10

ISBN 92-95003-13-6

trieste - italy

the
abdus salam
international
centre
for theoretical
physics

united nations
educational, scientific
and cultural
organization

international atomic
energy agency

ictp lecture notes

2001 SUMMER SCHOOL ON PARTICLE PHYSICS

2002

editors
A. Masiero
G. Senjanovic
A.Yu. Smirnov
G. Thompson

www.ictp.it — ictp home > publications & printing services > LNS archive > Vol. 10 ToC

ICTP Lecture Notes Series, Volume 10
(ISBN 92-95003-13-6) - *June 2002*

2001 Summer School on Particle Physics

Editors: *A. Masiero* (SISSA, Italy), *G. Senjanovic* (ICTP, Italy), *A.Yu Smirnov* (ICTP, Italy) and *G. Thompson* (ICTP, Italy)

Contents

Back to LNS Volume 10 main page

ICTP
Publications & Printing
Strada Costiera, 11
Trieste, Italy
pub_off@ictp.it

Introduction

The aim of this school was to give a panoramic view on the field of particle physics with its achievements and problems, successes and failures.

The standard model of the electroweak and strong interactions is in perfect shape. Physics of the standard model and its precision tests have been extensively discussed during the school.

What is next? Do we have a "standard model" of physics beyond the standard model? In this connection the status of low scale supersymmetry, supersymmetric Grand Unification and various flavor symmetries has been presented. Discovery of neutrino masses and mixing is probably the first experimental manifestation of new physics.

Do we have a viable alternative of the (TeV scale) SUSY and GUT? Models with large, or infinite, or wrapped extra dimensions, the bulk-brane scenarios (widely discussed in series of lectures) may give some answers to this question.

Is non-commutative field theory relevant for particle physics? Are the tools we have at hand enough to solve problems of particle physics? Is something fundamentally important missed in our approaches? These, and many other questions, were among the hot topics of the school.

In this volume we publish four courses of lectures given by leading experts in the fields which represent two main areas of the research mentioned above: Physics of the standard model and Physics beyond the standard model.

Both basic and advanced topics are presented in the lectures on non-perturbative QCD and quark-gluon plasma. First results from heavy ion collider RHIC are discussed. Important recent progress in particle physics is related to operation of the B-factories. This subject is covered in lectures on B-physics and CP-violation.

Physics beyond the standard model is represented by lectures on Grand Unification with emphasis on explanation of fermion masses, in particular neutrino masses and mixing, and on predictions for proton decay. Another course is devoted to the fascinating subject: physics of non-commutative field theories.

In conclusion, we wish to thank all participants: lecturers, students and our staff, for their invaluable contribution to the success of the school.

A.Yu. Smirnov
June, 2002

B Physics and CP Violation[*]

H. Quinn[†]

Stanford Linear Accelerator Center, Stanford University,
Stanford, California, USA

Lectures given at the
Summer School on Particle Physics
Trieste, 18 June - 6 July 2001

LNS0210001

[*] Work supported by the Department of Energy under contract number DE–AC03–76SF00515.

[†] quinn@slac.stanford.edu

Abstract

These lectures provide a basic overview of topics related to the study of CP Violation in B decays. In the first lecture, I review the basics of discrete symmetries in field theories, the quantum mechanics of neutral but flavor-non-trivial mesons, and the classification of three types of CP violation [1]. The actual second lecture which I gave will be separately published as it is my Dirac award lecture and is focussed on the separate topic of strong CP Violation. In Lecture 2 here, I cover the Standard Model predictions for neutral B decays, and in particular discuss some channels of interest for CP Violation studies. Lecture 3 reviews the various tools and techniques used to deal with the hadronic physics effects. In Lecture 4, I briefly review the present and planned experiments that can study B decays. I cannot teach all the details of this subject in this short course, so my approach is instead to try to give students a grasp of the relevant concepts and an overview of the available tools. The level of these lectures is introductory. I will provide some references to more detailed treatments and current literature, but this is not a review article so I do not attempt to give complete references to all related literature. By now there are some excellent textbooks that cover this subject in great detail [1]. I refer students to these for more details and for more complete references to the original literature.

Contents

1 Lecture 1: Preliminaries: Symmetries, Hermiticity, Rephasing Invariance

We begin with the basics of symmetries in Lagrangian Field Theory. Physicists use the term symmetry to denote an invariance of the Lagrangian, and thus of the associated equations of motion, under some change of variables. Such changes can be local, that is coordinate dependent, or global; and they can be a continuous set or a discrete set of changes. The value of such symmetries lies in the simplification they achieve by limiting possible terms in the Lagrangian and by their relationship to conservation laws and the conserved quantum numbers that then characterize physical states. The invariance may be with respect to coordinate redefinitions, as in the case of Lorentz Invariance, or field redefinitions, as in the case of gauge invariance. The particular invariances of interest to us in these lectures are the global discrete invariances known as C, P, and T. These are charge conjugation or C (replacement of a field by its particle-antiparticle conjugate), parity or P (sign reversal of all spatial coordinates), and time reversal or T (sign reversal of the time coordinate, which reverses the role of in and out states). Table 1 shows the effect of these operations on a Dirac spinor field ψ, and Table 2 summarizes the effect of the particular combination CP on some quantities that appear in a gauge theory Lagrangian. In Table 2, the symbol $(-1)^\mu$ denotes a factor +1 for $\mu = 0$ and -1 for $\mu = 1, 2, 3$.

Table 1: The operation of P,C, and T on a Dirac spinor field

$$\begin{aligned} P\psi(t,x)P &= \gamma^0\psi(t,-x) \ , \\ T\psi(t,x)T &= -\gamma^1\gamma^3\psi(-t,x) \ , \\ C\psi(t,x)C &= -i(\overline{\psi}(t,x)\gamma^0\gamma^2)^T \end{aligned}$$

When constructing a field theory we always require locality, the symmetries of Lorentz Invariance, and hermiticity of $\mathcal{L}$. That is sufficient to make any field theory automatically also invariant under the product of operations CPT. In many theories, for example for QED with fermion masses included, the combination CP, and thus also T are also separately automatic. This is the reason why the experimental discovery that CP is not an exact symmetry of nature caused such a stir. All the field theories that had been studied

Table 2: The effect of a CP transformation on various quantities

term	$\overline{\psi}_i\psi_j$	$i\overline{\psi}_i\gamma^5\psi_j$	$\overline{\psi}_i\gamma^\mu\psi_j$	$\overline{\psi}_i\gamma^\mu\gamma^5\psi_j$
CP-transformed term	$\overline{\psi}_j\psi_i$	$-i\overline{\psi}_j\gamma^5\psi_i$	$-(-1)^\mu\overline{\psi}_j\gamma^\mu\psi_i$	$-(-1)^\mu\overline{\psi}_j\gamma^\mu\gamma^5\psi_i$
term	H	A	$W^{\pm\mu}$	∂_μ
CP-transformed term	H	$-A$	$-(-1)^\mu W^{\mp\mu}$	$(-1)^\mu\partial_\mu$

up to that time had automatic CP conservation. So we need to examine how CP non-conservation manifests itself, and then ask what theories will give such effects.

CP non-conservation shows up, for example, as a rate difference between two processes that are the CP conjugates of one-another. How can such a rate difference appear? Consider a particle decay for which two different terms in the Lagrangian (two different Feynman diagrams) give possible contributions. The amplitude for such a process can be written as

$$A = A(A \to B) = g_1 r_1 e^{i\phi_1} + g_2 r_2 e^{i\phi_2} \ . \tag{1}$$

Here g_1 and g_2 are two different, possibly complex, coupling constants in the theory. The transition amplitudes corresponding to each coupling are written as $re^{i\phi}$ to emphasize that they too can have both a real part or magnitude and a phase or absorptive part. The physical source of this phase is that there may be multiple real intermediate states which can contribute to the process in question via rescattering effects. In the jargon of the field the phases ϕ are called strong phases because the rescattering effects among the various coupled channels are dominated by strong interactions. These phases are the same for a process and its CP conjugate because the CP-related sets of intermediate states must contribute the same absorptive part to the two processes. The phases of the coupling constants are often called weak phases because, in the Standard Model, the relevant complex couplings are in the weak interaction sector of the theory. When we look at the amplitude for the CP conjugate process we find

$$\overline{A} = A(\overline{A} \to \overline{B}) = g_1^* r_1 e^{i\phi_1} + g_2^* r_2 e^{i\phi_2} \ . \tag{2}$$

Note that the phases of the coupling constants change sign between any process and its CP conjugate process, while the strong phases, which arise from absorptive parts in the amplitudes, do not.

So now let us calculate the CP-violating difference in rates for these two processes. With a little algebra we find

$$|A|^2 - |\overline{A}|^2 = 2r_1r_2\mathrm{Im}g_1g_2^* \sin(\phi_1 - \phi_2) \ . \tag{3}$$

This shows that the effect will vanish if the two coupling constants can be made relatively real. In addition it depends on the difference of strong phases in the two amplitude contributions, and vanishes if this quantity is zero. Such a CP violation in the comparison of two CP-related decay rates is often called direct CP violation. I prefer the more descriptive term CP violation in the decay amplitudes. Whatever you choose to call it, this effect is characterized by the condition $|\overline{A}/A| \neq 1$. It is obvious that in any process where there is only a single contributing term in the decay amplitude the phase of the coupling constant is irrelevant and $|\overline{A}/A| = 1$ is automatic. You need two different couplings contributing, with non-zero relative phase of the two couplings to see any CP violation.

This statement applies for al types of CP violation. The phase of any single complex coupling in a Lagrangian is not a physically meaningful quantity. In general it can be redefined, and even made to vanish by simply redefining some field or set of fields by appropriate phase factors. But such rephasing of fields can never change the relative phase between two couplings (or products of couplings) that contribute to the same process. Both contributing terms must involve the same nett set of fields, and hence both change in the same way under any rephasings of those fields. These rephasing-invariant quantities are the physically meaningful phases in any Lagrangian, the existence of such a quantity signals the possibility of CP violation.

The second feature we note is that the CP-violating rate difference in Eq. (3) also depends on a difference of strong phases. Typically, this makes it difficult to calculate. Strong phases are, in general, long-range strong interaction physics effects, not amenable to perturbative calculation. One of the things that makes the decays of neutral but flavored mesons particularly interesting is that there we find other types of CP-violation effects where the role played here by the strong phases is replaced by other coupling constant phases, those relevant to the processes that mix the meson with its CP (and thus also flavor) conjugate meson. In such a case we may be able to relate a measured CP violation directly to phase-differences in the Lagrangian couplings, with no need to calculate any strong-interaction quantities. Only in the case of neutral but flavor non-trivial mesons can such mixing-dependent effects occur.

We have seen that only a theory with two coupling constants that are not relatively real can give CP violation. Thus we only can have CP violation in a theory where there is some set of couplings for which rephasing of all fields cannot remove all phases. CP conservation is automatic for any theory for which the most general form of the Lagrangian allows all complex phases to be removed by rephasing of some set of fields. Let us examine a few of the terms that occur in the QED Lagrangian to see why CP conservation is automatic in that theory. For the gauge coupling terms we have, after requiring hermiticity

$$gA^{\mu}\overline{\psi}\gamma_{\mu}\psi + g^{*}A_{\mu}\overline{\psi}\gamma^{\mu}\psi \ . \tag{4}$$

Thus hermiticity clearly makes the QED gauge coupling real, $(g + g^{*})$, because the term it multiplies is itself a hermitian quantity. After imposing hermiticity you will find that the fermion mass term must take the form

$$Re(m)\overline{\psi}\psi + i\mathrm{Im}(m)\overline{\psi}\gamma_5\psi \tag{5}$$

for any complex m. Hermiticity alone does not require that the fermion mass be real, but it does require that the imaginary part multiplies a factor of γ_5. But a chiral rephasing of the fermion field $\psi \to e^{i\phi\gamma_5}\psi$ can be made. This does not change the kinetic or gauge coupling terms at all. In QED, one can always choose the angle ϕ in this rotation in such a way that it makes m a real quantity. This tells us that, in such a theory, the phase of m is not a physically meaningful quantity. Hence the theory is indeed automatically CP conserving for any choice of m. (It is merely for convenience that we always choose to write QED with real particle masses; it is unnecessary to include additional parameters that you know are irrelevant to complicate your calculations.) Tomorrow we will see that this same rephasing is not so innocuous in QCD, and how this leads to the strong CP problem [2].

Given these examples you may be beginning to wonder how we ever get a CP violating coupling into a Lagrangian field theory. That is the question that puzzled everyone in 1964. The trick is to have a sufficient number of different terms in the Lagrangian involving the same set of fields. For example imagine a theory with multiple flavors of fermions and multiple scalar fields. In such a theory there can be Yukawa couplings of the form $Y_{ijk}\phi_k\overline{\psi}_i\psi_j$. Hermiticity then requires only that we also have a term $Y^{*}_{ijk}\phi^{*}_{k}\overline{\psi}_j\psi_i$ in the Lagrangian. Note that this is a different product of fields from the original term, so hermiticity does not disallow phases for the various Y_{ijk} in such a theory. But we still must ask whether we can make every such coupling real, by systematically redefining the phases of the various fields. That depends on the details of the theory. As we add more fields of a given type, either fermions or scalars, the number of possible coupling terms grows more rapidly than total number of fields. With enough fields of the each type there will be more couplings that there are possible phase redefinitions, and then not all couplings can be made real by rephasing the fields.

We can always make all couplings real by imposing CP invariance as a postulate, but it no longer an automatic feature of the theory. It turns out that the Standard Model with only one Higgs doublet and only two fermion generations has automatic CP invariance; all possible couplings can be made simultaneously real (ignoring for now the issue of strong CP-violation via a QCD-theta parameter). Adding one more generation of fermions or adding an additional Higgs doublet with no further symmetries imposed opens up the possibility of CP violating couplings [3]. The three generation Standard Model with a single Higgs doublet has only one CP-violating parameter, that is only one independent phase difference survives after as many couplings as possible are made real by field rephasing. This means that all CP-violating effects in this theory are related. That is what makes it so interesting to test the pattern of CP violation in B decays. Here there are many different channels in which possible CP-violating effects may be observed. In the Standard Model there are predicted relationships between these effects, and between CP violating effects and the values of other CP-conserving Standard Model parameters. Thus the patterns of the B decays, as well as their relationships to the observed CP violation in K-decays, provide ways to test for the effects of physics beyond the Standard Model. Such effects can disrupt the predicted Standard Model relationships between the different measurements.

1.1 Quantum Mechanics of Neutral Mesons

We now we turn to a general discussion of the physics of flavored neutral mesons, those made from different quark and antiquark types of the same charge. These are the K, D, B_d and B_s mesons, which we denote generically by M^0. (I use the notation B_d as a reminder of the quark content, even though the official name of this particle is simply B^0.) There is a beautiful quantum mechanical story here. In each case there are two CP-conjugate flavor eigenstates, $M^0 = \overline{q}q'$ and $\overline{M}^0 = \overline{q}'q$. In general $CPM^0 = e^{i\xi}\overline{M}^0$. The phase ξ is convention dependent and can be altered by redefining one or other of the quark fields by a phase. In much of the literature on this subject the convention $\xi = 0$ is chosen without comment, but elsewhere $\xi = \pi$ is used. Physical results are convention independent, but only as long as you consistently use the same convention. You can get into trouble if you combine formulae taken from two different sources without first checking that both are using the same convention. From this point on I will use the convention $\xi = 0$; if you want to see the equations with arbitrary phase factors explicitly displayed, go to the textbooks [1].

Let us for the moment assume that CP is a symmetry of our theory. What does this tell us about the neutral mesons? It says that the physical propagation-eigenstates of the system, that is the particles which propagate with a distinct mass (and lifetime), must be eigenstates of CP. These are the combinations $(M^0 \pm \overline{M}^0)/\sqrt{2}$. Particles produced by the strong interactions are produced as flavor eigenstates. This means initially one always has a coherent superposition of the two CP eigenstates. Then as time goes on, because of the difference in masses of these two states, their relative phases change. Thus, if both states are long-lived enough, the flavor composition oscillates. However there is also a difference in lifetime of the two CP eigenstates. If this is large then eventually the shorter-lived eigenstate decays away. Once one of the two mass eigenstates has decayed the other combination dominates, terminating the flavor oscillation and giving essentially a fixed admixture from that time on (in vacuum). For the kaon system the difference in lifetime is large compared to the difference in mass, so one does not talk about kaon oscillation, but rather about long-lived and short-lived states. Conversely for B_d the mass difference is large compared to the width difference, and one can discuss either oscillating flavor states, or, discuss the same phenomena in the language of mass eigenstates, $B_{H=\text{heavy}}$ and $B_{L=\text{light}}$. For the B_s both the mass and lifetime differences must be both be considered in analyzing the evolution of states. For the D mesons, in contrast, the mass and width differences are both small in the Standard model. Thus both mass eigenstates decay before any significant oscillation occurs. These particles are thus typically described in terms of flavor eigenstates. Experimental searches for evidence of mixing (mass or width differences) for the D^0 states are another way to seek non-Standard Model physics effects, since the effect as predicted in the standard Model is small [4].

Notice that the peculiar phenomenon of oscillating particles, here and in the neutrino case as well, occurs only if you insist on describing the process in terms of flavor eigenstates. The more physical description is to use the mass eigenstates as the things you call particles (as we do for the quarks themselves). Then all that changes with time is the proportion of the two eigenstates that are present, because of their different half-lives, and the relative phase of the two states, because of their different masses.

Now let us review the story of CP for neutral K mesons. The flavor quantum number strangeness is conserved in strong interactions. Strangeness-changing weak decays are suppressed by the Cabibbo factors $\tan(\theta_{\text{Cabibbo}})$ compared to strangeness conserving $u <-> d$ transitions. This first fact means strange mesons are typically pair produced, the second that they are relatively long lived. The assumption of CP-conservation in neutral Kaon decays "explains" the observation of the two very different half-lives for neutral kaons. If CP were exact, then only the CP-even state, $K_{\text{even}} = (K^0 + \overline{K}^0)/\sqrt{2}$, can decay to two pions, since a spin zero neutral state of two pions can only be CP-even. (By Bose statistics, it can have no I=1 part.) Three-pion final states can be either CP-even or CP-odd. But the phase space for the three pion decay of a neutral kaon is quite small compared to that for two pions. This predicts two very different half-lives for the two CP-eigenstates. They are different, in fact, by more than a factor of ten.

This successful picture was challenged in 1964 by the discovery by Christensen, Cronin, Fitch and Turlay [5], that the long-lived (and hence putatively CP-odd) kaon state did indeed sometimes decay into the CP-even two pion state. This result immediately shows that CP-invariance is violated. Comparison of the rates for charged and neutral pions further showed that the violation is principally in the fact that the mass eigenstate does not have a unique CP. This result was initially very puzzling. Until then almost any field theory that had been considered as a realistic physical theory had automatic CP conservation once the other desired symmetries of were imposed. Now, however, we know that the three generation Standard Model in its most general form includes one CP-violating parameter in the matrix of weak couplings, which is called the CKM matrix (for Cabibbo, Kobayashi and Maskawa). Thus CP violation *per se* is no longer a puzzle, but rather a natural part of the Standard Model. What we do not yet know is whether the Standard Model correctly describes the CP-violation found in nature. Exploration of that question is a major goal of the B-physics program.

Any theory for physics beyond the Standard Model will have, in general, possible additional CP-violating parameters. Any further fields, such as any additional Higgs fields, can introduce further CP-violating couplings. Such effects may then enter into B decay physics. For example, in many models additional Higgs particles lead to additional contributions to B^0-$\overline{B}^0$ mixing. This in turn gives possible deviations from the patterns predicted by the Standard Model for CP-violation in B decays. One of the motivations to search for such effects is that it is not possible to fit the observed matter-antimatter imbalance (or rather the consequent matter to radiation balance) of the Universe with the CP-violation in the quark mixing matrix as the only such effect [6]. (This failure suggests that there must be additional sources of CP-violation beyond those in the quark coupling matrix of the Standard Model, but does not require that any such effects will be apparent in B decays.)

Even with no other new particles, an extension of the Standard Model to include neutrino masses now appears to be needed. Then the weak couplings of the neutrino mass eigenstates are given by a CKM-like matrix. This introduces the possibility of further CP-violating parameters. Indeed if the neutrinos have Majorana type masses there are more CP-violating parameters in this matrix than in the quark case [7]. These parameters will be very difficult to determine and they play essentially no role in B physics. However they may have played an important role in the early universe, giving the matter-antimatter imbalance via leptogensis [8]. I will not discuss neurtrino masses further in these lectures.

As I will discuss tomorrow [2], once there is any CP violation in the Standard Model theory it becomes a problem to understand how it happens that CP is conserved in the strong interaction sector of the theory. Experiment tells us this is so to very high accuracy, chiefly via the upper limit on the electric dipole moment of the neutron. This result tells us that, far as the CP-violating effects that we want to explore in B decays go, we can ignore strong CP violation. So apart from tomorrow's Dirac lecture, I will not discuss it further in this series of talks.

1.2 General Formalism for Neutral Mesons with CP Violation

Once we know that CP is not a symmetry of our theory we must allow a more general form for the two mass eigenstates of neutral but flavored mesons. In the following I use the convention that these two states are defined to be M_H

and M_L where the H and L stand for heavy and light, which really means heavier and less heavy, since the mass difference may indeed be quite tiny.

I define the two eigenstates to be

$$M_H = pM^0 + q\overline{M}^0 \qquad M_L = pM^0 - q\overline{M}^0, \tag{6}$$

where $|p|^2 + |q|^2 = 1$. Note that this equation is again convention dependent, I have not specified a sign or phase for q, but I have defined the more massive state to be the one with a plus sign before q. In combination with my convention that $CPM^0 = \overline{M}^0$ this makes the phase of q a meaningful quantity. (Be aware however that, once again, other conventions are also used in the literature.)

The quantity q/p is determined from the mass and mixing matrix for the two-meson system, $\mathcal{M} = M + i\Gamma$. This matrix is written in the basis of the two flavor eigenstates. Note that both M and Γ are complex 2×2 matrices, M is hermitian and Γ is anti-hermitian. The off-diagonal (or mixing) elements are calculated from Feynman Diagrams that can convert one flavor eigenstate to the other. In the Standard Model these are dominated by the one loop box diagrams, shown in Fig. 1. Actual calculation of such quantities will be discussed in later lectures, for now we simply note that they exist. Then

$$q/p = \frac{\Delta M - i/2\Delta\Gamma}{2(M_{12} - i/2\Gamma_{12})} = \frac{M_{12} - i/2\Gamma_{12}}{2(\Delta M - i/2\Delta\Gamma)}\ . \tag{7}$$

Notice that the two mass eigenstates of this mixed system do not have to be orthogonal, in fact in general they will not be so, unless $|q/p| = 1$.

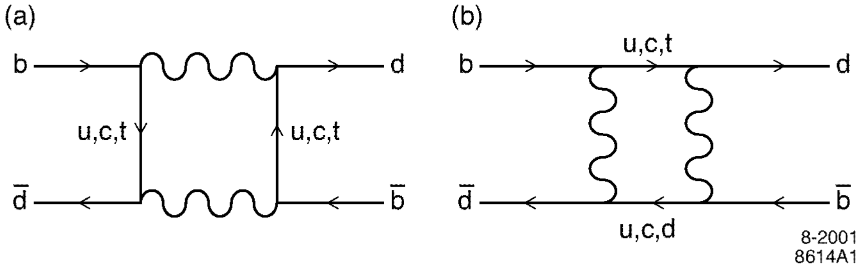

Figure 1: Leading Diagrams for $B\overline{B}$ Mixing in the Standard Model

1.3 The Three Types of *CP* Violation

In the above discussion we have already mentioned two possible ways that CP violation can occur. The first was CP violation in the decay, or direct

CP violation, which requires that two CP-conjugate processes to have differing absolute values for their amplitudes. A second possibility, seen for example in K decays, occurs if $|q/p| \neq 1$. It is very clear in this case that no choice of phase conventions can make the two mass eigenstates be CP eigenstates. This is generally called CP-violation in the mixing. As we will see later, in decays of the neutral mesons to a CP-eigenstate f, there is a third possibility. This can occur even when both the ratio of amplitudes and the quantity q/p have absolute value 1. The CP violation effects in such decays will be shown to depend only on the deviations from unity of the parameter $\lambda_f = (q/p)A(\overline{B}^0 \to f)/A(B^0 \to f)$. The third option is CP violation in the interference between decays to f with and without mixing. This effect is proportional to the imaginary part of λ_f and thus can be non-zero even when the absolute value satisfies $|\lambda_f| = 1$. Decays where this latter condition is true are particularly interesting. In such cases one can interpret any observed asymmetry as a direct measurement of some difference of phases of CKM matrix elements, with no theoretical uncertainties. We will see this in more detail in the next lecture.

2 Lecture 2: Standard Model Predictions for CP Violations in B Decays

2.1 CKM Unitarity

The CKM matrix of quark weak couplings has been discussed in some detail in previous lecture series in this school. It can be written, in the Wolfenstein parameterization [9], as

$$\begin{aligned} V &= \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \\ &\simeq \begin{pmatrix} 1-\lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\lambda^2/2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + O(\lambda^4)\ . \end{aligned} \tag{8}$$

In the previous lecture I talked about the ability to remove, or move, a complex phase of a coupling by redefining the phase of any field involved. This parameterization corresponds to a particular choice of phase convention which eliminates as many phases as possible and puts the one remaining, possibly large, complex phase in the matrix elements V_{ub} and V_{td}.

In this convention the upper right off-diagonal elements define the parameters. The parameterization is a convenient way to make the unitarity of the matrix explicit, up to higher order corrections in powers of $\lambda \equiv V_{us}$. (The higher order terms may also have phases, as required by the unitarity relationships, but bring in no new independent phase parameters.) The quantity λ is essentially the sine of the Cabibbo angle. It is a small number, of order 0.2. Wolfenstein's parameterization uses powers of λ is a convenient way to keep track of the relative sizes of the terms in the matrix. The other independent magnitude parameters A and $\rho^2 + \eta^2$ are known to be roughly of order unity. There is no theory behind which powers of λ enter each term. The Wolfenstein parameterization simply summarizes the observations in a neat way. The fact that V_{cb} and V_{ub} are both small (of order λ^2 and λ^3 respectively in Wolfenstein's parameterization) is responsible for the relatively long lifetimes of B-mesons (and b-containing baryons too). This is a fortunate property; it is essential to the feasibility of most B-physics experiments because it allows us to identify B decays by the spatial separation of the decay vertex from the production point. It is an observational fact, not a theoretical prediction.

Independent of the parameterization used, in the three generation Standard Model the CKM matrix must be unitary. This leads to a number of relationships among its elements of the form [(row)*x(column)]=0. Examples are

$$\begin{array}{rcll} V_{ud}V_{us}^* + V_{cd}V_{cs}^* + V_{td}V_{ts}^* & = & 0 & \quad a \\ V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* & = & 0 & \quad b \\ V_{ub}V_{ud}^* + V_{cb}V_{cd}^* + V_{tb}V_{td}^* & = & 0 & \quad c\,. \end{array} \tag{9}$$

In the Wolfenstein parameterization the relationship that arises from unitarity can be used to express the diagonal and lower left hand elements of the matrix in terms of the upper right elements, to any desired order in λ. The form given above drops terms of order λ^4 and above.

It is a trivial fact that any relationship of the form of a sum of three complex numbers equal to zero can be drawn as a closed triangle in the complex plane. Hence these, and the other similar relationships, are referred to as the Unitarity Triangle relationships. The fact that there is only one independent CP-violating quantity in the CKM matrix can be expressed in phase-convention-invariant form by defining the quantity J, called the

Jarlskog invariant for Cecilia Jarlskog who first pointed out this form [10],

$$\mathrm{Im} V_{ij} V_{kl} V^*_{il} V^*_{kj} = J \Sigma^3_{m,n=1} \epsilon_{ikm} \epsilon_{jln} \tag{10}$$

where i, j, k, l run over the values $1, 2, 3$ and ϵ_{ijk} takes the value $+1$ if the three indices are all different and in cyclic order, and -1 if they are all different and in anti-cyclic order, but is zero if any two are the same. All the unitarity triangles have the same area, $J/2$. This area shrinks to zero if the CP-violating phase differences in the matrix vanish.

Notice however that, while the triangles have the same area, the three examples given above are triangles of very different shapes. Triangle a has two sides of order λ and one of order λ^5. It would be very difficult to measure the area using such a triangle. Triangle b is a little better, but still a has one small angle, its larger sides are of order λ^2 while its small side is of order λ^4 giving an angle of order λ^2. Finally triangle c is the most interesting, because it has all three sides of order λ^3 so all three angles are *a priori* of comparable and large magnitude. The price one pays is that all the sides are small, but this is not as serious as the problem of measuring an asymmetry proportional to a very small angle. This triangle is the one most often discussed in relation to B-meson decays. Since these angles are large one expects some channels in both B_d and B_s decays with order 1 CP-violating asymmetries .

2.2 Fixing the Parameters

The triangle is conventionally drawn by dividing all sides by $V_{cb} V^*_{cd}$, which gives a triangle with base of unit length whose apex is the point (ρ, η) in the complex plane. Prior to considering the asymmetry measurements we can try to determine the shape of this triangle from measurements of CP-conserving quantities which fix the sides, plus the measured CP violation in K-decays. Notice that this information is already sufficient (in principle) to over constrain the set of parameters.

The quantity V_{cb} is determined from B decays to charmed final states, V_{ub} from final states with no charm, while measurements of the B_d and B_s mass differences constrain V_{td}. The CP violation in $K \to \pi\pi$ gives an allowed band for the apex of the triangle. In each case there is both an experimental uncertainty in the measurement and a theoretical uncertainty in the relationship between the measured quantity and the theoretical parameter(s). The theoretical uncertainties dominate. They are typically not statistical in nature, but rather have to do with the part of the calculation which involves models or approximations needed to allow for strong interaction physics effects. There is a large literature by now on the topic of how best to combine the various measurement and deal with both statistical and theoretical uncertainties [11].

New measurements from Belle and BaBar on a CP asymmetry in B-decays constraining the angle at the lower left of the triangle have recently been announced [12]. This is one measurement where the theoretical uncertainties are very small, so the constraint will improve as the statistics of the measurement improve for some time to come. So far all the various results give a consistent picture; the Standard Model fits the data. This means that, within the ranges of the various theoretical uncertainties, there is a region of possible choices for the Lagrangian parameters that are consistent with all data.

One hope of many physicists involved in the large effort in B physics is that at some point some measurements will give discrepant answers for some Standard Model parameters or predictions. This would be evidence for physics beyond the Standard Model, and cause for much excitement in the physics community. If results for some set of measurements should begin to look discrepant, then the question of the statistical significance of the discrepancy will be much debated, as different treatments of theoretical uncertainties will give different conclusions on this point.

Let us examine one of these quantities in a little more detail to see how the theoretical uncertainties arise. In each case there is a mix of weak interaction and short-distance strong-interaction physics, which both are perturbatively calculable and long range strong-interaction physics which is not perturbatively calculable. Tomorrow's lecture will introduce some of the methods that are used to deal with (or avoid) possible long-range strong interaction effects. Here I simply want to show how such effects can enter. Consider the question of the mass difference between the two mass eigenstates for B_d. The two one-loop diagrams given in Fig. 1 are the dominant contribution to this effect. Each loop-diagram can have either a t-, c-, or u-quark for each of the two internal quark lines. Calculation of the matrix element of these diagrams between a B^0 and a $\overline{B}^0$ meson would give $M_{12} + i\Gamma_{12}/2$.

The diagrams can be written as a local four-quark operator multiplied by a calculable coefficient which includes CKM factors. I will write the quark-propagator and coupling dependent part of this coefficient schematically as

$$Q = |V_{td}V_{tb}^*D_t + V_{tcd}V_{cb}^*D_c + V_{ud}V_{ub}^*D_u|^2 \tag{11}$$

where the D_q factors are the quark propagators. This expression is schematic because in writing it as a perfect square I ignored the differences in the momenta of the two quark lines in the diagram (which are typically small, $\mathcal{O}(m_b/m_W)$, compared to the loop momentum itself).

Notice that if all the quarks had equal mass then $D_t = D_c = D_u$ and the unitarity condition Eq. (10c) would say that this factor Q vanishes. Indeed we can use this condition to rewrite the expression as

$$Q = |V_{td}V_{tb}^*(D_t - D_u) + V_{cd}V_{cb}^*(D_c - D_u)|^2. \tag{12}$$

Because of the two W-propagators the loop integral is dominated by momenta of order M_W, which is large compared to either the c or u quark masses. Thus the two quark propagators in the second term of Eq. (12) above essentially cancel one-another, so the term is suppressed by a factor of order $(M_c^2 - M - u^2)/m_W^2$. Thus the mass difference is effectively proportional to the square of the coefficient of the remaining term, which $|V_{td}|^2$ (since V_{tb} is 1 up to order $|\lambda|^4$). (Note that this argument also shows why the mixing matrix is small in the D-meson case. There the three propagators are the down-type quarks, all three of which have masses that are small compared to M_W, so the Unitarity cancellations suppress the entire effect. Furthermore the contribution of the most-massive quark in this case, the b-quark, is Cabibbo-suppressed, further reducing the effect.)

To find the value of this V_{td} by measuring the B meson mass differences we need to know the matrix element of the four quark operator between the B^0 and $\overline{B}^0$ meson states. This is where the long-distance hadronic physics sneaks into the problem, this matrix element depends on the form of the B wavefunction, including all effects of soft gluons. The best available method to determine it is to use lattice QCD calculation [13].

A measurement of the mass difference of the two B_d mass eigenstates thus gives a measurement of V_{td} with a theoretical uncertainty that is dominated by the theoretical uncertainty in the lattice determination of the relevant four-quark matrix element. The result is usually written as some "known" factors times $B_B f_B^2$. (The "known" factors include quark masses, which are actually not so well-known and must be carefully defined.) Here the factor f_b^2 is the vacuum to one meson matrix element of the axial current which arises in the naive approximation to the matrix element obtained by splitting the four-quark operator into two-quark terms and inserting the vacuum state between them. This is known as the vacuum-insertion approximation. The quantity B_B is simply the correction factor between that approximate answer and the true answer. It can be estimated in various model calculations. The lattice calculation does not need to make this subdivision, it directly calculates the full matrix element. However the result is often quoted in terms of the B_B and f_B parameters. Lattice methods can also directly calculate the latter. Eventually f_b will be measured and that will provide a separate test of the lattice calculation.

Once there is a good measurement of the B_s mass difference the ratio $\Delta m_b/\Delta m_s$ will provide a better determination of V_{td} via the ratio V_{td}/V_{ts}. This mass ratio is relatively free of theoretical uncertainties, as most of these cancel in the ratio of matrix elements. The matrix elements for the B_d and the B_s mesons are similar. Only a small correction due to the difference of the s and d quark masses remains. The uncertainty in this correction gives a relatively small theoretical uncertainty in V_{td}. At present only a lower limit for the B_s mass difference is known; even this gives an important constraint (upper limit) on the range of V_{td}.

2.3 Time Evolution of the B States and Time-Dependent Measurements

Now I turn to the topic of decays of neutral B mesons. What can we measure and what does it tell us? To discuss this we need to understand the time evolution of state which at time $t = 0$ is known to be a pure B^0 meson. This means that at t=0 we have

$$B(t=0) = (B_H + B_L)/2p \ . \tag{13}$$

Since the two mass states evolve with different time-dependent exponential prefactors we find

$$B(t) = g_+(t)B^0 + (q/p)g_-(t)\overline{B}^0 \tag{14}$$

where the functions $g_\pm$ are just the sums and differences of the exponential mass and lifetime factors

$$\begin{aligned} g_\pm &= [e^{(-iM_H t - \Gamma_H t/2)} \pm e^{(-im_L t - \Gamma_L t/2)}]/2 \\ &= e^{-iMt-\Gamma t/2}[e^{(-i\Delta M - \Delta\Gamma/2)/2} \pm e^{(i\Delta M + \Delta\Gamma/2)/2}]/2 \ . \end{aligned} \tag{15}$$

Here we introduce the notation M and Γ for the average mass and width and ΔM and $\Delta\Gamma$ for the differences between the two sets of eigenvalues. In the case of B_d the width difference is small compared to the mass difference (and to the width itself) so to a good approximation we can neglect $\Delta\Gamma$. Then the expressions for the $g_\pm$ simplify in an obvious way. For B_s it is likely that the width difference is comparable to the mass difference and the full expressions must be used.

The time-dependent state that is a pure $\overline{B}^0$ at $t = 0$ can likewise be written in terms of these same functions

$$\overline{B}(t) = (p/q)g_-(t)B^0 + g_+(t)\overline{B}^0. \tag{16}$$

It is now straightforward to derive the time-dependent rate to reach a particular CP eigenstate final state f with CP quantum number η_f. It is given by

$$|A(B(t) \to f|^2 = |A(B^0 \to f)|^2[|g_+(t)|^2 + |\lambda_f g_-(t)|^2 + 2Re[g_+^*(t)g_-(t)\lambda_f]] \tag{17}$$

where the quantity

$$\lambda_f = (q/p)\frac{A(\overline{B} \to f)}{A(B \to f)} = \eta_f(q/p)\frac{A(\overline{B} \to \overline{f})}{A(B \to f)}. \tag{18}$$

In the second equality here we have used the fact that f is a CP eigenstate, $CPf = \overline{f} = \eta_f f$ where $\eta_f = \pm 1$, to write the ratio of amplitudes in a form that shows explicitly that one amplitude is simply the CP conjugate of the other.

The CP-violating asymmetry between the rates is defined to be

$$a(t) = \frac{|A(\overline{B}(t) \to \overline{f})|^2 - |A(B(t) \to f)|^2}{|A(\overline{B}(t) \to \overline{f})|^2 + |A(B(t) \to f)|^2} \ . \tag{19}$$

(Note once again you must beware of conventions, some of the literature defines the asymmetry with the opposite sign.)

If $\Delta\Gamma/\Gamma$ can be neglected, which is a very good approximation for B_d decays, then $|q/p| = 1$ and the asymmetry takes the form

$$a(t) = -[(1 - |\lambda_f|^2)\cos(\Delta M t) + 2Im\lambda_f \sin(\Delta M t)]/(1 + |\Lambda_f|^2) \ . \tag{20}$$

As promised previously, this relationship shows that the CP-violating effects measure properties of λ_f, in particular its magnitude and imaginary part.

(In the more general case the expressions are somewhat more complicated and depend also on the width difference.) In particular, if only the third type of CP violation is present, namely if in addition to $|q/p| = 1$ we have $|\overline{A}/A| = 1$ so that $|\lambda_f| = 1$, then this expression simplifies to

$$a(t) = -Im\lambda_f \sin(\Delta Mt)] \ . \tag{21}$$

The argument of λ depends simply on weak phases, so that

$$Im\lambda_f = \eta_f \sin(2\phi_{\text{mixing}} - 2\phi_{\text{decay}}) \ . \tag{22}$$

Here $2\phi_{\text{mixing}}$ is the phase of q/p and $2\phi_{\text{decay}}$ is the phase of $A(\overline{B} \to \overline{f})/A(B \to f)$ while η_f is the CP quantum number of the state f. These phases are each given by some combination of CKM matrix-element phases. While each of them separately can be changed by changes in phase convention (rephasing of quark fields) the difference is convention independent, as must be so for any physically measurable quantity. Thus the asymmetry directly measures the phase differences between particular CKM matrix elements with no uncertainties introduced by our inability to calculate strong interaction physics effects such as the magnitude or strong phase of an amplitude. These strong interaction effects all cancel exactly when $|\lambda_f|$ is 1.

2.4 CP Eigenstate Channels for $b \to c\overline{c}s$

There are many possible channels to investigate. The interest lies not just in one measurement but in whether the pattern of CP-violating asymmetries fits the predictions of the Standard Model. What channels should we study? We need a final state of definite CP. In general for a multibody final state even when the particle content is CP-self conjugate there will be an admixture of CP-even and CP-odd contributions because of different possible orbital angular momenta among the particles. The simplest way to get a definite CP final state is to require that the B decay to a two-body or quasi-two body final state with only one allowed orbital angular momentum. (Quasi-two-body here simply means a two-body state with one or two unstable particles, such as a $\rho\pi$ or $\rho\rho$. The actual observed final state is then three or four pions.) Given that the B has spin zero, the final state has a unique orbital angular momentum between the pair of particles if (and only if) at least one of the two particles has spin zero. For quasi-two body states where both particles have non-zero spin but at least one of them is unstable one can possibly separate out the CP-even and CP-odd final state contributions using an angular analysis of the distribution of secondary decay products [14]. The price is that, in general, a larger data sample is needed to achieve the same accuracy on the CP asymmetry measurement.

Note that the Feynman diagram structure is the same for all channels with the same quark content. Results from multiple channels can sometimes be combined to improve statistical accuracy. For example for the quark decay $b \to c\bar{c}s$ the B^0 decay channels $J/\psi K_S, \psi' K_S, \eta_c K_S$ $J/\psi K_L, \psi' K_L, \xi_c K_L$ (*etc.*) all depend on the same set of quark diagrams. For the $b \to u\bar{u}d$ (and $d\bar{d}d$)quark content there are likewise many channels: $\pi\pi, \rho\pi, \rho\rho$, *etc.* (The last of these needs angular analysis.)

Let us then examine what the predicted CP asymmetry is in each of these two cases. We begin with the modes such as $B \to J/\psi K_s$. These have been called the golden modes for analyzing CP violation in B decay. For once we have a situation where the mode for which the theoretical analysis is straightforward is also one with good experimental accessibility. One still needs a large sample of B decays because the branching fraction to these channels is not large. (In B decays there are so many open channels that branching fractions are small and smaller: the "large" modes occur at the few percent level; $J/\psi K_S$ and similar modes are about a tenth of a percent; a "rare" mode in this game has a branching fraction a few times 10^{-5}.)

First we need a little terminology. We use the term spectator quark for the quark other than the b-type quark (or antiquark) that is present in the initial B meson, since it is generally not involved in the b-decay diagram. There are two topologies of weak decay Feynman diagram that can contribute to B decays to leading order in the weak interactions. These are called "tree" and "penguin" diagrams and are shown in Fig. 2. A tree diagram is one where the W-boson creates or connects to a different quark line from the line that starts out as the b-quark. I thus also include any annihilation diagram or any diagram where the W-boson connects to the spectator quark as part of what I call the tree amplitude. Whenever such a diagram is allowed it will enter with the same CKM factors as the other tree diagram processes. A penguin diagram is a loop-diagram where the W reconnects to the quark line from which it was emitted. Then a hard gluon is emitted from the quark line in the loop, and either makes a pair or is absorbed by the spectator quark.

When higher order strong interaction rescattering effects are included the distinction between tree and penguin diagrams becomes blurred. However, it is useful (and standard) to start out by describing processes in this

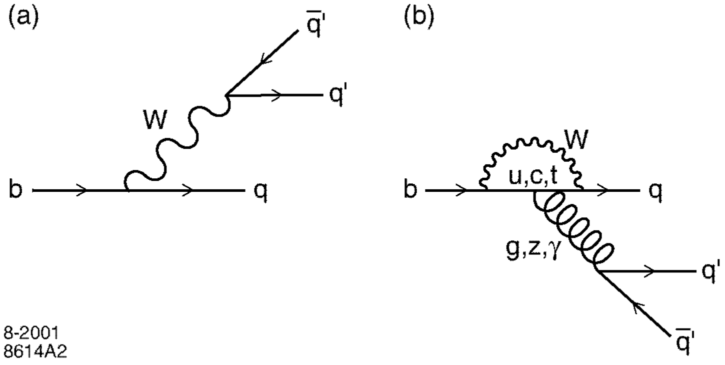

Figure 2: The (a) tree and (b) penguin weak decay processes at the quark level.

language as it allows us to identify all the relevant CKM factors, and the operators which they multiply. As we will shortly see, that is the essence of the story. Eventually we will group terms not by the diagrams, but by the CKM factors. That grouping is not blurred by any subsequent strong interactions. The language tree and penguin persists, but the "tree contribution", in my terminology will be taken to include not only the tree diagrams (including those that involve the spectator in the weak vertex), but also that part of the contribution from the penguin diagrams that has the same CKM factor as the tree diagrams. Obviously, if one wants to try to calculate the size of the contribution to the amplitude one must keep track of each diagram separately, but if we are only concerned with whether there is more than one CKM structure in the significant contributions we can lump together all the terms with a given CKM factor.

The cleanest cases theoretically are those where we can make a prediction without knowing anything about the sizes of the amplitudes because we are looking at a ratio of rates where these cancel to a good approximation. The CP-violating asymmetry in channels arising from quark transition $b \to c\bar{c}s$ in a B_d meson is just this type. The tree diagram has a CKM factor $V_{cb}^* V_{cs}$. Any time that penguin diagrams contribute to an amplitude there are three terms, corresponding to the three different up-type quarks that inside the loop. Thus we can write the b to s penguin amplitude P in the form

$$\begin{aligned} P &= V_{tb}^* V_{ts} f(m_t) + V_{cb}^* V_{cs} f(m_c) + V_{ub}^* V_{us} f(m_u) \\ &= V_{cb}^* V_{cs}[f(m_c) - f(m_t)] + V_{ub}^* V_{us}[f(m_u) - f(m_t)] \end{aligned} \qquad (23)$$

where the $f(m_q)$ is some function of the quark mass. In the second expression I have once again used the Unitarity relationship Eq. (10c) to rewrite the three terms in P in terms of two independent CKM factors. Notice that the first of these is the same as that for the tree term, so for this discussion we call that contribution part of the "tree amplitude". The remaining term is CKM suppressed by an additional factor of λ^2. The two differences of quark-mass-dependent factors are expected to be comparable in magnitude. Furthermore, ignoring CKM factors, the penguin graph contribution is expected to be suppressed by about 0.3 compared to the tree graph, because it is a loop graph and has an additional hard gluon. This means the suppressed second term in Eq. (23) is negligible (a few percent) compared to the "tree amplitude" which here is the sum of the tree term and the dominant penguin term.

