nabil Ghodbane	ghodban@lyohp5.in2p3.fr
Simona Giovannella	simona.giovannella@lnf.infn.it
Edda Gschwendtner	edda.gschwendtner@cern.ch
Leanne Guy	leanne.guy@cern.ch
Allan Hansen	aghansen@nbi.dk
Lilith Haroyan	lilit@crd3.yerphi.am
Stephan Heising	stephan.heising@cern.ch
Fabrice Hubaut	hubaut@cppm.in2p3.fr
Petra Huentemeyer	petra@bolte.desy.de
Anna Kaczmarska	kaczmars@chall.ifj.edu.pl
Hans-Christian Kaestli	kaestli@particle.phys.ethz.ch
Joanna Kiryluk	kiryluk@fuw.edu.pl
Katarzyna Klimek	kklimek@indyk3.ifj.edu.pl
Jukka Kokkonen	jukka.kokkonen@cern.ch
Henning Krueger	h.krueger@mpi-hd.mpg.de
Alexander Kupco	kupco@hp02.troja.mff.cuni.cz
Igor Lebedev	lebedev@satsun.sci.kz
Giovanna Lehmann	giovanna.lehmann@cern.ch
Sami Lehti	sami.lehti@helsinki.fi
Lena Leinonen	leinonen@physto.se
Roman Malina	malina@cern.ch
Dominique Mangeol	dominique.mangeol@cern.ch
Clemmens Mannert	mannert@mppmu.mpg.de
Victoria Martin	victoria.martin@cern.ch
Federica Mazzucato	federica.mazzucato@cern.ch
Adrian Mirea	mirea@cppm.in2p3.fr
Natalia Molokanova	molokan@sunse.jinr.ru
Remi Mommsen	remigius.mommsen@cern.ch
Danielle Moraes	danielle@if.ufrj.br
Basil Moshous	basil@mppmu.mpg.de
Pedro Movilla Fernandez	pedro@physik.rwth-aachen.de
Alexandra Muijs	alexandra.j.m.muijs@cern.ch
Niko Neufeld	niko.neufeld@cern.ch
Hakan Oztuk	ggunes@pamuk.cc.cu.edu.tr

Raquel Paiva	raquel.paiva@cern.ch
Carmen Palomares	carmen.palomares@cern.ch
Rocio Perez-Ochoa	rocio.perez-ochoa@cern.ch
Rok Pestotnik	rok.pestotnik@ijs.si
Enrico Piozzo	enrico.piozzo@cern.ch
Oleg Piskunov	o.piskunov@gsi.de
Boris Popov	popov@nusun.jinr.dubna.su
Michael Ruh	michael.ruh@desy.de
Vadim Rusu	vadim@hienc.ifa.ro
Stefan Schmitt	stefan.schmitt@cern.ch
Sergei Shmatov	shmatov@lhe.jinr.ru
Laurent Simard	simard@hep.saclay cea.fr
Jürgen Simon	juergen.simon@mpi-hd.mpg.de
Roman Slepnev	slepnev@sunhe.jinr.ru
Alexander Somov	somov@ifh.de
Albert Sotnikov	asp@cv.jinr.ru
Stojan Stojnev	stojan@heph.phys.uni-sofia.bg
Dieter Striegel	striegel@mppmu.mpg.de
Serge Sushkov	svs@hep.by
Fredrik Tegenfeldt	fredrik.tegenfeldt@cern.ch
Michael Thiergen	thiergen@hpfrsc.physik.uni-freiburg.de
Alessia Tricomi	alessia.tricomi@cern.ch
Philippe Vannerem	philippe.vannerem@cern.ch
Maria Varanda	mjesus@lip.pt
Alexey Vasiljev	vasiljev@inp.nsk.su
Sophie Versille	versille@in2p3.fr
Dmitry Vishnevsky	dma@cv.jinr.ru
Jan Visser	janvs@nikhef.nl
Sandor Voros	voros@cern.ch
Helge Voss	helge.voss@cern.ch
Martin Wegner	wegner@physik.rwth-aachen.de
Thomas Ziegler	thomas.ziegler@cern.ch