Thus we have an amplitude that effectively has only a single CKM coefficient and hence one overall weak phase. This then ensures $|\overline{A}/A| = 1$, which means there is no decay-type (direct) CP violation. (You will recall we needed two terms with different weak phases to get such an effect.) Remember too that for B_d we expect $|q/p| = 1$ to a good approximation. Thus we have a case where $|\lambda_f| = 1$ and the measured asymmetry arises purely from the interference of decay before and after mixing. We find

$$a_{J/\psi K_S} = -Im(\lambda_{J/\psi K_S})\sin(\Delta M t) = \sin(2\beta)\sin(\Delta M t)\ . \qquad (24)$$

Here the quantity β is the lower left-hand angle in the standard B physics Unitarity triangle (also sometimes called ϕ_1). (The minus sign disappears because $\eta_f = -1$ for $f = J/\psi K_S$.) Thus this asymmetry directly measure the phase of a rephasing-invariant combination of CKM elements.

Furthermore all the channels in the $c\bar{c}s$ list above measure the same asymmetry, up to an overall sign, the η_f factor of the channel in question. For example K_S and K_L are states of opposite CP, as are the ψ and η_c. Care must be taken to include the correct η_f factor for each state in combining the results. One can also include a state such as $J/\psi K^*$ provided the K^* decays to a flavor-blind combination such as $K_S\pi^0$, and angular analysis is used to separate CP-even and CP-odd contributions.

One can apply this same diagrammatic analysis to the decays $b \to c\bar{c}s$ in a B_s meson. This gives a prediction for channels such as $J/\psi\phi$ that the CP asymmetry is zero in the Standard Model, as the B_s mixing term is dominated by CKM factors with the same weak phase as this decay. Thus, in the Standard Model, only the CKM suppressed penguin terms which we neglected above can give CP violating asymmetries here, so at most a few

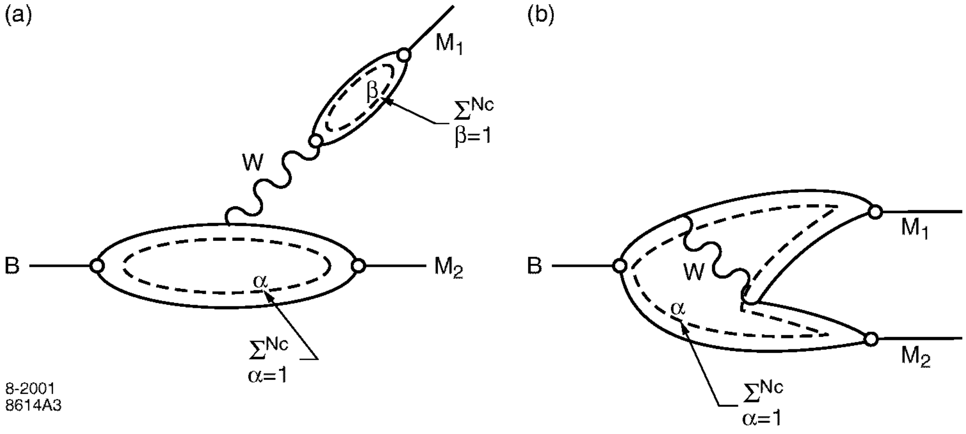

Figure 3: Possible two-meson tree-diagram decay processes showing color-flow loops as dotted lines. These are called (a) color-allowed tree contribution, and (b) color suppressed tree contribution.

percent asymmetry is expected. Such predictions of small or vanishing asymmetries give another way to examine the patterns of the Standard Model. Any theory of new physics effects which give additional mixing contributions could destroy the cancellation of mixing phase and decay phase which makes this asymmetry small in the Standard Model. However to interpret such a result one indeed needs some calculation of decay amplitudes, in order to quantify more precisely how big the "few percent" Standard Model asymmetry could be.

The trick of rewriting the sum of three penguin terms as two terms using the Unitarity relationships is a generally useful tool. In any channel one then has at most two CKM factors to consider. The next step is to get a rough estimate of the relative size of the two terms. This becomes important when $|\overline{A}/A| \neq 1$.

2.5 Some further *B* Physics Jargon

The B physics jargon distinguishes contributions by three attributes, because these three things give a first estimate of how big the contribution is. The first size factor is whether the diagram is tree or penguin. The penguin is suppressed relative to the tree because it is a loop diagram and because it involves a factor of α_{strong} at a scale of order m_b due to the hard gluon, together this makes for a suppression factor of order about 0.3, all else being equal. The next size factor is the powers of the Wolfenstein parameter λ in the associated CKM factors. All B-decay amplitudes have at least two powers of λ. Amplitudes with higher powers are called CKM-suppressed. The third size factor is the color flow pattern that forms the particular final state of interest. Diagrams where a quark-antiquark pair produced by a W finish up in the same meson are called color-allowed, because this pair is produced in the requisite color-singlet combination. In terms of color-flow diagrams there are two independent color-flow loops as shown in Fig. 3(a). When the quark and antiquark produced by the W end up in different final mesons the diagram is called color-suppressed (Fig. 3(b)). There is then only a single color-flow loop so that diagram is expected to be of the order of $1/N_c$ smaller than the corresponding color-allowed diagram.

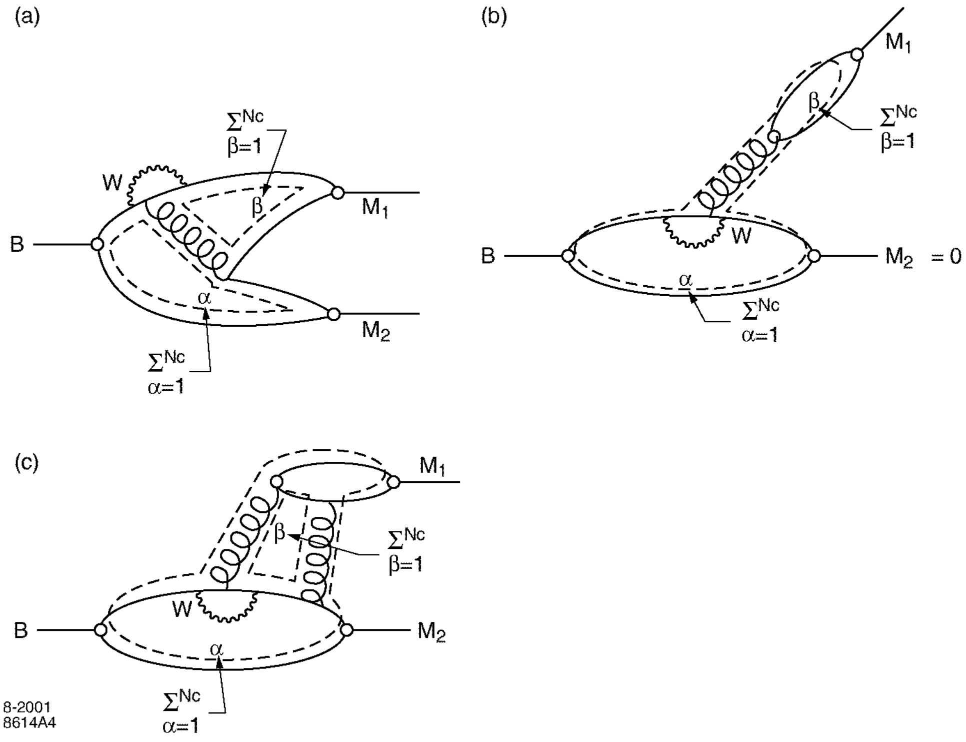

Figure 4: Possible penguin-type two-meson decay processes showing color-flow loops as dotted lines. These are called (a) color allowed penguin, (b) naive color suppressed penguin process, vanishes exactly, and (c) allowed diagram with additional gluon for so-called color-suppressed penguin process. (It has two color flow loops as does the "color-allowed", but an additional α_{qcd} factor.)

For penguin diagrams color suppression, if it works at all, works the other way around. Diagrams where the quark and antiquark from the gluon end up in two different mesons, Fig. 4(a), are color allowed, and indeed can be seen to have two-color-flow loops just as do the tree color-allowed contributions. Diagrams where the flavor-structure says the quark and antiquark produced by the hard gluon must be in the same meson are called color suppressed. In Fig. 4(b) there is only one color loop. However in this diagram the gluon makes a color singlet object. But a gluon is a color-octet state. Taken literally, the diagram vanishes. A second gluon must be exchanged here. If we were to count the extra gluon as a hard gluon, there would be an additional suppression factor of α_{strong}, but no $1/N_C$, because we would again see two color loops, Fig. 4(c). However the second gluon is not necessarily hard, so the relevant scale for the α_{strong} is not large. In some estimates these contributions are treated as $1/N_C$ suppressed terms, but there is no good argument that justifies this counting. As you can see from these arguments, the naive color-counting is not a very reliable measure of the relative strengths of the two types of penguin contributions. QCD-improved operator-product expansion calculations at leading order in Λ/m_b [15, 16, 17] can be made. These treat the color factors correctly. We will return to this approach at later, in Lecture 3. However there is a large literature of estimates that use the language of color-allowed and color-suppressed contributions, so it is important to know how these terms arose and how they are used.

All these size-counting factors are generally used to give first estimates of the order of magnitude of the various contributions. Clearly a more serious calculation can significantly change the relative sizes. The kinematics of the different diagrams are different. The matrix elements of the various operators are different. Indeed there is an interplay between the wave function of the mesons and the counting factors discussed above which in the end determines the size of an amplitude. Powers of Λ_{QCD}/m_b can arise from the wavefunction for particular kinematic configurations relative to others. Higher-order hard QCD effects can be systematically included, but the soft hadronization part of the calculation needs some additional input, either from a model or from some other measurement.

2.6 Another Sample Channel

Now let us look at one more set of channels to see what happens when this size counting says two CKM factors can occur with comparable coefficients. The case I choose to examine is the decay $B_d \to \pi^+\pi^-$. At the quark level this process is governed by decays $b \to u\overline{u}d$. You can readily find from the diagrams of Fig. 2 that there are both tree and penguin contributions for this quark content. The tree diagrams have a CKM factor $V_{ub}^*V_{ud}$. For the penguin contributions we can again use unitarity to rewrite the three different intermediate quark contributions as a sum of two terms. In this case all three CKM coefficients are of the same magnitude. I choose to eliminate $V_{cb}^*V_{cd}$ because then the second penguin term (the one that does not have the same weak phase as the tree term) has the same weak phase as the mixing term in the Standard Model. Then only one difference of CKM phases will enter my eventual formulae for the asymmetry. However we cannot ignore the second penguin term. The only thing that makes it small compared to the "tree amplitude" (which includes the first penguin term as well as the contribution from the tree diagram) is the fact it is a penguin loop. That is not sufficient to completely discard it.

So here we have a situation where there can be $|\overline{A}/A| \neq 1$ effects. We must use Eq. (20) to interpret the the measured asymmetry. One would like to extract from the measurement the CKM phase difference between mixing and tree decay contribution (which in this case is $\alpha \equiv \pi - \beta - \gamma$). One can measure two quantities, $|\lambda_f|$ from the coefficient of $\cos(\Delta M t)$, and $\mathrm{Im}\lambda_f$ from the coefficient of $\sin(\Delta M t)$.

However three unknown quantities enter in the expressions for λ_f in such a case. These are the relative weak phase of mixing and the tree decay amplitude α, and both the absolute value ratio, r, and the relative strong phase, δ of the penguin and tree terms. We can write

$$\lambda_f = e^{-2i\alpha}\frac{1 + re^{i(\delta+\alpha)}}{1 + r^{i(\delta-\alpha)}} \ . \tag{25}$$

Here the phase $\alpha = \pi - \gamma - \beta$ is the angle at the top vertex of the standard B-physics unitarity triangle; it is the difference between the weak phases of the mixing and that of the tree contribution to the decay. Obviously, knowledge of both the real and imaginary parts of λ_f is not enough to fix all three quantities. So we cannot extract a value of α from this asymmetry measurement alone. (Note, however that for very small r the expression simplifies so that the measurement of $\mathrm{Im}\lambda$ determines $\sin 2\alpha$.) We must use further theory or measurement inputs (or both) to determine α if r is not small. (A note of warning here, one often sees the statement that one tests the Standard Model by testing the relationship $\alpha = \pi - \beta - \gamma$ between the angles in the triangle. The relationship is a definition. The tests of the Standard Model are tests of whether one finds the same result for the two independent angles, usually chosen to be β and γ, using a variety of independent ways to measure them.)

Note also that the ratio, $re^{i\delta}$, of the tree to the penguin amplitudes will be different for the different channels with the same quark content. The kinematics of the tree and penguin diagrams are different, and so are the wave functions for forming a π or a ρ, for example. Thus, unlike the $c\bar{c}s$ decays, we cannot simply combine channels to improve statistical accuracy. Instead we must devise methods to remove the dependence on the additional parameters; these methods are different for each set of final state particles.

For the $\pi\pi$ case there are two ways to proceed. One is to rely on isospin symmetry and isospin-related channels to give the needed additional information. The second is to develop methods to calculate these various amplitudes more reliably. This may also involve using relationships to other channels where the tree and penguin amplitudes enter with different relative strengths because of different CKM structure. For example by using measurements on $K\pi$ channels as well with those from $\pi\pi$ channels one can gain some information on the size of the penguin amplitude which dominates the decay in the former case. One can then use SU(3) symmetry to relate that to the size of the penguin in the $\pi\pi$ case. Eventually such methods can much reduce the theoretical uncertainty in the extraction of the CKM parameter γ, or equivalently $\alpha = \pi - \beta - \gamma$. Tomorrow I will discuss both of these approaches in a little more detail.

The set of all possible B decays can be summarized by reviewing all possible b-quark decays and the channels to which they can contribute. A little care must be applied to this logic, as strong rescattering can turn one quark-antiquark combination into another, one must include this possibility in a full treatment. For example in any channel involving a π^0 or ρ^0 meson the penguin diagrams for $b \to d\bar{d}d$ must be added to the diagrams for $b \to u\bar{u}d$. I refer you to the table in the Particle Data Book review on this topic [18] that summarizes the quark decays and gives the CKM factors that enter for each (after using the Unitarity trick to get two terms only.) Any time you start thinking about a specific process you will find you want this information. You can rederive it readily by drawing the allowed quark diagrams and investigating their CKM factors.

3 Lecture 3. Theorist's Tools for *B*-physics

Today's lecture will briefly introduce a number of theoretical tools for calculating B decay processes. There are only a few examples of measurements for which we do not need to know the relative magnitude of various contributions to the decay amplitudes in order to relate the measurement to some parameters in the theory. We would like to go further and interpret the multitude of other measurements that are possible because of the many different B-decay channels. To do this we must devise methods to calculate or relate amplitudes. The available calculational methods all involve some mix of systematic expansion in powers of one or more small parameters, lattice calculation of matrix elements of operators, relationships based on symmetries of the strong interactions such as isospin and SU(3) flavor symmetry, and some input for transition matrix elements and or quark distribution functions. These last can be calculated reliably only in certain limits and in general require models and approximations. Alternately one can measure some of these quantities in one set of processes and use the measured values as input in the interpretation of other measurements.

This lecture will give a general picture of the toolkit of approaches, what each tool is, and how it can be used. There will not be time here to teach the details of any of the methods. This lecture summarizes a large body of theoretical work. I will not attempt to reference all the relevant papers, but will include references to some current papers as examples of the type of work now underway. I apologize in advance to the many whose papers I do not mention.

There are two small parameters in this game, namely Λ_{QCD}/m_b and $\alpha_{\text{strong}}(m_b)$. Here m_b is the mass of the b-quark and Λ_{QCD} is the scale that defines the running of the strong interaction coupling. The detailed definition of each of these quantities is fraught with technical problems, but there is a clear physical meaning for the rough size of these parameters. Λ_{QCD} is related to the inverse size of a typical hadron while the b-quark mass can be characterized as roughly the same scale as the mass of a B meson (up to corrections of order Λ_{QCD}/m_b). The strong coupling $\alpha_s(m_b)$ scales as a logarithm of Λ_{QCD}/m_b; we treat it as a separate small parameter because we can count powers of this parameter separately from the powers of Λ_{QCD}/m_b; they arise in different ways.

The fact that Λ_{QCD}/m_b is indeed quite small leads to a simple intuitive picture of a B meson at rest. It is an essentially static b quark with the light quark forming a cloud around it. The light-quark distribution is sometimes called the brown muck, because we cannot reliably calculate the details of it. However we do know that certain properties are rigorously true in the limit $m_b \to \infty$. For example in that limit the wavefunction does not depend on the spin orientation of the b-quark and hence is the same for a spin 0 B meson and a spin 1 B^*. A second way in which the large mass of the b-quark simplifies the problem is that any gluon that carries off a significant fraction of the b-quark mass is a hard gluon that can be treated perturbatively; it introduces the small parameter $\alpha_{\text{strong}}(m_b)$.

In addition to these expansions there is another part of the picture that is true because m_b/M_W is small. This means that weak decays of the b-quark are essentially local four-quark effects. Thus the B meson decay can, to a reasonable approximation, be thought of as proceeding in two stages: a b-quark decays and then the remnants hadronize to give the final state under study. It is this second stage, the hadronization, that introduces all the uncertainties into the calculations. We have good methods for applying QCD to things like jet-formation for well-separated high momentum quarks, but a B decay does not give us large enough quark momenta to use this formalism reliably. Further, we want to know amplitudes for specific few-body (quasi-two-body) final states (states of definite CP). Most likely these arise when the four quarks that are present after the b decay are not well-separated (so even if the B mass were much larger a jet calculation would not provide the answer). We cannot calculate these amplitudes completely from first principles. So my purpose in this lecture is to review the tools that we do have and how they can be used to minimize the theoretical uncertainty on the extraction of the desired quantities, such as CKM parameters, from experiment.

3.1 Operator Product Expansion

The operator product expansion is a way to formalize the separation of hard or short-distance physics from soft or long-distance physics. It begins by rewriting the Feynman diagrams into the form of local operators, defined at a given scale, with calculable, scale-dependent coefficients.

First we look at all the tree and penguin Feynman diagrams for the weak decay of the b-quark. Each can be written as a sum of four quark operators with definite coefficients at the scale M_W. This is the leading order operator product expansion. There are actually two types of penguin diagrams, those I mentioned earlier that involve a gluon, and a second set called electroweak penguins that involve a photon or a Z particle emitted from the loop. These last give an additional set of four-quark operators. At first glance one might guess that the electroweak penguin contributions are very small, with α_{QED} replacing the α_{strong} of the gluon case. However it turns out there is a part of the Z-penguin contribution which is enhanced by a factor M_t^2/M_W^2 and so there are cases where these terms can be important too.

Each class of diagrams corresponds to a distinct set of four quark operators at leading order. When hard QCD corrections are included, one must introduce a new scale into the problem, which is the hard-soft separation scale μ that defines which gluons are absorbed into the new scale-dependent operator coefficients and which are defined to be included in the scale-dependent matrix elements of operators. In addition, these corrections can mix the operators, and thereby blur the distinction between tree and penguin contributions. Thus the labels of each operator as being tree or penguin type is a leading order distinction only. However they are usually listed in that way as it is a useful way to keep track of which operator arises with which CKM coefficients. In addition, if a hard gluon connects the weak decay vertex to the spectator quark this can also introduce additional local operators that involve six quark fields, again with calculable coefficients that begin at order $\alpha_s(m_b)$.

One must choose the μ-scale that separates hard and soft physics. In principle no physics depends on this choice. In practice if one makes approximations for the matrix elements one does not usually get the correct scale-dependence in their values. So results do to some extent depend on the choice of scale. This dependence is minimized by doing higher order QCD calculations, but in general is not fully removed even with that laborious step.

Each four-quark operator takes the form

$$\mathcal{O}_n = \overline{b}\Gamma_{n1}q^i\overline{q}^j\Gamma_{n1}q^k \tag{26}$$

where each Γ_{ni} denote a specific combination of gamma matrices and QCD color structure and the q^i denote the relevant quark flavor (and color) content. The details of the color and flavor flow in the diagram can be read off once these operators are written. I do not include here the detailed list nor any discussion of the coefficients. That is available many places [1]; my point here is not to discuss this well-developed technical subject, but rather to talk about the additional steps between writing down an operator and its coefficient and calculating an amplitude for any particular channel.

The matrix elements of the operators between the initial B state and the final set of mesons are where hadronic physics enters the game. Our methods for calculating that physics are limited. We can however use information that we do have about symmetries of the strong interactions, for example, to tell us about the ratios of matrix elements that occur in different decays.

3.2 The Factorization Approximation

The simplest approach to the problem, for example for calculation of a color-allowed tree diagram, is to approximate the matrix element in a two-hadron decay as the product of the transition matrix element of a two-quark weak current between the B meson and one final state meson (that can be measured in a semileptonic decay), times the matrix element for the W to create the second meson, which is also measured elsewhere. This approach is called factorization, (or sometimes "naive factorization") because it factorizes the four-quark hadronic operator matrix element into a product of two two-quark matrix elements. This idea can be generalized to divide any four-quark operator into two two-quark operators, which can either be extracted from experiment or estimated using models for the quark distribution functions of the mesons. The approximation neglects any effect of interactions between the two mesons in the final state, effects known as final state interactions.

Now we know that two mesons (for a concrete example think of two pions) colliding at the energy corresponding to a B-mass certainly do interact. So at first glance you may think this approximation has no reason to be accurate. It is certainly not rigorously true, except in a few special cases. However it is motivated by a reasonable physical picture, usually attributed to Bjorken [19] (although in this reference he says the argument is common knowledge).

The idea is that the weak decay is a very local process which converts one quark to three. Only for the kinematic configuration where two of these quarks (or rather one quark and one antiquark) go off essentially together, with the third one recoiling in the opposite direction, is there any significant probability that the system will hadronize as a two-body final state. (All other configurations are assumed to make multi-body final states, for example by fragmentation of the four final-state quarks.) In the special case that gives two-body states the quark and anti-quark that travel together start out much closer together in the transverse direction than the size of a typical hadron. They get quite far from the region containing the other quark and the "brown muck" of the spectator quark before they evolve into the hadronic-sized meson that is observed. They must start out in a color-singlet state to form such a meson. In a local color-singlet configuration (small compared to a meson) the strong interactions must cancel. So initially there are no strong interactions because the pair is in a local color-singlet configuration. Later there is no strong interaction because the two mesons are well-separated and strong interactions are a short-range phenomenon.

The justification of the factorization approximation, as described above, applies for a tree diagram with no direct involvement of the other valence quark of the B meson quark in the weak decay vertex. More generally one can try to factorize any four quark operator (possibly after making a Fierz rearrangement to group the relevant quark fields as flavor-flow dictates they must be grouped to form the mesons of interest). One then uses other measurements, or possibly lattice calculations, to fix the two two-quark matrix elements. In the case of a color-suppressed contribution, or one arising from a penguin diagram the flavor-flow does not automatically match two color-singlet quark pairings. However, if a color-singlet meson is to be formed then there must be a color-singlet piece of the amplitude, and for this piece the factorization argument applies.

In some processes the flavor content of the final state allows a contribution either from annihilation (in the case of a charged B meson) or from exchange of a W between the two initial state valence quarks (for neutral B's). Both processes are suppressed in the heavy quark limit by the quark-mass dependence of the wave-function at the origin (the B to vacuum transition matrix element of a local two-quark current). These contributions are typically neglected in rough estimates of two-hadron decay rates.

Despite all the caveats, the factorization approximation is generally used to make first guess estimates of the sizes of various partial rates. To determine the reliability of this calculation one must look more carefully at what is being done here. I mentioned previously that the operator coefficients can be calculated with hard QCD corrections taken into account. This introduces a scale dependence into their definition, the scale of the separation between hard and soft corrections in QCD. This is not a physical scale, but an arbitrarily chosen one, so the true answer cannot depend on it. Any scale-dependence in the coefficients must be compensated by cancelling scale-dependence in the matrix elements. But when we use measurement of a semi-leptonic process to determine the matrix element there is no reference to any hard-soft division scale; the measured quantity is scale independent. So we clearly have a problem, even in the best cases, factorization cannot be quite correct.

The naive way to deal with this problem is to say it is reasonable to pick a scale somewhere between $m_b/2$ and $2m_b$ since the mass of the b-quark sets the typical momentum scale for the quarks arising from its decay. One then asks how the quantity in question varies as one changes the scale within this range and uses this variation to assign a central value and a theoretical uncertainty to the result. While this seems quite a plausible approach there is no way to be sure it is right. The problem is alleviated somewhat, though not completely removed, when higher order QCD calculations of the operator coefficients are used. It can only be dealt with correctly when a consistent treatment of higher order matrix elements is used, along with the higher order coefficients. Any finite order calculation, however, will typically have some residual scale-dependence problems.

The issue of determining the theoretical uncertainty, that is the reasonable range of values of a theoretical estimate, is one to which we will return again and again in this lecture. Our ability to test the Standard Model by comparing its predictions with experiment depends on our ability to determine how big the uncertainties in our theoretical calculation are. A clean result is one where we know that these uncertainties are very small, or at least where we know very well how big they can be. But more often than not we find a part of the calculation is not so clean. The methods of determining the possible range of the predictions of the Standard Model are all too often subjective and ill-defined. Theorists continue to work to remove such ambiguities, and to find those measurements, or sets of measurements, for which they are minimal. This is an important task.

3.3 Heavy Quark Limit Relationships between *B* and *D* Mesons

One powerful technique for dealing with B decays is use the fact that the b-quark mass is large compared to the QCD scale and to calculate quantities in terms of a power series expansion in that ratio. If one also treats the charm quark as heavy compared to the QCD scale then one has an even more powerful set of relationships. Then to leading order in Λ_{QCD}/m_q the distribution of the light quark in a heavy-light meson is independent of the spin orientation or the mass of the heavy quark. This means it is the same for a B or a B^* or a D or a D^* meson. This is a very important statement because it gives us at least one limit in which we know the transition matrix element between a B and a D or D^* meson.

Consider for example the semi-leptonic decay $B^0 \to D^* \ell \nu$. In the kinematic limit where the D^* is at rest in the B rest frame the wave-function overlap is 1. There is a small but calculable QCD correction to the unit wave-function overlap. Then there are the corrections to the heavy-quark limit relationships, which in this case turn out to be quadratic in Λ_{QCD}/m_q. This is reasonably small even for the charm quark. This means that we can, in principle, use a measurement of this quantity to extract the CKM matrix element V_{cb} with very little theoretical uncertainty. The only problem is that the configuration where this relationship holds is, as I said, a kinematic limit. That means that the rate vanishes at that point! One must measure the rate as a function of q^2, and use an extrapolation to extract the quantity of interest. The extrapolation requires some knowledge about the behavior of the form factor as one goes away from the perfect-overlap situation, and that introduces some theoretical uncertainty into the answer for V_{cb}. However as more data is collected one can measure the rate ever closer to the end point, thereby reducing the sensitivity to the extrapolation.

There are some other technical issues that appear in this problem. One interesting one that crops up here, and in other problems too, is the choice of the definition of the quark mass m_b (or m_c). If you remember from muon decay, the semileptonic decay rate for a fermion (here the b-quark) goes like the fifth power of the mass of the decaying particle. Thus any uncertainty in the definition of the quark mass translates into a huge uncertainty in the predicted rate. But it is even worse than this. If you try to define the quark mass as the mass at the pole of the quark propagator this definition is scale dependent and even diverges as the scale is reduced (known as the renormalon problem). Clearly this is an unphysical effect, because you chose an unphysical definition of the quark mass. The problem is to find a definition that avoids this problem and leads to a well-controlled result. This can indeed be done. The full discussion of how one does it is beyond the scope of this lecture. I merely warn you that you can get into trouble by blithely assuming you know what someone means when they write m_b. This quantity cannot be directly measured. It is dependent on definition convention and on renormalization scale. As you compare results of different calculations you must always be aware of the conventions and definitions that have been used. Otherwise you will not be able to interpret and apply the results correctly.

3.4 QCD-Improved Factorization

The word picture explanation of factorization is to some extent confirmed by explicit calculation of QCD corrections up to order α_S and at leading order in Λ/m_q. It is found that the color-singlet nature of the meson leads to cancellation of the soft-gluon exchange between the two final-state mesons. In general, particularly for processes dominated by penguin or color-suppressed diagrams, there are found to be additional contributions which cannot be described by the simple factorization of a four-quark operator, but rather add to the picture a local six-quark operator. They arise because of a hard-gluon exchange between the so-called spectator quark (now no longer just a spectator) and another quark within the same meson. The matrix elements of this operator can be approximated as the a product of three valence-quark-distribution functions, one for each meson (one initial and two final) times the hard coefficient which begins in order $\alpha_s(m_b)$. Uncertainties arise from limitations on our knowledge of the quark distribution functions.

One has to be careful here when matching the calculated hard-quark coefficient with measured transition matrix elements and form factors. The scale-dependence matching must be done correctly. One must also ensure that one is not double counting contributions of hard quarks that are effectively inside one of the measured quantities. But these are technical problems that can be dealt with correctly.

This treatment is known as qcd-improved factorization [15]. Here the term factorization is used for the factorization of the hard and soft physics. This form of factorization has been demonstrated to work for the leading order in Λ/m_b and one order in $\alpha_s(m_b)$ corrections to the leading diagrams. The actual Λ/M_b power counting is dependent on the assumptions about quark distribution functions; it assumes they vanish as a power of x at their end-point. As the calculation includes all gluon energy scales it is argued that all final state interactions are included in the formalism. The question remains as to whether this argument applies to all orders. It has been proven true to all orders in α_s and leading order in Λ/m_q for the special case of a $D\pi$ final state with flavor such that the spectator quark in the B ends up in the D and the charm quark is treated as a heavy quark in the Λ/m_q power counting [20].

It turns out that the numerical results depend quite sensitively on the details of input assumptions on the quark distribution functions [16, 17]. A variant of the approach making quite different, and indeed additional, assumptions about the quark distribution function end-point behavior gets numerically very different results [17]. The second approach is called perturbative QCD by its proponents. It is claimed in this approach that the entire result is perturbatively calculable. While these claims are open to question [21], one can simply regard the results of this work as the output of a set of ansaetze for the distribution functions. The results raise issues that have contributed important points to the discussion. One is the question of exactly how small some of the (Λ/m_b)-suppressed contributions are in actuality. The annihilation-graph contribution, for example, is found to be significant, even though formally suppressed.

The sensitivity of results to inputs is unfortunate. It means that even these more sophisticated calculations leave us with some significant theoretical uncertainties. The best one can do to quantifying these uncertainties is to see how much the results change when one varies over some reasonable set of assumptions for the various inputs such as quark distribution functions and transition matrix elements. But how do you decide what is a reasonable range? As the existing debates show, in many cases this comes down to some subjective choices, not all rigorously decidable! (Some choices are, however, quite clearly unreasonable and should be excluded from discussion, for example a calculation that sets the scale of transverse momenta in a hadron at $k_\perp^2 = \Lambda m_b$, or a form-factor model that does not fit a rigorous theoretical limit relationship.) As data and calculations for multiple channels are obtained it is likely that we will develop a better understanding of such issues, and a more consistent view of what range of assumptions are reasonable will emerge. Meanwhile it is very important that any calculation reported should include an honest estimate of its uncertainties, and a clear explanation of the assumptions made and the ranges of input variables that were included in obtaining this estimate.

3.5 Isospin

Another useful tool for extracting clean results for strong decay amplitudes is the symmetries of the strong interactions. The best of these, in that it most close to a true symmetry of the hadronic decays, is Isospin symmetry. I find I must explain this symmetry from scratch for current students. It is a piece of old fashioned physics knowledge which is not always taught in modern courses. Isospin is a symmetry under interchange of u and d quark flavors. It is called "iso", because atoms which differ by such an interchange (originally by replacing a neutron by a proton or vice versa) are called isomers because they have nearly equal mass, and "spin" because the two quarks form an SU(2) doublet and the mathematics of SU(2) is the familiar mathematics of spin doublets. Isospin has nothing to do with any angular momentum. Notice also that I do not here mean the weak isospin (so called because it is yet another SU(2)); the isospin doublet is truly u with d, not with some admixture of d,s, and b.

Isospin is, quite obviously, broken by electromagnetic effects since these distinguish quark charges, and it is also broken by quark masses. Now the up and down quark mass are nowhere near the same, the ratio $(m_u - m_d)/(m_u + m_d)$ is not a small number. So why is Isospin ever a good symmetry? The answer is that in many cases, (including most but not all hadron decays) the relevant scale with which to compare the quark mass difference is not the quark mass sum but the hadron mass scale. That scale is set either by Λ_{QCD} or by some heavy quark mass. Then the corrections to isopin-based predictions are small. One must be careful, however, to look out for the cases where the effect is one that is "chirally enhanced" that is where the sum of up and down masses does appear in the denominator. (A similar issue may also arise when making a heavy-quark expansion; terms that behave like $\Lambda^2_{QCD}/m_b(m_u + m_d)$, though formally suppressed in the large m_b limit, are not always numerically negligible.)

How does isospin help clarify B decay processes? Its chief value is that it allows us to make an experimental separation of some tree and QCD-penguin type contributions. In some processes these have different isospin structure, as well as having different CKM structure. Let us take the example of B decaying to two pions. First let us look at the final states, two pions in a spin zero state. A pion has isopin 1. Naively there are three possible isospins for the two-pion states, 0, 1 and 2. However Bose statistics says the overall state must be even under pion interchange. Since the spin zero spatial state is even, the isopin state must be even too. This eliminates the $I = 1$ possibility. Now let us examine the quark decays. The tree $b \to u\overline{u}d$ contribution contains both $\Delta I = 1/2$ and $\Delta I = 3/2$ contributions. These combine with the spectator quark to contribute to the $I = 0$ and $I = 2$ final states respectively. But a gluon is an isosinglet particle—it has no isospin. Hence the $b \to d$ QCD penguin graph is purely $\Delta I = 1/2$ and contributes only to the $I = 0$ final state. (In quark language the gluon makes $u\bar{u}+d\bar{d}$.) We can use measurements of several isospin-related channels (Here $B^0 \to \pi^+\pi^-$, $B^0 \to \pi^0\pi^0$ and $B^+ \to \pi^+\pi^0$ and their CP conjugates) to isolate the $I = 2$ contribution [22]. Then we have found a pure tree process, which thus depends on only one weak phase (up to small corrections from electroweak penguin effects.) Thus the isospin analysis gives us a way to separate out the dependence on α, the difference of the weak phase of the mixing and the weak phase of the tree diagram, without having to calculate the relative strength of the penguin and tree contributions.

The theoretical uncertainty that we found in the previous lecture in trying to extract the CKM parameter α from the asymmetry in $B \to \pi^+\pi^-$ decays can then be much reduced. If, in addition to measuring that time-dependent asymmetry in that channel, one also measures the rates for the isospin related channels, one has, in principle, enough information to determine $\sin(2\alpha)$. Unfortunately, the $\pi^0\pi^0$ rate is expected to be small, so that it may be some time before the experimental uncertainties of this approach are small enough that the result is actually improved by it. However even an upper bound on the neutral pion rate can provide useful constraints [23].

Electroweak penguin effects can also be considered in an isospin analysis, by writing the isospin structure of the Z-boson decay. However, since this decay has isospin 1 as well as isospin 0 parts, there is a $\Delta I = 3/2, I_{\text{final}} = 2$ contribution, and this cannot be separated from the tree term via any multichannel analysis. This results in some residual theoretical uncertainty in the extraction of α, but it is significantly smaller than that from the gluonic penguin contribution without isospin analysis.

A similar situation makes isospin analysis useless in separating tree and penguin parts for $b \to c\bar{c}d$ channels such as D^+D^-. Here both the tree and penguin contributions are pure $\Delta I = 1/2$, so there is no way to distinguish them via their isospin structure.

3.6 SU(3) Symmetry

One can get further relationships between different processes if one extends the idea of isospin to the full flavor SU(3), which treats the three lightest quarks as a degenerate triplet. In particular the subgroup of SU(3) known as U-spin under which the down and strange quarks are a doublet gives lots of interesting relationships between amplitudes [24]. As with any approximate method, the challenge here is to estimate the size of possible corrections from symmetry breaking effects, that is to estimate the theoretical uncertainty in the predictions. One can distinguish three different types of SU(3) breaking effects. First there are kinematic factors that occur because of the different quark (and hence different meson) masses give different phase space factors. These may be large but can be well-estimated and lead to small theoretical uncertainties for any given set of channels. Second there are the factors of F_π (or f_π) versus the similar factors for the kaon. These are measured numbers so, where a vector or pseudoscalar meson is directly produced by a W, they again lead to no significant uncertainties. However when the local operator that produces the light meson is not an axial current then the corresponding ratio is not so well determined. Calculations often use the known ratio of F (or f) factors to estimate the SU(3) breaking in such cases also, but now the uncertainty is not so well-controlled. Finally there are cases where the prediction depends also on assuming an SU(3) relationship between the phases of decay amplitudes. Results sensitive to this assumption may have a larger theoretical uncertainty.

The application of SU(3)symmetry can allow one to use measured penguin-dominated amplitudes such as $B \to K\pi$ to constrain the penguin contribution to a tree-dominated amplitude such as $B \to \pi\pi$. This provides a collection of additional approaches to fix the CKM parameter γ from the combined $\pi\pi$ and $K\pi$ data [25].

Another value of both Isospin and SU(3) relationships is that they provide a window to search for effects of physics beyond the Standard Model. There are a number of cases where possible new physics effects do not respect the relationships predicted by these symmetries [26]. Tests of these relationships may then provide a window for new physics.

3.7 Lattice Calculations

Perhaps the best way to include hadronic physics and QCD effects in a calculation of the matrix element of any operator is to use lattice QCD methods. Methods to treat heavy-light mesons on the lattice have been developed and are steadily improving. There are a number of cases where this method will eventually yield theoretical predictions with well controlled errors. Lattice calculation is particularly useful for quantities such as the B-mixing matrix element which is a one-particle to one-particle transition, or f_B, which is a one-particle to vacuum transition. For one particle to multiparticle transitions (where multi here means two or more) the problem of including final state interactions is not solved by lattice calculations. These calculations are performed in Euclidean space-time and require analytic continuation to give the actual physical result. The uncertainties introduced by this step are difficult to quantify and can be large.

There are basically four sources of uncertainties in lattice of calculations of the one-particle to one-particle (or one to zero-particle) matrix elements. The first is the statistical reliability of the Monte-Carlo treatment. This is simply a matter of doing enough calculation, and is very well understood. Second there are the extrapolations and scale-matching to match the finite-volume, finite-lattice-spacing parameters and results with the infinite-volume continuum quantities. Again the process is highly developed and for the most part in good control. Third are the methods of handling the heavy quark on the lattice, which are also now quite well-developed. The critical last ingredient in this progression is for the lattice calculation to be "unquenched". This means that the lattice allows the development of virtual light quark-antiquark loops. Such calculations require significantly more computer time than the corresponding "quenched calculation" which suppresses quark-loop effects. Unquenched calculations are beginning to appear, for example for the matrix element that is relevant to the mixing between B and $\overline{B}$ mesons. There then remains some extrapolation in the light quark masses and in the number and degeneracies of the light quarks. The prospect is that all sources of uncertainty can be investigated, and that, at least for some of the critical quantities, the lattice will eventually provide the most accurate and well-controlled estimates of the matrix elements. Well-controlled here means that the uncertainty in the estimate can be reliably constrained.

3.8 Quark-Hadron Duality

Even with all these methods we are again and again confronted with data that cannot be interpreted without further input. We are reduced to using models, or to making further assumptions. One commonly used assumption goes under the name of "quark-hadron duality". This is the assumption that if I can calculate a quantity, such as an inclusive rate, at the quark level then that calculation must also give the correct answer at the hadronic level. In a situation where we can average over a range of energies one can indeed prove that this must be true for certain averages, for example the energy-averaged total cross-section for electron-positron collisions to produce hadrons. On the other hand it is clear that if we look in detail at any process the quark result, calculated at low order in QCD, can not reproduce all the details of the hadronic spectrum correctly. In particular, thresholds or end-points of spectra are different for quarks and for mesons. Perturbative quark calculations know nothing about resonance masses, at least not in any fixed-order calculation.

In a B decay we cannot average over energies, the energy of the decay is set by the B mass. Even so it is popularly believed that inclusive B decays can be well-described using the assumption of quark hadron duality. At the quark level we can calculate the b-quark decay. Now we assume that gives the inclusive meson decay correctly, because, if the quark has decayed it must hadronize to something. The level of assurance with which one can make an estimate for the corrections to this approximation varies with the process. For inclusive semi-leptonic decays integrating over lepton momenta provides integration over a range of hadron invariant mass. This can be expected to reduce the corrections. It has thus been argued that these are very small in the inclusive semileptonic case [27].

The demands of realistic measurements can also dilute the power of quark-hadron duality. Consider for example inclusive semi-leptonic decays of B mesons to hadrons that contain no charm. In principle the measurement of this total rate can be used to extract a value for the CKM parameter V_{ub}, if we can calculate the expected rate. We assume quark-hadron duality gives an accurate result for the full inclusive rate, by the arguments given above. However in any experimental measurement, we must make some kinematic restriction in order to exclude backgrounds coming from the much larger rate of decays to hadrons containing charm quarks. This introduces dependence on details of the spectrum, rather than just a particular integral of it.

There is more than one way to choose the kinematic cut: one can for example restrict the electron momentum to be large enough that charm production is excluded; or one can restrict the hadronic invariant mass to be small enough to exclude charm. Because of the unseen neutrino these restrictions are not identical. Each keeps some fraction of the total rate. To extract V_{ub} we must know what that fraction is. But to calculate that fraction we are looking at details of the spectrum for which the use of a quark-level calculation may not be so safe. Recent work has suggested using some combination of cuts on hadron mass and on lepton invariant mass (which requires neutrino reconstruction). A carefully chosen combination can minimize sensitivity to the spectrum end-point details. One can also make some tests as to the stability of the result as the cut prescription is varied [28, 29].

3.9 Models and Other Approximations

In many other channels, even once one uses QCD-improved factorization calculations one needs to know a meson-meson transition matrix and/or quark distribution functions for both initial and final state particles to calculate a rate. Lattice calculation, or measurement in a semi-leptonic decay, can be used to fix the transition matrix element. In certain cases one obtains self-consistent quark distribution functions using light-cone QCD arguments. Or one can parameterize these distributions, for example by their moments, and use some set of measurements to fix the set of parameters that dominate an effect (making sure that such parameters are indeed carefully and consistently defined in both processes).

Finally one can simply resort to making models for the unknown quantities. One can using rigorous limits obtained from QCD sum rules [30] and from the heavy quark limit to constrain the models and reduce the number of independent inputs needed. However this is not sufficient to remove all model dependence of the results. There are often still large (and not well-constrained) uncertainties that arise in this stage of the calculation.

3.10 Summary

For two-body hadronic decays even QCD-improved calculations require some input of transition matrix elements and quark distribution functions for the mesons in question in order to calculate amplitudes. These input quantities can sometimes be constrained by symmetries. Rigorous limits for some can be derived for example from the heavy quark limit and from QCD (*e.g.* the QCD sum rule methods). Some of the quantities of interest can eventually be accurately calculated on the lattice. Some can be measured in semileptonic processes. Data on a great variety of decays will help refine our understanding. This process has already begun. Data from CLEO and from the two asymmetric B factories gives us much to study, and will continue to do so.

Our ability to see whether different measurements yield consistent or inconsistent values for the Standard Model parameters is only as good as our ability to constrain the theoretical uncertainties in a reliable fashion. As one applies any method to a multitude of channels one can learn from experience what accuracy is obtained and refine the method on the basis of that experience. Because there are indeed many possible quasi-two-body

B decays this process will eventually improve our ability to constrain the theoretical uncertainty of a given calculational method. To achieve this ability it is important for theorists to be as precise and as honest as possible about the sensitivity of any results to input assumptions or models, and to explore this sensitivity in some detail. Only in this way can we find those sets of measurements which truly give us sensitive tests of the Standard Model.

4 Lecture 4. Experiments to Measure *B* Decays

In this last lecture I will review how one goes about studying these questions experimentally. Even though you (in this audience) are mostly theory students, it is important that you have some idea of how the measurements are made. The aim of the game is to make multiple measurements that can check Standard Model predictions in a redundant fashion. There are a number of ways that physics from beyond the Standard Model could show up. One could find inconsistent results for a particular Standard Model parameter (or set of parameters) when determining the same parameters by multiple independent methods. One could find a large CP-violating asymmetry in a mode for which the Standard model predicts a small or vanishing effect. One could find decay modes that are predicted to be rare present at a rate different from that expected or with a pattern of isospin or SU(3) symmetry violations that cannot be accommodated within the theoretical uncertainty of Standard Model predictions. Each of these possibilities requires ongoing work on both the theory front, to reduce theoretical uncertainties, and the experimental one, to make all the suggested measurements. I will focus on B decay experiments, but rare K-decay results also contribute to the picture, as do the existing results on CP-violation in K decays.

4.1 Tagging *B* Flavor

Up until now we have talked about various decays of an individual B meson as if we knew what meson we had at time $t = 0$. The flavor conservation of strong and electromagnetic interactions means that one produces a b-quark and an anti-b-quark in the same event. In general one has no *a priori* knowledge of which type of neutral B meson was formed at production. One must use other properties of the total event in order to determine whether one had a B^0 or $\overline{B}^0$ meson at production (or at some other known time). This process is called tagging. For example one can tag a B meson when another B meson in the same event decays in such a way that its b-flavor is identifiable. An example of a tag is a semileptonic decay; the charge of the lepton then identifies whether it came from the weak decay of a b or a $\overline{b}$ quark. The tagging possibilities and efficiencies are quite different in e^+e^- collisions and in hadronic collisions, but the requirement for tagging is common to both types of experiments.

In principle almost every event has some tagging information. Often this information is not precise. For example consider the lepton-charge tag suggested above. If the b-quark decays hadronically to a c-quark which then decays semileptonically then the detected lepton comes from the decay of the c instead of that of the b. Assuming it came from the b will give a wrong sign tag. The spectrum of such secondary-decay leptons is different from that of the primary ones. One can use such additional information to improve the correctness of the tag. However the two spectra overlap, so there will still be cases where there is an ambiguity. Only a probability for each tag-type can be determined. Each type of tag event thus has two properties that must be understood, its efficiency, ϵ, and the wrong tag fraction, w associated with it. Some methods have very high purity but low efficiency, others with much higher efficiency may have lower purity. The measure of tagging quality that eventually determines how well we can measure a CP-violating asymmetry is the product $\epsilon(1-2w)^2$. We will see below how this comes about. Both the efficiency and the wrong tag fraction are determined by a combination of Monte Carlo modelling of events and measurements, for example from samples of doubly tagged events. A significant systematic uncertainty in the result for any asymmetry arises from the uncertainty in determining the wrong tag fraction. Since that determination is at least in part data driven, this uncertainty will decrease as data samples increase.

4.2 e^+e^- Collisions

In an electron-positron collider the most efficient way to produce B^0 mesons is to tune the energy to the Υ_{4s}, since that large resonant peak in event rate is just above threshold to decay into either a B^+ and a B^- or into a B^0 and a $\overline{B}^0$. Hence the Υ_{4s} decays essentially 50% to each of these states. Furthermore, the two neutral mesons are produced in a coherent state which, even though both particles are oscillating as described previously, remains exactly one B^0 and one $\overline{B}^0$ until such time as one of the particles decays. For studies of CP-violation this turns out to be either a disaster or a very useful property depending on the design of your collider.

To observe CP violation we must look for decays where one of the two neutral B's decays in a way that identifies its flavor, so that it gives a good tag, and the other decays to the CP eigenstate of interest for the study. Then we examine the decay rate as a function of the time, t, between the tagging decay (defined to occur at $t = 0$) and the CP-eigenstate decay. When the tag is a $\overline{B}^0$ this means that the particle which decayed to the CP eigenstate is known to have been a B^0 at time $t = 0$ (or, for $t < 0$, to be that combination which would have evolved to be a B^0 at time $t = 0$). We denote this state as $B^0(t)$. Its decay rate as a function of time is given by

$$R(B^0(t) \to f) = |A(B^0 \to f)|^2 e^{-\Gamma|t|}[1+(1-|\lambda_f|^2)\cos(\Delta mt)+Im\lambda_f \sin(\Delta mt)] \tag{27}$$

where once again $\lambda_f = (q/p)[A(\overline{B}^0 \to f)/A(B^0 \to f)]$. In this equation and all following discussion of B_d decays we neglect $\Delta\Gamma$, and, equivalently, assume $|q/p| = 1$. (The corresponding formulae for B_s decays are a little more complicated as this approximation cannot be used in that case, you can find them in the textbooks [1].) Likewise, the rate when the tagging decay is a B^0 is

$$\begin{aligned} R(\overline{B}^0(t) \to f) \quad = \quad & |A(B^0 \to f)|^2 e^{-\Gamma|t|}[|\lambda_f|^2 + (|\lambda_f|^2 - 1)\cos(\Delta mt) \\ & -Im\lambda_f \sin(\Delta mt)] \end{aligned} \tag{28}$$

Notice that if we were to integrate over all times, $-\infty \leq 0 \leq \infty$ the term proportional to $\sin(\Delta Mt)$ would integrate to zero. This would destroy our sensitivity to the CP-violating quantity $\text{Im}\lambda_f$. We must measure the asymmetry between B tags and $\overline{B}$ tags as a function of time to avoid this cancellation. For a symmetric electron positron collider running at the Υ_{4s} this is essentially impossible. (This is the disaster referred to above.) The two B mesons are produced with small momenta. Even with the best detectors one cannot accurately measure the difference in distance from the collision point of the two decays. Indeed the size of the beam-beam interaction region is typically sufficient to destroy any possibility of resolving this difference. Hence cannot measure the time-difference between the decays. Pier Oddone suggested an idea that allowed B factories to be built to tackle CP violation [31]. The idea was to build two storage rings with different energies and collide the electrons and positrons so that the Υ_{4s}, and likewise the pair of B's to which it decays, are produced moving, with a significant relativistic gamma-factor. Then the physical separation of the decay vertices of the two B's is increased via the time dilation of the decay half-life. (A decay vertex is the point from which the tracks of the particles produced in the decay diverge.) In this case one can indeed, using a precision tracking device known as a vertex detector, resolve the two decay vertices and measure their separation with a resolution that is small compared to the average separation. Furthermore, since any transverse motion of the B mesons is small compared to the overall center-of-mass momentum, the distance between the decays (in the higher-energy beam direction) gives a good measure of the time between them. The uncertainty in the production point due to beam size is irrelevant for this measurement, as we are not concerned with time from production, but only the time between the two decays. Thus the initial coherent state gives a beautiful prediction for a measurable time-dependent asymmetry. The experiment has many internal cross checks that can be made to confirm that the effect is seen as predicted. For a detailed discussion of the physics capabilities of such a facility see for example the BaBar Physics Book, which is available via the web [32].

To see how the tagging efficiency affects the result consider how the measured asymmetry is related to the actual asymmetry. The total number of events that we count as B-tagged events is $\epsilon(N_B(1-w)+N_{\overline{B}}w)$ where N_B and $N_{\overline{B}}$ are the actual numbers of B and $\overline{B}$ events produced. Likewise the total count of $\overline{B}$ events is $\epsilon(N_B w + N_{\overline{B}}(1-w))$. Thus the measured asymmetry is

$$a_{\text{meas}} = (1-2w)\frac{(N_B - N_{\overline{B}})}{N_B + N_{\overline{B}})} = (1-2w)a_{\text{true}} \tag{29}$$

where a_{true} is the true asymmetry. In addition the total number of events included in the result scales with ϵ, the tagging efficiency, since only tagged events can be used. Since statistical accuracy grows like the square root of the number of events, the accuracy of the measurement is proportional to the square root of epsilon. Combining these two facts gives you an understanding of the earlier statement that the quality measure for tagging is $\epsilon(1-2w)^2$. This is sometimes called the effective tagging efficiency.

Both asymmetric B factory projects, one at SLAC [33] and the other at KEK [34]), have succeeded spectacularly in building and operating a two-storage-ring facility together with a detector and computer system capable of detecting and recording all the relevant details of millions of $B\overline{B}$ events. Interesting data from these facilities is now beginning to be reported and will continue over the next several years to yield new insights. See the websites of the BaBar [35] and Belle [36] experiments for details.

In addition to measuring CP-violating asymmetries these facilities are also compiling and analyzing large data samples for a variety of B_d decays. Together with measurements from the symmetric B factory at Cornell [37] and its detector CLEO [38], this data will considerably refine our ability to measure the CP-conserving parameters and to test theoretical calculations. I have talked in previous lectures about the uncertainties that plague many theoretical calculation methods, and in particular about the difficulty in quantifying these uncertainties. As data on multiple modes accumulates we can refine our understanding of the accuracy of various approaches by comparison with this data.

4.3 Proton Colliders

Because the B-factory machine's are optimized to run at the Υ_{4s} they are below the threshold to produce any B_s mesons. In principle they could do so by running at the Υ_{5s}. The smaller peak height of this resonance, together with the fact that it has many possible decay channels combine to make the production rate for $B_s\overline{B}_s$ pairs significantly lower than that for B_d at the Υ_{4s}. The machines would have to be be re-optimized to run at this higher energy, which itself is not a simple change. All these factors combine to make it unlikely that this will be attempted any time soon, while there is still so much to learn about the B_d decays. So for measurements of B_s decays, and also for those of baryons containing b-quarks, we need to look elsewhere, to hadron colliders. For the time being that means the Fermilab TeVatron [39], eventually it will also mean LHC [40] at CERN.

At a hadron collider the b and $\overline{b}$ quarks hadronize independently and each B meson is part of a large jet of many particles. Many more B's are produced in high energy hadron-hadron collisions than in an electron-positron B factory. Hadronic collisions also produce many other types of events, with yet higher cross-sections. Thus, for these experiments, it is critical to devise ways to identify B-events fast enough to trigger the system to record the event. The trigger is typically two charged tracks emerging from a B-decay vertex that is separated from the beam-beam collision region. The design of the trigger and its efficiency is a very important and challenging feature of these experiments. The triggering requirements restrict the decay channels that can be studied in a hadronic environment. The methods and efficiencies for tagging the flavor of the produced B are also quite different in the hadronic case than in the electron-positron B factory environment. The tagging particle may be a charged B or a baryon, or it may be deduced from properties of the leading particles in the jet containing the neutral B. Furthermore, since the two b-quark (or antiquark) containing particles are not in a coherent state, the time evolution of the CP-study particle (and also the tagging particle if it is a neutral B-meson) starts at production time. There are a number of interesting quantities that can only be studied in a hadron facility, others where the two types of machines are competitive, and some where the electron-positron machines have unique capabilities. Both approaches are needed to gather all the information we would like to have.

An example of a quantity where hadron collider results will be important is the determination of the side V_{td} of the unitarity triangle. Currently this quantity is determined by measuring the B_d mass difference. However there is a significant theoretical uncertainty that arises when relating the measurement to the parameter V_{td}. Much of this uncertainty would be removed by a measurement of the B_s mass difference as well as that for B_d. The ratio of the two mass differences gives V_{td}/V_{ts} with relatively controlled theoretical uncertainties. If the value predicted by the Standard Model is correct this measurement can be done at Fermilab in the CDF experiment, probably within the next couple of years.

There has been a detailed study of the opportunities for B physics in Run II at Fermilab [41]. The CDF [42] and D-Zero [43] detectors have just completed upgrades and are beginning to take data, including some B-physics-triggered data. In addition a new experiment,known as BTeV, with a detector optimized for B-physics capability, is planned [44]. At CERN there is also such an experiment planned, known as LHCB [45]. These detectors will give expanded B physics capability and perhaps allow some rare modes to be studied, with branching fractions that are too small to measure in the current experiments. (After my talk I was told there is also a study underway of a possible future B experiment at HERA, a follow-up to the HERA-B experiment [46] using a wire target in the proton beam of that e-p collider.) Another future option is an intense Z-production facility at a linear collider, where study of $Z \to b\bar{b}$ decays can yield useful additional possibilities.) All in all, the problem has many aspects. The complementarity of the different experiments will allow a rich program of measurements. Eventually we will have a clear picture of whether the pattern of results matches the Standard Model or requires some physics beyond the Standard Model to describe the data.

4.4 Some Final Remarks

As theorists search for ways to extract interesting information from B decays they will often describe desired measurements that are beyond present capabilities. This is not new. When Bigi and Sanda [47] first talked about CP-violation in B decays we did not know the B lifetime, so the measurements that they proposed seemed out of reach. Sometimes nature is kind and the numbers work out better than present knowledge suggests. Sometimes clever technical ideas, such as the asymmetric e^+e^- collider, extend our experimental reach. Improvements in the technology of particle tracking and particle identification have been essential in the B factory experiments and will continue to be so for BTeV and LHCB. The history of discovery in science continues because measurements deemed impossible in one era become feasible with new developments. Likewise new developments on the theory side, such as new techniques for unquenched lattice calculations are important, as they allow more measurements to be interpreted with good control of theoretical uncertainties.

To conclude this lecture series I would like to remind you that the aim of the game in studying CP is to examine this least-explored corner of the Standard Model in two ways. The first is to pin down the value of the remaining Standard Model parameters. The second is to test whether multiple measurements give consistent answers, both for the parameters and for other Standard Model predictions. The hope is that any discrepancy will be a clue to the nature of physics beyond the Standard Model, physics that can, for example, change the relative phase of a mixing amplitude compared to a decay amplitude. Indirect searches for new physics, such as these B physics probes, are a blunt instrument. Many extensions of the Standard Model may predict similar effects, for example additional contributions to the mixing. The challenge to theorists is to reduce theoretical uncertainties to the point that we sharpen that instrument enough to see the effects if they are there, rather than losing them in the ranges of possible answers given by our poor control of hadronic physics effects. This work is well begun, but there is more to do. I hope some of the students here will make interesting contributions to it in the near future.

References

[1] *CP* VIOLATION. By Gustavo Castelo Branco, Luis Lavoura, Joao Paulo Silva. Oxford Univ. Press, 1999. 511p. (The International Series of Monographs on Physics, Vol. 103) QCD161:B721:1999. *CP* VIOLATION. By I. I. Bigi and A. I. Sanda. Cambridge Univ. Press, 2000. 382p. (Cambridge Monographs on Particle Physics, Nuclear Physics, and Cosmology, Vol. 9)QCD161:B54:2000. The Babar Physics Book SLAC-Report-504

[2] H. Quinn, 2000 Dirac Medal Lecture, ICTP Trieste July 3, 2001.

[3] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49**, 652 (1973). S. Weinberg, Phys. Rev. Lett. **37**, 657 (1976).

[4] See, for example, H. Georgi, Phys. Lett. B **297**, 353 (1992) [arXiv:hep-ph/9209291].

[5] J. H. Christenson, J. W. Cronin, V. L. Fitch and R. Turlay, Phys. Rev. Lett. **13**, 138 (1964).

[6] See for example P. Huet and E. Sather, Phys. Rev. D **51**, 379 (1995) [arXiv:hep-ph/9404302].

[7] R. Svoboda [Super-Kamiokande Collaboration], Nucl. Phys. Proc. Suppl. **98**, 165 (2001). Q. R. Ahmad *et al.* [SNO Collaboration], solar neutrinos at the Sudbury Neutrino Observatory," nucl-ex/0106015. Y. Fukuda *et al.* [Super-Kamiokande Collaboration], Phys. Rev. Lett. **81**, 1562 (1998) [hep-ex/9807003].

[8] See, for example, W. Buchmuller, arXiv:hep-ph/0107153.

[9] L. Wolfenstein, Phys. Rev. Lett. **51**, 1945 (1983).

[10] C. Jarlskog, Phys. Rev. Lett. **55**, 1039 (1985).

[11] M. Ciuchini *et al.*, JHEP **0107**, 013 (2001) [hep-ph/0012308]. A. Hocker, H. Lacker, S. Laplace and F. Le Diberder, LAL-01-14, see also http://www.slac.stanford.edu/ laplace/ckmfitter.html

[12] B. Aubert *et al.* [BaBar Collaboration], Phys. Rev. Lett. **87**, 091801 (2001) [hep-ex/0107013]. K. Abe *et al.* [Belle Collaboration], Phys. Rev. Lett. **87**, 091802 (2001) [hep-ex/0107061].

[13] See, for example, N. Yamada and S. Hashimoto [JLQCD collaboration] [arXiv:hep-ph/0104136] and references contained therein.

[14] I. Dunietz, H. R. Quinn, A. Snyder, W. Toki and H. J. Lipkin, Phys. Rev. D **43**, 2193 (1991).

[15] M. Beneke, G. Buchalla, M. Neubert and C. T. Sachrajda, Phys. Rev. Lett. **83**, 1914 (1999) [hep-ph/9905312].

[16] M. Beneke, G. Buchalla, M. Neubert and C. T. Sachrajda, Nucl. Phys. B **606**, 245 (2001) [hep-ph/0104110].

[17] Y. Y. Keum, H. Li and A. I. Sanda, Phys. Rev. D **63**, 054008 (2001) [hep-ph/0004173].

[18] H. Quinn and A. I. Sanda, Eur. Phys. J. C **15** (2000) 626 Web version at pdg.lbl.gov/

[19] J. D. Bjorken, Nucl. Phys. Proc. Suppl. **11**, 325 (1989).

[20] C. W. Bauer, D. Pirjol and I. W. Stewart, hep-ph/0107002.

[21] S. Descotes-Genon and C. T. Sachrajda, arXiv:hep-ph/0109260.

[22] M. Gronau and D. London, Phys. Rev. Lett. **65**, 3381 (1990).

[23] Y. Grossman and H. R. Quinn, Phys. Rev. D **58**, 017504 (1998) [arXiv:hep-ph/9712306]. M. Gronau, D. London, N. Sinha and R. Sinha, Phys. Lett. B **514**, 315 (2001) [arXiv:hep-ph/0105308].

[24] See, for example, M. Gronau, O. F. Hernandez, D. London and J. L. Rosner, Phys. Rev. D **52**, 6356 (1995) [hep-ph/9504326].

[25] A. J. Buras and R. Fleischer, Eur. Phys. J. C **11**, 93 (1999) [hep-ph/9810260].

[26] See, for example, Y. Grossman, M. Neubert and A. L. Kagan, JHEP **9910**, 029 (1999) [hep-ph/9909297].

[27] I. I. Bigi and N. Uraltsev, hep-ph/0106346.

[28] C. W. Bauer, Z. Ligeti and M. Luke, hep-ph/0107074.

[29] A. K. Leibovich, I. Low and I. Z. Rothstein, Phys. Lett. B **513**, 83 (2001) [hep-ph/0105066].

[30] See, for example, M. Shifman TASI Lectures 1995 in "QCD and Beyond" World Scientific 1995.

[31] P. Oddone in Proceedings of the UCLA Workshop: Linear Collider $B\overline{B}$ Factory Conceptual Design, D. Stork ed. p243 (1987)

[32] SLAC Report 504 (1998) www.slac.stanford.edu/pubs/slacreports/slac-r-504.html

[33] www.slac.stanford.edu/accel/pepii/home.html

[34] www-acc.kek.jp/WWW-ACC-exp/KEKB/KEKB-home.html

[35] www.slac.stanford.edu/BFROOT/www/Public/index.html

[36] http://bsunsrv1.kek.jp/

[37] w4.lns.cornell.edu/public/CESR/

[38] w4.lns.cornell.edu/public/CLEO/

[39] adcon.fnal.gov/userb/www/tevatron/

[40] lhc.web.cern.ch/lhc/

[41] Talks and a preliminary draft of the report can be found at www-theory.fnal.gov/people/ligeti/Brun2/

[42] www-cdf.fnal.gov/

[43] www-d0.fnal.gov/

[44] www-btev.fnal.gov/btev.html

[45] lhcb.cern.ch/

[46] www-hera-b.desy.de

[47] I. I. Bigi and A. I. Sanda, Nucl. Phys. B **193**, 85 (1981). A. B. Carter and A. I. Sanda, Phys. Rev. D **23**, 1567 (1981).

Nonperturbative QCD and Quark-Gluon Plasma

Edward V. Shuryak*

Department of Physics and Astronomy, State University of New York,
Stony Brook, USA

Lectures given at the
Summer School on Particle Physics
Trieste, 18 June - 6 July 2001

LNS0210002

*shuryak@dau.physics.sunysb.edu

Abstract

This is a brief written version of 5 lectures made at 2001 ICTP Summer School on High Energy Physics in Triest. The lectures provide an overview of what we have learned about QCD vacuum, hadrons and hot/dense hadronic matter during the last 2 decades. Last two lectures contain discussion of heavy ion physics. We focus on the first surprising results from new heavy ion collider, RHIC, as well as recent development toward understanding of the old problem of “soft pomeron” in high energy hadronic collisions and its connection to new heavy ion data.

Contents

1 Introduction

1.1 An outline

In these lectures there are not so many formulae: I tried to clarify the main physics point instead, then jump over years of development to the main questions debated today, and show few recent examples. Systematic discussion of such vast range of subjects need a book[1], not short lecture notes. Technical description of instantons can be found in review [2], and the correlation functions in [4].

We will start in Lecture 1 with the QCD *vacuum structure*, in Lecture 2 we then proceed to the *hadronic structure*, discuss *phases* of hot/dense QCD in Lecture 3, and consider *high energy collisions* of heavy ions and hadrons in lectures 4 and 5, respectively.

The main line in all discussion would be a systematic use of semiclassical methods, specifically the instantons. The reasons for that are: (i) They are the only truly non-perturbative effects understood by now; (ii) They lead to large and probably even dominant effects in many cases; (iii) Due to progress during the last decade, we have near-quantitative theory of instanton effects, solved numerically to *all orders* in the so called 't Hooft interaction.

Although we still do not understand confinement, its companion problem - chiral symmetry breaking in the QCD vacuum - is now understood to a significant degree. Not only we have simple qualitative understanding of where those quasi-zero modes of the Dirac operator come from, but we can calculate their density, space-time shape and eventually **QCD correlation functions** with surprising accuracy. So, in a way, the problem of hadronic structure is nearly solved for light-quark hadrons[2].

As we will see below, although the high density and temperature domain can be understood in the (re-summed) perturbation theory, the boundaries of the QCD phases is a matter of non-perturbative physics. I will argue that this phase diagram can also be understood based on the instanton framework. The of three basic phases of QCD: (i) hadronic phase, (ii) Quark-Gluon Plasma (QGP) and (iii) Color Superconductor (CS) phases appear as a balance between three basic pairing channels, being (i) attraction in scalar colorless $\bar{q}q$ channel; (ii) instanton-antiinstanton pairing induced by light quark exchanges; and (iii) attraction in scalar but colored qq channels.

[1] I wrote such a book in mid-80's, [1], and now am working at its new edition.

[2] Medium-heavy-quark ones, such as $\bar{c}c, \bar{b}b$ do care about confining potential, while (hypothetical) extremely heavy quarkonia would need only the Coulomb forces.

The last part deals with high energy collisions of hadrons and heavy ions: those are related to the rest of the lectures since this is how we try to access another QCD phase, the Quark-Gluon Plasma, experimentally. We will discuss first results coming from RHIC, show that matter produced seems to behave macroscopically (namely, hydrodynamically) with proper Equation of State. We will also try to connect rapid onset of QGP equilibration with existing ideas based on perturbative and non-perturbative mechanisms. We will argue that tunneling dynamics described by instantons not only play role in vacuum, but in collisions as well. In this case, however, quantum paths describing the process can transfer from Euclidean to Minkowski space, crossing the so called "turning states" on the way. A sphaleron known in electroweak theory is one of them, and we will argue that those states are physically produced in high energy collisions.

1.2 Scales of QCD

Let me start with an introductory discussion of various "scales" of non-perturbative QCD. The major reason I do this is the following: some naive simplistic ideas we had in the early days of QCD, in the 70's, are still alive today. I would strongly argue against the picture of non-perturbative objects as some structure-less fields with typical momenta of the order of $p \sim \Lambda_{QCD} \sim (1\,fm)^{-1}$. In the mid-70's people considered hadrons to be structure-less "bags" filled with near-massless perturbative quarks, with mild non-perturbative effects appearing at its boundaries and confining them at the scale of 1 fm.

One logical consequence of this picture would be applicability of the derivative expansion of the non-perturbative fields or Operator Product Expansion (OPE), the basis of the QCD sum rules. However, after the first successful applications of the method [5] rather serious problems[7] have surfaced. All spin-zero channels (as we will see, those are the ones directly coupled to instantons) related with quark or gluon-based operators alike, indicate unexpectedly large non-perturbative effects and deviate from the OPE predictions at very small distances.

It provided a very important lesson: *the non-perturbative fields form structures with sizes significantly smaller than 1 fm* and local field strength much larger than Λ^2. Instantons are one of them: in order to describe many of these phenomena in a consistent way one needs instantons of small size [6] $\rho \sim 1/3\,fm$. We have direct confirmation of it from the lattice, but not real understanding of why there are no large-size instantons.

Furthermore, the instanton is not the only such small-scale gluonic object. We also learned from the lattice-based works that QCD flux tubes (or confining strings) also have small radius, only about $r_{string} \approx 1/5\, fm$. So, all hadrons (and clearly the QCD vacuum itself) have a *substructure*, with "constituent quarks" generated by instantons connected by such flux tubes.

Clearly this substructure should play an important role in hadronic physics. We would like to know why the usual quark model has been so successful in spectroscopy, and why so little of exotic states have been seen. Also, high energy hadronic collisions must tell us a lot about substructure, since the famous Pomeron also belongs to a list of those surprisingly small non-perturbative objects.

At the opposite end of the spectrum, people have found that QCD seem to have also surprisingly *small energy/momentum scale*, several times lower than Λ. It was found that behavior of the so called "quenched" and true QCD is very different, but only if the quark mass is below some scale of the order of 20-50 MeV. As we will see below, this surprising low scale has been explained by properties of the instanton ensemble.

2 Lecture 1. The QCD vacuum

2.1 Chiral symmetry breaking and instantons

Let me start around 1961, when the ideas about chiral symmetry and what it may take to break it spontaneously have appeared. The NJL model [13] was the first microscopic model which attempted to derive dynamically the properties of chiral symmetry breaking and pions, starting from some *hypothetical 4-fermion interaction.*

$$L_{NJL} = G(\vec{\pi}^2 + \sigma^2) \tag{1}$$

where π, σ denote the corresponding scalar isovector and scalar isoscalar currents.

Let me make few comments about it.

(i) It was the first bridge between the BCS theory of superconductivity and quantum field theory, leading the way to the Standard Model. It first showed that the vacuum can be truly nontrivial, a superconductor of a kind, with the mass gap Δ=330-400 MeV, known as "constituent quark mass".

(ii) The NJL model has 2 parameters: the strength of its 4-fermion interaction G and the cutoff $\Lambda \sim .8 - 1GeV$. The latter regulates the loops (the model is non-renormalizable, which is OK for an effective theory) and is directly the "chiral scale" we are discussing. We will relate Λ to the typical instanton size ρ, and G to a combination $n\rho^2$ of the size and density of instantons.
(iii) One non-trivial prediction of the NJL model was a the mass of the scalar is $m_\sigma \approx 2m_{const.quark}$. Because this state is the P-wave in non-relativistic language, it means that there is strong attraction which is able to compensate exactly for rotational kinetic energy. For decades simpler hadronic models failed to get this effect, and even now spectroscopists still argue that this (40-year-old!) result is incorrect. However, lattice results in fact show that it is exactly right and theoretically understood by instantons. Moreover, the phenomenological sigma meson is being revived now, so possibly it will even get back to its proper place in Particle Data Table, after decades of absence.

Let me now jump to instantons. We will show below that they generate quite specific 4-fermion 't Hooft interaction [12] (for 2-flavor theory: for pedagogical reasons we ignore strange quarks altogether now). Furthermore, its Lagrangian includes the NJL one, but it also has 2 new terms:

$$L_{tHooft} = G(\vec{\pi}^2 + \sigma^2 - \eta^2 - \vec{\delta}^2) \qquad (2)$$

with isoscalar pseudoscalar η and isovector scalar $\vec{\delta}$. T'Hooft's minus sign is crucial here: it shows that the axial U(1) symmetry (e.g. rotation of sigma into eta) is *not* a symmetry. That is why η (actuallyη' if strangeness is included) is *not* massless Goldstone particle like a pion.

The most important next development happened in 1980's: it has been shown in [6, 14] that instanton-induced interaction does break *spontaneously* the $SU(N_f)$ chiral symmetry. Unlike the NJL model, the instanton-induced interaction has a natural cut-off parameter ρ, and the coupling constants are not free parameters, but determined by a physical quantity, the instanton density. That eventually allowed to solve in all orders in 't Hooft interaction, and get quantitative results, see [2].

2.2 General things about the instantons

I would omit from this paper general things about the instantons, well covered elsewhere. Let me just briefly mention that the topologically-nontrivial

4d solution was found by Polyakov and collaborators in[8], and soon it was interpreted as semi-classical tunneling between topologically non-equivalent vacua. The name itself was suggested by t Hooft, meaning “existing for an instant”. Formally, instantons appear in the context of the semi-classical approximation to the (Euclidean) QCD partition function

$$Z = \int DA_\mu \, \exp(-S) \prod_f^{N_f} \det \left(\not{D} + m_f \right), \tag{3}$$

$$S = \frac{1}{4g^2} \int d^4x \, G^a_{\mu\nu} G^a_{\mu\nu}. \tag{4}$$

Here, S is the gauge field action and the determinant of the Dirac operator $\not{D} = \gamma_\mu(\partial_\mu - iA_\mu)$ accounts for the contribution of fermions. In the semi-classical approximation, we look for saddle points of the functional integral (3), i.e. configurations that minimize the classical action S. This means that saddle point configurations are solutions of the classical equations of motion.

These solutions can be found using the identity

$$S = \frac{1}{4g^2} \int d^4x \left[\pm G^a_{\mu\nu} \tilde{G}^a_{\mu\nu} + \frac{1}{2} \left(G^a_{\mu\nu} \mp \tilde{G}^a_{\mu\nu} \right)^2 \right], \tag{5}$$

where $\tilde{G}_{\mu\nu} = 1/2 \epsilon_{\mu\nu\rho\sigma} G_{\rho\sigma}$ is the dual field strength tensor (the field strength tensor in which the roles of electric and magnetic fields are reversed). Since the first term is a topological invariant (see below) and the last term is always positive, it is clear that the action is minimal if the field is (anti) self-dual

$$G^a_{\mu\nu} = \pm \tilde{G}^a_{\mu\nu}. \tag{6}$$

The action of a self-dual field configuration is determined by its topological charge

$$Q = \frac{1}{32\pi^2} \int d^4x \, G^a_{\mu\nu} \tilde{G}^a_{\mu\nu}. \tag{7}$$

From (5), we have $S = (8\pi^2|Q|)/g^2$. For finite action configurations, Q has to be an integer. The instanton is a solution with $Q = 1$ [8]

$$A^a_\mu(x) = \frac{2\eta_{a\mu\nu} x_\nu}{x^2 + \rho^2}, \tag{8}$$

where the 't Hooft symbol $\eta_{a\mu\nu}$ is defined by

$$\eta_{a\mu\nu} = \begin{cases} \epsilon_{a\mu\nu} & \mu, \nu = 1, 2, 3, \\ & \delta_{a\mu} \nu = 4, \\ & -\delta_{a\nu} \mu = 4. \end{cases} \tag{9}$$

and ρ is an arbitrary parameter characterizing the size of the instanton. This original instanton has its non-trivial topology at large distances, but if we are to consider instanton ensemble, its another form, the so called *singular gauge* on is needed

$$A^a_\mu(x) = \frac{2\overline{\eta}_{a\mu\nu} x_\nu \rho^2}{(x^2+\rho^2)x^2}, \tag{10}$$

because in this case the non-trivial topology is at the point singularity.

The classical instanton solution has a number of degrees of freedom, known as collective coordinates. In addition to the size, the solution is characterized by the instanton position z_μ and the color orientation matrix R^{ab} (corresponding to color rotations $A^a_\mu \to R^{ab}A^b_\mu$). A solution with topological charge $Q=-1$ can be constructed by replacing $\eta_{a\mu\nu} \to \overline{\eta}_{a\mu\nu}$, where $\overline{\eta}_{a\mu\nu}$ is defined by changing the sign of the last two equations in (9).

The physical meaning of the instanton solution becomes clear if we consider the classical Yang-Mills Hamiltonian (in the temporal gauge, $A_0 = 0$)

$$H = \frac{1}{2g^2}\int d^3x\,(E_i^2 + B_i^2), \tag{11}$$

where E_i^2 is the kinetic and B_i^2 the potential energy term. The classical vacua corresponds to configurations with zero field strength. For non-abelian gauge fields this limits the gauge fields to be "pure gauge" $A_i = iU(\vec{x})\partial_i U(\vec{x})^\dagger$. Such configurations are characterized by a topological winding number n_W which distinguishes between gauge transformations U that are not continuously connected.

This means that there is an infinite set of classical vacua enumerated by an integer n. Instantons are tunneling solutions that connect the different vacua. They have potential energy $B^2 > 0$ and kinetic energy $E^2 < 0$, their sum being zero at any moment in time. Since the instanton action is finite, the barrier between the topological vacua can be penetrated, and the true vacuum is a linear combination $|\theta\rangle = \sum_n e^{in\theta}|n\rangle$ called the theta vacuum. In QCD, the value of θ is an external parameter. If $\theta \neq 0$ the QCD vacuum breaks CP invariance. Experimental limits on CP violation require[3] $\theta < 10^{-9}$.

[3] The question why θ happens to be so small is known as the "strong CP problem". Most likely, the resolution of the strong CP problem requires physics outside QCD and we will not discuss it any further.

The rate of tunneling between different topological vacua is determined by the semi-classical (WKB) method. From the single instanton action one expects

$$P_{tunneling} \sim \exp(-8\pi^2/g^2). \tag{12}$$

The factor in front of the exponent can be determined by taking into account fluctuations $A_\mu = A_\mu^{cl} + \delta A_\mu$ around the classical instanton solution. This calculation was performed in a classic paper by 't Hooft [12]. The result is

$$dn_I = \frac{0.47\exp(-1.68N_c)}{(N_c-1)!(N_c-2)!}\left(\frac{8\pi^2}{g^2}\right)^{2N_c}\exp\left(-\frac{8\pi^2}{g^2(\rho)}\right)\frac{d^4z d\rho}{\rho^5}, \tag{13}$$

where $g^2(\rho)$ is the running coupling constant at the scale of the instanton size. Taking into account quantum fluctuations, the effective action depends on the instanton size. This is a sign of the conformal (scale) anomaly in QCD. Using the one-loop beta function the result can be written as $dn_I/(d^4z) \sim d\rho\rho^{-5}(\rho\Lambda)^b$ where $b = (11N_c/3) = 11$ is the first coefficient of the beta function. Since b is a large number, small size instantons are strongly suppressed. On the other hand, there appears to be a divergence at large ρ. In this regime, however, the perturbative analysis based on the one loop beta function is not applicable.

2.3 *Zero Modes and the $U(1)_A$ anomaly*

In the last section we showed that instantons interpolate between different topological vacua in QCD. It is then natural to ask if the different vacua can be physically distinguished. This question is answered most easily in the presence of light fermions, because the different vacua have different axial charge. This observation is the key element in understanding the mechanism of chiral anomalies.

Anomalies first appeared in the context of perturbation theory [9, 10]. From the triangle diagram involving an external axial vector current one finds that the flavor singlet current which is conserved on the classical level develops an anomalous divergence on the quantum level

$$\partial_\mu j_\mu^5 = \frac{N_f}{16\pi^2}G^a_{\mu\nu}\tilde{G}^a_{\mu\nu}. \tag{14}$$

This anomaly plays an important role in QCD, because it explains the absence of a ninth goldstone boson, the so called $U(1)_A$ puzzle.

The mechanism of the anomaly is intimately connected with instantons. First, we recognize the integral of the RHS of (14) as $2N_f Q$, where Q is the topological charge. This means that in the background field of an instanton we expect axial charge conservation to be violated by $2N_f$ units. The crucial property of instantons, originally discovered by 't Hooft, is that the Dirac operator has a zero mode $i\not{D}\psi_0(x) = 0$ in the instanton field. For an instanton in the singular gauge, the zero mode wave function is

$$\psi_0(x) = \frac{\rho}{\pi}\frac{1}{(x^2+\rho^2)^{3/2}}\frac{\gamma\cdot x}{\sqrt{x^2}}\frac{1+\gamma_5}{2}\phi \tag{15}$$

where $\phi^{\alpha m} = \epsilon^{\alpha m}/\sqrt{2}$ is a constant spinor, which couples the color index α to the spin index $m = 1, 2$. Note that the solution is left handed, $\gamma_5\psi_0 = -\psi_0$. Analogously, in the field of an anti-instanton there is a right handed zero mode.

We can now see how axial charge is violated during tunneling. For this purpose, let us consider the Dirac Hamiltonian $i\vec{\alpha}\cdot\vec{D}$ in the field of the instanton. The presence of a 4-dimensional normalizable zero mode implies that there is one left handed state that crosses from positive to negative energy during the tunneling event. This can be seen as follows: In the adiabatic approximation, solutions of the Dirac equation are given by

$$\psi_i(\vec{x},t) = \psi_i(\vec{x},t=-\infty)\exp\left(-\int_{-\infty}^{t} dt'\,\epsilon(t')\right). \tag{16}$$

The only way we can have a 4-dimensional normalizable wave function is if ϵ_i is positive for $t \to \infty$ and negative for $t \to -\infty$. This explains how axial charge can be violated during tunneling. No fermion ever changes its chirality, all states simply move one level up or down. The axial charge comes, so to say, from the "bottom of the Dirac sea".

2.4 *The effective interaction between quarks*

Proceeding from pure glue theory to QCD with light quarks, one has to deal with the much more complicated problem of quark-induced interactions. Indeed, on the level of a single instanton we can not even understand the presence of instantons in full QCD. The reason is again related to the existence of zero modes. In the presence of light quarks, the tunneling rate is proportional to the fermion determinant, which is given by the product of the eigenvalues of the Dirac operator. This means that (as $m \to 0$) the tunneling amplitude vanishes and individual instantons cannot exist!

This result is related to the anomaly: During the tunneling event, the axial charge of the vacuum changes, so instantons have to be accompanied by fermions. The tunneling amplitude is non-zero only in the presence of external quark sources, because zero modes in the denominator of the quark propagator can cancel against zero modes in the determinant. Consider the fermion propagator in the instanton field

$$S(x,y) = \frac{\psi_0(x)\psi_0^+(y)}{im} + \sum_{\lambda\neq 0}\frac{\psi_\lambda(x)\psi_\lambda^+(y)}{\lambda + im} \tag{17}$$

where $i\not{D}\psi_\lambda = \lambda\psi_\lambda$. For N_f light quark flavors the instanton amplitude is proportional to m^{N_f}. Instead of the tunneling amplitude, let us calculate a $2N_f$-quark Green's function $\langle\prod_f \bar\psi_f(x_f)\Gamma\psi_f(y_f)\rangle$, containing one quark and antiquark of each flavor. Performing the contractions, the amplitude involves N_f fermion propagators (17), so that the zero mode contribution involves a factor m^{N_f} in the denominator.

The result can be written in terms of an effective Lagrangian [12]. It is a non-local $2N_f$-fermion interaction, where the quarks are emitted or absorbed in zero mode wave functions. The result simplifies if we take the long wavelength limit (in reality, the interaction is cut off at momenta $k > \rho^{-1}$) and average over the instanton position and color orientation. For $N_f = 1$ the result is [12, 15]

$$\mathcal{L}_{N_f=1} = \int d\rho\, n_0(\rho)\left(m\rho - \frac{4}{3}\pi^2\rho^3\bar{q}_R q_L\right), \tag{18}$$

where $n_0(\rho)$ is the tunneling rate. Note that the zero mode contribution acts like a mass term. For $N_f = 1$, there is only one chiral $U(1)$ symmetry, which is anomalous. This means that the anomaly breaks chiral symmetry and gives a fermion mass term. This is not true for more than one flavor. For $N_f = 2$, the result is

$$\begin{aligned}\mathcal{L}_{N_f=2} = \int d\rho\, n_0(\rho)\Big[\prod_f\left(m\rho - \frac{4}{3}\pi^2\rho^3\bar{q}_{f,R}q_{f,L}\right) \qquad\qquad (19)\\ + \frac{3}{32}\left(\frac{4}{3}\pi^2\rho^3\right)^2\left(\bar{u}_R\lambda^a u_L\bar{d}_R\lambda^a d_L - \bar{u}_R\sigma_{\mu\nu}\lambda^a u_L\bar{d}_R\sigma_{\mu\nu}\lambda^a d_L\right)\Big].\end{aligned}$$

One can easily check that the interaction is $SU(2)\times SU(2)$ invariant, but $U(1)_A$ is explicitly broken. This Lagrangian is of the type first studied by Nambu and Jona-Lasinio [13] and widely used as a model for chiral symmetry breaking and as an effective description for low energy chiral dynamics. It can be transformed to the form discussed above when we compared it to NJL interaction.

2.5 *The quark condensate in the mean field approximation*

We showed in the last section that in the presence of light fermions, tunneling can only take place if the tunneling event is accompanied by N_f fermions which change their chirality. But in the QCD vacuum, chiral symmetry is broken and the quark condensate $\langle\bar{q}q\rangle = \langle\bar{q}_L q_R + \bar{q}_R q_L\rangle$ is non-zero. This means that there is a finite amplitude for a quark to change its chirality and we expect the instanton density to be finite.

For a sufficiently dilute system of instantons, we can estimate the instanton density in full QCD from the expectation value of the $2N_f$ fermion operator in the effective Lagrangian (19). Using the factorization assumption [5], we find that the factor $\prod_f m_f$ in the instanton density should be replaced by $\prod_f m_f^*$, where the effective quark mass is given by

$$m_f^* = m_f - \frac{2}{3}\pi^2\rho^2\langle\bar{q}_f q_f\rangle. \tag{20}$$

This shows that if chiral symmetry is broken, the instanton density is finite in the chiral limit.

This obviously raises the question whether the quark condensate itself can be generated by instantons. This question can be addressed using several different techniques (for a review, see [2, 3]). One possibility is to use the effective interaction (19) and to calculate the quark condensate in the mean field (Hartree-Fock) approximation. This correspond to summing the contribution of all "cactus" diagrams to the full quark propagator. The result is a gap equation [14]

$$\int \frac{d^4k}{(2\pi)^4}\frac{M^2(k)}{k^2+M^2(k)} = \frac{N}{4N_cV}, \tag{21}$$

which determines the constituent quark mass $M(0)$ in terms of the instanton density (N/V). Here, $M(k) = M(0)k^2\varphi'^2(k)/(2\pi\rho)$ is the momentum dependent effective quark mass and $\varphi'(k)$ is the Fourier transform of the zero mode profile [14]. The quark condensate is given by

$$\langle\bar{q}q\rangle = -4N_c\int \frac{d^4k}{(2\pi)^4}\frac{M(k)}{M^2(k)+k^2}. \tag{22}$$

Using our standard parameters $(N/V) = 1\,\mathrm{fm}^{-4}$ and $\rho = 1/3$ fm, one finds $\langle\bar{q}q\rangle \simeq -(255\,\mathrm{MeV})^3$ and $M(0) = 320$ MeV. Parametrically, $\langle\bar{q}q\rangle \sim (N/V)^{1/2}\rho^{-1}$ and $M(0) \sim (N/V)^{1/2}\rho$. Note that both quantities are not proportional to (N/V), but to $(N/V)^{1/2}$. This is a reflection of the fact that spontaneous breaking of chiral symmetry is not a single instanton effect, but involves infinitely many instantons.

A very instructive way to study the mechanism for chiral symmetry breaking at a more microscopic level is by considering the distribution of eigenvalues of the Dirac operator. A general relations that connects the spectral density $\rho(\lambda)$ of the Dirac operator to the quark condensate was given by Banks-Casher relation

$$\langle\bar{q}q\rangle = -\pi\rho(0). \tag{23}$$

This result is analogous to the Kondo formula for the electrical conductivity. Just like the conductivity is given by the density of states at the Fermi surface, the quark condensate is determined by the level density at zero virtuality λ. For a disordered, random, system of instantons the zero modes interact and form a band around $\lambda = 0$. As a result, the eigenstates are delocalized and chiral symmetry is broken. On the other hand, if instantons are strongly correlated, for example bound into topologically neutral molecules, the eigenvalues are pushed away from zero, the eigenstates are localized and chiral symmetry is unbroken. As we will see below, precisely which scenario is realized depends on the parameters of the theory, like the number of light flavors and the temperature. Of course, for "real" QCD with two light flavors at $T = 0$, we expect chiral symmetry to be broken. This is supported by numerical simulations of the partition function of the instanton liquid, see [2].

2.6 The Qualitative Picture of the Instanton Ensemble

Using basically such expressions and the known value of the quark condensate it was pointed out in[6] that all would be consistent only if the typical instanton size happened to be significantly smaller than their separation[4], $R = n^{-1/4} \approx 1fm$, namely $\rho_{\max} \sim 1/3\,\mathrm{fm}$.

In Fig.(1) one can see lattice data on instanton size distribution, obtain by cooling of the original gauge fields. Similar distribution can also be

[4] Derived in turn from the gluon condensate and the topological susceptibility.

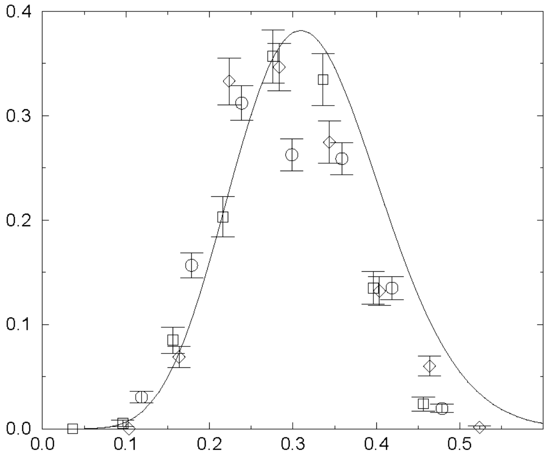

Figure 1: The instanton density $dn/d\rho d^4z$, [fm^{-5}] versus its size ρ [fm]. The points are from the lattice work [11], for this theory, with β=5.85 (diamonds), 6.0 (squares) and 6.1 (circles). Their comparison should demonstrate that results are rather lattice-independent. The line corresponds to one of the proposed expression $\sim exp(-2\pi\sigma\rho^2)$.

obtained from fermionic lowest Dirac eigenmodes: in this case no "cooling" is needed.

Let me now show another evidence for this value of the instanton size, taken from the pion form-factor calculated[16] in the instanton model. In Fig.(2) we show how the experimentally measured pion size correlates with the input mean instanton size: one can see from it that the value .35 fm is a clear winner.

With my current student, Pietro Faccioli, we are now working on the pion form-factor at larger momentum transfer, and have found that the agreement between the instanton-induced contribution and the monopole fit continues to at least $Q^2 \sim 10GeV^2$. At higher momentum transfers, the instanton term must die out, leaving the (probably undetectably small) perturbative asymptotics.

In summary, the following qualitative picture of the QCD vacuum have emerged:

1. Since the instanton size is significantly smaller than the typical separation R between instantons, $\rho/R \sim 1/3$, the vacuum is fairly dilute. The

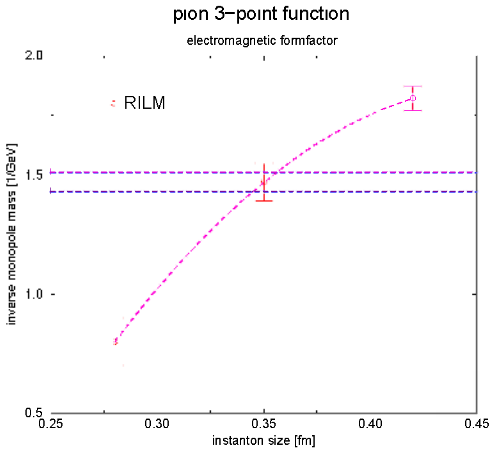

Figure 2: The fitted parameter M of the pion form-factor $ff \sim M^2/(Q^2+M^2)$ versus the inputed instanton size.

fraction of spacetime occupied by strong fields is only a few percent.