LIST OF POSTERS

A. Chernyak, A. Vasiljev	CMD-2 and SND Experiments at VEPP-2M e^+e^- Collider in Novosibirsk, Russia
A. Somov	Prospects of Measuring Dilepton Production in the Upsilon Mass Region at HERA-B
N. Molokanova	Inclusive ϕ Meson Production in Neutron-Carbon Interactions at Serpukhov Accelerator (EXCHARM Experiment)
M. Beckman, M. Ruh	Extraction of Polarized Quark Distributions at HERMES
F. Mazzucato	Search for New Physics in the Photonic Channel
A. Cervera Vallaneuva, J. Kokkonen	A Silicon Target (STAR) for Neutrino Physics
M. Ellis	Study of Charm Production by Neutrinos
A. Tricomi	b and c Quark Forward-Backward Asymmetries (ALEPH Experiment)
G. Bruno, M. Cwiok	Resistive Plate Chambers for the CMS Detector
A. Biguzzi, R. Coxe, J. Dubbert, H. Voss	LEP 2 Standard Model and WW Physics at OPAL
T. Ziegler	Cosmics and Colour-Reconnection Effects in ALEPH
H. Krüger, J. Simon	Measurement of Hyperon Charge Radii at SELEX (E781)
J.C. Andress, S. Versille	The BaBar Experiment
E. Gschwendtner	Background Measurements in H6 for the ATLAS Muon Spectrometer
S. Lehti, J. Kokkonen, J. Tuominiemi	$H_{\text{SUSY}} \rightarrow \tau\tau \rightarrow e\mu$
G. Lehmann, R. Mommsen	ATLAS – Trigger and DAQ
S. Cabrera-Urbana	Measurement of the b Quark Mass at LEP1 with the DELPHI Detector
C. Palomares	WW Production in L3 Detector
M. Adams	The Muon Pretrigger System of the HERA-B Experiment
O. Bouhali	Development and Tests of MSGCs for the CMS Forward Tracker
E. Piotto	New Measurement of V_{ub} with the DELPHI Detector at LEP
S. Heising	Pixel Detectors at Collider Experiments
D. Moraes	Updated Measurement of τ Polarisation in $\tau \rightarrow \pi (K) \nu_\tau$
J. Visser	Silicon Detector Systems at HERMES
P. Movilla Fernandez	Measurements of α_s with JADE Data and Studies of Power Corrections
D. Buirra-Clark	ESPI Measurements for ATLAS Inner Detector
B. Moshous	Si Detector Modules for the HERA-B Experiment
T. da Silva	Study of the Decay $J \rightarrow 3h \pi^0 \nu_\tau$ Using the DELPHI Detector at CERN
P. Hubaut	The ANTARES Project
M. Aleksa	ATLAS Drift Tube Performance in a High Background Environment

M.R. Guedes Paiva	Search for Anomalous Higgs Boson Couplings at LEP2 with the DELPHI Detector
E.J. Buis	The LHC1 Pixel Detector Studied in a 120 GeV Pion Test Beam
I. Ariño, J.A. Bastos, R. Pestotnik	First Results from the HERA-B Rich Detector
V. Rusu	Hadronic Calorimeter at LHCb
P. Hüntemeyer	Electroweak Charm Physics at LEP1
L. Guy	B_s^0 Mixing Measurements with the ATLAS Detector
S. Vörös	Hunting the Quark-Gluon Plasma with the WA98 Experiment at 158 A GeV
D. Mangeol	Analysis of the Shape of the Charged Particle Multiplicity Distribution Using Higher-Order Momenta, H_q , at 91.2 GeV with the L3 Detector
J. Koryluk	Measurements of the Spin Structure of the Nucleon from the SMC
L. Simard	Measurement of the W Mass at the DELPHI Experiment
O. Deschamps	SM Higgs Boson Search in the $HZ \rightarrow b\bar{b}q\bar{q}$ Channel at LEP2 with the ALEPH Detector
C. Biscarat	In situ Calibration of the ATLAS Hadron Calorimeter Using Isolated Hadrons
L. Leinonen	Measurement of the A_0^0 Lifetime
S. Schmitt, M. Thiergen	LEP1 B Physics Studies with the OPAL Detector
M. Varanda	The Tile Hadron Calorimeter for the ATLAS Experiment
F. Tegenfeldt	A Measurement of A_b Inclusive Branching Ratio at the DELPHI Experiment
B. Popov	The NOMAD Experiment at CERN
L. Haroyan	The GAMMA Experiment
S. Shmatov	Study Simulation of Global Characteristics in Ultrarelativistic Domain
N. Ghodbane	Supersymmetry: Helicity Effects in Production and Decay of SUSY Particle Phases in the MSSM
V. Vladimir	Research of Single-Spin Asymmetries in Inclusive Hadron Production in Interactions of Vector-Polarized Deuterons with Nuclei Using the Set-up 'Marusia'
H. Fox, V. Martin	Measuring Direct CP Violation at NA48