2. The fields inside the instanton are very strong, $G_{\mu\nu} \gg \Lambda^2_{QCD}$. This means that the semi-classical approximation is valid, and the typical action is large

$$S_0 = 8\pi^2/g^2(\rho) \sim 10-15 \gg 1. \tag{24}$$

Higher order corrections are proportional to $1/S_0$ and presumably small.

3. Instantons retain their individuality and are not destroyed by interactions. From the dipole formula, one can estimate

$$|\delta S_{int}| \sim (2-3) \ll S_0. \tag{25}$$

4. Nevertheless, interactions are important for the structure of the instanton ensemble, since

$$\exp|\delta S_{int}| \sim 20 \gg 1. \tag{26}$$

This implies that interactions have a significant effect on correlations among instantons, and the instanton ensemble in QCD is not a dilute gas but an *interacting liquid.*

The aspects of the QCD vacuum for which instantons are most important are those related to light fermions. Their importance in the context of chiral symmetry breaking is related to the fact that the Dirac operator has a chiral zero mode in the field of an instanton. These zero modes are localized quark states around instantons, like atomic states of electrons around nuclei. At finite density of instantons those states can become collective, like atomic states in metals. The resulting de-localized state corresponds to the wave function of the quark condensate.

Direct tests of all these ideas on the lattice are possible. One may have a look at the lowest eigenmodes and see if they are related to instantons or something else (monopoles, vortices...) by identifying their shapes - 4d bumps (lines or 2-d sheets) respectively. So far, only bumps (that is the instantons) were seen.

One may also test how locally chiral are the lowest eigenmodes. Just recently lattice practitioners learned how to get very accurate massless fermions on the lattice, and in a variety of ways: the domain wall method, the "perfect" actions or just empirically improved ones based on Wilson-Ginsparg relation. Let me refer to just few papers [25] which discuss those results, confirming the instanton model in its central prediction, that the majority of lowest eigenvectors of the QCD vacuum are made of instanton zero modes.

Let me now explain about the *lowest QCD scale* generated by instantons, mentioned above. The width of the *zero mode zone* of states is of the order of root-mean-square matrix element of the Dirac operator $< I|/D|J >\sim \rho^2/R^3$. Here states I,J are some instanton and anti-instanton zero modes, rho is the instanton size and $R \sim n^{-1/4} \approx 1 fm$ is the distance between their centers. Note small factor $(\rho/R)^2 \sim 1/10$ here. The Dirac eigenvalues from the zone have similar magnitude. Now, the eigenvalues enter together with quark mass m: and so only when this quark mass is smaller than this scale we start seeing the physics of the zero mode zone. In particular, for *quenched* QCD (or instanton liquid) there is no determinant and the zone states have rather wrong spectrum. However, only if the quark mass is small compared to its width we start observing the difference. Only recently lattice practitioners were able to do so: indeed, quenched QCD results at small m start deviating from the correct answers quite drastically.

2.7 Interacting instantons

In the QCD partition function there are two types of fields, gluons and quarks, and so the first question one addresses is *which integral to take first.*

(i) One way is to eliminate *gluonic* degrees of freedom first. Physical motivation for this may be that gluonic states are heavy and an effective fermionic theory should be better suited to derive an effective low-energy fermionic theory. It is a well-trodden path and one can follow it to the development of a similar four-fermion theory, the NJL model. One can do simple mean field or random field approximation (RPA) diagrams, and find the mean condensate and properties of the Goldstone mesons[14]. The results for Color Super-conductors at high density reported below are done with the same technique as well. But nevertheless, not much can really be done in such NJL-like approach. In fact, multiple attacks during the last 40 years at the NJL model *beyond the mean field* basically failed. In particular, one might think that since baryons are states with three quarks, and one may wonder if using quasi-local four-fermion Lagrangians for the three body problem is a solvable quantum mechanical problem, and one can at least tell if nucleons are or are not bound in NJL. In fact it is not: the results depend strongly on subtleties of how the local limit for the interaction is defined, and there is no clear answer to this question. Other notorious attempts to sum more complicated diagrams deal with the possible modification of the the chiral condensate. Some works even claim that those diagrams destroy it *completely*!

Going from NJL to instantons improves the situation enormously: the shape of the form-factor is no longer a guess (it is provided by the shape of zero modes) and one can in principle evaluate any particular diagram. However *summing them all up* still seems like an impossible task.

(ii) The solution to this problem was found. For that one has to follow the opposite strategy and do the $fermion$ integral first. The first step is simple and standard: fermions only enter quadratically, leading to a fermionic determinant. In the instanton approximation, it leads to the Interacting Instanton Liquid Model, defined by the following partition function:

$$Z = \sum_{N_+,\,N_-} \frac{1}{N_+!N_-!} \int \prod_i^{N_++N_-} [d\Omega_i \; d(\rho_i)] \exp(-S_{\rm int}) \prod_f^{N_f} \det(\hat{D}+m_f)\,, \quad (27)$$

describing a system of pseudo-particles interacting via the bosonic action and the fermionic determinant. Here $d\Omega_i = dU_i \, d^4 z_i \, d\rho_i$ is the measure in color orientation, position and size associated with single instantons, and $d(\rho)$ is the single instanton density $d(\rho) = dn_{I,\bar{I}}/d\rho dz$.

The gauge interaction between instantons is approximated by a sum of pure two-body interaction $S_{\rm int} = \frac{1}{2}\sum_{I\neq J} S_{\rm int}(\Omega_{IJ})$. Genuine three body effects in the instanton interaction are not important as long as the ensemble is reasonably dilute. Implementation of this part of the interaction (quenched simulation) is quite analogous to usual statistical ensembles made of atoms.

As already mentioned, quark exchanges between instantons are included in the fermionic determinant. Finding a diagonal set of fermionic eigenstates of the Dirac operator is similar to what people are doing, e.g., in quantum chemistry when electron states for molecules are calculated. The difficulty of our problem is however much higher, because this set of fermionic states should be determined for *all* configurations which appear during the Monte-Carlo process.

If the set of fermionic states is however limited to the subspace of instanton zero modes, the problem becomes tractable numerically. Typical calculations in the IILM involved up to N$\sim$ 100 instantons (+anti-instantons): which means that the determinants of $N \times N$ matrices are involved. Such determinants can be evaluated by an ordinary workstation (and even PC these days) so quickly that a straightforward Monte Carlo simulation of the IILM is possible in a matter of minutes. On the other hand, expanding the determinant in a sum of products of matrix elements, one can easily identify the sum of all closed loop diagrams up to order N in the 't Hooft interaction. Thus, in this way one can actually take care of about 100 factorial diagrams!

3 Lecture 2. Hadronic Structure and the QCD correlation functions.

3.1 Correlators as a bridge between hadronic and partonic worlds

Consider two currents separated by a *space − like* distance x (which can be considered as the spatial distance, or an Euclidean time) and introduce correlation functions of the type

$$K(x) =< T(J(x)J(0)) > \tag{28}$$

with $J(x) = \bar{\psi}(x)\Gamma\psi(x)$. The matrix Γ contains γ_μ for vector currents, γ_5 for the pseudoscalar or 1 for the scalars, etc, and also a flavor matrix, if needed.

We will start with isovector vector and axial currents, and then discuss 4 scalar-pseudoscalar channels: π (P=-1, I=1), σ or f_0 (P=+1,I=0), η (P=-1,I=0) and δ or a_0 (P=+1,I=1).

In a (relativistic) field theory, correlation functions of gauge invariant local operators are the proper tool to study the spectrum of the theory. The correlation functions can be calculated either from the physical states (mesons, baryons, glueballs) or in terms of the fundamental fields (quarks and gluons) of the theory. In the latter case, we have a variety of techniques at our disposal, ranging from perturbative QCD, the operator product expansion (OPE), to models of QCD and lattice simulations. For this reason, correlation functions provide a bridge between hadronic phenomenology on the one side and the underlying structure of the QCD vacuum on the other side.

Loosely speaking, hadronic correlation functions play the same role for understanding the forces between quarks as the NN scattering phase shifts did in the case of nuclear forces. In the case of quarks, however, confinement implies that we cannot define scattering amplitudes in the usual way. Instead, one has to focus on the behavior of gauge invariant correlation functions at short and intermediate distance scales. The available theoretical and phenomenological information about these functions was recently reviewed in [4].

In all cases at small x we expect $K(x) \approx K_0(x)$ where the latter corresponds to just $free$ propagation of (about massless) light quarks. The zeroth order correlators are all just $K_0(x) = 12/(\pi^4 x^6)$, basically the square of the massless quark propagator.

The first deviations due to non-perturbative effects can be studied using Wilsonian Operator Product Expansion (OPE) in ref[5]. For all scalar and pseudoscalar channels the resulting first correction is

$$\frac{K(x)}{K_0(x)} = 1 + \frac{x^4}{384} < (gG)^2 > +... \tag{29}$$

The "gluon condensate" is assumed to be made out of a soft vacuum field, and therefore all arguments can be simply taken at the point $x = 0$. The so-called *standard* value of the "gluon condensate" appearing here was estimated previously from charmonium sum rules:

$$< (gG)^2 >_{SVZ} \approx .5\, GeV^4 \tag{30}$$

Thus, the OPE suggests the following scale, at which the correction becomes equal to the first term:

$$x_{OPE} = (384/ < (gG)^2 >_{SVZ})^{1/4} \approx 1.0\, fm \tag{31}$$

This seems to be completely consistent with the approximation used. However, as Novikov, Shifman, Vainshtein and Zakharov soon noticed[7], this (and other OPE corrections) completely failed to describe all the $J^P = O^{\pm}$ channels: we return to this issue after we consider vectors and axials.

3.2 Vector and axial correlators

The information available on vector correlation functions from experimental data on $e^+e^- -> hadrons$, the OPE and other exact results was reviewed in [4]. Since then, however, new high statistics measurement of hadronic τ decays $\tau \to \nu_\tau +$hadrons have been done. For definiteness, we use results of one of them, ALEPH experiment at CERN [17, 18].

The vector and axial-vector correlation functions are $\Pi_V(x) = \langle j^a_\mu(x) j^a_\mu(0)\rangle$ and $\Pi_A(x) = \langle j^{5\,a}_\mu(x) j^{5\,a}_\mu(0)\rangle$. Here, $j^a_\mu(x) = \bar{q}\gamma_\mu\frac{\tau^a}{2}q$, $j^{5\,a}_\mu(x) = \bar{q}\gamma_\mu\gamma_5\frac{\tau^a}{2}q$ are the isotriplet vector and axial-vector currents. The Euclidean correlation functions have the spectral representation [4]

$$\Pi_{V,A}(x) = \int ds\, \rho_{V,A}(s) D(\sqrt{s}, x), \tag{32}$$

where $D(m,x) = m/(4\pi^2 x)K_1(mx)$ is the Euclidean coordinate space propagator of a scalar particle with mass m. We shall focus on the linear combinations $\Pi_V + \Pi_A$ and $\Pi_V - \Pi_A$. These combinations allows for a clearer separation of different non-perturbative effects. The corresponding spectral functions $\rho_V \pm \rho_A$ measured by the ALEPH collaboration are shown in Fig. 3. The errors are a combination of statistical and systematic ones (below we use them conservatively, as pure systematic): the main problem seems to be separation into V and A of channels with Kaons, which may affect $V - A$ at $s > 2\, GeV$ at 10% level. None of our conclusions are sensitive to it.

In QCD, the vector and axial-vector spectral functions must satisfy chiral sum rules. Assuming that $\rho_V - \rho_A = 0$ at above $s > m_\tau^2$, and using ALEPH data below it, one finds that all 4 of the sum rules are satisfied within the experimental uncertainty, but the central values differ significantly from the chiral predictions [17]. In general, both functions are expected to have oscillations of decreasing amplitude, and putting $\rho_V - \rho_A$ to zero at arbitrary

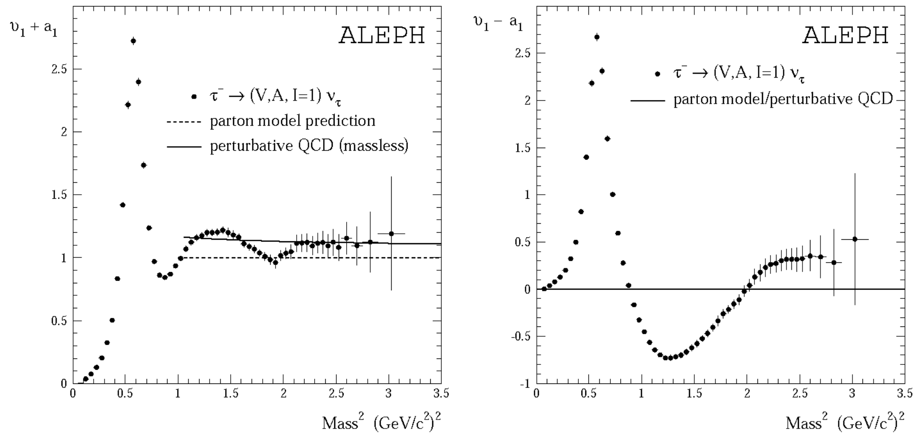

Figure 3: Spectral functions $v(s) \pm a(s) = 4\pi^2(\rho_V(s) + \rho_A(s))$ extracted by the ALEPH collaboration from *tau* lepton hadronic decays.

point imply appearance of spurious dimension $d = 2, 4$ operators in the correlation functions at small x. Therefore, we have decided to terminate the data above a specially tuned point, $s_0 = 2.5\,\mathrm{GeV}^2$, enforcing all 4 chiral sum rules. (The reader should however be aware of the fact that we have, in effect, slightly moved the data points in the small x region within the error band.) Finally we add the pion pole contribution (not shown in Fig. 3), which corresponds to an extra term $\Pi^{\pi}_{A}(x) = f^2_{\pi} m^2_{\pi} D(m_{\pi}, x)$. The resulting correlation functions $\Pi_V(x) \pm \Pi_A(x)$ are shown in Figs. 4.

We begin our analysis with the combination $\Pi_V - \Pi_A$. This combination is sensitive to chiral symmetry breaking, while perturbative diagrams, as well as gluonic operators cancel out.

In Fig. 4 we compare the measured correlation functions with predictions from the instanton liquid model (in its simplest form, random instanton liquid with parameters n, ρ fixed in [6] and discussed above).

The agreement of the instanton prediction with the measured $V - A$ correlation is impressive: it extends all the way from short to large distances. At distances $x > 1.25$ fm both combinations are dominated by the pion contribution while at intermediate x the ρ, ρ' and a_1 resonances contribute.

We shall now focus our attention on the $V + A$ correlation function. The unique feature of this function is the fact that the correlator remains close to free field behavior for distances as large as 1 fm. This phenomenon

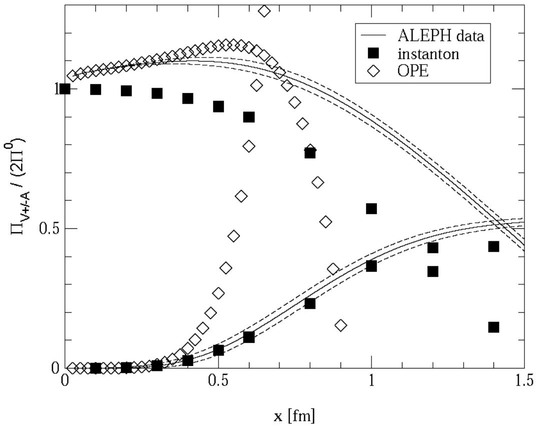

Figure 4: Euclidean coordinate space correlation functions $\Pi_V(x) \pm \Pi_A(x)$ normalized to free field behavior. The solid lines show the correlation functions reconstructed from the ALEPH spectral functions and the dotted lines are the corresponding error band. The squares show the result of a random instanton liquid model and the diamonds the OPE fit described in the text.

was referred to as "super-duality" in [4]. The instanton model reproduces this feature of the $V+A$ correlator. We also notice that for small x the deviation of the correlator in the instanton model from free field behavior is small compared to the perturbative $O(\alpha_2/\pi)$ correction. This opens the possibility of precision studies of the pQCD contribution. But before we do so, let us compare the correlation functions to the OPE prediction

$$\begin{aligned}\frac{\Pi_V(x)+\Pi_A(x)}{2\Pi_0(x)} &= 1+\frac{\alpha_s}{\pi}-\frac{1}{384}\langle g^2G^2_{\mu\nu}\rangle x^4 \\ &\quad -\frac{4\pi^3}{81}\alpha_s(x)\langle\bar{q}q\rangle\log(x^2)x^6+\ldots \qquad (33)\end{aligned}$$

Note that the perturbative correction is attractive, while the power corrections of dimension $d=4$ and $d=6$ are repulsive. Direct instantons also induce an $O(x^4)$ correction $1-\frac{\pi^2}{12}\left(\frac{N}{V}\right)x^4+\ldots$, which is consistent with the OPE because in a dilute instanton liquid we have $\langle g^2G^2\rangle = 32\pi^2(N/V)$. This term can indeed be seen in the instanton calculation and causes the correlator to drop below 1 at small x. It is possible to extract the value of Λ_{QCD} (we find $\alpha_s(m_\tau) = 0.35$) and even clear indication of running coupling. It is only possible to do because the non-perturbative corrections (represented by instantons) are basically cancelling each other to very high degree, in V+A channel.

Why is it happening? The first order in 't Hooft is indeed absent, due to chirality mismatch. There is no general theoretical reason why all non-perturbative of higher order should also do so: but ALEPH data used wrongly hint that they actually do so.

3.3 Spin-zero correlation functions

Now we will see cases which are completely opposite to those just considered: the instanton-induced effects would be large. Furthermore, the 4 channels actually show completely different non-perturbative deviation from K_0 at small x: half of them (π, σ) deviate upward, and another pair (η, δ) deviate downward.

But let me first demonstrate that the OPE scale determined above cannot be right. All we have to do is to evaluate the strength of the pion contribution to the correlator in question:

$$K_\pi(x) = \frac{\lambda_\pi^2}{4\pi^2 x^2} \tag{34}$$

The coupling constant is defined as $\lambda_\pi =< 0|J(0)|\pi >$ and the rest is nothing more than the scalar massless propagator[5]. Because both the pion term and the gluon condensate correction happen to be $1/x^2$, let us compare the coefficients. Ideal matching would mean they are about the same

$$\lambda_\pi^2 \approx \frac{< (gG)^2 >_{SVZ}}{8\pi^2} \tag{35}$$

The r.h.s. is about 0.0063 GeV^4. However, phenomenology tells us that (unlike the better known coupling to the axial current f_π) the coupling λ_π is surprisingly large[6]. The l.h.s. of this relation is actually $\lambda_\pi^2 = (.48\, GeV)^4 =$

[5] We can ignore the pion mass at the distances in question. We also ignore contributions of other states, which can only add positively to the correlator and made disagreement only worse.

[6] The reason for that is the the pion is rather compact and also the wave function is concentrated at its center, so that its value at $r = 0$ is large. We return to this point in the discussion of the "instanton liquid" model.

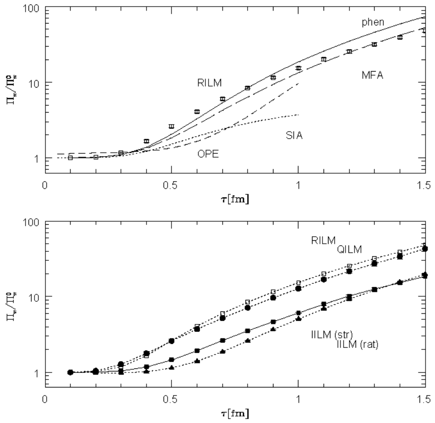

Figure 5: Pion correlation function in various approximations and instanton ensembles. In the top figure we show the phenomenological expectation (solid), the OPE (dashed), the single instanton (dash-dotted) and mean field approximations (dashed) as well as data in the random instanton ensemble. In the bottom figure we compare different instanton ensembles, random (open squares), quenched (circles) and interacting (streamline: solid squares, ratio ansatz solid triangles).

$0.053\,GeV^4$, about *10 times larger* than the r.h.s. It means much larger non-perturbative effect is needed to explain the deviation from the perturbative behavior.

Now, let us see why is it so. The instanton effects in spin-0 channels are in these cases much larger because effect of 't Hooft interaction appears in those cases in the first order. Furthermore, since it its flavor structure is non-diagonal $(\bar{u}u)(\bar{d}d)$ the correlator of two π^0 currents $(\bar{u}\gamma_5 u - \bar{d}\gamma_5 d)$ have it with opposite sign as compared to the correlator of η' currents $(\bar{u}\gamma_5 u + \bar{d}\gamma_5 d)$. What it means is that instantons are as attractive in the pion channel as they are repulsive in the η' case. The situation is reversed in the scalar channels: the isoscalar sigma is attractive and isovector is repulsive.

Full results from versions of the instanton liquid model for pion correlators are shown in fig.5. Different versions of the model (mentioned in figures below as IILM(rat) etc) differ by a particular ansatz for the gauge field used, from which the interaction is calculated. Note also, that these figures contain also a curve marked "phen": this is what the correlator actually looks like, according to phenomenology.

We simply show a few results of correlation functions in the different instanton ensembles (see original refs in[2]). Some of them (like vector and axial-vector ones) turned out to be easy: nearly any variant of the instanton

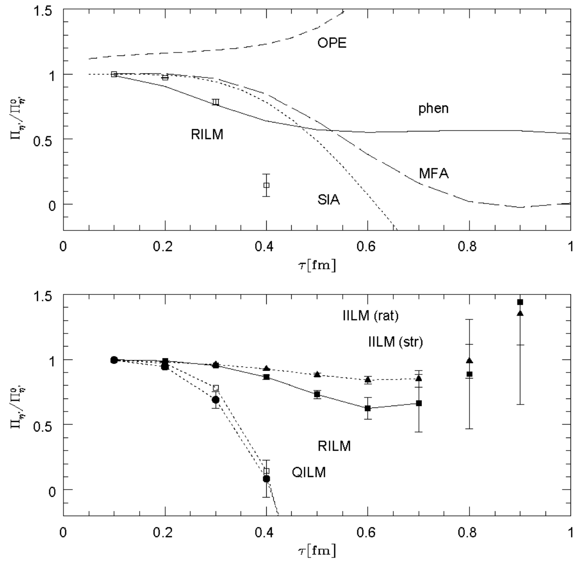

Figure 6: Eta prime meson correlation functions. The various curves and data sets are labeled as in Fig. 5. Note that random instanton liquid model (RILM) and quenched version (no fermionic determinant, only bosonic interactions) predict η' correlator to go negative. The same unphysical behavior has been found on the lattice.

model can reproduce the (experimentally known!) correlators well. Some of them are sensitive to details of the model very much: two such cases are shown in Figs. 5-6. The pion correlation functions in the different ensembles are qualitatively very similar. The differences are mostly due to different values of the quark condensate (and the physical quark mass) in the different ensembles. Using the Gell-Mann-Oaks-Renner relation, one can extrapolate the pion mass to the physical value of the quark masses. The results are consistent with the experimental value in the streamline ensemble (both quenched and unquenched), but clearly too small in the ratio ansatz ensemble. This is a reflection of the fact that the ratio ansatz ensemble is not sufficiently dilute.

The situation is drastically different in the η' channel. Among the ~ 40 correlation functions calculated in the random ensemble, only the η' and the isovector-scalar δ were found to be completely unacceptable. The correlation function decreases very rapidly and becomes *negative* at $x \sim 0.4$ fm. This behavior is incompatible even with a normal spectral representation. The interaction in the random ensemble is too repulsive, and the model "over-explains" the $U(1)_A$ anomaly.

The results in the unquenched ensembles (closed and open points) significantly improve the situation. This is related to dynamical correlations between instantons and anti-instantons (topological charge screening). The single instanton contribution is repulsive, but the contribution from pairs is attractive. Only if correlations among instantons and anti-instantons are sufficiently strong are the correlators prevented from becoming negative. Quantitatively, the δ and $\eta_{\rm ns}$ masses in the streamline ensemble are still too heavy as compared to their experimental values. In the ratio ansatz, on the other hand, the correlation functions even show an enhancement at distances on the order of 1 fm, and the fitted masses are too light. This shows that the η' channel is very sensitive to the strength of correlations among instantons.

In summary, pion properties are mostly sensitive to global properties of the instanton ensemble, in particular its diluteness. Good phenomenology demands $\bar{\rho}^4 n \simeq 0.03$, as originally suggested in[6]. The properties of the ρ meson are essentially independent of the diluteness, but show sensitivity to $\bar{I}I$ correlations. These correlations become crucial in the η' channel.

3.4 Baryonic correlation functions

The existence of a strongly attractive interaction in the pseudoscalar quark-antiquark (pion) channel also implies an attractive interaction in the scalar quark-quark (diquark) channel. This interaction is phenomenologically very desirable, because it immediately explains why the nucleon is light, while the delta (S=3/2,I=3/2) is heavy.

The so called Ioffe currents (with no derivatives and the minimum number of quark fields) are local operators which can excite states with nucleon quantum numbers. Those with positive parity and spin 1/2 can also be represented in terms of scalar and pseudoscalar diquarks

$$\eta_{1,2} = (2,4)\left\{\epsilon_{abc}(u^a C d^b)\gamma_5 u^c \mp \epsilon_{abc}(u^a C\gamma_5 d^b)u^c\right\}. \tag{36}$$

Nucleon correlation functions are defined by $\Pi^N_{\alpha\beta}(x) = \langle\eta_\alpha(0)\bar{\eta}_\beta(x)\rangle$, where α,β are the Dirac indices of the nucleon currents. In total, there are six different nucleon correlators: the diagonal $\eta_1\bar{\eta}_1$, $\eta_2\bar{\eta}_2$ and off-diagonal $\eta_1\bar{\eta}_2$ correlators, each contracted with either the identity or $\gamma\cdot x$. Let us focus on the first two of these correlation functions (for more detail, see[2] and references therein).

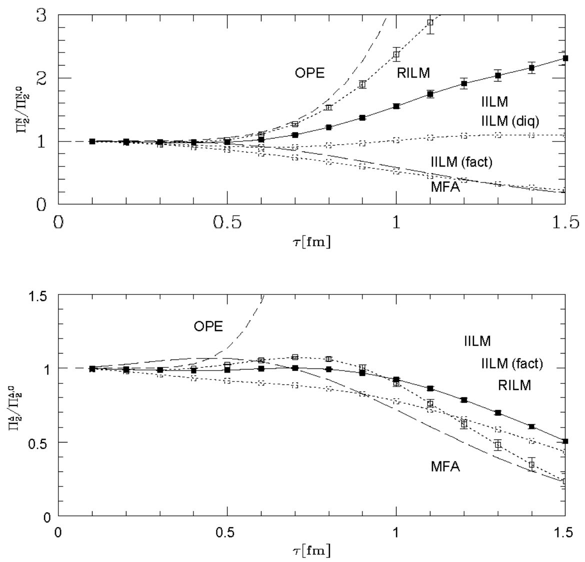

Figure 7: Nucleon and delta correlation functions Π_2^N and Π_2^Δ. Curves labeled as in Figs.on mesonic correlators.

The correlation function Π_2^N in the interacting ensemble is shown in Fig. 7. The fact that the nucleon in IILM is actually bound can also be demonstrated by comparing the full nucleon correlation function with that of three non-interacting quarks (the cube of the average propagator). The full correlator is significantly larger than the non-interacting one.

There is a significant enhancement over the perturbative contribution which is nicely described in terms of the nucleon contribution. Numerically, we find[7] $m_N = 1.019$ GeV. In the random ensemble, we have measured the nucleon mass at smaller quark masses and found $m_N = 0.96 \pm 0.03$ GeV. The nucleon mass is fairly insensitive to the instanton ensemble. However, the strength of the correlation function depends on the instanton ensemble. This is reflected by the value of the nucleon coupling constant, which is smaller in the IILM. In[19] we studied all six nucleon correlation functions. We showed that all correlation functions can be described with the same nucleon mass and coupling constants.

The fitted value of the threshold is $E_0 \simeq 1.8$ GeV, indicating that there is little strength in the "three quark continuum" (dual to higher resonances in

[7] Note that this value corresponds to a relatively large current quark mass $m = 30$ MeV.

the nucleon channel). A significant part of this interaction was traced down to the strongly attractive *scalar diquark* channel. The nucleon (at least in IILM) is a strongly bound diquark, plus a loosely bound third quark. The properties of this diquark picture of the nucleon continue to be disputed by phenomenologists. We will return to diquarks in the next section, where they will become Cooper pairs of Color Super-conductors.

In the case of the Δ resonance, there exists only one independent Ioffe current, given (for the Δ^{++}) by

$$\eta_\mu^\Delta = \epsilon_{abc}(u^a C\gamma_\mu u^b)u^c. \tag{37}$$

However, the spin structure of the correlator $\Pi^\Delta_{\mu\nu;\alpha\beta}(x) = \langle \eta^\Delta_{\mu\alpha}(0)\bar{\eta}^\Delta_{\nu\beta}(x)\rangle$ is much richer. In general, there are ten independent tensor structures, but the Rarita-Schwinger constraint $\gamma^\mu \eta^\Delta_\mu = 0$ reduces this number to four.

The mass of the delta resonance is too large in the random model, but closer to experiment in the unquenched ensemble. Note that,similar to the nucleon, part of this discrepancy is due to the value of the current mass. Nevertheless, the delta-nucleon mass splitting in the unquenched ensemble is $m_\Delta - m_N = 409$ MeV, larger but comparable to the experimental value 297 MeV. It mostly comes from the *absent scalar diquarks* in Δ channel.

4 Lecture 3. The Phases of QCD

4.1 The Phase Diagram

In this section we discuss QCD in extreme conditions, such as finite temperature/density. Let me first emphasize why it is interesting and instructive to do. It is not simply to practice once again the semi-classical or perturbative methods similar to what have been done before in vacuum. What we are looking for here are *new phases* of QCD (and related theories), namely new self-consistent solutions which differ qualitatively from what we have in the QCD vacuum.

One such phase occurs at high enough temperature $T > T_c$: it is known as Quark Gluon Plasma (QGP). It is a phase understandable in terms of basic quark and gluon-like excitations[38], without confinement and with unbroken chiral symmetry in the massless limit[8]. One of the main goals

[8] It does not mean though, that it is a simple issue to understand even the high-T limit of QCD, related to non-perturbative 3d dynamics.

of heavy ion program, especially at new the dedicated Brookhaven facility RHIC, is to study transitions to this phase.

Another one, which has been getting much attention recently, is the direction of finite density. Very robust Color Superconductivity was found to be the case here. Let me also mention one more frontier which has not yet attracted sufficient attention: namely a transition (or many transitions?) as the number of light flavors N_f grows. The minimal scenario includes a transition from the usual hadronic phase to a more unusual QCD phase, the $conformal$ one, in which there are no particle-like excitations and correlators are power-like in the infrared. Even the position of the critical point is unknown. The main driving force of these studies is the intellectual challenge it provides.

The QCD phase diagram as we understand it now is shown in Fig 8(a), in the baryonic chemical potential μ (normalized per quark, not per baryon) and the temperature T plane. Some part of it is old: it has the hadronic phase at small values of both parameters, and QGP phase at large T,μ.

The phase transition line separating them most probably does not really start at $T = T_c, \mu = 0$ but at an “endpoint” E, a remnant of the so called QCD tricritical point which QCD has in the chiral (all quarks are massless) limit. Although we do not know where it is[9], we hope to find it one day in experiment. The proposed ideas rotate around the fact that the order parameter, the VEV of the sigma meson, is at this point truly massless, and creates a kind of “critical opalecence”. Similar phenomena were predicted and then indeed observed at the endpoint of another line (called M from multi-fragmentation), separating liquid nuclear matter from the nuclear gas phase.

The large-density (and low-T) region looks rather different from what was shown at conferences just a year ago: two new Color Super-conducting phases appear there. Unfortunately heavy ion collisions do not cross this part of the phase diagrams and so it belongs to neutron star physics.

Above I mentioned an approach to high density starting from the vacuum. One can also work out in the opposite direction, starting from very large densities and going down. Since the electric part of one-gluon exchange is screened, and therefore the Cooper pairs appear due to magnetic forces. It is interesting by itself, as a rare example: one has to take care of *time delay effects* of the interaction. The result is indefinitely growing gaps at

[9] Its position is very sensitive to the precise value of the strange quark mass m_s

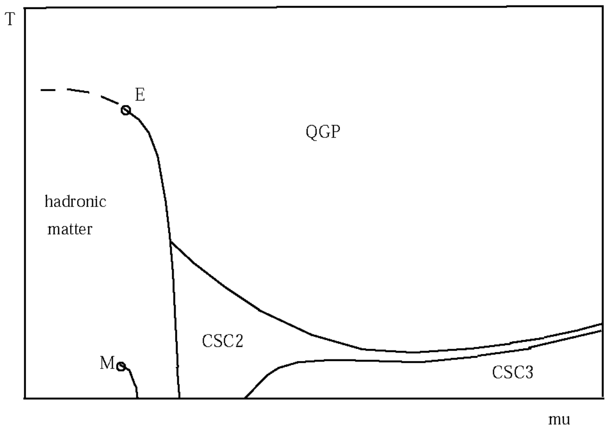

Figure 8: Schematic phase diagram of QCD, in temperature T- baryon chemical potential μ plane. E and M show critical endpoints of first order transitions: M (from multi-fragmentation) is that for liquid-gas transition in nuclear matter. The color superconducting phases, CSC2 and CSC3 are explained in the text.

large $\mu > 10GeV$, as [34] $\Delta \sim \mu exp(-\frac{3\pi^2}{\sqrt{2}g(\mu)})$.

4.2 Finite Temperature transition and Large Number of Flavors

There is no place here to discuss in detail the rather extensive lattice data available now, and I only mention some results related to instantons. In the vacuum a quasi-random set of instantons leads to chiral symmetry breaking and quasi-zero modes: but what in the same terms does the high-T phase look like?

The simplest solution would be just *suppression* of instantons at $T > T_c$, and at some early time people thought this is what actually happens. However, it should not be like this because the Debye screening which is killing them only appears at $T = T_c$. Lattice data works have also found no depletion of the instanton density up to $T = T_c$.

On the other hand, the absence of the condensate and quasi-zero modes implies that the "liquid" is now broken into finite pieces. The simplest of them are pairs, or the instanton-anti-instanton molecules. This is precisely what instanton simulations have found[2], see fig.9. Whether it is indeed so on the lattice is not yet clear: nice molecules were located, but the evidence for the molecular mechanism of chiral restoration is still far from being convincing. (No alternative I am aware of have been so far proposed, though.)

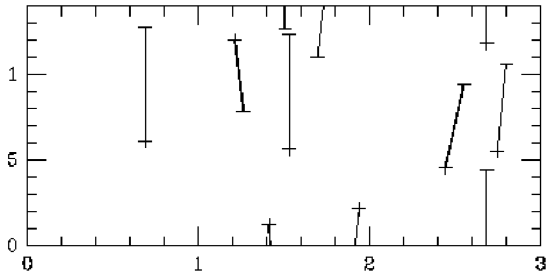

Figure 9: Typical configuration from instanton liquid simulation, at $T > T_c$. Lines indicate the direction in which quark propagators are the largest. Clear pairing of instantons and instantons are observed: the pairs tend to have the same spatial position and being separated mostly by Euclidean time.

The results of IILM simulations with variable number of flavors $N_f = 2, 3, 5$[10] flavors with equal masses can be summarized as follows. For $N_f = 2$ there is a second order phase transition which turns into a line of first order transitions in the $m - T$ plane for $N_f > 2$. If the system is in the chirally restored phase ($T > T_c$) at $m = 0$, we find a discontinuity in the chiral order parameter if the mass is increased beyond some critical value. Qualitatively, the reason for this behavior is clear. While increasing the temperature increases the role of correlations caused by fermion determinant, increasing the quark mass has the opposite effect. We also observe that increasing the number of flavors lowers the transition temperature. Again, increasing the number of flavors means that the determinant is raised to a higher power, so fermion induced correlations become stronger. For $N_f = 5$ we find that the transition temperature drops to zero and the instanton liquid has a chirally symmetric ground state, provided the dynamical quark mass is less than some critical value. Studying the instanton ensemble in more detail shows that in this case, all instantons are bound into molecules.

Unfortunately, little is known about QCD with large numbers of flavors from lattice simulations. There are data by the Columbia group for $N_f = 4$. The most important result is that chiral symmetry breaking effects were found to be drastically smaller as compared to $N_f = 0, 2$. In particular, the mass splittings between chiral partners such as $\pi - \sigma$, $\rho - a_1$, $N(\frac{1}{2}^+) - N(\frac{1}{2}^-)$,

[10] The case $N_f = 4$ is omitted because in this case it is very hard to determine whether the phase transition happens at $T > 0$.

extrapolated to $m = 0$ were found to be 4-5 times smaller. This agrees well with what was found in the interacting instanton model: more work in this direction is certainly needed.

4.3 High Density and Color Superconductivity

Although the idea of color superconductivity originates from 70's, the field of high density QCD was in the dormant state for long time till two papers [20, 21] (posted on the same day) in 1998 have claimed gaps about 100 times larger than previously thought. The field is booming since, as one can see from about 250 citations in 2 years those papers got.

Then-Princeton group (Alford-Rajagopal-Wilczek) have been thinking about different pairings from theory perspective, but our (Stony Brook) team (Rapp,Schafer,ES,Velkovsky) had started from the impressive qq pairing phenomenon found theoretically [19] in the instanton liquid model *inside the nucleon.* As explained above, we have found it to be, roughly speaking, a small drop of CS matter, made of one Cooper pair of sort (the *ud scalar diquark*) and one massive quark[11] . T.Schafer heroically attempted numerical simulations of the instanton liquid model at finite μ: although he was not very successful[12] he found out strange "polymers" made of instantons connecting by 2 through going quark lines. It take us some time to realize we see paths of condensed diquarks! It was like finding superconductivity by watching electrons moving on your computer screen.

The main point I would like to emphasize here is that the qq pairing of such diquarks have in fact deep dynamical roots: it follows from the same basic dynamics as the "superconductivity" of the QCD vacuum, the chiral (χ-)symmetry breaking. These spin-isospin-zero diquarks are related to pions, as we will see below.

The most straightforward argument for deeply bound diquarks came from the bi-color ($N_c = 2$) theory: in it the scalar diquark is degenerate with pions. By continuity from $N_c = 2$ to 3, a trace of it should exist in real QCD[13].

Instantons create the following amusing *triality*: there are three attractive channels which compete: (i) the instanton-induced attraction in $\bar{q}q$ chan-

[11] As opposed to Δ (decuplet) baryons, which is a small drop of "normal" quark matter, without scalar diquarks.

[12] for the same reason as lattice people cannot do it: the fermionic determinant is not real.

[13] Instanton-induced interaction strength in diquark channel is $1/(N_c - 1)$ of that for $\bar{q}\gamma_5 q$ one. It is the same at $N_c = 2$, zero for large N_c, and is exactly in between for $N_c = 3$.

nel leading to χ-symmetry breaking. (ii) The instanton-induced attraction in qq which leads to color superconductivity. (iii) The *light-quark-induced* attraction of $\bar{I}I$, which leads to pairing of instantons into "molecules" and a Quark-Gluon Plasma (QGP) phase without *any* condensates.

At very high density we also can find *arbitrarily dilute instanton liquid*, as shown recently in [35]. The reason it cannot exist in vacuum or high T is that if instanton density goes below some critical value, the cannot be any condensate. (The system then breaks into instanton molecules or other clusters and chiral symmetry is restored.) However at high density the superconducting condensate can be created perturbatively as well (we mentioned it above) and there is no problem. The dilute instantons interact by exchanging very light η' (which would be massless without instantons): one can calculate effective Lagrangian, theta angle dependence etc.

Bi-color QCD: a very special theory One reason it is special (well known to to the lattice community): its fermionic determinant is *real* even for non-zero μ, which makes simulations possible. However the major interest in this theory is related the so called *Pauli-Gursey symmetry.* We have argued above that pions and diquarks appear at the same one-instanton level, and are so to say brothers. In bi-color QCD they becomes identical twins: due to the additional symmetry mentioned the diquarks are *degenerate* with mesons.

In particular, chiral symmetry breaking is done like this $SU(2N_f) \rightarrow Sp(2N_f)$, and for $N_f = 2$ the coset $K = SU(4)/Sp(4) = SO(6)/SO(5) = S^5$. Those 5 massless modes are pions plus the scalar diquark S and its antiparticle $\bar{S}$.

Vector diquarks are degenerate with vector mesons, etc. Therefore, the scalar-vector splitting is in this case about twice the constituent quark mass, or about 800 MeV. It should be compared to binding in the "real" $N_c = 3$ QCD of only 200-300 MeV, and to zero binding in the large-N_c limit.

The corresponding sigma model describing this χ-symmetry breaking was worked out in[20]: for further development see[22]. As argued in [20], in this theory the critical value of the transition to Color Superconductivity is simply $\mu = m_\pi/2$, or zero in the chiral limit. The diquark condensate is just a rotated $< \bar{q}q >$ one, and the gap is the constituent quark mass. Recent lattice works [26] display it in great detail, building confidence for other cases.

New studies reveal possible new crystalline phases. These phases still have somewhat debatable status, so I have not indicated them on the phase diagram.

Once again, there were two papers submitted by chance on the same day. The "Stony Brook" team[23] have found that a "chiral crystal" with oscillating $< \bar{q}q(x) >$ (similar to Overhouser spin waves in solid state) can compete with the BCS 2-flavor superconductor at its onset, or $\mu \approx 400\,MeV$. The proper position of this phase is somewhere in between the hadronic phase (with constant $< \bar{q}q >$) and color superconductor.

The "MIT group"[24] have looked at the oscillating superconducting condensate $< qq(x) >$, following earlier works on the so called LOFF phase in usual superconductors. They have found that it is appearing when the difference between Fermi momenta of different quark flavors become comparable to the gap. The natural place for it on the phase diagram is close to the line at which color superconductivity disappears because the gap goes to zero.

5 Lecture 4.High Energy Collisions of Heavy Ions

5.1 The Little Bang: AGS, SPS and now the RHIC era

Let me start with brief comparison of these two magnificent explosions: the Big Bang versus the Little Bang, as we call heavy ion collisions.

The expansion law is roughly the Hubble law in both, $v(r) \sim r$ although strongly anisotropic in the Little Bang. The Hubble constant tells us the expansion rate today: similarly radial flow tells us the final magnitude of the transverse velocity. The acceleration history is not really well measured. For Big Bang people use distance supernovae, we use Ω^- which does not participate at the late stages to learn *what was the velocity earlier*. Both show small dipole (quadrupole or elliptic for Little Bang) components which has some physics, and who knows maybe we will see higher harmonics fluctuations later on, like in Universe. As we will discuss below, in both cases the major puzzle is how this large entropy has been actually produced, and why it happened so early.

The major lessons we learned from AGS experiments ($E_{LAB} = 2 - 12AGeV$) are:
(i) Strangeness enhancement over simple multiple NN collisions appear from very low energies, and heavy ion collisions quickly approach nearly ideal chemical equilibrium of strangeness.
(ii) "Flows" of different species, in their radial,directed and elliptical form, are in this energy domain driven by collective potentials and absorptions:

they are not really flows in hydro sense. All of them strongly diminish by the high end of the AGS region, demonstrating the onset of "softness" of the EoS. Probably it is some precursor of the QCD phase transition.

Several important lessons came so far from CERN SPS data:
(i) Much more particle ratios have been measured there: overall those show surprisingly good degree of chemical equilibration: the chemical freeze-out parameters are tantalizingly close to the QGP phase boundary.
(ii) Dileptons show that radiation spectral density is very different in dense matter compared to ideal hadronic gas. The most intriguing data are CERES finding of "melting of the ρ", which seem to be transformed into a wide continuum reaching down to invariant masses as low as 400 MeV. It puts in doubt "resonance gas" view of hadronic matter at these conditions. Intermediate mass dileptons studied by NA50 can be well described by thermal radiation with QGP rates.
(iii) The impact parameter of J/ψ and ψ' suppression in PbPb collisions studied by NA50 collaboration shows rather non-trivial behavior. More studies are needed, including especially measurements of the open charm yields, to understand the origin and magnitude of the suppression.

However, during last several months those discussions have been overshadowed by a list of news from RHIC, Relativistic Heavy Ion Collider at Brookhaven National Laboratory. It had its first run in summer 2000 and reported recently at Quark Matter 2001 conference [27]: many details are discussed in Prof.M.Gylassy's lectures.

A brief summary is as follows. These results have shown that heavy ions collisions (AA) at these energies significantly differ *both* from the pp collisions at high energies and the AA collisions at lower (SPS/AGS) energies. The main features of these data are quite consistent with the Quark-Gluon Plasma (QGP) (or Little Bang) scenario, in which entropy is produced promptly and subsequent expansion is close to adiabatic expansion of equilibrated hot medium.

(Let me mention here two other pictures of the heavy ion production, discuss prior to appearance of these data. One is the *string picture*, used in event generators like RQMD and UrQMD: they predicted effectively very soft EoS and elliptic flow decreasing with energy. The other one is *pure minijet scenario*, in which most secondaries would come from independently fragmenting minijets. If so, there are basically no collective phenomena whatsoever.)

Already the very first multiplicity measurements reported by PHOBOS collaboration [47] have shown that particle production per participant nucleon is no longer constant, as was the case at lower (SPS/AGS) energies. This new component may be due to long-anticipated $pQCD$ processes, leading to perturbative production of new partons. Unlike high p_t processes resulting in visible jets, those must be undetectable *"mini-jets"* with momenta $\sim 1-2\,GeV$. Production and decay of such *mini-jets* was discussed in Refs [48], also this scenario is the basis of widely used event generator HIJING [46]. Its crucial parameter is the *cutoff scale* p_{min}: if fitted from pp data to be 1.5-2 GeV, it leads to predicted mini-jet multiplicity $dN_g/dy \sim 200$ for central AuAu collisions at $\sqrt(s) = 130\,AGeV$. If those fragment independently into hadrons, and are supplemented by "soft" string-decay component, the predicted total multiplicity was found to be in good agreement with the first RHIC multiplicity data. Because partons interact perturbatively, with their scattering and radiation being strongly peaked at small angles, their equilibration is expected to be relatively long [49]. However, new set of RHIC data reported in [27] have provided serious arguments *against* the mini-jet scenario, and point toward quite rapid entropy production rate and early QGP formation.

(i) If most of mini-jets fragment independently, there is no *collective phenomena* such as transverse flow related with the QGP pressure. However, it was found that those effects are very strong at RHIC. Furthermore, STAR collaboration have observed very robust *elliptic flow* [37], which is in perfect agreement with predictions of hydrodynamical model [43, 42] assuming equilibrated QGP with its full pressure $p \approx \epsilon/3$ above the QCD phase transition. This agreement persists to rather peripheral collisions, in which the overlap almond-shaped region of two nuclei is only a couple fm thick. STAR and PHENIX data on spectra of identified particles, especially $p, \bar{p}$, indicate spectacular radial expansion, also in agreement with hydro calculations [43, 42]. (ii) Spectra of hadrons at large p_t, especially the π^0 spectra agree well with HIJING for peripheral collisions, but show much smaller yields for central ones, with rather different, (exponential-shaped) spectra. It means long-anticipated *"jet quenching"* at large p_t is seen for the first time, with a surprisingly large suppression factor $\sim 1/5$. Keeping in mind that jets originating from the surface outward cannot be quenched, the effect seem to be as large as it can possibly be. For that to happen, the outgoing high-p_t jets should propagate through matter with parton population larger than the abovementioned minijet density predicted by HIJING.

(iii) Curious interplay between collective and jet effects have also been studied by STAR collaboration, in form of elliptic asymmetry parameter $v_2(p_t)$. At large transverse momenta $p_t > 2\,GeV$ the data depart from hydro predictions and levels off. When compared to predictions of jet quenching models worked out in [50], they also indicate gluon multiplicity several times larger than HIJING prediction, and are even consistent with its maximal possible value evaluated from the final entropy at freeze-out, $(dN/dy)_\pi \sim 1000$.

5.2 Collective flows and EoS

If we indeed have produced excited matter (rather than just a bunch of partons which fly away and fragment independently), we expect to see certain collective phenomena. Ideally, those should be quantitatively reproduced by relativistic hydrodynamics which is basically just local energy-momentum conservation plus the EoS we know from the lattice and models.

The role of the QCD phase transition in matter expansion is significant. QCD lattice simulations [40] show approximately 1st order transition. Over a wide range of energy densities $e = .5 - 1.4\,GeV/fm^3$ the temperature T and pressure p are nearly constant. So the ratio of pressure to energy density, p/e, decreases till a minimum at particular energy density $e_{sp} \approx 1.4\,GeV/fm^3$, known as the *softest point* [41]. Near e_{sp} small pressure gradient can not effectively accelerate the matter and the evolution stagnates. However when the initial energy density is well above the QCD phase transition region, $p/e \approx 1/3$, and this pressure drives the collective motion. The energy densities reached at time $\sim 1fm/c$ at SPS($\sqrt{s}_{NN} = 17\,GeV$) and RHIC ($\sqrt{s}_{NN} = 130\,GeV$) are about 4 and 8 GeV/fm^3, respectively. We found that at RHIC conditions we are in the latter regime, and matter accelerates to $v \sim .2c$ $before$ entering the soft domain. Therefore by freeze-out this motion changes the spatial distribution of matter dramatically: e.g. as shown in [36] the initial almond-shape distribution 10 fm/c later looks like two separated shells, with a little "nut" in between.

The simplest way to see hydro expansion is in spectra of particles: on top of chaotic thermal distributions $\sim exp(-m_t/T), m_t^2 = p_t^2 + m^2$ one expect to see additional broadening due to hydro outward motion. This effect is especially large if particles are heavy, since flow with velocity v add momentum mv.

Derek Teaney [43] have developed a comprehensive Hydro-to-Hadrons (H2H) model combines the hydrodynamical description of the initial QGP/

mixed phase ($e > .5GeV/fm^3$) stages, where hadrons are not appropriate degrees of freedom, with a hadronic cascade RQMD for the hadronic stage. In this way, we can include different EoS displaying properties of the phase transition, and also incorporate complicated final state interaction at freeze-out. The set of EoS used is shown in Fig.10.

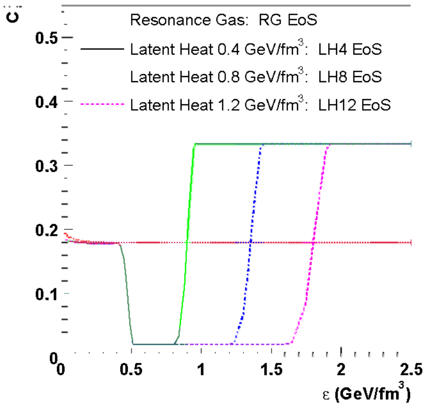

Figure 10: The EoSs in form squared speed of sound $c^2 = dp/d\epsilon$ with variable Latent Heats $.4GeV/fm^3$, $.8GeV/fm^3$... labeled as LH4, LH8,..versus the energy density.

Radial flow is usually characterized by the slope parameter T: each particle spectra are fitted to the form $dN/dp_t^2dy \sim exp(-m_t/T), m_t^2 = p_t^2 + m^2$.Although we denoted the slope by T, it is *not* the temperature: it incorporates random thermal motion and collective transverse velocity. The SPS NA49 slope parameters for pion and protons are shown in Fig. 11(a). Parameter T grows with particle multiplicity due to increased velocity of the radial flow. Furthermore, the rate of growth depends on the EoS: the softer it is, the less growth. The SPS NA49 data correspond to two data points (our fits to spectra) favor the (relatively stiff) LH8 EoS. (Details of the fit, discussion of the b-dependence etc see in [43].) It is very important to get these parameters for RHIC, especially for heavy secondaries like nucleons and hyperons.

For $non-central$ collisions the overlap region in the transverse plane has an elliptic, "almond", shape, and larger pressure gradient force matter to expands preferentially in the direction of the impact parameter [39]. Compared to radial flow, the elliptic flow is formed earlier, and therefore it measures the early pressure. The *elliptic flow* is quantified experimentally by measuring

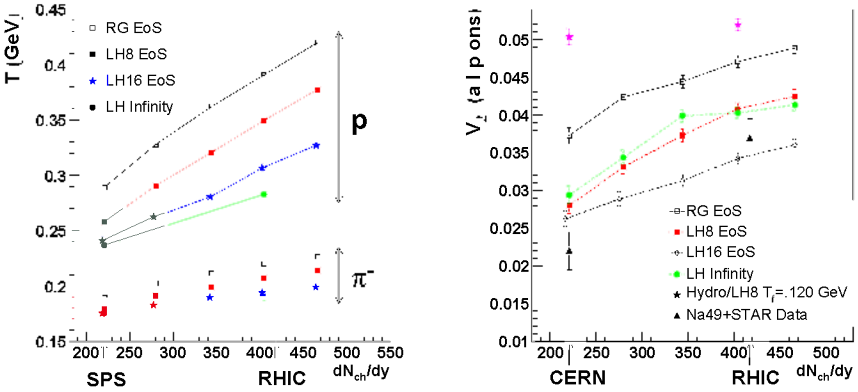

Figure 11: The transverse mass slope T (a) and elliptic flow parameter V_2 (b) versus midrapidity (y=0) charged purticle multiplicity, for AuAu collisions with b—6.

the azimuthal distributions of the produced particles and calculating the elliptic flow parameter $V_2 = \langle cos(2\phi)\rangle$ where ϕ angle is measured with respect to the impact parameter direction, around the beam axis. It appears due to the elliptic *spatial* deformation of the overlap region in the nucleus-nucleus collision, quantified by its eccentricity $\epsilon_2 =< y^2 - x^2 > / < x^2 + y^2 >$, usually calculated in Glauber model. Since the effect (v_2) is proportional to the cause (ϵ_2), the ratio v_2/ϵ_2 does not have strong dependence on the impact parameters b, and this ratio is often used for comparison. (We would not do that below, in the detailed comparison to data, because $\epsilon_2(b)$ is not directly measured.

In figure 11(b) the elliptic flow of the system is plotted as a function of charged particle multiplicity at an impact parameter of 6 fm. Before discussing the energy dependence, let us quantify the magnitude of elliptic flow at the SPS. Ideal relativistic hydrodynamics used in earlier works [39, 42] generally over-predicts elliptic flow by about factor 2. Example of such kind is indicated by a star in figure 11(b): it is our hydro result (with LH8 EoS) which has been followed hydrodinamically till very late stages, the freeze-out temperature $T_f = 120\, MeV$. By switching to hadronic cascade at late stages, we have more appropriate treatment of resonance decays and re-scattering rate, and so one can see that it significantly reduces V_2, to the range much closer to the data points.

One might thing that one can also do that by simply taking *softer* EoS, e.g. increasing the latent heat. However, it only happens till LH16 and then v_2 start even slightly increase again. The explanation of this non-monotonous behavior is the interplay of the initial "QGP push" for stiffer EoS, with longer time for hadronic stage available for softer EoS. We cannot show here details, but it turns out that a given (experimental) V_2 value can correspond to *two different solutions*, one with earlier push and another with the later expansion dominating. Coincidentally, STAR data point happen to be right at the onset of such a bifurcation, close to LH16. The *multiplicity dependence* of V_2 appears simple from figure 11(b): all curve show growth with about the same rate. Note however that such growth of V_2 from SPS to RHIC (first predicted in [44] where our first preliminary results has been shown) Is not shared by most other models. In particular, *string-based* models like UrQMD predicts a decrease by a factor of ≈ 2 [45]. It happens because strings produce no transverse pressure and so the effective EoS is super-soft at high energies. Models based on *independent parton scattering and decay* (such as HIJING) also predict basically vanishing (or slightly negative)[46] V_2.

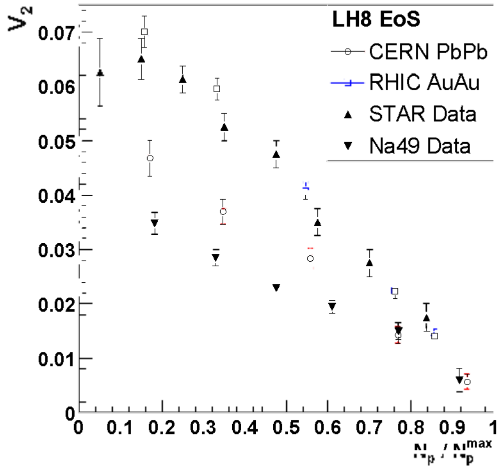

Figure 12: v2 versus impact parameter b, described experimentally by the number of participant nucleons, for RHIC STAR and SPS NA49 experiments. Both are compared to our results. for EoS LH8.

In Fig.12 we show how our results compare with data as a function of impact parameter. One can see that the agreement becomes much better at RHIC. Furthermore, one may notice that deviation from linear dependence we predict becomes visible at SPS for more peripheral collisions with $N_p/N_p^{max} < 0.6$ or so, while at RHIC only the most peripheral point, with $N_p/N_p^{max} = 0.05$ show such deviation. This clearly shows that hydrodynamical regime in general works much better at RHIC.

In summary, the flow phenomena observed at RHIC are stronger than at SPS. It is in complete agreement with the QGP scenario. All data on elliptic and radial flow can be nicely reproduced by the H2H model. Furthermore, we are able to restrict the EoS, to those with the latent heat about .8 GeV/fm^3.

5.3 How QGP happened to be produced/equilibrated so early?

One possible solution to the puzzle outlined above can be a *significantly lower cutoff scale* in AA collisions, as compared to $p_{min} = 1.5 - 2\,GeV$ fitted from the pp data. That increases perturbative cross sections, both due to smaller momenta transfer and larger coupling constant. As I argued over the years, the QGP is a new phase of QCD which is *qualitatively different* from the QCD vacuum: therefore the cut-offs of pQCD may have entirely different values and be determined by different phenomena. Furthermore, since QGP is a plasma-like phase which screens itself perturbatively [38], one may think of a cut-offs to be determined *self-consistently* from resummation of perturbative effects. These ideas known as *self-screening* or *initial state saturation* were discussed in Refs. [49]. Although the scale in question grows with temperature or density, *just above* T_c it may actually be *smaller* than the value 1.5-2 GeV we observe in the vacuum. Its first experimental manifestation may be dropping of the so called "duality scale" in the observed dilepton spectrum, see discussion in [52].

Another alternative to explain large gluon population at RHIC would be an existence of more rapid multi-gluon production processes. Let us consider an alternative *non−perturbative* scenario based entirely on non-perturbative processes involving *instantons* and *sphalerons* [51]. But before we do that, we have to take a look at hadronic collisions and briefly review few recent papers on the subject.

6 Lecture 5. Instanton-induced effects in high energy collisions

6.1 Why all hadronic cross sections grow with energy?

At $s > 10^3 GeV^2$ hadronic cross sections as $\bar{p}p, pp, \pi p, Kp$, γN and even $\gamma\gamma$ slowly grow with the collision energy s, approximately as $\sigma \sim s^{\Delta}$. This behavior can be parameterized by Regge phenomenology, with the leading role plaied by the so called *soft Pomeron.* We cannot describe here its long history, starting from Pomeranchuck and Gribov in 1960's. Phenomenologically it is still in very good shape. where a supercritical pole with the intercept $\Delta \sim 0.08$. Below TeV energies such growth can be well described by a simple logarithmically growing term

$$\sigma_{hh'}(s) = \sigma_{hh'}(s_0) + log(s/s_0) X_{hh'} \Delta + ... \tag{38}$$

and we will concentrate on its origin, ignoring both the higher powers of log(s) and other, decreasing, Regge terms. We will use those two parameters from PDG-2000 recent fits, the intercept and its coefficient in $pp, \bar{p}p$ collisions, $\Delta = \alpha(0) - 1 = 0.093(2)$, $X_{NN} = 18.951(27)\ mb$.

The physical origin of constant and logarithmically growing parts of the are different. The former can be explained by prompt color *exchanges*, as suggested by Low and Nussinov long ago. It nicely correlates with flux tube picture of the final state.

The *growing* part of the cross section cannot be generated by t-channel color exchanges and is associated with processes promptly producing some objects, with log(s) coming from the longitudinal phase space. In pQCD it is *gluon* production, by processes like the one shown in Fig.13(a). If iterated in the t-channel in ladder-type fashion, the result is approximately a BFKL pole [53]. Although the power predicted is much larger than Δ mentioned, it seem to be consistent with much stronger growth seen in hard processes at HERA: thus it is therefore sometimes called the "hard pomeron".

At this point I has been frequently asked: why is it so difficult to understand the growth of hadronic cross section, if HERA data shows spectacular increase of the gluonic number at small x? Shouldn't all these gluons collide with each other and naturally generate such growth?

The issue is not that simple, and the first thing to do at this point is to remind the reader about the scales involved. At large Q^2 we resolve the partons and see these magnificent rise toward the small x indeed: but high energy hadron collisions do not proceed at such scale. In fact the scale is "semi-hard" $Q^2 \sim 1\,GeV$, as the Pomeron slope indicate. If we now go back to analysis of HERA/SLAC data and try to extract gluon density at this scale, we will not find a significant growth. What it indicates, is that all these multiple gluons actually add up into some coherent fields at such scale, which we do not yet understand.

The second issue has to do with the mutual screening of all these gluons. If the effective size of the hadron would not grow with energy, any number of interaction can only produce constant cross section of a black disk, without growth.

The physical origin of cross section growth remains an outstanding open problem: neither the perturbative resummations nor many non-perturbative models are really quantitative. It is hardly surprising, since scale at which soft Pomeron operates (as seen e.g. from the Pomeron slope $\alpha'(0) \approx 1/(2\,GeV)^2$) is also the "substructure scale" mentioned above.

There are basically three distinct approaches:

(i) *Minijet-based models* use familiar formulae from pQCD [48]. They are well-tested in the domain of hard jets, but their application at the semi-hard scale is a drastic extrapolation. All of these models assume the existence of a non-perturbative momentum cutoff, p_{cutoff}, in order to render pQCD results finite. This cutoff is left unexplained, treated as a purely phenomenological parameter, and all results depend greatly on its value.

(ii) *Instanton-based* dynamics, to be discussed below, have only recently been applied to high-energy scattering [28, 32, 29] and use insights obtained a decade ago in electroweak theory [33]. Particularly relevant for this work are the first two references, in which the growing part of the hadron-hadron cross sections is ascribed to multi-gluon production via instantons.

(iii) The *Color Glass Condensate*, a classical Weitzecker-Williams field of gluons carried by interacting hadrons, can be excited to produce prompt gluons [58]. This is another example of a weakly-coupled system involving non-perturbative gauge field configurations.

For long time people have constructed multi-peripheral models with ladders made of hadrons. Recent story started with Kharzeev and Levin[30] who kept t-channel gluons but tried to substitute the gluonic "rungs" of the BFKL ladder by those with a pair of pions, or sigma meson, to increase the cross section. They used the gg-$\pi\pi$ non-perturbative vertices known from the low energy theorem. Their estimated value for Δ was close to Δ_{phen}. Introducing *instantons* into the problem, I re-analyzed [31] the contribution of the colorless scalar channel generated by operator $G^2_{\mu\nu}$, using the gg-$\pi\pi$ and gg$-$*scalar* $-$ *glueball* couplings determined previously from the calculation of appropriate Euclidean correlators, see [2]. The result turns out to reduce those of the KL paper, with $\Delta \approx 0.05$ only, and pions and glueball contributions being roughly equal.

6.2 "Soft" Pomeron from instantons

We put "soft" in quotation marks here because we do not entirely agree with this terminology. It is now clear that the Pomeron itself is a small object, with its size represented by the slope of its trajectory, $\alpha'(t=0) \approx 1/(4\,\mathrm{GeV}^2)$. The scale involved, 0.1 fm, is much smaller than hadronic radii, and so the Pomeron exchanges should in fact be treated on the level of individual partons, appropriately defined at the intermediate momentum scale of 1-2 GeV. For lack of a better standard term, we will refer to it as the *semi-hard* scale.

More precisely, we will not consider the nature of the soft Pomeron in full either. The leading Regge pole, if it exists, is the analog of a single bound state appearing (in t-channel),as a result of a rather different interactions[14]. Although existence of such pole is an attractive possibility, no general principles demand it to be true in real QCD.

Recent application of the instanton-induced dynamics to this problem have been discussed in several papers [54]. Especially relevant for this Letter are two last works which use insights obtained a decade ago in discussion of instanton-induced processes in electroweak theory [33], and the growing part of the hh cross sections were ascribed to multi-gluon production via instantons, see Fig.13(b). Among qualitative features of this theory is the explanation of why no odderon appears (instantons are SU(2) objects, in which quarks and antiquarks are not really distinct), an explanation of the small power Δ (it is proportional to "instanton diluteness parameter" $n\rho^4$ mentioned above), the small size of the soft Pomeron (governed simply by small size of instantons $\rho \sim 1/3\,fm$). Although instanton-induced amplitudes contain small "diluteness" factor, there is no extra penalty for production of new gluons: thus one should expect instanton effects to exceed perturbative amplitudes of sufficiently high order. This generic idea is also behind the present work, dealing with prompt multi-gluon production.

[14] For example, J/ψ is definite charmonium state, which appears as a result of an interplay of both perturbative and confining potentials.

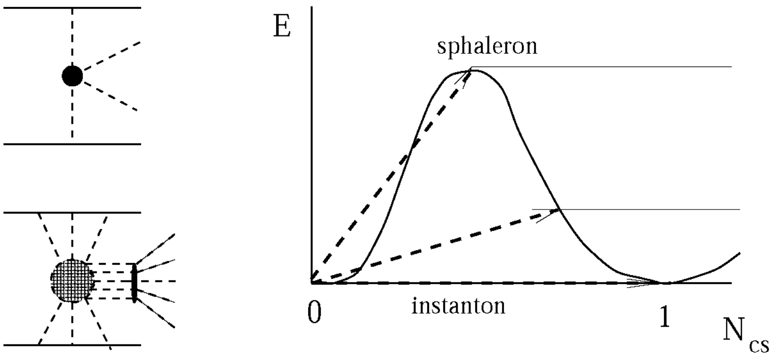

Figure 13: (a) A typical inelastic perturbative process (two t-channel gluons collide, producing a pair of gluons) to (b) non-perturbative inelastic process, incorporating collisions of few t-channel gluons with the instanton (the shaded circle), resulting in multi-gluon production. The bottom of the figure (c) shows the same process, but in a quantum mechanical way. The energy of Yang-Mills field versus the Chern-Simons number N_{cs} is a periodic function, with zeros at integer points. The *instanton* (shown by the lowest dashed line) is a transition between such points. However if some nonzero energy is deposited into the process during transition, the virtual path (the dashed line) leads to a *turning points*, from which starts the real time motion outside the barrier (shown by horizontal solid lines). The maximal cross section corresponds to the transition to the top of the barrier, called the *sphaleron*.

Technical description of the process can be split into two stages. The first (at which one evaluates the probability) is the motion *under the barrier*, and it is described by Euclidean paths approximated by instantons. Their interaction with the high energy colliding partons results in some energy deposition and subsequent motion *over the barrier*. Furthermore, the intermediate stage of the process (shown by the horizontal dashed lines) indicate *coherence* of the outgoing gluons: they are first produced in the form of specific gluomagnetic field configuration, the *turning states* at the figure above, which we study right now [60].

The top point is known as the *sphaleron*[15] configuration [55], first found in the context of electroweak theory. Intensive studies of the instanton-induced processes also were done in this context in early 1990's, driven basically by possible observability of baryon number violating processes in electroweak theory[33]. The so called "holy grail function" showed that processes with multiple quanta production indeed lead to growing cross section, reaching its maximum at the sphaleron mass and then decreasing. However,

[15] Which means "ready to fall" in Greek.

since in electroweak theory the maximal cross section has been found to be still very far from observability, the interest to this direction have mostly disappeared around 1993 or so.

At this second, Minkowski, stage the action is real, and the factor $exp(iS)$ does not affect the probability, and we only need to consider it for final state distributions. The sphaleron mass in QCD is

$$M_{sph} \approx \frac{30}{g^2(\rho)\rho} \sim 2.5\, GeV \tag{39}$$

Since those field configurations are close to classically unstable saddle point at the top of the barrier, they roll downhill and develop gluoelectric fields. When both become weak enough, solution can be decomposed into perturbative gluons. This part of the process can also be studied directly from classical Yang-Mills equation: for electroweak sphalerons it has been done in Refs[56], calculation for its QCD version is in progress [60]. While rolling, the configurations tend to forget the initial imperfections (such as a non-spherical shapes) since there is only one basic instability path downward: so the resulting fields should be nearly perfect spherical expanding shells. Electroweak sphalerons decay into approximately 51 W,Z,H quanta, of which only about 10% are Higgs bosons, which carry only 4% of energy. Ignoring those, one can estimate mean gluon multiplicity per sphaleron decay, by simple re-scaling of the coupling constants: the result gives 3-4 gluons. Although this number is not large, it is important to keep in mind that they appear as a coherent expanding shell of strong gluonic field.

In [59] we have tried to formulate a phenomenological model which would reasonably well describe describe data on various hadronic processes. In particular, we have shown that with the cross section of "sphaleron production" (per unit rapidity) by two "effective quarks" [16] being $\sigma_{qq} = 1.69 * 10^{-3} fm^2$, one can understand data about the energy growth of $NN, \pi N, \gamma N, \gamma\gamma$ cross sections. Furthermore, one can understand the effective power of the energy dependence versus impact parameter b, in pp collisions, see comparison in Fig. 14.

The next issue we address is whether the instanton approach can explain difference is the growing parts for different hadrons. To check that we need first to get the number of "relevant partons" for the nucleon, pion, and photon are summarized below in Table 1. The references given in the table are

[16] Those are defined as the number of quarks plus twice the number of gluons. Their number is evaluated from the structure functions. If integrated above the value of the Feynman $x \approx 0.01$ we get about 12 effective quarks/nucleon.

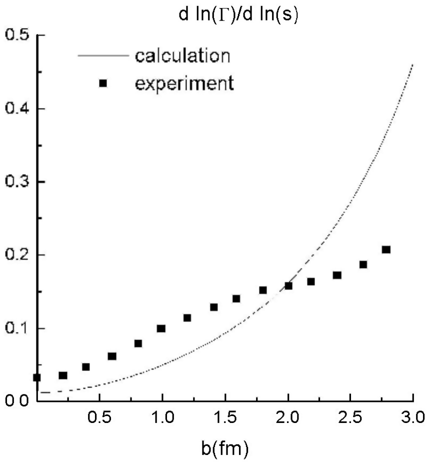

Figure 14: Effective power of the s-dependence of NN cross section, $\Delta(b)$, as a function of the impact parameter, b. Its decrease at small b is a consequence of "shadowing" of assumed instanton generated growing cross section by the ordinary color exchanges (with large but s-independent cross section). The squares marked "experiment" originate from the total and elastic amplitudes, those are taken from Kopeliovich et al. The line is our model. The agreement is not spectacular, but reasonable for a parameter-free model.

revised GRV parton distributions evaluated at next-to-leading order (NLO), taken at the scale of $Q^2 = 1$ GeV, which are then integrated over interval $x = [0.01, 1.0]$.

In principle, with more accurate parameterizations, we might try to test parton additivity by separately extracting, from the data of the *growing* part of hadronic cross section, the contributions of qq, qg, and gg to semi-hard processes. This was attempted, but with the accuracy at hand the differences between taking quarks and gluons is negligible. We are therefore forced to make a model-dependent assumption about their relative magnitude.

The instanton model [59] leads to a simple rule: changing a quark to a gluon result in extra Casimir factor 2. (It is different from the usual SU(3) Casimir scaling because we deal with SU(2) instanton fields, basically, although derivation is rather involved.) Therefore one can simply take the effective number of partons to be N_q+2N_g, where N_q and N_g are the numbers of quarks and gluons, respectively, taken from Table 1. This leaves us with only one unknown: the growing part of the qq cross section.

Combining the parton content with this simple recipe, one obtains the ratios of cross sections which may be compared to the coefficients of $\ln(s)$

Proton, with NLO structure functions GRV
$N_g = 4.10$
Valence $N_u = 1.70$
Valence $N_d = 0.84$
Sea $N_{u+d} = 1.16$
Pion, with NLO structure functions from GRV
$N_g = 3.1$
Valence $N_{u+\bar{d}} = 1.8$
Sea $N_{u+d} = 0.48$
Photon, with NLO structure functions from GRV
$N_g = 1.9\ \alpha$
$N_u = N_{\bar{u}} = 0.87\ \alpha$
$N_d = N_{\bar{d}} = 0.30\ \alpha$

Table 1: Partonic content of scattered particles (α is the fine structure constant).

extracted from experiment. The results, summarized in Table 2, are reasonable, but cannot be taken as precise since shadowing corrections have not been considered here.

Ratio	Computed	Part.Data Group
$\frac{1}{\alpha}\frac{X_{\gamma N}}{X_{NN}}$	0.50	0.43
$\frac{X_{\pi N}}{X_{NN}}$	0.73	0.63
$\frac{1}{\alpha}\frac{X_{\gamma\gamma}}{X_{\gamma N}}$	0.69	0.68

Table 2: Cross Section ratios as computed in the text and reported by the Particle Data Group.

After detailed study of shadowing in pp, we determined the quark-quark cross section to be $\sigma_{qq} = 1.69 \times 10^{-3}$ fm^2. We are now able to calculate the rising parts of total cross sections for other hadrons, and our precitions for $p\pi$, $p\gamma$, and $\gamma\gamma$ are given in Table 3. We find reasonable agreement between these numbers and the data, having fixed only one free parameter, c.

6.3 Instanton-induced production in heavy ion collisions

It has been suggested in [51] that if sphaleron-type object are copiously produced they may significantly increase the entropy produced and speed up

	Calculated	Part.Data Gr.
$X_{p\pi}$	0.132	0.111
$X_{p\gamma}$	5.65×10^{-4}	5.51×10^{-4}
$X_{\gamma\gamma}$	1.72×10^{-6}	1.45×10^{-6}

Table 3: Coefficients $X_{ij} = d\sigma_{ij}^{tot}/d\ln(s)$ in fm^2 for different hadronic constituents.

the equilibration process, as compared to mini-jet based scenarios considered previously.

For symmetric, central AA collisions of two nuclei we use the simplest model, one of two spheres with homogeneously distributed partons. The total parton number is AN_q, with $N_q \approx 12$ being the number of "effective quarks" (quarks number plus twice gluons number) per nucleon[17].

The total number of qq collisions in this case is easily obtained from the follwing geometric integral:

$$\begin{aligned} N_{coll} &= 8\pi\sigma_{qq}n_q^2 \int_0^R dr_t r_t \left(R^2 - rt^2\right) \\ &= 3^{4/3}2^{-5/3}\pi\sigma_{qq}n_q^2 \left(\frac{AN_q}{\pi n_q}\right)^{4/3}, \end{aligned} \tag{40}$$

where the quark density is determined by the nuclear density to be $n_q = N_q \times 0.16$ fm^{-3}.

With $A = 197$ (gold) and the value for the quark-quark cross section extracted above, $\sigma_{qq} = 1.69 \times 10^{-3}$ fm^2, we have the following production rate per unit rapidity of sphaleron-like clusters:

$$\frac{dN_{coll}}{dy} \approx 76.5\,, \tag{41}$$

a number somewhat smaller than estimated in Ref. [51].

Each cluster will in turn decay into a number of quarks and gluons. Simply scaling of the couplings from the studies of sphaleron decay in electroweak theory leads to about 3.5 gluons per cluster, with 0-6 quarks (up to a complete set of light quark-antiquark pairs, $\bar{u}u\bar{d}d\bar{s}s$). As an average we tentatively take 3.5 gluons and 2.5 quarks, the latter obtained by applying a factor of one half for the suppression of strange quarks and another one

[17] Of course, the clustering of partons into "constituent quarks" and nucleons increases the number of collisions, but we will ignore such correlations for now.

half to account for the possibly change in Chern-Simons number. This yields an average of six partons per cluster, or in central $AuAu$ collisions at RHIC about $76.5 \times 6 = 460$ partons per rapidity from sphaleron production. This is roughly *one half* the maximal possible value, $dN_{partons}/dy \sim dN_{hadrons}/dy \sim$ 1000, inferred experimentally from the final entropy limitations.

This result is in good agreement with phenomenological studies of the energy and impact-parameter dependence of multiplicity [57], which have deduced that the contribution to multiplicity which scales as the number of parton collisions generates about half of the total, when calculated from the standard Glauber model and using the experimental nuclear density distribution for a gold nucleus. In this picture, the ~ 500 hadrons per unit rapidity are then a result of prompt production from QCD sphalerons.

7 Brief Summary

In Lecture 1 we have discussed QCD vacuum, and concluded that the most important part of quark vacuum states are those with very small Dirac eigenvalues, made of *collectivized instanton zero modes.* Those form the quark condensate and in general dominate quark propagators at not-too-small distances.

In Lecture 2 we have studied the Euclidean correlation functions, the best bridge between theory, experiment and numerical experiment (lattice). Phenomenology of correlators is based on hadronic phenomenology, but is more directly related to quark motion and interaction. Dramatic instanton effects has been discussed, and some examples of truly quantitative description of (the tau decay) data has been shown. Again, keeping quasi-zero modes in all propagators does the job.

in Lecture 3 we learned that the QCD vacuum is not the only phase this theory may have. At least three directions are known, leading to quite different phases, and nearly all phase boundaries can be explained with instantons. At *high T* the instantons and anti-instantons form closed pairs with the top.charge zero: this restores chiral symmetry and lead to semi-perturbative phase known as *Quark-Gluon plasma.* At *high density* or chemical potentials, quark matter with intricate set of *Color superconducting phases* appear. In this case instantons and even one-gluon exchanges (at very high densities) create quark-quark Cooper pairs, which condense. Those play the role of composite Higgs scalar, and is in many respect similar to the Standard Model.

In Lecture 4 we discussed recent progress in heavy ion physics, devoted to experimental production of Quark-Gluon Plasma. New facility, RHIC, has just started, with many puzzling results. We have seen that its first data already show a spectacular explosion, driven by the predicted "QGP push". Many more puzzling phenomena, such as apparent jet disappearance in central collisions, are also discussed.

Lecture 5 was based on more recent material, it is attempts to explain old Pomeron phenomenology in terms of instanton-induced dynamics. The main lesson is that glue can be produced, from unphysical Euclidean paths to physical Minkowski evolution, in forms of the static magnetic "turning states", the relatives of the sphaleron. It was also conjectured that those objects are important for explaining RHIC data, and puzzling rapid production/equilibration of the QGP.

Acknowledgments. It is a pleasure to thank the organizers of the school, Alexei Smirnov and Goran Senjanovic, for their kind invitation and help during the school. It was indeed a very successful school, where people (I, at least) have learned a lot. This work was supported in parts by the US-DOE grant DE-FG-88ER40388.

References

[1] E.V. Shuryak, "The QCD Vacuum, Hadrons and the Superdense Matter," *SINGAPORE, WORLD SCIENTIFIC (1988) 401p*

[2] T. Schafer and E.V. Shuryak, Rev.Mod.Phys. 70 323 1998 hep-ph/9610451

[3] CHIRAL SYMMETRY BREAKING BY INSTANTONS. D. Diakonov, In *Varenna 1995, Selected topics in nonperturbative QCD* 397-432. e-Print Archive: hep-ph/9602375

[4] E.V. Shuryak, Rev. Mod. Phys. 651 1993

[5] M.A. Shifman, A.I. Vainshtein and V.I. Zakharov, Phys. Lett. 76B 471 1978

[6] E.V. Shuryak, Nucl. Phys. B203 93,116,140 1982

[7] V.A. Novikov, M.A. Shifman, A.I. Vainshtein and V.I. Zakharov, Nucl. Phys. B191 301 1981

[8] A.A. Belavin, A.M. Polyakov, A.S. Schvartz and Yu.S. Tyupkin, Phys. Lett. B59:85-87, 1975

[9] S.L. Adler, *Phys. Rev.*, 177:2426, 1969

[10] J.S. Bell and R. Jackiw, *Nuo. Cim.*, A60:47, 1969

[11] A. Hasenfratz and C. Nieter, Instanton Content of the SU(3) Vacuum, hep-lat/9806026

[12] G. 't Hooft, Phys. Rev. D14, 3432, 1976

[13] Y. Nambu and G. Jona-Lasinio, *Phys. Rev.* 122 345 1961

[14] D. Diakonov and V.Yu. Petrov, Nucl. Phys. B272:457, 1986

[15] M.A. Shifman, A.I. Vainshtein and V.I. Zakharov, Nucl. Phys. B165:45, 1980

[16] A. Blotz and E.V. Shuryak, Phys. Rev. D55:4055-4065, 1997

[17] R. Barate *et al.*, [ALEPH Collaboration], Z. Phys. **C76**, 15 (1997)

[18] R. Barate *et al.*, [ALEPH Collaboration], Eur. Phys. J. **C4**, 409 (1998)

[19] T. Schäfer, E.V. Shuryak, and J.J.M. Verbaarschot, Nucl. Phys. B412 143 1994

[20] R. Rapp, T. Schäfer, E.V. Shuryak and M. Velkovsky, Phys. Rev. Lett. 81 53 1998

[21] M. Alford, K. Rajagopal and F. Wilczek, Phys. Lett. B422 247 1998

[22] J.B. Kogut, M.A. Stephanov and D. Toublan, hep-ph/9906346

[23] R. Rapp, E.V. Shuryak and I. Zahed, Phys. Rev. D63:034008, 2001: hep-ph/0008207

[24] M. Alford, J. Bowers and K. Rajagopal, Phys. Rev. D**63**, 074016, 2001, [hep-ph/0008208]

[25] C. Gattringer, M. Gockeler, C.B. Lang, P.E. Rakow and A. Schafer, hep-lat/0108001; T. DeGrand, hep-lat/0106001; T. Blum *et al.*, hep-lat/0105006; R.G. Edwards and U.M. Heller, hep-lat/0105004; I. Hip, T. Lippert, H. Neff, K. Schilling and W. Schroers, hep-lat/0105001; T. DeGrand and A. Hasenfratz, hep-lat/0103002

[26] S. Hands, I. Montvay, M. Oevers, L. Scorzato and J. Skullerud, Nucl. Phys. Proc. Suppl. **94**, 461, 2001, [hep-lat/0010085]

[27] Proceedings of Quark Matter 2001, Stony Brook Jan. 2001, Nucl. Phys. A, in press

[28] E.V. Shuryak and I. Zahed, hep-ph/0005152; Phys. Rev. D, in press

[29] M. Nowak, E.V. Shuryak and I. Zahed, Soft Pomeron from Interacting Instantons, in progress

[30] D. Kharzeev and E. Levin, BNL-NT-99-8, Dec 1999, 12pp. hep-ph/991221

[31] E. Shuryak, `hep-ph 0001189`

[32] D. Kharzeev, Y. Kovchegov and E. Levin, hep-ph/0007182

[33] A. Ringwald, Nucl. Phys. B330 (1990) 1; O. Espinosa, Nucl. Phys. B343 (1990) 310; V.V. Khoze, A. Ringwald, Phys. Lett. B259:106-112, 1991; V.I. Zakharov, Nucl. Phys. B353 (1991) 683; M. Maggiore and M. Shifman, Phys. Rev. D46:3550-3564, 1992

[34] D.T. Son, Phys.Rev. D59 094019 1999; hep-ph/9812287

[35] D.T. Son, M.A. Stephanov and A.R. Zhitnitsky, hep-ph/0103099

[36] D. Teaney and E.V. Shuryak, Phys. Rev. Lett. **83**, 4951 (1999); [nucl-th/9904006]

[37] K.H. Ackermann *et al.*, [STAR Collaboration], "Elliptic flow in Au + Au collisions at s(N N)**(1/2) = 130-GeV," nucl-ex/0009011

[38] E. Shuryak, Phys. Rep. **61** , 71 (1980); Phys. Lett. **78B**, 150 (1978); Sov. J. Nucl. Phys. **28**, 408 (1978)

[39] J.Y. Ollitrault, Phys. Rev. **D46**, 229 (1992); Phys. Rev. **D48**, 1132 (1993)

[40] F. Karsch, E. Laermann, A. Peikert, Phys. Lett. B478:447-455, 2000: hep-lat/0002003

[41] C.M. Hung and E. Shuryak, Phys. Rev. **C57**, 1891 (1998)

[42] P.F. Kolb, J. Sollfrank and U. Heinz, hep-ph/0006129

[43] D. Teaney, J. Lauret, and E.V. Shuryak, in progress

[44] E.V. Shuryak, Invited talk at 14th International Conference on Ultrarelativistic Nucleus-Nucleus Collisions (QM 99), Torino, Italy, 10-15 May 1999. Nucl. Phys. A661:119-129, 1999; hep-ph/9906443

[45] M. Bleicher and H. Stocker, hep-ph/0006147

[46] Xin-Nian Wang, Miklos Gyulassy, Phys. Rev. D44:3501-3516, 1991

[47] B.B. Back *et al.*, [PHOBOS Collaboration], Phys. Rev. Lett. **85**, 3100 (2000) [hep-ex/0007036]

[48] J.P. Blaizot and A.H. Mueller, Nucl. Phys. **B289**, 847 (1987) Kajantie, P.V. Landshoff and J. Lindfors, Phys. Rev. Lett. **59**, 2527 (1987)

[49] T.S. Biro, E. van Doorn, B. Muller, M.H. Thoma and X.N. Wang, Phys. Rev. C **48**, 1275 (1993) [nucl-th/9303004]; L. Xiong and E. Shuryak, Phys. Rev. C **49**, 2203 (1994) [hep-ph/9309333]; R. Baier, A.H. Mueller, D. Schiff and D.T. Son, hep-ph/0009237

[50] M. Gyulassy, I. Vitev and X.N. Wang, Phys. Rev. Lett. **86**, 2537 (2001) [nucl-th/0012092]

[51] E.V. Shuryak, Phys. Lett. B **515**, 359 (2001)

[52] R. Rapp and J. Wambach, "Vector mesons in medium and dileptons in heavy-ion collisions", nucl-th/0001014.

[53] E. Kuraev, L. Lipatov and V. Fadin, Sov. Phys. JETP **45** (1977) 199; I. Balitsky and L. Lipatov, Sov. J. Nucl. Phys. **28** (1978) 822; L. Lipatov, Sov. Phys. JETP **63** (1986) 904

[54] D. Kharzeev, Y. Kovchegov and E. Levin, `hep-ph/0007182`; E. Shuryak and I. Zahed, Phys. Rev. **D62** (2000) 085014, `hep-ph/0005152`; M.A. Nowak, E.V. Shuryak and I. Zahed, hep-ph/0012232; Phys. Rev. D, in press

[55] N. Manton, Phys. Rev. D28 (1983) 2019; F.R. Klinkhamer and N. Manton, Phys. Rev. D30 (1984) 2212

[56] J. Zadrozny, Phys. Lett. B **284**, 88 (1992); M. Hellmund and J. Kripfganz, Nucl. Phys. B **373**, 749 (1992)

[57] D. Kharzeev and M. Nardi, Phys. Lett. B **507**, 121 (2001)

[58] L.D. McLerran and R. Venugopalan, Phys. Rev. D **59**, 094002 (1999); [arXiv:hep-ph/9809427]. A. Krasnitz, Y. Nara and R. Venugopalan, Phys. Rev. Lett. **87**, 192302 (2001), [arXiv:hep-ph/0108092]

[59] G. Carter, D. Ostrovsky and E. Shuryak,Instanton-induced Semi-hard Parton Interactions and phenomenology of High Energy Hadron Collisions, hep-ph/0112036

[60] G. Carter, D. Ostrovsky and E. Shuryak, in preparation

Confronting the Conventional Ideas of Grand Unification with Fermion Masses, Neutrino Oscillations and Proton Decay[*]

Jogesh C. Pati[†]

Department of Physics, University of Maryland, College Park, USA[‡]
and
Stanford Linear Accelerator Center, Menlo Park, USA

Lectures given at the
Summer School on Particle Physics
Trieste, 18 June - 6 July 2001

LNS0210003

[*] Invited talk presented at the International Summer School held at ICTP, Trieste (June, 2001) and at WHEPP-7 Conference, Allahabad, India (January, 2002). This is an updated version of the talk presented at the Erice School (September, 2000), hep-ph/0106082

[†] pati@physics.umd.edu

[‡] Permanent address.

Abstract

It is noted that one is now in possession of a set of facts, which may be viewed as the *matching pieces of a puzzle*; in that all of them can be resolved by just one idea - that is grand unification. These include: (i) the observed family-structure, (ii) quantization of electric charge, (iii) meeting of the three gauge couplings, (iv) neutrino oscillations; in particular the mass squared-difference $\Delta m^2(\nu_\mu - \nu_\tau)$ (suggested by SuperK), (v) the intricate pattern of the masses and mixings of the fermions, including the smallness of V_{cb} and the largeness of $\theta^{\mathrm{osc}}_{\nu_\mu \nu_\tau}$, and (vi) the need for B–L as a generator to implement baryogenesis (via leptogenesis). All these pieces fit beautifully together within a single puzzle board framed by supersymmetric unification, based on SO(10) or a string-unified G(224)-symmetry. The two notable pieces of the puzzle still missing, however, are proton decay and supersymmetry.

A concrete proposal is presented, within a predictive SO(10)/G(224)-framework, that successfully describes the masses and mixings of all fermions, including the neutrinos - with eight predictions, all in agreement with observation. Within this framework, a systematic study of proton decay is carried out, which (a) pays special attention to its dependence on the fermion masses, including the superheavy Majorana masses of the right-handed neutrinos, and (b) limits the threshold corrections so as to preserve natural coupling unification. The study updates prior work by Babu, Pati and Wilczek, in the context of both MSSM and its (interesting) variant, the so-called ESSM, by allowing for improved values of the matrix elements and of the short and long-distance renormalization effects. It shows that a conservative upper limit on the proton lifetime is about (1/3 - 2)×10^{34} years, with $\overline{\nu}K^+$ being the dominant decay mode, and quite possibly μ^+K^0 and $e^+\pi^0$ being prominent. This in turn strongly suggests that an improvement in the current sensitivity by a factor of five to ten (compared to SuperK) ought to reveal proton decay. Otherwise some promising and remarkably successful ideas on unification would suffer a major setback. For comparison, some alternatives to the conventional approach to unification pursued here are mentioned at the end.

Contents

1 Introduction

The standard model of particle physics, based on the gauge symmetry SU$(2)_L \times$U$(1)_Y \times$ SU$(3)_C$ [1, 2] is in excellent agreement with observations, at least up to energies of order 100 GeV. Its success in turn constitutes a triumph of quantum field theory, especially of the notions of gauge invariance, spontaneous symmetry breaking, and renormalizability. The next step in the unification-ladder is associated with the concept of "grand unification", which proposes a unity of quarks and leptons, and simultaneously of their three basic forces: weak, electromagnetic and strong [3, 4, 5]. This concept was introduced on purely aesthetic grounds, in fact *before* any of the empirical successes of the standard model was in place. It was realized in 1972 that the standard model judged on aesthetic merits has some major shortcomings [3, 4]. For example, it puts members of a family into five scattered multiplets, assigning rather peculiar hypercharge quantum numbers to each of them, without however providing a compelling reason for doing so. It also does not provide a fundamental reason for the quantization of electric charge, and it does not explain why the electron and proton possess exactly equal but opposite charges. Nor does it explain the co-existence of quarks and leptons, and that of the three gauge forces—weak, electromagnetic and strong—with their differing strengths.

The idea of grand unification was postulated precisely to remove these shortcomings. It introduces the notion that quarks and leptons are members of one family, linked together by a symmetry group G, and that the weak, electromagnetic and strong interactions are aspects of one force, generated by gauging this symmetry G. The group G of course inevitably contains the standard model symmetry $G(213) = SU(2)_L \times U(1)_Y \times SU(3)_C$ as a subgroup. Within this picture, the observed differences between quarks and leptons and those between the three gauge forces are assumed to be low-energy phenomena that arise through a spontaneous breaking of the unification symmetry G to the standard model symmetry G(213), at a very high energy scale $M \gg 1$ TeV. As a *prediction* of the hypothesis, such differences must then disappear and the true unity of quarks and leptons and of the three gauge forces should manifest at energies exceeding the scale M.

The second and perhaps the most dramatic prediction of grand unification is proton decay. This important process, which would provide the window to view physics at truly short distances ($< 10^{-30}$ cm), is yet to be seen. Nevertheless, as I will stress in this talk, there has appeared over the years an impressive set of facts, favoring the hypothesis of grand unification which in turn suggest that the discovery of proton decay should be imminent. These include:

(a) The observed family structure: The five scattered multiplets of the standard model, belonging to a family, neatly become parts of a whole (*a single multiplet*), with their weak hypercharges predicted by grand unification, precisely as observed. It is hard to believe that this is just an accident. Realization of this feature calls for an extension of the standard model symmetry G(213) $=\mathrm{SU}(2)_L\times\mathrm{U}(1)_Y\times\mathrm{SU}(3)^C$ *minimally* to the symmetry group G(224) $=\mathrm{SU}(2)_L\times\mathrm{SU}(2)_R\times$ $\mathrm{SU}(4)^C$ [3], which can be extended further into the simple group SO(10) [6], but not SU(5) [4]. The G(224) symmetry in turn introduces some additional attractive features (see Section 2), including especially the right-handed (RH) neutrinos (ν_R's) accompanying the left-handed ones (ν_L's), and B–L as a local symmetry. As we will see, both of these features now seem to be needed, on empirical grounds, to understand neutrino masses and to implement baryogenesis.

(b) Quantization of electric charge and the fact that $Q_{\mathbf{electron}} = -Q_{\mathbf{proton}}$: Grand Unification provides compelling reasons for both of these facts.

(c) Meeting of the gauge couplings: Such a meeting is found to occur at a scale $M_X \approx 2 \times 10^{16}$ GeV, when the three gauge couplings are extrapolated from their values measured at LEP to higher energies, in the context of supersymmetry [7]. This dramatic phenomenon provides a strong support in favor of the ideas of both grand unification and supersymmetry [8]. Both of these features in turn may well emerge from a string theory [9] or M-theory [10] (see discussion in Section 3).

(d) $\boldsymbol{\Delta m^2(\nu_\mu - \nu_\tau) \sim (1/20eV)^2}$: The recent discovery of atmospheric neutrino-oscillation at SuperKamiokande [11] suggests a value $\Delta m^2(\nu_\mu\nu_\tau) \sim (1/20\ \mathrm{eV})^2$. It has been argued (see e.g. Ref. [12]) that precisely such a magnitude of $\Delta m^2(\nu_\mu\nu_\tau)$ can be understood very simply by utilizing the SU(4)-color relation $m(\nu_\tau)_{\mathrm{Dirac}} \approx m_{\mathrm{top}}$ and the SUSY unification scale M_X, noted above (See Section 4).

(e) Some intriguing features of fermion masses and mixings: These include: (i) the "observed" near equality of the masses of the b-quark and the τ-lepton at the unification-scale (i.e. $m_b^0 \approx m_\tau^0$) and (ii) the observed largeness of the ν_μ-ν_τ oscillation angle ($\sin^2 2\theta^{\mathrm{osc}}_{\nu_\mu\nu_\tau} \geq 0.92$) [11], together with the smallness of the corresponding quark mixing parameter $V_{cb}(\approx 0.04)$ [13]. As shown in recent work by Babu, Wilczek and me [14], it turns out that these features and more can be understood remarkably well (see discussion in Section 5) within an economical and predictive SO(10)-framework based on a minimal Higgs system. The success of this framework is in large part due simply to the group-structure of SO(10). For most purposes, that of G(224) suffices.

(f) Baryogenesis: To implement baryogenesis [15] successfully, in the presence of electroweak sphaleron effects [16], which wipe out any baryon excess generated at high temperatures in the (B–L)-conserving mode, it has become apparent that one would need B–L as a generator of the underlying symmetry in four dimensions, whose spontaneous violation at high temperatures would yield, for example, lepton asymmetry (leptogenesis). The latter in turn is converted to baryon-excess at lower temperatures by electroweak sphalerons. This mechanism, it turns out, yields even quantitatively the right magnitude for baryon excess [17]. The need for B–L, which is a generator of SU(4)-color, again points to the need for G(224) or SO(10) as an effective symmetry near the unification-scale M_X.

The success of each of these six features (a)–(f) seems to be non-trivial. Together they make a strong case for both *the conventional ideas on supersymmetric grand unification* and simultaneously for the G(224)/SO(10)-route to such unification, as being relevant to nature at short distances $\leq (10^{16}\ \text{GeV})^{-1}$, in four dimensions.[1] However, despite these successes, as long as proton decay remains undiscovered, the hallmark of grand unification—that is *quark-lepton transformability*—would remain unrevealed.

The relevant questions in this regard then are: What is the predicted range for the lifetime of the proton—in particular an upper limit—within the empirically favored route to unification mentioned above? What are the expected dominant decay modes within this route? Are these predictions compatible with current lower limits on proton lifetime mentioned above, and if so, can they still be tested at the existing or possible near-future detectors for proton decay?

Fortunately, we are in a much better position to answer these questions now, compared to a few years ago, because meanwhile we have learnt more about the nature of grand unification, and also there have been improved evaluations of the relevant matrix elements and short and long-distance renormalization effects. As noted above (see also Section 2 and Section 4),

[1] For comparison, some alternative attempts, including those based on the ideas of (a) large extra dimensions, and (b) unification occurring only in higher dimensions, are mentioned briefly in Section 6 G.

the neutrino masses and the meeting of the gauge couplings together seem to select out the supersymmetric G(224)/SO(10)-route to higher unification. The main purpose of my talk here will therefore be to address the questions raised above, in the context of this route. For the sake of comparison, however, I will state the corresponding results for the case of supersymmetric SU(5) as well.

My discussion will be based on a recent study of proton decay by Babu, Wilczek and me [14], an update presented in the Erice talk [18], and a subsequent update of the same as presented here. Relative to other analyses, this study has four distinctive features:

(i) It systematically takes into account the link that exists between proton decay and the masses and mixings of all fermions, including the neutrinos.

(ii) In particular, in addition to the contributions from the so-called "standard" $d = 5$ operators [19] (see Section 6), it includes those from a *new* set of $d = 5$ operators, related to the Majorana masses of the RH neutrinos [20]. These latter are found to be generally as important as the standard ones.

(iii) As discussed in the Appendix, the work also restricts GUT-scale threshold corrections, so as to preserve naturally coupling unification, in accord with the observed values of the three gauge couplings.

(iv) Finally, the present update incorporates recently improved values of the matrix elements, and the short and long-distance renormalization effects, which significantly enhance proton decay rate.

Each of these features turn out to be *crucial* to gaining a reliable insight into the nature of proton decay. Our study shows that the inverse decay rate for the $\overline{\nu}K^+$-mode, which is dominant, is less than about 1.2×10^{31} years for the case of MSSM embedded in minimal SUSY SU(5), and that it is less than about 10^{33} years for the case of MSSM embedded in SO(10). These upper bounds are obtained by making generous allowance for uncertainties in the matrix element and the SUSY-spectrum. Typically, the lifetime should of course be less than these bounds.

Proton decay is studied also for the case of the extended supersymmetric standard model (ESSM), that has been proposed a few years ago [21] on several grounds, based on the issues of (a) an understanding of the inter-family mass-hierarchy, (b) removing the mismatch between MSSM and string-unification scales, and (c) dilaton-stabilization (see Section 6 and the appendix). This case adds an extra pair of vector-like families at the TeV-scale, transforming as $\mathbf{16}+\overline{\mathbf{16}}$ of SO(10), to the MSSM spectrum. While the case of ESSM is fully compatible with both neutrino-counting at LEP and precision electroweak tests, it can of course be tested directly at the LHC through a search for the vectorlike fermions. Our study shows that, with the inclusion of only the "standard" $d = 5$ operators (defined in Section 6), ESSM, embedded in SO(10), can quite plausibly lead to proton lifetimes in the range of $10^{33} - 10^{34}$ years, for nearly central values of the parameters pertaining to the SUSY-spectrum and the matrix element. Allowing for a wide variation of the parameters, owing to the contributions from both the standard and the neutrino mass-related $d = 5$ operators (discussed in Section 6), proton lifetime still gets bounded above by about 2×10^{34} years, for the case of ESSM, embedded in SO(10) or a string-unified G(224).

For either MSSM or ESSM, embedded in G(224) or SO(10), due to contributions from the new operators, the μ^+K^0-mode is found to be prominent, with a branching ratio typically in the range of 10-50%. By contrast, minimal SUSY SU(5), for which the new operators are absent, would lead to branching ratios $\leq 10^{-3}$ for this mode. It is stressed that the $e^+\pi^0$-mode induced by gauge boson-exchange, in either SUSY SU(5) or SUSY SO(10), could have an inverse decay rate as short as about $(1 - 2) \times 10^{34}$ years.

Thus our study of proton decay, correlated with fermion masses, strongly suggests that discovery of proton decay should be imminent. Allowing for the possibility that the proton lifetime may well be closer to the upper bound stated above, a next-generation detector providing a net gain in sensitivity in proton decay-searches by a factor of 5–10, compared to SuperK, would certainly be needed not just to produce proton-decay events, but also to clearly distinguish them from the background. It would of course also be essential to study the branching ratios of certain sub-dominant but crucial decay modes, such as the μ^+K^0 and $e^+\pi^0$. The importance of such improved sensitivity, in the light of the successes of supersymmetric grand unification, is emphasized at the end.

2 Advantages of the Symmetry G(224) as a Step to Higher Unification

As mentioned in the introduction, the hypothesis of grand unification was introduced to remove some of the conceptual shortcomings of the standard model (SM). To illustrate the advantages of an early suggestion in this regard, consider the five standard model multiplets belonging to the electron-family as shown:

$$\begin{pmatrix} u_r & u_y & u_b \\ d_r & d_y & d_b \end{pmatrix}_L^{\frac{1}{3}} ; \begin{pmatrix} u_r & u_y & u_b \end{pmatrix}_R^{\frac{4}{3}} ; \begin{pmatrix} d_r & d_y & d_b \end{pmatrix}_R^{-\frac{2}{3}} ; \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L^{-1} ; (e^-)_R^{-2} . \tag{1}$$

Here the superscripts denote the respective weak hypercharges Y_W (where $Q_{em} = I_{3L} + Y_W/2$) and the subscripts L and R denote the chiralities of the respective fields. If one asks: how one can put these five multiplets into just one multiplet, the answer turns out to be simple and unique. As mentioned in the introduction, the minimal extension of the SM symmetry G(213) needed, to achieve this goal, is given by the gauge symmetry [3]:

$$\mathrm{G(224)} = \mathrm{SU(2)}_L \times \mathrm{SU(2)}_R \times \mathrm{SU(4)}^C . \tag{2}$$

Subject to left-right discrete symmetry ($L \leftrightarrow R$), which is natural to G(224), all members of the electron family become parts of a single left-right self-conjugate multiplet, consisting of:

$$F^e_{L,R} = \begin{bmatrix} u_r & u_y & u_b & \nu_e \\ d_r & d_y & d_b & e^- \end{bmatrix}_{L,R} . \tag{3}$$

The multiplets F^e_L and F^e_R are left-right conjugates of each other and transform respectively as (2,1,4) and (1,2,4) of G(224); likewise for the muon and the tau families. Note that the symmetries $\mathrm{SU(2)}_L$ and $\mathrm{SU(2)}_R$ are just like the familiar isospin symmetry, except that they operate on quarks and well as leptons, and distinguish between left and right chiralities. The left weak-isospin $\mathrm{SU(2)}_L$ treats each column of F^e_L as a doublet; likewise $\mathrm{SU(2)}_R$ for F^e_R. The symmetry SU(4)-color treats each row of F^e_L *and* F^e_R as a quartet; *thus lepton number is treated as the fourth color.* Note also that postulating either SU(4)-color or $\mathrm{SU(2)}_R$ forces one to introduce a right-handed neutrino (ν_R) for each family as a singlet of the SM symmetry. *This requires that there be sixteen two-component fermions in each family, as opposed to fifteen for the SM.* The symmetry G(224) introduces an elegant charge formula:

$$Q_{em} = I_{3L} + I_{3R} + \frac{B - L}{2} \tag{4}$$

expressed in terms of familiar quantum numbers I_{3L}, I_{3R} and -B–L, which applies to all forms of matter (including quarks and leptons of all six flavors, gauge and Higgs bosons). Note that the weak hypercharge given by $Y_W/2 =$

$I_{3R} + \frac{B-L}{2}$ is now completely determined for all members of the family. The values of Y_W thus obtained precisely match the assignments shown in Eq. (1). Quite clearly, the charges I_{3L}, I_{3R} and B–L, being generators respectively of $SU(2)_L$, $SU(2)_R$ and $SU(4)^c$, are quantized; so also then is the electric charge Q_{em}.

In brief, the symmetry G(224) brings some attractive features to particle physics. These include:

(i) Unification of all 16 members of a family within one left-right self-conjugate multiplet;

(ii) Quantization of electric charge, with a reason for the fact that $Q_{\text{electron}} = -Q_{\text{proton}}$

(iii) Quark-lepton unification (through SU(4) color);

(iv) Conservation of parity at a fundamental level [3, 22];

(v) Right-handed neutrinos ($\nu_R' s$) as a compelling feature; and

(vi) B–L as a local symmetry.

As mentioned in the introduction, the two distinguishing features of G(224)—i.e. the existence of the RH neutrinos and B–L as a local symmetry—now seem to be needed on empirical grounds. Furthermore, SU(4)-color provides simple relations between the masses and mixings of quarks and leptons, while $SU(2)_L\times SU(2)_R$ relates the mass-matrices in the up and down sectors. As we will see in Sections 4 and 5, these relations are in good accord with observations.

Believing in a complete unification, one is led to view the G(224) symmetry as part of a bigger symmetry, which itself may have its origin in an underlying theory, such as string theory. In this context, one may ask: Could the effective symmetry below the string scale in four dimensions (see Section 3) be as small as just the SM symmetry G(213), even though the latter may have its origin in a bigger symmetry, which lives only in higher dimensions? I will argue in Section 4 that the data on neutrino masses and the need for baryogenesis provide an answer to the contrary, suggesting that it is the *effective symmetry in four dimensions, below the string scale, which must minimally contain either* G(224) *or a close relative* $G(214) = SU(2)_L\times I_{3R}\times SU(4)^C$.

One may also ask: does the effective four dimensional symmetry have to be any bigger than G(224) near the string scale? In preparation for an answer to this question, let us recall that the smallest simple group that contains the SM symmetry G(213) is SU(5) [4]. It has the virtue of demonstrating how the main ideas of grand unification, including unification of the gauge couplings, can be realized. However, SU(5) does not contain G(224) as a subgroup. As such, it does not possess some of the advantages listed above. In particular, it does not contain the RH neutrinos as a compelling feature, and B–L as a local symmetry. Furthermore, it splits members of a family (not including ν_R) into two multiplets: $\overline{\mathbf{5}} + \mathbf{10}$.

By contrast, the symmetry SO(10) has the merit, relative to SU(5), that it contains G(224) as a subgroup, and thereby retains all the advantages of G(224) listed above. (As a historical note, it is worth mentioning that these advantages had been motivated on aesthetic grounds through the symmetry G(224) [3], and *all* the ideas of higher unification were in place [3, 4, 5], before it was noted that G(224) [isomorphic to SO(4)×SO(6)] embeds nicely into SO(10) [6]). Now, *SO(10) even preserves the 16-plet family-structure of G(224) without a need for any extension.* By contrast, if one extends G(224) to the still higher symmetry E_6 [23], the advantages (i)–(vi) are retained, but in this case, one must extend the family-structure from a 16 to a 27-plet, by postulating additional fermions. In this sense, there seems to be some advantage in having the effective symmetry below the string scale to be minimally G(224) [or G(214)] and maximally no more than SO(10). I will compare the relative advantage of having either a string-derived G(224) or a string-SO(10), in the next section. First, I discuss the implications of the data on coupling unification.

3 The Need for Supersymmetry: MSSM versus String Unifications

It has been known for some time that the precision measurements of the standard model coupling constants (in particular $\sin^2\theta_W$) at LEP put severe constraints on the idea of grand unification. Owing to these constraints, the non-supersymmetric minimal SU(5), and for similar reasons, the one-step breaking minimal non-supersymmetric SO(10)-model as well, are now excluded [24]. But the situation changes radically if one assumes that the standard model is replaced by the minimal supersymmetric standard model (MSSM), above a threshold of about 1 TeV. In this case, the three gauge couplings are found to meet [7], to a very good approximation, barring a few percent discrepancy which can be attributed to threshold corrections (see Appendix). Their scale of meeting is given by

$$M_X \approx 2 \times 10^{16}\,\text{GeV} \quad (\text{MSSM or SUSY SU(5)})\ . \tag{5}$$

This dramatic meeting of the three gauge couplings, or equivalently the agreement of the MSSM-based prediction of $\sin^2\theta_W(m_Z)_{\rm th} = 0.2315 \pm 0.003$ [25] with the observed value of $\sin^2\theta_W(m_Z) = 0.23124 \pm 0.00017$ [13], provides a strong support for the ideas of both grand unification and supersymmetry, as being relevant to physics at short distances $\lesssim (10^{16}\,\text{GeV})^{-1}$.

In addition to being needed for achieving coupling unification there is of course an independent motivation for low-energy supersymmetry—i.e. for the existence of SUSY partners of the standard model particles with masses of order 1 TeV. This is because it protects the Higgs boson mass from getting large quantum corrections, which would (otherwise) arise from grand unification and Planck scale physics. It thereby provides at least a technical resolution of the so-called gauge-hierarchy problem. *In this sense low-energy supersymmetry seems to be needed for the consistency of the hypothesis of grand unification.* Supersymmetry is of course also needed for the consistency of string theory. Last but not least, as a symmetry linking bosons and fermions, it is simply a beautiful idea. And it is fortunate that low-energy supersymmetry can be tested at the LHC, and possibly at the Tevatron, and the proposed NLC.

The most straightforward interpretation of the observed meeting of the three gauge couplings and of the scale M_X, is that a supersymmetric grand unification symmetry (often called GUT symmetry), like SU(5) or SO(10), breaks spontaneously at M_X into the standard model symmetry G(213), and that supersymmetry-breaking induces soft masses of order one TeV.

Even if supersymmetric grand unification may well be a good effective theory below a certain scale $M \gtrsim M_X$, it ought to have its origin within an underlying theory like the string/M theory. Such a theory is needed to unify all the forces of nature including gravity, and to provide a good quantum theory of gravity. It is also needed to provide a rationale for the existence of flavor symmetries (not available within grand unification), which distinguish between the three families and can resolve certain naturalness problems including those associated with inter-family mass hierarchy. In the context of string or M-theory, an alternative interpretation of the observed meeting of the gauge couplings is however possible. This is because, even if the effective symmetry in four dimensions emerging from a higher dimensional string theory is non-simple, like G(224) or even G(213), string theory can still ensure familiar unification of the gauge couplings at the string scale. In this case, however, one needs to account for the small mismatch between the MSSM unification scale M_X (given above), and the string unification scale, given by $M_{st} \approx g_{st} \times 5.2 \times 10^{17}$ GeV $\approx 3.6 \times 10^{17}$ GeV (Here we have put $\alpha_{st} = \alpha_{GUT}(\mathrm{MSSM}) \approx 0.04$) [26]. Possible resolutions of this mismatch have been proposed. These include: (i) utilizing the idea of *string-duality* [27] which allows a lowering of M_{st} compared to the value shown above, or alternatively (ii) the idea of the so-called "Extended Supersymmetric Standard Model" (ESSM) that assumes the existence of two vector-like families, transforming as $(\mathbf{16} + \overline{\mathbf{16}})$ of SO(10), with masses of order one TeV [21], in addition to the three chiral families. The latter leads to a semi-perturbative unification by raising α_{GUT} to about 0.25-0.3. Simultaneously, it raises M_X, in two loop, to about $(1/2 - 2) \times 10^{17}$ GeV. (Other mechanisms resolving the mismatch are reviewed in Ref. [28]). In practice, a combination of the two mechanisms mentioned above may well be relevant. [2]

While the mismatch can thus quite plausibly be removed for a non-GUT string-derived symmetry like G(224) or G(213), a GUT symmetry like SU(5) or SO(10) would have an advantage in this regard because it would keep the gauge couplings together between M_{st} and M_X (even if $M_X \sim M_{st}/20$), and thus not even encounter the problem of a mismatch between the two scales. A supersymmetric four dimensional GUT-solution [like SU(5) or SO(10)], however, has a possible disadvantage as well, because it needs certain color triplets to become superheavy by the so-called doublet-triplet splitting mechanism (see Section 6 and Appendix), in order to avoid the problem of rapid proton decay. However, no such mechanism has emerged yet, in string theory, for the GUT-like solutions [29]. [3]

Non-GUT string solutions, based on symmetries like G(224) or G(2113) for example, have a distinct advantage in this regard, in that the dangerous color triplets, which would induce rapid proton decay, are often naturally

[2] I have in mind the possibility of string-duality [27] lowering M_{st} for the case of semi-perturbative unification in ESSM (for which $\alpha_{st} \approx 0.25$, and thus, without the use of string-duality, M_{st} would have been about 10^{18} GeV) to a value of about $(1\text{-}2)\times10^{17}$ GeV (say), and semi-perturbative unification [21] raising the MSSM value of M_X to about 5×10^{16} GeV$\approx M_{st}(1/2$ to $1/4)$ (say). In this case, an intermediate symmetry like G(224) emerging at M_{st} would be effective only within the short gap between M_{st} and M_X, where it would break into G(213). Despite this short gap, one would still have the benefits of SU(4)-color that are needed to understand neutrino masses (see Section 4), and to implement baryogenesis via leptogenesis. At the same time, since the gap is so small, the couplings of G(224), unified at M_{st} would remain essentially so at M_X, so as to match with the "observed" coupling unification, of the type suggested in Ref. [21].

[3] Some alternative mechanisms for doublet-triplet splitting, and for suppression of the $d = 5$ proton decay operators have been proposed in the context of higher dimensional theories. These will be mentioned briefly in Section 6 G.

projected out for such solutions [30, 31]. Furthermore, the non-GUT solutions invariably possess new "flavor" gauge symmetries, which distinguish between families. These symmetries are immensely helpful in explaining qualitatively the observed fermion mass-hierarchy (see e.g. Ref. [31]) and resolving the so-called naturalness problems of supersymmetry such as those pertaining to the issues of squark-degeneracy [32], CP violation [33] and quantum gravity-induced rapid proton decay [34].

Weighing the advantages and possible disadvantages of both, it seems hard at present to make a priori a clear choice between a GUT versus a non-GUT string-solution. As expressed elsewhere [35], it therefore seems prudent to keep both options open and pursue their phenomenological consequences. Given the advantages of G(224) or SO(10) in the light of the neutrino masses (see Sections 2 and 4), I will thus proceed by assuming that either a suitable four dimensional G(224)-solution [with the scale M_X being close to M_{st} (see footnote 2)], or a realistic four-dimensional SO(10)-solution (with the desired mechanism for doublet-triplet splitting) emerges effectively from an underlying string theory, at the "conventional" string-scale $M_{st} \sim 10^{17}$-10^{18} GeV, and that the G(224)/SO(10) symmetry in turn breaks spontaneously at the conventional GUT-scale of $M_X \sim 2 \times 10^{16}$ GeV (or at $M_X \sim 5 \times 10^{16}$ GeV for the case of ESSM, as discussed in footnote 2) to the standard model symmetry G(213). The extra dimensions of string/M-theory are assumed to be tiny with sizes $\leq M_X^{-1} \sim 10^{-30}$ cm, so as not to disturb the successes of GUT. In short, I assume that essentially *the conventional (good old) picture of grand unification, proposed and developed sometime ago [3, 4, 5, 6, 7], holds as a good effective theory above the unification scale M_X and up to some high scale $M \lesssim M_{st}$, with the added presumption that it may have its origin from the string/M-theory.* Such a picture seems to be directly motivated on observational grounds such as those based on (a) coupling unification (discussed above), (b) neutrino masses including the (mass)^2-difference of the ν_μ-ν_τ system and the near maximal ν_μ-ν_τ oscillation angle (see discussions in the next sections), and (c) the fact that spontaneous violation of B–L local symmetry at high temperatures, seems to be needed to implement baryogenesis via leptogenesis.[4]

We will see that with the broad assumption mentioned above, an economical and predictive framework emerges, which successfully accounts for

[4] Alternative scenarios such as those based on TeV-scale large extra dimensions [36], though intriguing, do not seem to provide simple explanations of these features: (a), (b) and (c). They will be mentioned briefly in Section 6 G.

a host of observed phenomena pertaining to the masses and the mixings of all fermions, including neutrinos. It also makes some crucial testable predictions for proton decay. I next discuss the implications of the mass of ν_τ, or rather of $\Delta m^2(\nu_\mu \nu_\tau)$, as revealed by the SuperK data.

4 $\Delta m^2(\nu_\mu \nu_\tau)$: Evidence In Favor of the G(224) Route

One can obtain an estimate for the mass of ν_L^τ in the context of G(224) or SO(10) by using the following three steps (see e.g. Ref. [12]):

(i) Assume that B−L and I_{3R}, contained in a string-derived G(224) or SO(10), break near the unification-scale:

$$M_X \sim 2 \times 10^{16}\,\mathrm{GeV}\,, \qquad (6)$$

through VEVs of Higgs multiplets of the type suggested by string-solutions—i.e. $\langle (1,2,4)_H \rangle$ for G(224) or $\langle \overline{\mathbf{16}}_H \rangle$ for SO(10), as opposed to $\mathbf{126}_H$ which seems to be unobtainable at least in weakly interacting string theory [37]. In the process, the RH neutrinos (ν_R^i), which are singlets of the standard model, can and generically will acquire superheavy Majorana masses of the type $M_R^{ij}\, \nu_R^{iT}\, C^{-1}\, \nu_R^j$, by utilizing the VEV of $\langle \overline{\mathbf{16}}_H \rangle$ and effective couplings of the form:

$$\mathcal{L}_M\,(SO(10)) \;=\; f_{ij}\; \mathbf{16}_i \cdot \mathbf{16}_j\; \overline{\mathbf{16}}_H\; \cdot \overline{\mathbf{16}}_H / M + h.c. \qquad (7)$$

A similar expression holds for G(224). Here $i, j \;=\; 1,2,3$, correspond respectively to e, μ and τ families. Such gauge-invariant non-renormalizable couplings might be expected to be induced by Planck-scale physics, involving quantum gravity or stringy effects and/or tree-level exchange of superheavy states, such as those in the string tower. With f_{ij} (at least the largest among them) being of order unity, we would thus expect M to lie between $M_{\mathrm{Planck}} \approx 2 \times 10^{18}$ GeV and $M_{\mathrm{string}} \approx 4 \times 10^{17}$ GeV. Ignoring for the present off-diagonal mixings (for simplicity), one thus obtains [5]:

$$M_{3R} \approx \frac{f_{33} \langle \overline{\mathbf{16}}_H \rangle^2}{M} \approx f_{33}\,(2 \times 10^{14}\,\mathrm{GeV})\,\rho^2\,(M_{\mathrm{Planck}}/M) \qquad (8)$$

[5] The effects of neutrino-mixing and of the more legitimate choice of $M = M_{string} \approx 4 \times 10^{17}$ GeV (instead of $M = M_{\mathrm{Planck}}$) on the values of $m(\nu_L^\tau)$ and of M_{3R} are considered in Ref. [14] and are reflected in our discussions in Section 5. The two effects together end up in yielding essentially the same mass for $m(\nu_L^\tau)$ as obtained within the simplified picture presented in this section, together with a value for $M_{3R} \approx$ (5-10) $\times\, 10^{14}$ GeV.

This is the Majorana mass of the RH tau neutrino. Guided by the value of M_X, we have substituted $\langle\overline{\mathbf{16}}_H\rangle = (2 \times 10^{16}\,\mathrm{GeV})\,\rho$,where we expect $\rho \approx 1/2$ to 2 (say).

(ii) Now using SU(4)-color and the Higgs multiplet $(\mathbf{2},\mathbf{2},\mathbf{1})_H$ of G(224) or equivalently $\mathbf{10}_H$ of SO(10), one obtains the relation $m_\tau(M_X) = m_b(M_X)$, which is known to be successful. Thus, there is a good reason to believe that the third family gets its masses primarily from the $\mathbf{10}_H$ or equivalently $(2,2,1)_H$ (see Section 5). In turn, this implies:

$$m(\nu^\tau_{\mathrm{Dirac}}) \approx m_{\mathrm{top}}(M_X) \approx (100\text{-}120)\,\mathrm{GeV}\ . \tag{9}$$

Note that this relationship between the Dirac mass of the tau-neutrino and the top-mass is special to SU(4)-color. It does not emerge in SU(5).

(iii) Given the superheavy Majorana masses of the RH neutrinos as well as the Dirac masses as above, the see-saw mechanism [38] yields naturally light masses for the LH neutrinos. For ν^τ_L (ignoring flavor-mixing), one thus obtains, using Eqs.(8) and (9),

$$m(\nu^\tau_L) \approx \frac{m(\nu^\tau_{\mathrm{Dirac}})^2}{M_{3R}} \approx [(1/20)\,\mathrm{eV}\,(1\text{-}1.44)/f_{33}\,\rho^2]\,(M/M_{\mathrm{Planck}})\ . \tag{10}$$

In the next section, we discuss the masses and mixings of all three neutrinos. As we will see, given the hierarchical masses of quarks and charged leptons and the see-saw mechanism, we naturally obtain $m(\nu^\mu_L) \sim (1/10)m(\nu^\tau_L)$. We are thus led to predict that $\Delta m^2(\nu_\mu\nu_\tau)_{th} \equiv |m^2(\nu^\tau_L) - m^2(\nu^\mu_L)|_{th} \approx m^2(\nu^\tau_L)_{th} =$ square of the RHS of Eq. (10). Now SuperK result strongly suggests that it is observing ν^μ_L-ν^τ_L (rather than ν^μ_L-ν_X) oscillation, with a $\Delta m^2(\nu_\mu\nu_\tau)_{obs} \approx 3 \times 10^{-3}$ eV^2. It seems *truly remarkable* that the expected magnitude of $\Delta m^2(\nu_\mu\nu_\tau)$, given to a very good approximation by the square of the RHS of Eq. (10), is just about what is observed at SuperK, if $f_{33}\,\rho^2(M_{\mathrm{Planck}}/M) \approx 1.3$ to 1/2. Such a range for $f_{33}\,\rho^2(M_{\mathrm{Planck}}/M)$ seems most plausible and natural (see discussion in Ref. [12]). Note that the estimate (10) crucially depends upon the supersymmetric unification scale, which provides a value for M_{3R}, as well as on SU(4)-color that yields $m(\nu^\tau_{\mathrm{Dirac}})$. *The agreement between the expected and the SuperK results thus clearly favors supersymmetric unification, and in the string theory context, it suggests that the effective symmetry below the string-scale should contain SU(4)-color.* Thus, minimally this effective symmetry should be either G(214) or G(224), and maximally as big as SO(10), if not E$_6$.

By contrast, if SU(5) is regarded as either a fundamental symmetry or as the effective symmetry below the string scale, there would be no compelling reason based on symmetry alone, to introduce a ν_R, because it is a singlet of SU(5). Second, even if one did introduce ν_R^i by hand, their Dirac masses, arising from the coupling $h^i\,\overline{\mathbf{5}}_i\langle\mathbf{5}_H\rangle\nu_R^i$, would be unrelated to the up-flavor masses and thus rather arbitrary [contrast with Eq. (9)]. So also would be the Majorana masses of the ν_R^i's, which are SU(5)-invariant, and thus can be even of order string scale . This would give extremely small values of $m(\nu_L^\tau)$ and $m(\nu_L^\mu)$ and thus of $\Delta m^2(\nu_\mu\nu_\tau)$, which would be in gross conflict with observation.

Before passing to the next section, it is worth noting that the mass of ν_τ or of $\Delta m^2(\nu_\mu\nu_\tau)$ suggested by SuperK, as well as the observed value of $\sin^2\theta_W$ (see Section 3), provide valuable insight into the nature of GUT symmetry breaking. They both favor the case of a *single-step breaking* (SSB) of SO(10) or a string-unified G(224) symmetry at a high scale of order M_X, into the standard model symmetry G(213), as opposed to that of a multi-step breaking (MSB). The latter would correspond, for example, to SO(10) [or G(224)] breaking at a scale M_1 into G(213), which in turn breaks at a scale $M_2 \ll M_1$ into G(213). One reason why the case of single-step breaking is favored over that of MSB is that the latter can accommodate but not really predict $\sin^2\theta_W$, whereas the former predicts the same successfully. Furthermore, since the Majorana mass of ν_R^τ arises arises only after B–L and I_{3R} break, it would be given, for the case of MSB, by $M_{3R} \sim f_{33}(M_2^2/M)$, where $M \sim M_{st}$ (say). If $M_2 \ll M_X \sim 2\times10^{16}$ GeV, and M $> M_X$, one would obtain too low a value ($\ll 10^{14}$ GeV) for M_{3R} [compare with Eq. (8)], and thereby too large a value for $m(\nu_L^\tau)$, compared to that suggested by SuperK. By contrast, the case of single-step breaking (SSB) yields the right magnitude for $m(\nu_\tau)$ [see Eq. (10)].

Thus the success of the results on $m(\nu_\tau)$ and thereby on $\Delta m^2(\nu_\mu\nu_\tau)$ discussed above not only favors the symmetry SO(10) or G(224) beging effective in 4D at a high scale, but also clearly suggests that B–L and I_{3R} break near the conventional GUT scale $M_X \sim 2\times10^{16}$ GeV, rather than at an intermediate scale $\ll M_X$. In other words, the observed values of both $\sin^2\theta_W$ and $\Delta m^2(\nu_\mu\nu_\tau)$ favor only *the simplest pattern of symmetry-breaking*, for which SO(10) or a string-derived G(224) symmetry breaks in one step to the standard model symmetry, rather than in multiple steps. It is of course only this simple pattern of symmetry breaking that would be rather restrictive as regards its predictions for proton decay (to be discussed in Section 6).

I next discuss the problem of understanding the masses and mixings of all fermions.

5 Understanding Aspects of Fermion Masses and Neutrino Oscillations in SO(10)

Understanding the masses and mixings of all quarks *in conjunction with* those of the charged leptons *and* neutrinos is a goal worth achieving by itself. It also turns out to be essential for the study of proton decay. I therefore present first a partial attempt in this direction, based on a quark-lepton unified G(224)/SO(10)-framework, which seems most promising [14]. A few guidelines would prove to be helpful in this regard. The first of these is motivated by the desire for economy [see (11)], and the rest (see below) by the data. In essence, we will be following (partly) a *bottom-up approach* by appealing to the data to provide certain clues as regards the pattern of the Yukawa couplings, and simultaneously a *top-down approach* by appealing to grand unification, based on the symmetry G(224)/SO(10), to restrict the couplings by the constraints of group theory. The latter helps to interrelate the masses and mixings of quarks with those of the charged leptons and the neutrinos. As we will see, it is these *interrelationships*, which permit predictivity, and are found to be remarkably successful. The guidelines which we adopt are as follows.

1) Hierarchy Through Off-diagonal Mixings: Recall earlier attempts [39] that attribute hierarchical masses of the first two families to mass matrices of the form:

$$M = \begin{pmatrix} 0 & \epsilon \\ \epsilon & 1 \end{pmatrix} m_s^{(0)}\,, \tag{11}$$

for the (d, s) quarks, and likewise for the (u, c) quarks. Here $\epsilon \sim 1/10$. The hierarchical patterns in Eq. (11) can be ensured by imposing a suitable flavor symmetry which distinguishes between the two families (that in turn may have its origin in string theory (see e.g. Ref [31]). Such a pattern has the virtues that (a) it yields a hierarchy that is much larger than the input parameter ϵ: $(m_d/m_s) \approx \epsilon^2 \ll \epsilon$, and (b) it leads to an expression for the Cabibbo angle:

$$\theta_c \approx \left| \sqrt{\frac{m_d}{m_s}} - e^{i\phi}\sqrt{\frac{m_u}{m_c}} \right|, \tag{12}$$

which is rather successful. Using $\sqrt{m_d/m_s} \approx 0.22$ and $\sqrt{m_u/m_c} \approx 0.06$, we see that Eq. (12) works to within about 25% for any value of the phase ϕ. Note that the square root formula (like $\sqrt{m_d/m_s}$) for the relevant mixing angle arises because of the symmetric form of M in Eq. (11), which in turn is ensured if the contributing Higgs is a 10 of SO(10). A generalization of the pattern in Eq. (11) would suggest that the first two families (i.e. the e and the μ) receive masses primarily through their mixing with the third family (τ), with $(1,3)$ and $(1,2)$ elements being smaller than the $(2,3)$; while $(2,3)$ is smaller than the $(3,3)$. We will follow this guideline, except for the modification noted below.

2) The Need for an Antisymmetric Component: Although the symmetric hierarchical matrix in Eq. (11) works well for the first two families, a matrix of the same form fails altogether to reproduce V_{cb}, for which it yields:

$$V_{cb} \approx \left| \sqrt{\frac{m_s}{m_b}} - e^{i\chi} \sqrt{\frac{m_c}{m_t}} \right| . \tag{13}$$

Given that $\sqrt{m_s/m_b} \approx 0.17$ and $\sqrt{m_c/m_t} \approx 0.0.06$, we see that Eq. (13) would yield V_{cb} varying between 0.11 and 0.23, depending upon the phase χ. This is too big, compared to the observed value of $V_{cb} \approx 0.04 \pm 0.003$, by at least a factor of 3. We interpret this failure as a *clue* to the presence of an antisymmetric component in M, together with symmetrical ones (so that $m_{ij} \neq m_{ji}$), which would modify the relevant mixing angle to $\sqrt{m_i/m_j}\, \sqrt{m_{ij}/m_{ji}}$, where m_i and m_j denote the respective eigenvalues.

3) The Need for a Contribution Proportional to B–L: The success of the relations $m_b^0 \approx m_\tau^0$, and $m_t^0 \approx m(\nu_\tau)^0_{\mathrm{Dirac}}$ (see Section 4), suggests that the members of the third family get their masses primarily from the VEV of a SU(4)-color singlet Higgs field that is independent of B–L. This is in fact ensured if the Higgs is a 10 of SO(10). However, the empirical observations of $m_s^0 \sim m_\mu^0/3$ and $m_d^0 \sim 3m_e^0$ [40] call for a contribution proportional to B–L as well. Further, one can in fact argue that understanding naturally the suppression of V_{cb} (in the quark-sector) together with an enhancement of $\theta^{\mathrm{osc}}_{\nu_\mu \nu_\tau}$ (in the lepton sector) calls for a contribution that is not only proportional to B–L, but also antisymmetric in the family space (this later feature is suggested already in item (2)). We show below how both of these requirements can be met in SO(10), even for a minimal Higgs system.

4) Up-Down Asymmetry: Finally, the up and the down-sector mass matrices must not be proportional to each other, as otherwise the CKM angles would all vanish. Note that the cubic couplings of a single 10_H with the fermions in the 16's will not serve the purpose in this regard.

Following Ref. [14], I now present a simple and predictive mass-matrix, based on SO(10), that satisfies *all four* requirements (1), (2), (3) and (4). The interesting point is that one can obtain such a mass-matrix for the fermions by utilizing only the minimal Higgs system, that is needed anyway to break the gauge symmetry SO(10). It consists of the set:

$$H_{\text{minimal}} = \{\mathbf{45}_H, \mathbf{16}_H, \overline{\mathbf{16}}_H, \mathbf{10}_H\}. \tag{14}$$

Of these, the VEV of $\langle \mathbf{45}_H \rangle \sim M_X$ breaks SO(10) into G(2213), and those of $\langle \mathbf{16}_H \rangle = \langle \overline{\mathbf{16}}_H \rangle \sim M_X$ break G(2213) to G(213), at the unification-scale M_X. Now G(213) breaks at the electroweak scale by the VEV of $\langle 10_H \rangle$ to $\mathrm{U}(1)_{em} \times \mathrm{SU}(3)^c$.

One might have introduced large-dimensional tensorial multiplets of SO(10) like $\overline{\mathbf{126}}_H$ and $\mathbf{120}_H$, both of which possess cubic level Yukawa couplings with the fermions. In particular, the coupling $\mathbf{16}_i\mathbf{16}_j(\mathbf{120}_H)$ would give the desired family-antisymmetric as well as (B–L)-dependent contribution. We do not however introduce these multiplets in part because there is a general argument suggesting that they do not arise at least in weakly interacting heterotic string solutions [37], and in part also because mass-splittings within such large-dimensional multiplets could give excessive threshold corrections to $\alpha_3(m_z)$ (typically exceeding 20%), rendering observed coupling unification fortuitous. By contrast, the multiplets in the minimal set (shown above) can arise in string solutions. Furthermore, the threshold corrections for the minimal set are found to be naturally small, and even to have the right sign, to go with the observed coupling unification [14] (see Appendix).

The question is: can the minimal set of Higgs multiplets [see Eq. (14)] meet all the requirements listed above? Now $\mathbf{10}_H$ (even several $\mathbf{10}$'s) cannot meet the requirements of antisymmetry and $(B\text{-}L)$-dependence. Furthermore, a single $\mathbf{10}_H$ cannot generate CKM-mixings. This impasse disappears, however, as soon as one allows for not only cubic, but also effective non-renormalizable quartic couplings of the minimal set of Higgs fields with the fermions. These latter couplings could of course well arise through exchanges of superheavy states (e.g. those in the string tower) involving renormalizable couplings, and/or through quantum gravity.

Allowing for such cubic and quartic couplings and adopting the guideline (1) of hierarchical Yukawa couplings, as well as that of economy, we are led to suggest the following effective lagrangian for generating Dirac masses and mixings of the three families [14] (for a related but different pattern, involving a non-minimal Higgs system, see Ref. [41]).

$$\begin{aligned}\mathcal{L}_{\mathbf{Yuk}} = {} & h_{33}\,\mathbf{16}_3\,\mathbf{16}_3\,\mathbf{10}_H + [h_{23}\,\mathbf{16}_2\,\mathbf{16}_3\,\mathbf{10}_H + a_{23}\,\mathbf{16}_2\,\mathbf{16}_3\,\mathbf{10}_H\,\mathbf{45}_H/M \\ & + g_{23}\,\mathbf{16}_2\,\mathbf{16}_3\,\mathbf{16}_H\,\mathbf{16}_H/M] + \{a_{12}\,\mathbf{16}_1\,\mathbf{16}_2\,\mathbf{10}_H\,\mathbf{45}_H/M \\ & + g_{12}\,\mathbf{16}_1\,\mathbf{16}_2\,\mathbf{16}_H\,\mathbf{16}_H/M\}. \qquad (15)\end{aligned}$$

Here, M could plausibly be of order string scale. Note that a mass matrix having essentially the form of Eq. (11) results if the first term $h_{33}\langle\mathbf{10}_H\rangle$ is dominant. This ensures $m_b^0 \approx m_\tau^0$ and $m_t^0 \approx m^0(\nu_{\rm Dirac})$. Following the assumption of progressive hierarchy (equivalently appropriate flavor symmetries [6]), we presume that $h_{23} \sim h_{33}/10$, while h_{22} and h_{11}, which are not shown, are assumed to be progressively much smaller than h_{23}. Since $\langle\mathbf{45}_H\rangle \sim \langle\mathbf{16}_H\rangle \sim M_X$, while $M \sim M_{st} \sim 10M_X$, the terms $a_{23}\langle\mathbf{45}_H\rangle/M$ and $g_{23}\langle\mathbf{16}_H\rangle/M$ can quite plausibly be of order $h_{33}/10$, if $a_{23} \sim g_{23} \sim h_{33}$. By the assumption of hierarchy, we presume that $a_{12} \ll a_{23}$, and $g_{12} \ll g_{23}$

It is interesting to observe the symmetry properties of the a_{23} and g_{23}-terms. Although $\mathbf{10}_H \times \mathbf{45}_H = \mathbf{10} + \mathbf{120} + \mathbf{320}$, given that $\langle\mathbf{45}_H\rangle$ is along B–L, which is used to implement doublet-triplet splitting (see Appendix), only **120** in the decomposition contributes to the mass-matrices. This contribution is, however, antisymmetric in the family-index and, at the same time, proportional to B–L. *Thus the a_{23} term fulfills the requirements of both antisymmetry and (B–L)-dependence, simultaneously* [7] . With only h_{ij} and a_{ij}-terms, however, the up and down quark mass-matrices will be proportional to each other, which would yield $V_{CKM} = 1$. This is remedied by the g_{ij} coupling, because, the $\mathbf{16}_H$ can have a VEV not only along its SM singlet

[6] Although no explicit string solution with the hierarchy in all the Yukawa couplings in Eq. (15)—i.e. in h_{ij}, a_{ij} and g_{ij}—exists as yet, one can postulate flavor symmetries of the type alluded to (e.g. two abelian U(1) symmetries), which assign flavor charges not only to the fermion families and the Higgs multiplets, but also to a few (postulated) SM singlets that acquire VEVs of order M_X. The flavor symmetry-allowed effective couplings such as $\mathbf{16}_2\mathbf{16}_3\mathbf{10}_H\left\langle S\right\rangle/M$ would lead to $h_{23} \sim \left\langle S\right\rangle/M \sim 1/10$. One can verify that the full set of hierarchical couplings shown in Eq. (15) can in fact arise in the presence of two such U(1) symmetries. String theory (at least) offers the scope (as indicated by the solutions of Refs. [31] and [30]) for providing a rationale for the existence of such flavor symmetries, together with that of the SM singlets. For example, there exist solutions with the top Yukawa coupling being leading and others being hierarchical (as in Ref. [31]).

[7] The analog of $\mathbf{10}_H \cdot \mathbf{45}_H$ for the case of G(224) would be $\chi_H \equiv (2,2,1)_H \cdot (1,1,15)_H$. Although in general, the coupling of χ_H to the fermions need not be antisymmetric, for a string-derived G(224), the multiplet $(1,1,15)_H$ is most likely to arise from an underlying 45 of SO(10) (rather than 210); in this case, the couplings of χ_H must be antisymmetric like that of $\mathbf{10}_H \cdot \mathbf{45}_H$.

component (transforming as $\tilde{\overline{\nu}}_R$) which is of GUT-scale, but also along its electroweak doublet component—call it $\mathbf{16}_d$—of the electroweak scale. The latter can arise by the the mixing of $\mathbf{16}_d$ with the corresponding doublet (call it $\mathbf{10}_d$) in the $\mathbf{10}_H$. The MSSM doublet H_d, which is light, is then a mixture of $\mathbf{10}_d$ and $\mathbf{16}_d$, while the orthogonal combination is superheavy (see Appendix). Since $\langle\mathbf{16}_d\rangle$ contributes only to the down-flavor mass matrices, but not to the up-flavor, the g_{23} and g_{12} couplings generate non-trivial CKM-mixings. *We thus see that the minimal Higgs system (as shown in Eq. (14)) satisfies* a priori *all the qualitative requirements (1)–(4), including the condition of $V_{CKM} \neq 1$.* I now discuss that this system works well even quantitatively.

With the six effective Yukawa couplings shown in Eq. (15), the Dirac mass matrices of quarks and leptons of the three families at the unification scale take the form:

$$U = \begin{pmatrix} 0 & \epsilon' & 0 \\ -\epsilon' & 0 & \epsilon+\sigma \\ 0 & -\epsilon+\sigma & 1 \end{pmatrix} m_U,$$

$$D = \begin{pmatrix} 0 & \epsilon'+\eta' & 0 \\ -\epsilon'+\eta' & 0 & \epsilon+\eta \\ 0 & -\epsilon+\eta & 1 \end{pmatrix} m_D,$$

$$N = \begin{pmatrix} 0 & -3\epsilon' & 0 \\ 3\epsilon' & 0 & -3\epsilon+\sigma \\ 0 & 3\epsilon+\sigma & 1 \end{pmatrix} m_U,$$

$$L = \begin{pmatrix} 0 & -3\epsilon'+\eta' & 0 \\ 3\epsilon'+\eta' & 0 & -3\epsilon+\eta \\ 0 & 3\epsilon+\eta & 1 \end{pmatrix} m_D. \qquad (16)$$

Here the matrices are multiplied by left-handed fermion fields from the left and by anti–fermion fields from the right. (U, D) stand for the mass matrices of up and down quarks, while (N, L) are the Dirac mass matrices of the neutrinos and the charged leptons. The entries $1, \epsilon$,and σ arise respectively from the h_{33}, a_{23} and h_{23} terms in Eq. (15), while η entering into D and L receives contributions from both g_{23} and h_{23}; thus $\eta \neq \sigma$. Similarly η' and ϵ' arise from g_{12} and a_{12} terms respectively. Note the quark-lepton correlations between U and N as well as D and L arise because of $\mathrm{SU}(4)^C$, while the up-down correlations between U and D as well as N and L arise because of $\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$. Thus, these correlations emerge just because of the symmetry property of G(224). The relative factor of -3 between quarks and leptons involving the ϵ entry reflects the fact that $\langle \mathbf{45_H} \rangle$ is proportional to (B–L), while the antisymmetry in this entry arises from the group structure of SO(10), as explained above[7]. As we will see, this ϵ-entry helps to account for (a) the differences between m_s and m_μ, (b) that between m_d and m_e, and most important, (c) the suppression of V_{cb} *together with* the enhancement of the ν_μ-ν_τ oscillation angle.

The mass matrices in Eq. (16) contain 7 parameters [8]: ϵ, σ, η, $m_D = h_{33}\,\langle 10_d \rangle$, $m_U = h_{33}\,\langle 10_U \rangle$, η' and ϵ'. These may be determined by using, for example, the following input values: $m_t^{\mathrm{phys}} = 174$ GeV, $m_c(m_c) = 1.37$ GeV, $m_s(1\ \mathrm{GeV}) = 110$–$116$ MeV [42], $m_u(1\ \mathrm{GeV}) \approx 6$ MeV and the observed masses of e, μ and τ, which lead to (see Ref. [14], for details):

$$\sigma \simeq 0.110\,,\ \eta \simeq 0.151\,,\ \epsilon \simeq -0.095\,,\ |\eta'| \approx 4.4 \times 10^{-3} \ \text{and}\ \epsilon' \approx 2 \times 10^{-4}$$

$$m_U \simeq m_t(M_U) \simeq (100\text{-}120)\,\mathrm{GeV}\,,\ m_D \simeq m_b(M_U) \simeq 1.5\,\mathrm{GeV}\,. \qquad (17)$$

Here, I will assume, only for the sake of simplicity, as in Ref. [14], that the parameters are real.[9] Note that in accord with our general expectations discussed above, each of the parameters σ, η and ϵ are found to be of order 1/10, as opposed to being [10] $O(1)$ or $O(10^{-2})$, compared to the leading (3,3)-element in Eq. (16). Having determined these parameters, we are led to a total of five predictions involving only the quarks (those for the leptons are listed separately):

$$m_b^0 \approx m_\tau^0(1 - 8\epsilon^2)\,;\quad \text{thus}\ m_b(m_b) \simeq (4.6\text{-}4.9)\,\mathrm{GeV} \qquad (18)$$

[8] Of these, $m_U^0 \approx m_t^0$ can in fact be estimated to within 20% accuracy by either using the argument of radiative electroweak symmetry breaking, or some promising string solutions (see e.g. Ref. [31]).

[9] Babu and I have recently studied supersymmetric CP violation within the G(224)/SO(10) framework, by using precisely the fermion mass-matrices as in Eq. (16). We have observed [33] that complexification of the parameters can lead to observed CP violation, without upsetting in the least the success of Ref. [14] (i.e. of the fermion mass-matrices of Eq. (16)) in describing the masses and mixings of all fermions, including neutrinos. Even with complexification the relative signs and the approximate magnitudes of the real parts of the parameters must be the same as in Eq. (17), to retain the success.

[10] This is one characteristic difference between our work and that of Ref. [41], where the (2,3)-element is even bigger than the (3,3).

$$|V_{cb}| \simeq |\sigma - \eta| \approx \left| \sqrt{m_s/m_b} \left|\frac{\eta + \epsilon}{\eta - \epsilon}\right|^{1/2} - \sqrt{m_c/m_t} \left|\frac{\sigma + \epsilon}{\sigma - \epsilon}\right|^{1/2} \right| \simeq 0.045 \tag{19}$$

$$m_d\,(1\mathrm{GeV}) \simeq 8\,\mathrm{MeV} \tag{20}$$

$$\theta_C \simeq \left| \sqrt{m_d/m_s} - e^{i\phi}\sqrt{m_u/m_c} \right| \tag{21}$$

$$|V_{ub}/V_{cb}| \simeq \sqrt{m_u/m_c} \simeq 0.07\,. \tag{22}$$

In making these predictions, we have extrapolated the GUT-scale values down to low energies using $\alpha_3(m_Z) = 0.118$, a SUSY threshold of 500 GeV and $\tan\beta = 5$. The results depend weakly on these choices, assuming $\tan\beta \approx$ 2-30. Further, the Dirac masses and mixings of the neutrinos and the mixings of the charged leptons also get determined. We obtain:

$$m^D_{\nu_\tau}(M_U) \approx 100\text{-}120\,\mathrm{GeV};\; m^D_{\nu_\mu}(M_U) \simeq 8\,\mathrm{GeV}, \tag{23}$$

$$\theta^\ell_{\mu\tau} \approx -3\epsilon + \eta \approx \sqrt{m_\mu/m_\tau}\left|\frac{-3\epsilon + \eta}{3\epsilon + \eta}\right|^{1/2} \simeq 0.437 \tag{24}$$

$$m^D_{\nu_e} \simeq [9\epsilon'^2/(9\epsilon^2 - \sigma^2)]\,m_U \simeq 0.4\,\mathrm{MeV} \tag{25}$$

$$\theta^\ell_{e\mu} \simeq \left|\frac{\eta' - 3\epsilon'}{\eta' + 3\epsilon'}\right|^{1/2} \sqrt{m_e/m_\mu} \simeq 0.85\sqrt{m_e/m_\mu} \simeq 0.06 \tag{26}$$

$$\theta^\ell_{e\tau} \simeq \frac{1}{0.85}\sqrt{m_e/m_\tau}\,(m_\mu/m_\tau) \simeq 0.0012\,. \tag{27}$$

In evaluating $\theta^\ell_{e\mu}$, we have assumed ϵ' and η' to be relatively positive.

Given the bizarre pattern of quark and lepton masses and mixings, it seems remarkable that the simple and economical pattern of fermion mass-matrices, motivated in part by the assumption of flavor symmetries[6] which distinguish between the three families and in large part by the group theory of G(224)/SO(10), gives an overall fit to all of them [Eqs. (18) through (22)] which is good to within 10%. This includes the two successful predictions on m_b and V_{cb} [Eqs.(18) and (19)]. Note that in supersymmetric unified theories, the "observed" value of $m_b(m_b)$ and renormalization-group studies suggest that, for a wide range of the parameter $\tan\beta$, m_b^0 should in fact be about 10-20% *lower* than m_τ^0 [43]. This is neatly explained by the relation: $m_b^0 \approx m_\tau^0(1-8\epsilon^2)$ [Eq. (18)], where exact equality holds in the limit $\epsilon \to 0$ (due to SU(4)-color), while the decrease of m_b^0 compared to m_τ^0 by $8\epsilon^2 \sim 10\%$ is precisely because the off-diagonal ϵ-entry is proportional to B–L [see Eq. (16)].

Specially intriguing is the result on $V_{cb} \approx 0.045$ which compares well with the observed value of $\simeq 0.04$. The suppression of V_{cb}, compared to the value of 0.17 ± 0.06 obtained from Eq. (13), is now possible because the mass matrices [Eq. (16)] contain an antisymmetric component $\propto \epsilon$. That corrects the square-root formula $\theta_{sb} = \sqrt{m_s/m_b}$ [appropriate for symmetric matrices, see Eq. (11)] by the asymmetry factor $|(\eta+\epsilon)/(\eta-\epsilon)|^{1/2}$ [see Eq. (19)], and similarly for the angle θ_{ct}. This factor suppresses V_{cb} if η and ϵ have opposite signs. The interesting point is that, *the same feature necessarily enhances the corresponding mixing angle $\theta^\ell_{\mu\tau}$ in the leptonic sector*, since the asymmetry factor in this case is given by $[(-3\epsilon+\eta)/(3\epsilon+\eta)]^{1/2}$ [see Eq. (24)]. This enhancement of $\theta^\ell_{\mu\tau}$ helps to account for the nearly maximal oscillation angle observed at SuperK (as discussed below). This intriguing correlation between the mixing angles in the quark versus leptonic sectors—*that is suppression of one implying enhancement of the other*—has become possible only because of the ϵ-contribution, which is simultaneously antisymmetric and is proportional to B–L. That in turn becomes possible because of the group-property of SO(10) or a string-derived G(224)[7].

Taking stock, we see an impressive set of facts in favor of having B–L as a gauge symmetry and in fact for the full SU(4)-color-symmetry. These include: (i) the suppression of V_{cb}, together with the enhancement of $\theta^\ell_{\mu\tau}$, mentioned above; (ii) the successful relation $m_b^0 \approx m_\tau^0(1-8\epsilon^2)$; (iii) the usefulness again of the SU(4)-color-relation $m(\nu^\tau_{\text{Dirac}})^0 \approx m_t^0$ in accounting for $m(\nu^\tau_L)$ (see Section 4); (iv) the agreement of the relation $|m_s^0/m_\mu^0| = |(\epsilon^2-\eta^2)/(9\epsilon^2-\eta^2)|$ with the data, in that the ratio is naturally *less than* 1, if $\eta \sim \epsilon$ [The presence of $9\epsilon^2$ in the denominator is because the off-diagonal entry is proportional to B–L.]; and finally (v), the need for (B–L)—as a local symmetry, to implement baryogenesis via leptogenesis, as noted in Section 1.

Turning to neutrino masses, while all the entries in the Dirac mass matrix N are now fixed, to obtain the parameters for the light neutrinos, one needs to specify those of the Majorana mass matrix of the RH neutrinos ($\nu_R^{e,\mu,\tau}$). Guided by economy and the assumption of hierarchy, we consider the following pattern [14]:

$$M_\nu^R = \begin{pmatrix} x & 0 & z \\ 0 & 0 & y \\ z & y & 1 \end{pmatrix} M_R \,. \tag{28}$$

As discussed in Section 4, the magnitude of $M_R \approx$ (5-10) $\times\, 10^{14}$ GeV can quite plausibly be justified in the context of supersymmetric unification[5] [e.g. by using $M \approx M_{st} \approx 4 \times 10^{17}$ GeV in Eq. (8)]. To the same extent, the magnitude of $m(\nu_\tau) \approx$ (1/10-1/30) eV, which is consistent with the SuperK value, can also be anticipated by allowing for $\nu_\mu - \nu_\tau$ mixing [see Ref. [14]]. Thus there are effectively three new parameters: x, y, and z. Since there are six observables for the three light neutrinos, one can expect three predictions. These may be taken to be $\theta^{\rm osc}_{\nu_\mu\nu_\tau}$, m_{ν_τ} [see Eq. (10)], and for example $\theta^{\rm osc}_{\nu_e\nu_\mu}$.

Assuming successively hierarchical entries as for the Dirac mass matrices, we presume that $|y| \sim 1/10, |z| \leq |y|/10$ and $|x| \leq z^2$. Now given that $m(\nu_\tau) \sim 1/20$ eV [as estimated in Eq. (10)], the MSW solution for the solar neutrino puzzle [44] suggests that $m(\nu_\mu)/m(\nu_\tau) \approx$ 1/8-1/20. With hierarchical neutrino masses, the higher value of the mass-ratio (like 1/8) holds only for the large angle MSW solution (see below). With the mass-ratio being in the range of 1/8-1/20, one obtains: $|y| \approx$ (1/17 to 1/21), with y having the same sign as ϵ [see Eq. (17)]. This solution for y obtains only by assuming that y has a hierarchical value $O(1/10)$ rather than $O(1)$. Combining now with the mixing in the μ-τ sector determined above [see Eq. (24)], one can then determine the ν_μ-ν_τ oscillation angle. The two predictions of the model for the neutrino-system are then:

$$m(\nu_\tau) \;\approx\; (1/10\,\text{-}\,1/30)\,\text{eV} \tag{29}$$

$$\theta^{\rm osc}_{\nu_\mu\nu_\tau} \;\simeq\; \theta^\ell_{\mu\tau} - \theta^\nu_{\mu\tau} \;\simeq\; \left(0.437 \,+\, \sqrt{\frac{m_{\nu_2}}{m_{\nu_3}}}\right) . \tag{30}$$

Thus,

$$\sin^2 2\theta^{\rm osc}_{\nu_\mu\nu_\tau} = (0.99, 0.975, 0.92, 0.87) \tag{31}$$

for

$$m_{\nu_2}/m_{\nu_3} = (1/8, 1/10, 1/15, 1/20)\,. \tag{32}$$

Both of these predictions are extremely successful.[11]

[11] In writing Eq. (31), the small angle approximation exhibited in Eq. (30) is replaced by the more precise expression, given in Eq. (12) of Ref. [14], with the further understanding that $\sqrt{m_\mu/m_\tau}$ appearing in Eq. (12) (of Ref. [14]) is replaced by the μ-τ mixing angle $\approx$ 0.437.

Note the interesting point that the MSW solution, and the requirement that $|y|$ should have a natural hierarchical value (as mentioned above), lead to y having the same sign as ϵ. Now, that (it turns out) implies that the two contributions in Eq. (30) must *add* rather than subtract, leading to an *almost maximal oscillation angle* [14]. The other factor contributing to the enhancement of $\theta^{\mathrm{osc}}_{\nu_\mu\nu_\tau}$ is, of course, also the asymmetry-ratio which increases $|\theta^{\ell}_{\mu\tau}|$ from 0.25 to 0.437 [see Eq. (24)]. We see that one can derive rather plausibly a large ν_μ-ν_τ oscillation angle $\sin^2 2\theta^{\mathrm{osc}}_{\nu_\mu\nu_\tau} \geq 0.92$, together with an understanding of hierarchical masses and mixings of the quarks and the charged leptons, while maintaining a large hierarchy in the seesaw derived neutrino masses ($m_{\nu_2}/m_{\nu_3} = 1/8$-$1/15$), all within a unified framework including both quarks and leptons. In the example exhibited here, the mixing angles for the mass eigenstates of neither the neutrinos nor the charged leptons are really large, in that $\theta^{\ell}_{\mu\tau} \simeq 0.437 \simeq 23^\circ$ and $\theta^{\nu}_{\mu\tau} \simeq (0.22\text{-}0.35) \approx (13\text{-}20.5)^\circ$, *yet the oscillation angle obtained by combining the two is near-maximal.* This contrasts with most works in the literature in which a large oscillation angle is obtained either entirely from the neutrino sector (with nearly degenerate neutrinos) or almost entirely from the charged lepton sector.

Small Versus Large Angle MSW Solutions

In considerations of ν_e-ν_μ and ν_e-ν_τ oscillation angles, tiny *intrinsic* non-diagonal Majorana masses $\sim 10^{-3}$ eV of the LH neutrinos leading to $\nu^e_L\nu^\mu_L$ and $\nu^e_L\nu^\tau_L$-mixings, which can far exceed those induced by the standard see-saw mechanism, can be rather important, especially for ν_e-ν_μ mixing. As explained below, such intrinsic masses can arise quite naturally through higher dimensional operators and can lead to the large angle MSW solution of the solar neutrino puzzle.

Let us first ignore the intrinsic Majorana masses of the LH neutrinos and include only those that arise through the standard see-saw mechanism, involving the superheavy Majorana masses of the RH neutrinos, with a pattern given, for example, by Eq. (28). Note that, while $M_R \approx (5\text{-}15) \times 10^{14}$ GeV and $y \approx -1/20$ are better determined, the parameters x and z can not be obtained reliably at present because very little is known about observables involving ν_e. Taking, for concreteness, $m_{\nu_e} \approx (10^{-5}\text{-}10^{-4})$ (1 to few)) eV and $\theta^{\mathrm{osc}}_{e\tau} \approx \theta^{\ell}_{e\tau} - \theta^{\nu}_{e\tau} \approx 10^{-3} \pm 0.03$ as inputs, we obtain: $z \sim (1\text{-}5) \times 10^{-3}$ and $x \sim (1 \text{ to few})(10^{-6}\text{-}10^{-5})$, in accord with the guidelines of $|z| \sim |y|/10$ and

$|x| \sim z^2$. This in turn yields: $\theta_{e\mu}^{\rm osc} \approx \theta_{e\mu}^{\ell} - \theta_{e\mu}^{\nu} \approx 0.06 \pm 0.015$. Note that the mass of $m_{\nu_\mu} \sim 3 \times 10^{-3}$ eV, that follows from a natural hierarchical value for $y \sim -(1/20)$, and $\theta_{e\mu}$ as above, go well with the small angle MSW explanation of the solar neutrino puzzle. In short the framework presented so far, that neglects intrinsic Majorana masses of the LH neutrinos altogether, generically tends to yield the small angle MSW solution.

As alluded to above, we now observe that small intrinsic non-seesaw masses of the LH neutrinos $\sim 10^{-3}$ eV, which could mix ν_{eL} and $\nu_{\mu L}$, can, however, arise quite naturally through higher dimensional operators in the superpotential of the form [12]: $W \supset \kappa_{12}\mathbf{16}_1\mathbf{16}_2\mathbf{16}_H\mathbf{16}_H\mathbf{10}_H\mathbf{10}_H/M_{\rm GUT}^3$. One can verify that such a term would lead to an *intrinsic* Majorana mixing mass term of the form $m_{12}^{(0)}\nu_L^e\nu_L^\mu$, with a strength given by $m_{12}^{(0)} \approx \kappa_{12}(\langle\mathbf{16}_H\rangle/M_{\rm GUT})^2(175 \text{ GeV})^2/M_{\rm GUT} \approx (1.5\text{-}6) \times 10^{-3} eV$, where we have put $\langle\mathbf{16}_H\rangle \approx (1\text{-}2)M_{\rm GUT}$ and $M_{\rm GUT} \approx 2\times10^{16}$ GeV. Such an intrinsic Majorana mixing mass $\sim 10^{-3}$ eV, though small, is still much larger than what one would get for the corresponding term from the standard see-saw mechanism. Now, as discussed above, the diagonal $(\nu_L^\mu\nu_L^\mu)$ mass-term, arising from the standard see-saw mechanism can naturally be of order $(3\text{-}8)\times10^{-3}$ eV (for $|y| \approx 1/20$ to $1/15$, say). In addition, the intrinsic contribution of the type mentioned above may in general also contribute to the diagonal $(\nu_L^\mu\nu_L^\mu)$ mass (depending upon flavor symmetries) which can be (few)$\times10^{-3}$ eV. Thus, taking the net values of $m_{22} \approx (6\text{-}7) \times 10^{-3}$ eV (say), $m_{12}^{(0)} \approx (3\text{-}4) \times 10^{-3}$ eV, and $m_{11}^{(0)} \lesssim (1\text{-}2) \times 10^{-3}$ eV, which are all very plausible, we obtain $m_{\nu_\mu} \approx (6\text{-}7) \times 10^{-3}$ eV, $m_{\nu_e} \sim 1 \times 10^{-3}$ eV, so that $\Delta m_{12}^2 \approx (3.6\text{-}5) \times 10^{-5}$ eV^2, and $\sin^2 2\theta_{12}^{\rm osc} \approx 0.6\text{-}0.7$. This goes well with the large angle MSW solution of the solar neutrino puzzle, which is now favored over the small angle solution by the SuperK data [45].

In summary, the intrinsic non-seesaw contribution to the Majorana masses of the LH neutrinos quite plausibly has the right magnitude for ν_e-ν_μ mixing, so as to lead to the rather large oscillation angle as mentioned above, in accord with the data. In contrast to the case of the ν_μ-ν_τ oscillation angle, however, given the smallness of the entries involving the first two families, the relatively large angle solution for $\nu_e - \nu^\mu$ oscillation may not be regarded as a firm prediction of the SO(10)/G(224)-framework presented here. It is

[12] Such a term can be induced in the presence of, for example, a singlet S and a ten-plet (denoted by $\hat{\mathbf{10}}$), both having GUT-scale masses, and possessing renormalizable couplings of the form $a_i\mathbf{16}_i\mathbf{16}_H\widehat{\mathbf{10}}$, $b\widehat{\mathbf{10}}\,\mathbf{10}_HS$, M_SSS and $\widehat{M}\widehat{\mathbf{10}}^2$. In this case, $\kappa_{12}/M_{\rm GUT}^3 = a_1a_2b/(\widehat{M}^2M_S)$.

nevertheless a very reasonable possibility.

It is worth noting that although the superheavy Majorana masses of the RH neutrinos cannot be observed directly, they can be of cosmological significance. The pattern given above and the arguments given in Section 3 and in this section suggests that $M(\nu_R^\tau) \approx$ (5-15) $\times 10^{14}$ GeV, $M(\nu_R^\mu) \approx$ (1-4) $\times 10^{12}$ GeV (for $|y| \approx 1/20$); and $M(\nu_R^e) \sim$ (1/2-10) $\times 10^9$ GeV (for $x \sim (1/2\text{-}10)10^{-6} > z^2$). A mass of $\nu_R^e \sim 10^9$ GeV is of the right magnitude for producing ν_R^e following reheating and inducing lepton asymmetry in ν_R^e decay into $H^0 + \nu_L^i$, that is subsequently converted into baryon asymmetry by the electroweak sphalerons [16, 17].

In summary, we have proposed an economical and predictive pattern for the Dirac mass matrices, within the SO(10)/G(224)-framework, which is remarkably successful in describing the observed masses and mixings of *all* the quarks and charged leptons. It leads to five predictions for just the quark- system, all of which agree with observation to within 10%. The same pattern, supplemented with a similar structure for the Majorana mass matrix, accounts for both the nearly-maximal ν_μ-ν_τ oscillation angle and a (mass)^2-difference $\Delta m^2(\nu_\mu\nu_\tau) \sim (1/20 \text{ eV})^2$, suggested by the SuperK data. Given this degree of success, it makes good sense to study proton decay concretely within this SO(10)/G(224)-framework. The results of this study [14, 18] are presented in the next section, together with an update.

Before turning to proton decay, it is worth noting that much of our discussion of fermion masses and mixings, including those of the neutrinos, is essentially unaltered if we go to the limit $\epsilon' \to 0$ of Eq. (28). This limit clearly involves:

$$m_u = 0\,, \;\; \theta_C \simeq \sqrt{m_d/m_s}\,, \;\; m_{\nu_e} = 0\,, \;\; \theta^\nu_{e\mu} = \theta^\nu_{e\tau} = 0$$

$$|V_{ub}| \simeq \sqrt{\frac{\eta - \epsilon}{\eta + \epsilon}}\sqrt{m_d/m_b}\,(m_s/m_b) \simeq (2.1)(0.039)(0.023) \simeq 0.0019\,. \quad (33)$$

All other predictions remain unaltered. Now, among the observed quantities in the list above, $\theta_C \simeq \sqrt{m_d/m_s}$ is a good result. Considering that $m_u/m_t \approx 10^{-5}$, $m_u = 0$ is also a pretty good result. There are of course plausible small corrections which could arise through Planck scale physics; these could induce a small value for m_u through the (1,1)-entry $\delta \approx 10^{-5}$. For considerations of proton decay, it is worth distinguishing between these two *extreme* variants which we will refer to as cases I and II respectively.

$$\text{Case I}: \quad \epsilon' \approx 2 \times 10^{-4}\,, \;\; \delta = 0$$

$$\text{Case II:} \qquad \delta \approx 10^{-5}\,, \;\; \epsilon' = 0\,. \tag{34}$$

It is worth noting that the observed value of $|V_{ub}| \approx 0.003$ favors a non-zero value of ϵ' ($\approx$ (1-2) $\times 10^{-4}$). Thus, in reality, ϵ' may not be zero, but it may lie in between the two extreme values listed above. In this case, the predicted proton lifetime for the standard $d = 5$ operators would be intermediate between those for the two cases, presented in Section 6.

6 Expectations for Proton Decay in Supersymmetric Unified Theories

6.1 Preliminaries

Turning to the main purpose of this talk, I present now the reason why the unification framework based on SUSY SO(10) or G(224), together with the understanding of fermion masses and mixings discussed above, strongly suggest that proton decay should be imminent.

Recall that supersymmetric unified theories (GUTs) introduce two new features to proton decay: (i) First, by raising M_X to a higher value of about 2×10^{16} GeV (contrast with the non-supersymmetric case of nearly 3×10^{14} GeV), they strongly suppress the gauge-boson-mediated $d = 6$ proton decay operators, for which $e^+\pi^0$ would have been the dominant mode (for this case, one typically obtains: $\Gamma^{-1}(p \to e^+\pi^0)|_{d=6} \approx 10^{35\pm1}$ years). (ii) Second, they generate $d = 5$ proton decay operators [19] of the form $Q_iQ_jQ_kQ_l/M$ in the superpotential, through the exchange of color triplet Higginos, which are the GUT partners of the standard Higgs(ino) doublets, such as those in the $\mathbf{5} + \overline{\mathbf{5}}$ of SU(5) or the 10 of SO(10). Assuming that a suitable doublet-triplet splitting mechanism provides heavy GUT-scale masses to these color triplets and at the same time light masses to the doublets (see e.g, the Appendix), these "standard" $d = 5$ operators, suppressed by just one power of the heavy mass and the small Yukawa couplings, are found to provide the dominant mechanism for proton decay in supersymmetric GUT [46, 47, 48, 49, 50].

Now, owing to (a) Bose symmetry of the superfields in $QQQL/M$, (b) color antisymmetry, and especially (c) the hierarchical Yukawa couplings of the Higgs doublets, it turns out that these standard $d = 5$ operators lead to dominant $\overline{\nu}K^+$ and comparable $\overline{\nu}\pi^+$ modes, but in all cases to highly suppressed $e^+\pi^0$, e^+K^0 and even μ^+K^0 modes. For instance, for minimal SUSY SU(5), one obtains (with $\tan\beta \leq 20$, say):

$$[\Gamma(\mu^+K^0)/\Gamma(\overline{\nu}K^+)]^{SU(5)}_{std} \sim [m_u/(m_c \sin^2\theta_c)]^2\, R \approx 10^{-3}\,, \tag{35}$$

where $R \approx 0.1$ is the ratio of the relevant $|\text{matrix element}|^2 \times$(phase space), for the two modes.

It was recently pointed out that in SUSY unified theories based on SO(10) or G(224), which assign heavy Majorana masses to the RH neutrinos, there exists a new set of color triplets and thereby very likely a *new source* of $d = 5$ proton decay operators [20]. For instance, in the context of the minimal set of Higgs multiplets [13] $\{\mathbf{45}_H, \mathbf{16}_H, \overline{\mathbf{16}}_H$ and $\mathbf{10}_H\}$ (see Section 5), these new $d = 5$ operators arise by combining three effective couplings introduced before:—i.e., (a) the couplings $f_{ij}\mathbf{16}_i\mathbf{16}_j\overline{\mathbf{16}}_H\overline{\mathbf{16}}_H/M$ [see Eq. (7)] that are required to assign Majorana masses to the RH neutrinos, (b) the couplings $g_{ij}\mathbf{16}_i\mathbf{16}_j\mathbf{16}_H\mathbf{16}_H/M$, which are needed to generate non-trivial CKM mixings [see Eq. (15)], and (c) the mass term $M_{16}\mathbf{16}_H\overline{\mathbf{16}}_H$. For the f_{ij} couplings, there are two possible SO(10)-contractions (leading to a 45 or a 1) for the pair $\mathbf{16}_i\overline{\mathbf{16}}_H$, both of which contribute to the Majorana masses of the RH neutrinos, but only the non-singlet contraction (leading to **45**), would contribute to $d = 5$ proton decay operator. In the presence of non-perturbative quantum gravity, one would in general expect the two contractions to have comparable strength. Furthermore, the couplings of **45′s** lying in the string-tower or possibly below the string-scale, and likewise of singlets, to the $\mathbf{16}_i \cdot \overline{\mathbf{16}}_H$-pair, would respectively generate the two contractions. It thus seems most likely that both contractions would be present, having comparable strength. Allowing for a difference between the relevant projection factors for ν_R masses versus proton decay, and also for the fact that both contractions contribute to the former, but only the non-singlet one (i.e. 45) to the latter, we would set the relevant f_{ij} coupling for proton decay to be $(f_{ij})_p \equiv (f_{ij})_\nu \cdot K$, where $(f_{ij})_\nu$ defined in Section 4 directly yields ν_R -masses [see Eq. (8)]; and K is a relative factor, which generically is expected to be of order unity. [14] As a plausible range, we will take $K \approx 1/5$ to 2 (say). In the presence of the non-singlet contraction, the color-triplet Higginos in $\overline{\mathbf{16}}_H$ and $\mathbf{16}_H$ of mass M_{16} can be exchanged between $\tilde{q}_i q_j$ and $\tilde{q}_k q_l$-pairs (correspondingly, for G(224), the color triplets would arise from $(1, 2, 4)_H$ and $(1, 2, \overline{4})_H$). This exchange generates a new set of $d = 5$ operators in the

[13] The origin of the new $d = 5$ operators in the context of other Higgs multiplets, in particular in the cases where $\mathbf{126}_H$ and $\overline{\mathbf{126}}_H$ are used to break B–L, has been discussed in Ref. [20].

[14] For the special case of $K = 0$ (which would arise if only the singlet-contraction of $(\mathbf{16}_i \cdot \overline{\mathbf{16}}_H)$ contributes), the new $d = 5$ operators shown in Eq. (36) would not, of course, contribute to proton decay.

superpotential of the form

$$W_{\rm new} \propto (f_{ij})_\nu \, g_{kl} K \, (\mathbf{16}_i \, \mathbf{16}_j) \, (\mathbf{16}_k \, \mathbf{16}_l) \, \langle \overline{\mathbf{16}}_H \rangle \, \langle \mathbf{16}_H \rangle / M^2 \; \times (1/M_{16}), \quad (36)$$

which induce proton decay. Note that these operators depend, through the couplings f_{ij} and g_{kl}, both on the Majorana and on the Dirac masses of the respective fermions. *This is why within SUSY SO(10) or G(224), if the generic case of $K \neq 0$ holds, proton decay gets intimately linked to the masses and mixings of all fermions, including neutrinos.*

6.2 Framework for Calculating Proton Decay Rate

To establish notations, consider the case of minimal SUSY SU(5) and, as an example, the process $\tilde{c}\tilde{d} \to \bar{s}\bar{\nu}_\mu$, which induces $p \to \overline{\nu}_\mu K^+$. Let the strength of the corresponding $d = 5$ operator, multiplied by the product of the CKM mixing elements entering into wino-exchange vertices, (which in this case is $\sin\theta_C \cos\theta_C$) be denoted by $\widehat{A}$. Thus (putting $\cos\theta_C = 1$), one obtains:

$$\begin{aligned} \widehat{A}_{\tilde{c}\tilde{d}}(SU(5)) \; &= \; (h^u_{22} \, h^d_{12}/M_{H_C}) \, \sin\theta_c \\ &\simeq \; (m_c m_s \, \sin^2\theta_C/v_u^2) \, (\tan\beta/M_{H_C}) \\ &\simeq \; (1.9 \times 10^{-8}) \, (\tan\beta/M_{H_C}) \qquad\qquad\qquad (37) \\ &\approx \; (2 \times 10^{-24} \, \mathrm{GeV}^{-1}) \, (\tan\beta/2) \, (2 \times 10^{16} \, \mathrm{GeV}/M_{H_C}), \end{aligned}$$

where $\tan\beta \equiv v_u/v_d$, and we have put $v_u = 174$ GeV and the fermion masses extrapolated to the unification-scale—i.e. $m_c \simeq 300$ MeV and $m_s \simeq 40$ MeV. The amplitude for the associated four-fermion process $dus \to \overline{\nu}_\mu$ is given by:

$$A_5(dus \to \overline{\nu}_\mu) \; = \; \widehat{A}_{\tilde{c}\tilde{d}} \times (2f) \qquad (38)$$

where f is the loop-factor associated with wino-dressing. Assuming $m_{\tilde{w}} \ll m_{\tilde{q}} \sim m_{\tilde{l}}$, one gets: $f \simeq (m_{\tilde{w}}/m_{\tilde{q}}^2)(\alpha_2/4\pi)$. Using the amplitude for $(du)(s\nu_\ell)$, as in Eq. (38), ($\ell = \mu$ or τ), and the recently obtained matrix element and renormalization effects (see below), one then obtains [48, 49, 50, 14, 18]:

$$\begin{aligned} \Gamma^{-1}(p \to \overline{\nu}_\tau K^+) \;\; &\approx \;\; (0.15 \times 10^{31}) \, \mathrm{years} \times (0.32/A_L)^2 \qquad (39) \\ &\times \left(\frac{0.93}{A_S}\right)^2 \left[\frac{0.014 \, \mathrm{GeV}^3}{\beta_H}\right]^2 \left[\frac{(1/6)}{(m_{\tilde{W}}/m_{\tilde{q}})}\right]^2 \\ &\left[\frac{m_{\tilde{q}}}{1.2 \, \mathrm{TeV}}\right]^2 \left[\frac{2 \times 10^{-24} \, \mathrm{GeV}^{-1}}{\widehat{A}(\overline{\nu})}\right]^2 . \end{aligned}$$

Here β_H denotes the hadronic matrix element defined by $\beta_H u_L(\vec{k}) \equiv \epsilon_{\alpha\beta\gamma}\left\langle 0|(d_L^\alpha u_L^\beta)u_L^\gamma|p,\vec{k}\right\rangle$. While the range β_H = (0.003-0.03) GeV^3 has been used in the past [49], given that one lattice calculation yields $\beta_H = (5.6 \pm 0.5) \times 10^{-3}$ GeV^3 [51], and a recent improved calculation yields $\beta_H \approx 0.014$ GeV^3 [52] (whose systematic errors that may arise from scaling violations and quenching are hard to estimate [52]), we will take as a conservative, but plausible, range for β_H to be given by (0.014 GeV^3)$(1/2-2)$. (Compare this with the range for $\beta_H = (0.006$ GeV$^3)(1/2-2)$ as used in Ref. [14]). A_S denotes the short-distance renormalization effect for the $d = 5$ operator which arises owing to extrapolation between the GUT and the SUSY-breaking scales [47, 49, 53]. The average value of $A_S = 0.67$, given in Ref. [49] for $m_t = 100$ GeV, has been used in most early estimates. For $m_t = 175$ GeV, one would, however, have $A_S \approx 0.93$ to 1.2 [53]. Conservatively, I would use $A_S = 0.93$; this would enhance the rate by a factor of two compared with previous estimates. A_L denotes the long-distance renormalization effect of the $d = 6$ operator due to QCD interaction that arises due to extrapolation between the SUSY breaking scale and 1 GeV [47]. Using the two-loop expression for A_L [54], together with the two-loop value for α_3, Babu and I find: $A_L \approx 0.32$, in contrast to $A_L \approx 0.22$, used in previous works[15]. In what follows, I would use $A_L \approx 0.32$. This by itself would also increase the rate by a factor of $(0.32/0.22)^2 \approx$2, compared to the previous estimates [47, 48, 49, 50, 14, 18]. Including the enhancements in both A_S and A_L, we thus see that the net increase in the proton decay rate solely due to new evaluation of renormalization effects is nearly a factor of four, compared to the previous estimates (including that in Ref. [14]).

Note that the familiar factors that appear in the expression for proton lifetime—i.e., M_{H_C}, $(1+y_{tc})$ representing the interference between the $\tilde{t}$ and $\tilde{c}$ contributions, and $\tan\beta$ (see e.g. Ref. [49] and discussion in the Appendix of Ref. [14])—are all effectively contained in $\widehat{A}(\overline{\nu})$. In Ref. [14], guided by the demand of naturalness (i.e. absence of excessive fine tuning) in obtaining the Higgs boson mass, squark masses were assumed to lie in the range of 1 TeV$(1/\sqrt{2} - \sqrt{2})$, so that $m_{\tilde{q}} \lesssim 1.4$TeV. Recent work, based on the notion

[15] In most previous works starting with Ref. [47] through [50], as well as in Refs. [14] and [18], the one-loop value of A_L was taken to be 0.22. It was, however, noted in Refs. [54] and [55] that there is a numerical error in the evaluation of the one-loop expression for A_L [47], and that the correct value for $A_L(\text{one}-\text{loop}) \approx 0.43$ (this remained unnoticed by most authors). The two-loop value for A_L (as stated above) is nearly 0.32, which is lower than 0.43 but higher that the previously used value of 0.22.

of focus point supersymmetry however suggests that squarks may be considerably heavier without conflicting with the demands of naturalness [56]. In the interest of obtaining a conservative upper limit on proton lifetime, we will therefore allow squark masses to be as heavy as about 2.4 TeV and as light as perhaps 600 GeV. [16]

Allowing for plausible and rather generous uncertainties in the matrix element and the spectrum we take:

$$\beta_H \;=\; (0.014\,\mathrm{GeV}^3)\,(1/2\text{-}2)$$

$$m_{\tilde{w}}/m_{\tilde{q}} \;=\; 1/6\,(1/2\text{-}2)\,, \quad \text{and} \quad m_{\tilde{q}} \;\approx\; m_{\tilde{\ell}} \;\approx\; 1.2\,\mathrm{TeV}\,(1/2\text{-}2)\,. \qquad (40)$$

Using Eqs. (39–40), we get:

$$\Gamma^{-1}(p \to \overline{\nu}_\tau K^+) \approx (0.15{\times}10^{31}\,\text{years})\,[\,2{\times}10^{-24}\,\mathrm{GeV}^{-1}/\widehat{A}(\overline{\nu}_\ell)\,]^2{\times}\{64\text{-}1/64\,\}\,. \qquad (41)$$

Note that the curly bracket would acquire its upper-end value of 64, which would serve towards maximizing proton lifetime, only provided all the uncertainties in Eq. (41) are stretched to the extreme so that $\beta_H = 0.007$ GeV^3, $m_{\tilde{W}}/m_{\tilde{q}} \approx 1/12$ and $m_{\tilde{q}} \approx 2.4$ TeV. This relation, as well as Eq. (39) are general, depending only on $\widehat{A}(\overline{\nu}_\ell)$ and on the range of parameters given in Eq. (40). They can thus be used for both SU(5) and SO(10).

[16] We remark that if the recently reported (g-2)-anomaly for the muon [57], together with reevaluation of the contribution from light by light-scattering [58], is attributed to supersymmetry [59], one would need to have extremely light s-fermions [i.e. $m_{\tilde{l}} \approx 200$ - 400 GeV (say) and correspondingly, for promising mechanisms of SUSY-breaking, $m_{\tilde{q}} \lesssim$ 300 – 600 GeV (say)], and simultaneously relatively large $\tan\beta(\approx$ 6-24). However, not worrying about grand unification, such light s-fermions, together with large or very large $\tan\beta$ would typically be in gross conflict with the limits on the edm's of the neutron and the electron, unless one can explain naturally the occurrence of minuscule phases ($\lesssim 1/200$ to 1/500) and/or large cancellation. Thus, if the $(g-2)_\mu$-anomaly turns out to be real, it may well find a non-supersymmetric explanation, in accord with the edm-constraints which ordinarily seem to suggest that squarks are (at least) moderately heavy ($m_{\tilde{q}} \gtrsim 0.6-1$ TeV, say), and $\tan\beta$ is not too large ($\lesssim$ 3 to 10, say). We mention in passing that the extra vector—like matter—specially a $16 + \overline{16}$ of SO(10)—as proposed in the so-called extended supersymmetric standard model (ESSM) [21, 60], with the heavy lepton mass being of order 200 GeV, can provide such an explanation [61]. Motivations for the case of ESSM, based on the need for (a) removing the mismatch between MSSM and string unification scales, and (b) dilaton-stabilization, have been noted in Ref. [21]. Since ESSM is an interesting and viable variant of MSSM, and would have important implications for proton decay, we will present the results for expected proton decay rates for the cases of both MSSM and ESSM in the discussion to follow.

The experimental lower limit on the inverse rate for the $\bar{\nu}K^+$ modes is given by Ref. [62],

$$\left[\sum_{\ell} \Gamma(p \to \overline{\nu}_\ell K^+)\right]^{-1}_{\rm expt} \geq 1.9 \times 10^{33}\,\text{years}\,. \qquad (42)$$

Allowing for all the uncertainties to stretch in the same direction (in this case, the curly bracket = 64), and assuming that just one neutrino flavor (e.g. ν_μ for SU(5)) dominates, the observed limit (Eq. (42)) provides an upper bound on the amplitude[17]:

$$\widehat{A}(\overline{\nu}_\ell) \, \leq \, 0.46 \times 10^{-24}\,\text{GeV}^{-1} \qquad (43)$$

which holds for both SU(5) and SO(10). Recent theoretical analyses based on LEP-limit on Higgs mass ($\gtrsim$ 114 GeV), together with certain assumptions about MSSM parameters (as in CMSSM) and/or constraint from muon g-2 anomaly [57] suggest that $\tan\beta \gtrsim 3$ to 5 [63]. In the interest of getting a conservative upper limit on proton lifetime, we will therefore use, as a conservative lower limit, $\tan\beta \geq 3$. We will however exhibit relevant results often as a function of $\tan\beta$ and exhibit proton lifetimes corresponding to higher values of $\tan\beta$ as well. For minimal SU(5), using Eqs. (37) and (43) and, conservatively $\tan\beta \geq 3$, one obtains a lower limit on M_{HC} given by:

$$M_{HC} \, \geq \, 13 \times 10^{16}\,\text{GeV} \;\; (\text{SUSY SU(5)})\;. \qquad (44)$$

At the same time, gauge coupling unification in SUSY SU(5) strongly suggests $M_{HC} \leq (1/2\text{-}1) \times 10^{16}$ GeV. (See Ref. [64] where an even more stringent upper bound on M_{HC} is suggested.) Thus we already see a conflict, in the case of minimal SUSY SU(5), between the experimental limit on proton lifetime on the one hand, and coupling unification and constraint on $\tan\beta$ on the other hand. To see this conflict another way, if we keep $M_{HC} \leq 10^{16}$ GeV (for the sake of coupling unification) we obtain from Eq. (37): $\widehat{A}(\text{SU(5)}) \geq 5.7 \times 10^{-24}\,\text{GeV}^{-1}(\tan\beta/3)$. Using Eq. (41), this in turn implies that

$$\Gamma^{-1}(p \to \overline{\nu}K^+) \, \leq \, 1.2 \times 10^{31}\,\text{years} \times (3/\tan\beta)^2 \quad (\text{SUSY SU(5)})\;. \qquad (45)$$

For $\tan\beta \geq 3$, a lifetime of 1.2×10^{31} years is thus a most conservative upper limit. In practice, it is unlikely that all the uncertainties, including these

[17] If there are sub-dominant $\overline{\nu}_i K^+$ modes with branching ratio R, the right side of Eq. (43) should be divided by $\sqrt{1+R}$.

in M_{HC} and $\tan\beta$, would stretch in the same direction to nearly extreme values so as to prolong proton lifetime. Given the experimental lower limit [Eq. (42)], we see that minimal SUSY SU(5) is already excluded by a large margin by proton decay-searches. This is in full accord with the conclusion reached by other authors (see especially Ref. [64]). We have of course noted in Section 4 that SUSY SU(5) does not go well with neutrino oscillations observed at SuperK.

Now, to discuss proton decay in the context of supersymmetric SO(10), it is necessary to discuss first the mechanism for doublet-triplet splitting. Details of this discussion may be found in Ref. [14]. A synopsis is presented in the Appendix.

6.3 Proton Decay in Supersymmetric SO(10)

The calculation of the amplitudes $\widehat{A}_{\text{std}}$ and $\widehat{A}_{\text{new}}$ for the standard and the new operators for the SO(10) model, are given in detail in Ref. [14]. Here, I will present only the results. It is found that the four amplitudes $\widehat{A}_{\text{std}}(\overline{\nu}_\tau K^+)$, $\widehat{A}_{\text{std}}(\overline{\nu}_\mu K^+)$, $\widehat{A}_{\text{new}}(\overline{\nu}_\tau K^+)$ and $\widehat{A}_{\text{new}}(\overline{\nu}_\mu K^+)$ are in fact very comparable to each other, within about a factor of two to five, either way. Since there is no reason to expect a near cancellation between the standard and the new operators, especially for both $\overline{\nu}_\tau K^+$ and $\overline{\nu}_\mu K^+$ modes, we expect the net amplitude (standard + new) to be in the range exhibited by either one. Following Ref. [14], I therefore present the contributions from the standard and the new operators separately.

One important consequence of the doublet-triplet splitting mechanism for SO(10) outlined briefly in the appendix and in more detail in Ref. [14] is that the standard $d = 5$ proton decay operators become inversely proportional to $M_{\text{eff}} \equiv [\lambda\langle \mathbf{45}_H\rangle]^2/\ M_{10'} \sim M_X^2/M_{10'}$, rather than to M_{H_C}. Here, $M_{10'}$ represents the mass of $\mathbf{10}'_H$, that enters into the D-T splitting mechanism through effective coupling $\lambda\mathbf{10}_H\mathbf{45}_H\mathbf{10}'_H$ in the superpotential [see Appendix, Eq. (A1)]. As noted in Ref. [14], $M_{10'}$ can be naturally suppressed (due to flavor symmetries) compared to M_X, and thus M_{eff} correspondingly larger than M_X by even one to three orders of magnitude. It should be stressed that M_{eff} does not represent the physical masses of the color triplets or of the other particles in the theory. It is simply a parameter of order $M_X^2/M_{10'}$. *Thus values of M_{eff}, close to or even exceeding the Planck scale, do not in any way imply large corrections from quantum gravity.* Now accompanying the suppression due to M_{eff}, the standard proton decay amplitudes for SO(10) possess an intrinsic enhancement as well, compared to those for SU(5), owing primarily due to differences in their Yukawa couplings for the up sector (see Appendix C of Ref. [14]). As a result of this enhancement, combined with the suppression due to higher values of M_{eff}, a typical standard $d = 5$ amplitude for SO(10) is given by (see Appendix C of Ref. [14])

$$\widehat{A}(\bar{\nu}_\mu K^+)^{SO(10)}_{std} \approx (h^2_{33}/M_{\text{eff}})(2 \times 10^{-5}),$$

which should be compared with $\widehat{A}(\bar{\nu}_\mu K^+)^{SU(5)}_{std} \approx (1.9 \times 10^{-8})(\tan\beta/M_{H_C})$ [see Eq. (37)]. Note, taking $h^2_{33} \approx 1/4$, the ratio of a typical SO(10) over SU(5) amplitude is given by $(M_{H_c}/M_{\text{eff}})(88)(3/\tan\beta)$. Thus the enhancement by a factor of about 88 (for $\tan\beta = 3$), of the SO(10) compared to the SU(5) amplitude, is compensated in part by the suppression that arises from M_{eff} being larger than M_{H_c}.

In addition, note that in contrast to the case of SU(5), the SO(10) amplitude does not depend *explicitly* on $\tan\beta$. The reason is this: if the fermions acquire masses only through the $\mathbf{10}_H$ in SO(10), as is well known, the up and down quark Yukawa couplings will be equal. By itself, it would lead to a large value of $\tan\beta = m_t/m_b \approx 60$ and thereby to a large enhancement in proton decay amplitude. Furthermore, it would also lead to the bad relations: $m_c/m_s = m_t/m_b$ and $V_{CKM} = 1$. However, in the presence of additional Higgs multiplets, in particular with the mixing of $(\mathbf{16}_H)_d$ with $\mathbf{10}_H$ (see Appendix and Section 5), (a) $\tan\beta$ can get lowered to values like 3-20, (b) fermion masses get contributions from both $\langle\mathbf{16}_H\rangle_d$ and $\langle\mathbf{10}_H\rangle$, which correct all the bad relations stated above, and simultaneously (c) the explicit dependence of $\widehat{A}$ on $\tan\beta$ disappears. It reappears, however, through restriction on threshold corrections, discussed below.

Although M_{eff} can far exceed M_X, it still gets bounded from above by demanding that coupling unification, as observed [18], should emerge as a natural prediction of the theory as opposed to being fortuitous. That in turn requires that there be no large (unpredicted) cancellation between GUT-scale threshold corrections to the gauge couplings that arise from splittings

[18] For instance, in the absence of GUT-scale threshold corrections, the MSSM value of $\alpha_3(m_Z)_{MSSM}$, assuming coupling unification, is given by $\alpha_3(m_Z)^{\circ}_{MSSM} = 0.125 \pm 0.13$ [7], which is about 5-8% higher than the observed value: $\alpha_3(m_Z)^{\circ}_{MSSM} = 0.118 \pm 0.003$ [13]. We demand that this discrepancy should be accounted for accurately by a net *negative* contribution from D-T splitting and from "other" threshold corrections [see Appendix, Eq. (A4)], without involving large cancellations. That in fact does happen for the minimal Higgs system $(45, 16, \overline{16})$ (see Ref. [14]).

within different multiplets as well as from Planck scale physics. Following this point of view, we have argued (see Appendix) that the net "other" threshold corrections to $\alpha_3(m_Z)$ arising from the Higgs (in our case $\mathbf{45}_H$, $\mathbf{16}_H$ and $\overline{\mathbf{16}}_H$) and the gauge multiplets should be negative, but conservatively and quite plausibly no more than about 10%, at the electroweak scale. This in turn restricts how big can be the threshold corrections to $\alpha_3(m_Z)$ that arise from (D-T) splitting (which is positive). Since the latter is proportional to $\ln(M_{\rm eff}\cos\gamma/M_X)$ (see Appendix), we thus obtain an upper limit on $M_{\rm eff}\cos\gamma$. For the simplest model of D-T splitting presented in Ref. [14] and in the Appendix [Eq. (A1)], one obtains: $\cos\gamma \approx (\tan\beta)/(m_t/m_b)$. An upper limit on $M_{\rm eff}\cos\gamma$ thus provides an upper limit on $M_{\rm eff}$ which is inversely proportional to $\tan\beta$. In short, our demand of natural coupling unification, together with the simplest model of D-T splitting, introduces an implicit dependence on $\tan\beta$ into the lower limit of the SO(10)-amplitude—i.e. $\hat{A}(SO(10)) \propto 1/M_{\rm eff} \geq [(\text{a quantity}) \propto \tan\beta]$. These considerations are reflected in the results given below.

Assuming $\tan\beta \geq 3$ and accurate coupling unification (as described above), one obtains for the case of MSSM, a conservative upper limit on $M_{\rm eff} \leq 2.7 \times 10^{18}$ GeV $(3/\tan\beta)$ (see Appendix and Ref. [14]). Using this upper limit, we obtain a lower limit for the standard proton decay amplitude given by

$$\hat{A}(\overline{\nu}_\tau K^+)_{std} \;\geq\; \begin{bmatrix} (7.8\times10^{-24}\,\text{GeV}^{-1})\,(1/6\text{ - }1/4) & \text{case I} \\ (3.3\times10^{-24}\,\text{GeV}^{-1})\,(1/6\text{ - }1/2) & \text{case II} \end{bmatrix} \begin{pmatrix} \text{SO(10)/MSSM, with} \\ \tan\beta\geq 3 \end{pmatrix}. \tag{46}$$

Substituting into Eq. (41) and adding the contribution from the second competing mode $\overline{\nu}_\mu K^+$, with a typical branching ratio $R \approx 0.3$, we obtain

$$\Gamma^{-1}(\overline{\nu}K^+)_{std} \;\leq\; \begin{bmatrix} (0.18\times10^{31}\,\text{years})\,(1.6\text{ - }0.7) \\ (0.4\times10^{31}\,\text{years})\,(4\text{ - }0.44) \end{bmatrix} \{64\text{ - }1/64\} \begin{pmatrix} \text{SO(10)/MSSM, with} \\ \tan\beta\geq 3 \end{pmatrix}. \tag{47}$$

The upper and lower entries in Eqs. (46) and (47) correspond to the cases I and II of the fermion mass-matrix with the *extreme values* of ϵ'—i.e. $\epsilon' = 2\times10^{-4}$ and $\epsilon' = 0$—respectively, (see Eq. (34)). The uncertainty shown inside the square brackets correspond to that in the relative phases of the different contributions. The uncertainty of $\{64 \text{ to } 1/64\}$ arises from that in β_H, $(m_{\tilde{W}}/m_{\tilde{q}})$ and $m_{\tilde{q}}$ [see Eq. (40)]. Thus we find that for MSSM embedded in SO(10), for the two extreme values of ϵ' (cases I and II) as mentioned above, the inverse partial proton decay rate should satisfy:

$$\begin{aligned} \Gamma^{-1}(p \to \overline{\nu} K^+)_{std} &\leq \begin{bmatrix} 0.20 \times 10^{31^{+2.0}_{-1.7}} \text{ years} \\ 0.32 \times 10^{31^{+2.4}_{-1.86}} \text{ years} \end{bmatrix} \\ &\leq \begin{bmatrix} 0.2 \times 10^{33} \text{ years} \\ 1 \times 10^{33} \text{ years} \end{bmatrix} \left(\begin{matrix} \text{SO(10)/MSSM, with} \\ \tan\beta \geq 3 \end{matrix} \right) . \end{aligned} \tag{48}$$

The central value of the upper limit in Eq. (48) corresponds to taking the upper limit on $M_{\text{eff}} \leq 2.7 \times 10^{18}$ GeV, which is obtained by restricting threshold corrections as described above (and in the Appendix) and by setting (conservatively) $\tan\beta \geq 3$. The uncertainties of matrix element, spectrum and choice of phases are reflected in the exponents. The uncertainty in the most sensitive entry of the fermion mass matrix—i.e. ϵ'—is incorporated (as regards obtaining an upper limit on the lifetime) by going from case I (with $\epsilon' = 2 \times 10^{-4}$) to case II ($\epsilon' = 0$). Note that this increases the lifetime by almost a factor of six. Any non-vanishing intermediate value of ϵ' would only shorten the lifetime compared to case II. In this sense, the larger of the two upper limits quoted above is rather conservative. We see that the predicted upper limit for case I of MSSM (with the extreme value of $\epsilon' = 2 \times 10^{-4}$) is lower than the empirical lower limit [Eq. (43)] by a factor of ten, while that for case II, i.e. $\epsilon' = 0$ (with all the uncertainties stretched as mentioned above) is about two times lower than the empirical lower limit.

Thus the case of MSSM embedded in SO(10) is already tightly constrained, to the point of being disfavored, by the limit on proton lifetime. The constraint is of course augmented especially by *our requirement of natural coupling unification* which prohibits accidental large cancellation between different threshold corrections[19] (see Appendix); and it will be even more severe, especially within the simplest mechanism of D-T splitting (as discussed in the Appendix), if $\tan\beta$ turns out to be larger than 5 (say). On the positive side, improvement in the current limit by a factor of even 2 to

[19] Other authors (see e.g., Ref. [65]) have considered proton decay in SUSY SO(10) by allowing for rather large GUT-scale threshold corrections, which do not, however, go well with our requirement of "natural coupling unification".

3 ought to reveal proton decay, otherwise the case of MSSM embedded in SO(10), would be clearly excluded.

6.4 The case of ESSM

Before discussing the contribution of the new $d = 5$ operators to proton decay, an interesting possibility, mentioned in the introduction (and in footnote 16), that would be especially relevant in the context of proton decay, if $\tan\beta$ is large, is worth noting. This is the case of the extended supersymmetric standard model (ESSM), which introduces an extra pair of vector-like families [$\mathbf{16} + \overline{\mathbf{16}}$ of SO(10)], at the TeV scale [21, 60]. Adding such complete SO(10)-multiplets would of course preserve coupling unification. From the point of view of adding extra families, ESSM seems to be the minimal and also the maximal extension of the MSSM, that is allowed in that it is compatible with (a) LEP neutrino-counting, (b) precision electroweak tests, as well as (c) a semi-perturbative as opposed to non-perturbative gauge coupling unification [21, 60]. [20] *The existence of two extra vector-like families of quarks and leptons can of course be tested at the LHC.*

Theoretical motivations for the case of ESSM arise on several grounds: (a) it provides a better chance for stabilizing the dilaton by having a semi-perturbative value for $\alpha_{\text{unif}} \approx 0.35\text{-}0.3$ [21], in contrast to a very weak value of 0.04 for MSSM; (b) owing to increased two-loop effects [21, 66], it raises the unification scale M_X to $(1/2{-}2)\times 10^{17}$ GeV and thereby considerably reduces the problem of a mismatch [28] between the MSSM and the string unification scales (see Section 3); (c) It lowers the GUT-prediction for $\alpha_3(m_Z)$ to (0.112–0.118) (in absence of unification-scale threshold corrections), which is in better agreement with the data than the corresponding value of (0.125–0.13) for MSSM; and (d) it provides a simple reason for inter-family mass-hierarchy [21, 60]. In this sense, ESSM, though less economical than MSSM, offers some distinct advantages.

In the present context, because of (b) and (c), ESSM naturally enhances the GUT-prediction for proton lifetime, in full accord with the data [62]. As explained in the appendix, the net result of these two effects—i.e. a raising of M_X and a lowering of $\alpha_3(m_Z)^{\circ}_{\text{ESSM}}$—is that for ESSM embedded in SO(10), $\tan\beta$ can span a wide range from 3 to even 30, and simultaneously the value

[20] For instance, addition of *two* pairs of vector-like families at the TeV-scale, to the three chiral families, would cause gauge couplings to become non-perturbative below the unification scale.

or the upper limit on $M_{\rm eff}$ can range from $(60 \text{ to } 6)\times 10^{18}$ GeV, in full accord with our criterion for accurate coupling unification discussed above.

As a result, in contrast to MSSM, ESSM allows for larger values of $\tan\beta$ (like 10 or 20), without needing large threshold corrections, and simultaneously without conflicting with the limit on proton lifetime.

To be specific, consider first the case of a moderately large $\tan\beta = 10$ (say), for which one obtains $M_{\rm eff} \approx 1.8 \times 10^{19}$ GeV, with the "other" threshold correction $-\delta_3'$ being about 5% (see Appendix for definition). In this case, one obtains:

$$\Gamma^{-1}(\overline{\nu}K^+)_{\rm std} \approx \left[\begin{array}{c}(1.6-0.7)\\(10-1)\end{array}\right]\{64-1/64\}\,(7\times 10^{31}\,\text{years}) \left(\begin{array}{c}\text{SO(10)/ESSM, with}\\ \tan\beta = 10\end{array}\right). \quad (49)$$

As before, the upper and lower entries correspond to cases I ($\epsilon' = 2\times 10^{-4}$) and II ($\epsilon' = 0$) of the fermion mass-matrix [see Eq. (34)]. The uncertainty in the upper and lower entries in the square bracket of Eq. (49) corresponds to that in the relative phases of the different contributions for the cases I and II respectively, while the factor {64-1/64} corresponds to uncertainties in the SUSY spectrum and the matrix element (see Eq. (40)).

We see that by allowing for an uncertainty of a factor of $(30-100)$ jointly from the two brackets proton lifetime arising from the standard operators would be expected to lie in the range of $(2.1-7)\times 10^{33}$ years, for the case of ESSM embedded in SO(10), even for a moderately large $\tan\beta = 10$. Such a range is compatible with present limits, but accessible to searches in the near future.

The other most important feature of ESSM is that, by allowing for larger values of $M_{\rm eff}$, especially for smaller values of $\tan\beta \approx 3$ to 5 (say), *the contribution of the standard operators by itself can be perfectly consistent with present limit on proton lifetime even for almost central or "median" values of the parameters pertaining to the SUSY spectrum, the relevant matrix element, ϵ' and the phase-dependent factor.*

For instance, for ESSM, one obtains $M_{\rm eff} \approx (4.5\times 10^{19}\text{GeV})(4/\tan\beta)$, with the "other" threshold correction - δ_3' being about 5% [see Appendix and Eq. (A6)]. Now, *combining* cases I ($\epsilon' = 2\times 10^{-4}$) and II ($\epsilon' = 0$), we see that the square bracket in Eq. (49) which we will denote by [S], varies from 0.7 to 10, depending upon the relative phases of the different contributions and the values of ϵ'. Thus as a "median" value, we will take $[S]_{\text{med}} \approx 2$ to 6. The curly bracket {64-1/64}, to be denoted by {C}, represents the uncertainty in the SUSY spectrum and the matrix element [see Eq. (40)]. Again as a "nearly central" or "median" value, we will take $\{C\}_{\text{med}} \approx 1/6$ to 6. Setting M_{eff} as above we obtain

$$\Gamma^{-1}(\bar{\nu}K^{+})^{\text{``median''}}_{\text{std}} \approx [S]_{\text{med}}\{C\}_{\text{med}}(0.45\times10^{33}\text{years})(4/\tan\beta)^{2}(\text{SO}(10)/\text{ESSM}). \tag{50}$$

Choosing a few sample values of the effective parameters [S] and {C}, with low values of $\tan\beta = 3$ to 5, the corresponding values of $\Gamma^{-1}(\bar{\nu}K^{+})$, following from Eq. (50), are listed below in Table 1.

Note that ignoring contributions from the new $d = 5$ operators for a moment[21], the entries in Table 1 represent *a very plausible range of values* for the proton lifetime, for the case of ESSM embedded in SO(10), with $\tan\beta \approx 3$ to 5 (say), *rather than upper limits for the same.* This is because they are obtained for "nearly central" or "median" values of the parameters represented by the values of [S] and {C}, as discussed above. For instance, consider the cases {C}=1 and {C}=1/2 respectively, which (as may be inferred from the table) can quite plausibly yield proton lifetimes in the range of (2 to 5)×10^{33} years Now {C}=1 corresponds, e.g., to $\beta_H = 0.014$ GeV^3 (the central value of Ref. [52]) $m_{\tilde{q}} = 1.2$ TeV and $m_{\tilde{W}}/m_{\tilde{q}} = 1/6$ [see Eq. (40)], while that of {C}=1/2 would correspond, for example, to $\beta_H = 0.014$ GeV^3, with $m_{\tilde{q}} \approx 710$ GeV and $m_{\tilde{W}}/m_{\tilde{q}} \approx 1/6$. *In short, for the case of ESSM, with low values of* $\tan\beta \approx$ *3 to 5 (say), squark masses can be well below 1 TeV, without conflicting with present limit on proton lifetime.* This feature is not permissible within MSSM embedded in SO(10).

Thus, confining for a moment to the standard operators only, if ESSM represents low-energy physics, and if $\tan\beta$ is rather small (3 to 5, say), we do not have to stretch the uncertainties in the SUSY spectrum and the matrix elements to their extreme values (in contrast to the case of MSSM) in order to understand why proton decay has not been seen as yet, and still can be optimistic that it ought to be discovered in the near future, with a lifetime $\leq 10^{34}$ years. The results for a wider variation of the parameters are listed in Table 2, where contributions of the new $d = 5$ operators are also shown.

It should also be remarked that if in the unlikely event, all the parameters

[21] As I will discuss in the next section, we of course expect the new $d = 5$ operators to be important and significantly influence proton lifetime (see e.g. Table 2). Entries in Table 1 could still represent the actual expected values of proton lifetimes, however, if the parameter K defined in 6.1 (also see 6.5) happens to be unexpectedly small ($\ll 1$).

(i.e. β_H, $(m_{\tilde{W}}/m_{\tilde{q}})$, $m_{\tilde{q}}$ and the phase-dependent factor) happen to be closer to their extreme values so as to extend proton lifetime, and if $\tan\beta$ is small (≈ 3 to 5, say) and at the same time the value of $M_{\rm eff}$ is close to its allowed upper limit (see Appendix), the standard $d = 5$ operators by themselves would tend to yield proton lifetimes exceeding even $(0.8 \text{ to } 2.5)\times 10^{34}$ years for the case of ESSM, (see Eq. (49) and Table 2). In this case (with the parameters having nearly extreme values), however, as I will discuss shortly, the contribution of the new $d = 5$ operators related to neutrino masses [see Eq. (36)], are likely to dominate and quite naturally yield lifetimes bounded above in the range of $(1 - 10) \times 10^{33}$ years (see Section 6.5 and Table 2). *Thus in the presence of the new operators, the range of* $(10^{33} - 10^{34})$ *years for proton lifetime is not only very plausible but it also provides a reasonable upper limit, for the case of ESSM embedded in SO(10).*

6.5 Contribution from the new d=5 operators

As mentioned in Section 6.1, for supersymmetric G(224)/SO(10), there very likely exists a new set of $d = 5$ operators, related to neutrino masses, which can induce proton decay [see Eq. (42)]. The decay amplitude for these operators for the leading mode (which in this case is $\bar{\nu}_\mu K^+$) becomes proportional to the quantity $P \equiv \{(f_{33})_\nu\langle\mathbf{16}_H\rangle/M\}h_{33}K/(M_{16}\tan\gamma)$, where $(f_{33})_\nu$ and h_{33} are the effective couplings defined in Eqs. (7) and (15) respectively, and M_{16} and $\tan\gamma$ are defined in the Appendix. The factor K, defined by $(f_{33})_p \equiv (f_{33})_\nu K$, is expected to be of order unity (see Section 6.1 for the origin of K). As a plausible range, we would take $K \approx 1/5$ to 2. Using $M_{16}\tan\gamma = \lambda'\langle\mathbf{16}_H\rangle$ (see Appendix), and $h_{33} \approx 1/2$ (given by top mass), one gets: $P \approx [(f_{33})_\nu/M](1/2\lambda')K$. Here M denotes the string or the Planck scale (see Section 4 and footnote 2); thus $M \approx (1/2 - 1)\times 10^{18}$ GeV; and λ' is a quartic coupling defined in the appendix. Validity of perturbative calculation suggests that λ' should not much exceed unity, while other considerations suggest that λ' should not be much less than unity either (see Ref. [14], Section 6 E). Thus, a plausible range for λ' is given by $\lambda' \approx (1/2 - \sqrt{2})$. (Note it is only the upper limit on λ' that is relevant to obtaining an upper limit on proton lifetime). Finally, from consideration of ν_τ mass, we have $(f_{33})_\nu \approx 1$ (see Section 4). We thus obtain: $P \approx (5\times 10^{-19}\text{GeV}^{-1})(1/\sqrt{2}$ to 4)K. Incorporating a further uncertainty by a factor of (1/2 to 2) that arises due to choice of the relative phases of the different contributions (see Ref.

[14]), the effective amplitude for the new operator is given by

$$\widehat{A}(\bar{\nu}_\mu K^+)_{\rm new} \approx (1.5 \times 10^{-24}{\rm GeV}^{-1})(1/2\sqrt{2} \text{ to } 8)K \tag{51}$$

Note that this new contribution is independent of $M_{\rm eff}$; *thus it is the same for ESSM as it is for MSSM, and it is independent of* $\tan\beta$. Furthermore, it turns out that the new contribution is also insensitive to ϵ'; thus it is nearly the same for cases I and II of the fermion mass-matrix. Comparing Eq. (51) with Eq. (46) we see that the new and the standard operators are typically quite comparable to one another. Since there is no reason to expect near cancellation between them (especially for both $\bar{\nu}_\mu K^+$ and $\bar{\nu}_\tau K^+$ modes), we expect the net amplitude (standard+new) to be in the range exhibited by either one. It is thus useful to obtain the inverse decay rate assuming as if the new operator dominates. Substituting Eq. (51) into Eq. (41) and allowing for the presence of the $\bar{\nu}_\tau K^+$ mode with an estimated branching ratio of nearly 0.4 (see Ref. [14]), one obtains

$$\Gamma^{-1}(\overline{\nu}K^+)_{\rm new} \approx (0.25 \times 10^{31} \text{ years}) [8\text{-}1/64] \{64\text{-}1/64\} (K^{-2} \approx 25 \text{ to } 1/4)\,. \tag{52}$$

The square bracket represents the uncertainty reflected in Eq. (51), while the curly bracket corresponds to that in the SUSY spectrum and matrix element (Eq. (40)). Allowing for the net uncertainty factor at the upper end, arising jointly from the *three brackets* in Eq. (52) to be 1000 to 4000 (say), which can be realized for plausible range of values of the parameters (see below), the new operators related to neutrino masses, by themselves, lead to a proton decay lifetime given by:

$$\Gamma^{-1}(\overline{\nu}K^+)^{\rm upper}_{\rm new} \approx (2.5\text{-}10) \times 10^{33}\,\text{years (SO(10) or string G(224))} \quad \text{(Indep. of } \tan\beta)\,. \tag{53}$$

The superscript "upper" corresponds to estimated lifetimes near the upper end. For instance, taking the curly bracket in Eq. (52) to be ≈ 8 to 16 (say) [corresponding for example, to $\beta_H = 0.010$ GeV^3, $(m_{\tilde{W}}/m_{\tilde{q}}) \approx 1/12$ and $m_{\tilde{q}} \approx (1 \text{ to } 1.4)(1.2 \text{ TeV})$], instead of its extreme value of 64, and setting the square bracket in Eq. (52) to be ≈ 6, and $K^{-2} \approx 20$, which are quite plausible, we obtain: $\Gamma^{-1}(\bar{\nu}K^+)_{\rm new} \approx (2.5 - 5) \times 10^{33}$ years; independently of $\tan\beta$, for both MSSM and ESSM. Proton lifetime for other choices of parameters, which lead to similar conclusion, are listed in Table 2.

It should be stressed that the standard $d = 5$ operators [mediated by the color-triplets in the 10_H of SO(10)] may naturally be absent for a string-derived G(224)-model (see e.g. Ref. [30] and [31]), but the new $d = 5$

operators, related to the Majorana masses of the RH neutrinos and the CKM mixings, should very likely be present for such a model, as much as for SO(10). These would induce proton decay [22]. *Thus our expectations for the proton decay lifetime [as shown in Eq. (53)] and the prominence of the μ^+K^0 mode (see below) hold for a string-derived G(224)-model, just as they do for SO(10).* For a string - G(224) - model, however, *the new d=5 operators would be essentially the sole source of proton decay*[21].

Nearly the same situation emerges for the case of ESSM embedded in G(224) or SO(10), with low $\tan\beta(\approx 3$ to 10, say), especially if the parameters (including β_H, $m_{\tilde{W}}/m_{\tilde{q}}$, $m_{\tilde{q}}$, the phase-dependent factor as well as M_{eff}) happen to be somewhat closer to their extreme values so as to extend proton lifetime. In this case, (that is for ESSM) as noted in the previous sub-section, the contribution of the standard $d = 5$ operators would be suppressed; and proton decay would proceed primarily via the new operators with a lifetime quite plausibly in the range of $10^{33} - 10^{34}$ years, as exhibited above.

6.6 The Charged Lepton Decay Modes ($p \to \mu^+K^0$ and $p \to e^+\pi^0$)

I now note a distinguishing feature of the SO(10) or the G(224) model presented here. Allowing for uncertainties in the way the standard and the new operators can combine with each other for the three leading modes i.e. $\overline{\nu}_\tau K^+$, $\overline{\nu}_\mu K^+$ and μ^+K^0, we obtain (see Ref. [14] for details):

$$B(\mu^+K^0)_{std+new} \approx [1\% \text{ to } 50\%]\ \kappa \quad (\text{SO(10) or string G(224)}) \tag{54}$$

where κ denotes the ratio of the squares of relevant matrix elements for the μ^+K^0 and $\overline{\nu}K^+$ modes. In the absence of a reliable lattice calculation for the $\bar{\nu}K^+$ mode, one should remain open to the possibility of $\kappa \approx 1/2$ to 1 (say). We find that for a large range of parameters, the branching ratio $B(\mu^+K^0)$ can lie in the range of 20 to 40% (if $\kappa \approx 1$). This prominence of the μ^+K^0 mode for the SO(10)/G(224) model is primarily due to contributions from the new $d = 5$ operators. This contrasts sharply with the minimal SU(5) model, in which the μ^+K^0 mode is expected to have a branching ratio of only about 10^{-3}. In short, prominence of the μ^+K^0 mode, if seen, would clearly show the relevance of the new operators, and thereby reveal the proposed link between neutrino masses and proton decay [20].

[22] In addition, quantum gravity induced $d = 5$ operators are also expected to be present at some level, depending upon the degree of suppression of these operators due to flavor symmetries (see e.g. Ref. [34]).

The $d = 5$ operators as described here (standard and new) would lead to highly suppressed $e^+\pi^0$ mode, for MSSM or ESSM embedded in SO(10). The gauge boson-mediated $d = 6$ operators, however, still give (using the recently determined matrix element $\alpha_H = 0.015 \pm 0.001$ GeV^3 [52]) proton decaying into $e^+\pi^0$ with an inverse rate:

$$\Gamma^{-1}(p \to e^+\pi^0)^{\mathrm{SO(10)/SU(5)}}_{\mathrm{MSSM}} \approx 10^{35\pm1}\text{years} \ . \tag{55}$$

This can well be as short as about 10^{34} years. For the case of ESSM embedded into SO(10) [or for an analogous case embedded into SU(5)], there are two new features. Considering that in this case, both α_{unif} and the unification scale M_X (thereby the mass M_V of the (X, Y) gauge bosons) are raised by nearly a factor of (6 to 7) and (2.5 to 5) respectively, compared to those for MSSM (see discussions in Section 6.4), and that the inverse decay rate is proportional to $(M_V^4/\alpha^2_{\mathrm{unif}})$, we expect

$$\Gamma^{-1}(p \to e^+\pi^0)^{\mathrm{SO(10)/SU(5)}}_{\mathrm{ESSM}} \approx (1 \text{ to } 17)\Gamma^{-1}(p \to e^+\pi^0)^{\mathrm{SO(10)/SU(5)}}_{\mathrm{MSSM}} \ . \tag{56}$$

The net upshot is that the gauge boson-mediated $d = 6$ operators can quite plausibly lead to observable $e^+\pi^0$ decay mode with an inverse decay rate in the range of 10^{34}-10^{35} years. For ESSM embedded in SO(10), there can be the interesting situation that both $\bar{\nu}K^+$ (arising from $d = 5$) and $e^+\pi^0$ (arising from $d = 6$) may have comparable rates, with proton having a lifetime $\sim$ (1/2-2) $\times 10^{34}$ years. It should be stressed that the $e^+\pi^0$-mode is the *common denominator* of all GUT models (SU(5), SO(10), etc.) which unify quarks and leptons and the three gauge forces. Its rate as mentioned above is determined essentially by the SUSY unification-scale, without the uncertainty of the SUSY-spectrum. I should also mention that the $e^+\pi^0$-mode is predicted to be the dominant mode in the flipped SU(5) × U(1)-model [67]. For these reasons, intensifying the search for the $e^+\pi^0$-mode to the level of sensitivity of about 10^{35} years in the next generation proton decay detector should be well worth the effort.

Before summarizing the results of this section, I note below a few distinctive features of the conventional approach adopted here compared to those of some alternatives.

6.7 Conventional Versus Other Approaches

In these lectures, as elaborated in Section 3, I have pursued systematically the consequences for fermion masses, neutrino oscillations *and* proton decay of the assumption that essentially the conventional picture of

SUSY grand unification [3, 4, 5, 6, 7] holds, providing a good effective theory in 4D between the conventional GUT-scale $M_X \sim 2 \times 10^{16}$ GeV (for ESSM, $M_X \sim$ (1/2-2) $\times\, 10^{17}$ GeV) and the conventional string scale $M_{\rm st} \sim$ (few to 10) $\times\, 10^{17}$ GeV. Believing in an underlying string/M-theory, and yet knowing that a preferred ground state of this theory is not yet in hand, the attitude, based on a bottom-up approach, has been to subject the assumed effective theory of grand unification to as many low-energy tests as possible, and to assess its soundness on empirical grounds. With this in mind, I have assumed that either a realistic 4D SO(10)-solution (with the desired mechanism of doublet-triplet splitting operating in 4D), or a suitable string-derived G(224)-solution (with $M_X \sim (1/2)M_{\rm st}$, see footnote 2) emerges effectively from an underlying string theory at the conventional string scale as mentioned above, and that the G(224)/SO(10) symmetry breaks into G(213) at the conventional GUT-scale M_X. The extra dimensions of string/M-theory are assumed to be tiny lying between the GUT-scale size $\sim M_X^{-1}$ and the string-size $M_{\rm st}^{-1}$, so as not to disturb the successes of GUT (see below). As mentioned before, this conventional picture of grand unification described above seems to be directly motivated on observational grounds such as those based on (a) coupling unification or equivalently the agreement between the observed and the predicted values of $\sin^2\theta_W$ (see Section 3), (b) neutrino masses including $\Delta m^2(\nu_\mu\text{-}\nu_\tau)$ and (c) the fact that spontaneous violation of B–L local symmetry seems to be needed to implement baryogenesis via leptogenesis [16, 17]. The relevance of the group theory of G(224)/SO(10)-symmetry for the 4D theory is further suggested by the success of the predictions of the masses and the mixings of all fermions including neutrinos; these include $m_b^0 \approx m_\tau^0$, $m(\nu_{\rm Dirac}^\tau) \approx m_t(M_X)$, and the smallness of $V_{cb} \approx 0.04$ correlated with the largeness of $\sin^2 2\theta^{\rm osc}_{\nu_\mu\nu_\tau} \approx 1$ (see Section 5).

In contrast to this conventional approach based on a presumed string-unified G(224) or an SO(10)-symmetry, there are several alternative approaches (scenarios) which have been proposed in the literature in recent years. Of importance is the fact that in many of these alternatives an attempt is made to strongly suppress proton decay, in some cases exclusively the $d = 5$ operators (though not necessarily the $d = 6$), invariably utilizing a higher dimensional mechanism. Each of these alternatives is interesting in its own right. However, it seems to me that the collection of successes mentioned above is not (yet) realized within these alternatives. For comparison, I mention briefly only a few, leaving out many interesting variants.

One such alternative is based on the idea of TeV-scale large extra dimensions [36]. Though most intriguing, it does not seem to provide simple explanations for (a) coupling unification, (b) neutrino-masses (or their $(\text{mass})^2$-differences) of the observed magnitudes[23], (c) a large (or maximal) ν_μ-ν_τ oscillation angle, and (d) baryogenesis via leptogenesis that seems to require violation of B–L at high temperatures. Within this scenario, quantum-gravity induced proton decay would ordinarily be extra rapid. This is prevented, for example, by assuming that quarks and leptons live in different positions in the extra dimension. It appears to me that this idea (introduced just to prevent proton decay) however, sacrifices the simple reason for the co-existence quarks and leptons that is provided by a gauge unification of matter within a family as in G(224) or SO(10).

There is an alternative class of attempts, carried out again in the context of higher dimensional theories, which, in contrast to the case mentioned above, assume that the extra dimensions ($d > 4$) are all small, lying between (or around) the conventional GUT and string scales. The approach of this class of attempts is rather close in spirit to that of the conventional approach of grand unification pursued here (see Section 3). As may be seen from the discussions below, they could essentially coincide with the string-unified G(224)-picture presented here if the effective symmetry in 4D, below the string (or compactification) scale, contains at least the G(224) symmetry.

Motivated by the original attempts carried out in the context of string theory [69] most of the recent attempts in the class mentioned above are made in the spirit of a bottom-up approach[24] to physics near the GUT and the string scales. They assume, following the spirit of the results of Ref. [69], and of analogous results obtained for the free fermionic formulation of string theory [70] (for applications based on this formulation, see e.g., Ref. [71], [30], [31] and [72]), that grand unification occurs, through symmetries like E_6, SO(10) or SU(5), only in some higher dimension ($d > 4$), and that the breaking of the unification gauge symmetry to some lower symmetry containing the standard model gauge group as well as doublet-triplet splitting occurs in the process of compactification. More specifically the latter two

[23] By placing the singlet (right-handed) neutrino in the bulk, for example, one can get a light Dirac neutrino [68] with a mass $m_\nu \approx \kappa v_{EW} M^*/M_{\rm Pl} \approx \kappa(2 \times 10^{-5}\ \text{eV})$, where $M^* \approx 1$ TeV, $M_{\rm Pl} \approx 10^{19}$ GeV (as in [68]), and κ is the effective Yukawa coupling. To get $m_\nu \sim 1/20$ eV (for SuperK), one would, however, need too large a $\kappa \sim 2 \times 10^3$ and/or too large a value for $M^* \gtrsim 100$ TeV; which would seem to face the gauge-hierarchy problem.

[24] This is of course also the case for the approach adopted here which is outlined in Section 3.

phenomena take place through either (a) Wilson lines [69], or (b) orbifolds [73] (for an incomplete list of recent attempts based on orbifold compactification, see e.g., Refs. [74, 75, 76, 77, 79, 80, 81]), or (c) essentially equivalently by a set of boundary conditions together with the associated GSO projections for the free fermionic formulation (see e.g., [30, 31, 71, 72]), or (d) discrete symmetries operating in higher dimensions [82].

Most of these attempts end up not only in achieving (a) doublet-triplet splitting by projecting out the relevant color triplets from the zero mode-spectrum in 4D, and (b) gauge symmetry breaking, as mentioned above, but also (c) suppressing strongly or eliminating the $d = 5$ proton decay operators. It should be mentioned, however, that in some of these attempts (see e.g., [75]), the mass of the X gauge boson is suggested to be lower than the conventional GUT-scale of 2×10^{16} GeV by about a factor of 3 to 8; correspondingly they raise the prospect for observing the $d = 6$ gauge boson mediated $e^{+}\pi^{0}$ mode, which is allowed in [75].

One crucial distinction between the various cases is provided by the nature of the effective gauge symmetry that is realized in 4D, below the string (or compactification) scale. References [74, 75, 76, 77, 78, 79] assume a supersymmetric SU(5) gauge symmetry in 5D, which is broken down to the standard model gauge symmetry in 4D through compactification. References [80] and [81], on the other hand, assume a supersymmetric SO(10) gauge symmetry in 6D and show (interestingly enough) that there are two 5D subspaces containing G(224) and SU(5)×U(1) subgroups respectively, whose intersection leads to $\mathrm{SU(3)\times SU(2)\times U(1)}_Y\times \mathrm{U(1)}_X$ in 4D, which contains B–L. While it is desirable to have B–L in 4D, consistent breaking of $\mathrm{U(1)}_X$ (or B–L) and generating desired masses of the right handed neutrinos, not to mention the masses and the mixings of the other fermions, is not yet realized in these constructions.

For comparison, it seems to me that at the very least B–L should emerge as a generator in 4D to implement baryogenesis via leptogenesis, and also to protect RH neutrinos from acquiring a string-scale mass. This feature is not available in models which start with SU(5) in 5D. Furthermore, the full SU(4)-color symmetry, which of course contains B–L, plays a crucial role in yielding not only $m_b^0 \approx m_\tau^0$ but also (a) $m(\nu_{\mathrm{Dirac}}^{\tau}) \approx m_t(M_X)$ that is needed to account for $m(\nu_\tau)$ or rather $\Delta m^2(\nu_\mu\text{-}\nu_\tau)$, in accord with observation (see Section 4), and (b) the smallness of V_{cb} together with the near maximality of $\sin^2 2\theta^{\mathrm{osc}}_{\nu_\mu\nu_\tau}$ (see Section 5). The symmetry $\mathrm{SU(2)}_L\times\mathrm{SU(2)}_R$ is also most useful in that it relates the masses and mixings of the up and the down sectors. Without such relations, we will not have the predictivity of the framework presented in Section 5.

In short, as mentioned before, certain intriguing features of the masses and mixings of all fermions including neutrinos, of the type mentioned above, as well as the need for leptogenesis, seem to strongly suggest that the effective symmetry below the string-scale in 4D should contain minimally the symmetry G(224) [or a close relative G(214)] and maximally SO(10). The G(224)/SO(10)-framework developed here has turned out to be the most predictive, in large part by virtue of its group structure and the assumption of minimality of the Higgs system. Given that it is also most successful so far, as regards its predictions, derivation of such a picture from an underlying theory, especially at least that based on an effective G(224)-symmetry[25] in 4D leading to the pattern of Yukawa couplings presented here remains a challenge.[26] Pending such a derivation, however, given the empirical support it has received so far, it makes sense to test the supersymmetric G(224)/SO(10)-framework, and thereby the *conventional picture of grand unification* on which it rests, thoroughly. There are two notable missing pieces of this picture. One is supersymmetry which will be probed at the LHC and a future NLC. The other, that constitutes the hallmark of grand unification, is proton decay. The results of this section on proton decay are summarized below.

6.8 Section Summary

Given the importance of proton decay, a systematic study of this process has been carried out within the supersymmetric SO(10)/G(224)-framework[27], with special attention paid to its dependence on fermion masses and thresh-

[25] For this case, following the examples of Refs. [30] and [31], the color triplets in the $\mathbf{10}_H$ of SO(10) would be projected out of the zero-mode spectrum, and thus the standard $d = 5$ operators which would have been induced by the exchange of such triplets would be absent, as in Refs. [74, 75, 76, 77, 78, 79, 80, 81, 82]. But, as long as the Majorana masses of the RH neutrinos are generated as in Section 4, the new neutrino-mass related $d = 5$ proton decay operators would generically be present (see Section 6 E).

[26] In this regard, three-generation solutions containing the G(224)-symmetry in 4D have been obtained in the context of the fermionic formulation of string theory in Ref. [30], within type-I string vacua with or without supersymmetry in [83, 84, 85] in the context of D-brane inspired models in [86], within type-I string-construction or string-motivated models obtained from intersecting D-branes (with G(224) breaking into G(213) at $M_X \sim M_{\rm st}$) in [87, 88], in string model with unification at the string scale in [89], and in other contexts (see e.g. [90] and [91]).

[27] As described in Sections 3, 4 and 5.

old effects. A representative set of results corresponding to different choices of parameters is presented in Tables 1 and 2. Allowing for the ESSM-variant, the study strongly suggests that an upper limit on proton lifetime is given by

$$\tau_{\text{proton}} \leq (1/3\text{-}2) \times 10^{34} \text{ years}, \tag{57}$$

with $\overline{\nu}K^+$ being the dominant decay mode, and quite possibly μ^+K^0 and $e^+\pi^0$ being prominent. Although there are uncertainties in the matrix element, in the SUSY-spectrum, in the phase-dependent factor, $\tan\beta$ and in certain sensitive elements of the fermion mass matrix, notably ϵ' (see Eq. (48) for predictions in cases I versus II), this upper limit is obtained, for the case of MSSM embedded in SO(10), by allowing for a generous range in these parameters and stretching all of them in the same direction so as to extend proton lifetime. In this sense, while the predicted lifetime spans a wide range, the upper limit quoted above, in fact more like 10^{33} years, is most conservative, for the case of MSSM (see Eq. (48) and Table 1). It is thus tightly constrained already by the empirical lower limit on $\Gamma^{-1}(\overline{\nu}K^+)$ of 1.9×10^{33} years to the point of being disfavored. For the case of ESSM embedded in SO(10), the standard $d = 5$ operators are suppressed compared to the case of MSSM; as a result, by themselves they can naturally lead to lifetimes in the range of $(1 - 10) \times 10^{33}$ years, for nearly central values of the parameters pertaining to the SUSY-spectrum and the matrix element (see Eq. (50) and Table 1). Including the contribution of the new $d = 5$ operators, and allowing for a wide variation of the parameters mentioned above, one finds that the range of $(10^{33} - 2 \times 10^{34})$ years for proton lifetime is not only very plausible but it also provides a rather conservative upper limit, for the case of ESSM embedded in either SO(10) or G(224) (see Section 6.5 and Table 2). Thus our study provides a clear reason to expect that the discovery of proton decay should be imminent for the case of ESSM, and even more so for that of MSSM. The implication of this prediction for a next-generation detector is emphasized in the next section.

7 Concluding Remarks

The preceding sections show that, but for two missing pieces—supersymmetry and proton decay—the evidence in support of grand unification is now strong. It includes: (i) the observed family-structure, (ii) quantization of electric charge, (iii) the meeting of the gauge couplings, (iv) neutrino-oscillations as observed at SuperK, (v) the intricate pattern of the masses and mixings of all fermions, including the neutrinos, and (vi) the need for B–L as a generator, to implement baryogenesis. Taken together, these not only favor grand unification but in fact select out a particular route to such unification, based on the ideas of supersymmetry, SU(4)-color and left-right symmetry. Thus they point to the relevance of an effective string-unified G(224) or SO(10)-symmetry in four dimensions, as discussed in Sections 3 and 4.

Based on a systematic study of proton decay within the supersymmetric SO(10)/G(224)-framework, that (a) allows for the possibilities of both MSSM and ESSM, and (b) incorporates the improved values of the matrix element and renormalization effects, I have argued that a conservative upper limit on the proton lifetime is about $(1/3\text{-}2)\times 10^{34}$ years.

So, unless the fitting of all the pieces listed above is a mere coincidence, it is hard to believe that that is the case, discovery of proton decay should be around the corner. In particular, as mentioned in the Introduction, one expects that candidate events should very likely be observed in the near future already at SuperK, if its operation is restored. However, allowing for the possibility that proton lifetime may well be near the upper limit stated above, a next-generation detector providing a net gain in sensitivity by a factor five to ten, compared to SuperK, would be needed to produce real events and distinguish them unambiguously from the background. Such an improved detector would of course be essential to study the branching ratios of certain crucial though (possibly) sub-dominant decay modes such as the μ^+K^0 and $e^+\pi^0$ as mentioned in Section 6.6.

The reason for pleading for such improved searches is that proton decay would provide us with a wealth of knowledge about physics at truly short distances ($< 10^{-30}$ cm), which cannot be gained by any other means. Specifically, the observation of proton decay, at a rate suggested above, with $\overline{\nu}K^+$ mode being dominant, would not only reveal the underlying unity of quarks and leptons but also the relevance of supersymmetry. It would also confirm a unification of the fundamental forces at a scale of order 2×10^{16} GeV. Furthermore, prominence of the μ^+K^0 mode, if seen, would have even deeper significance, in that in addition to supporting the three features mentioned above, it would also reveal the link between neutrino masses and proton decay, as discussed in Section 6. *In this sense, the role of proton decay in probing into physics at the most fundamental level is unique.* In view of how valuable such a probe would be and the fact that the predicted upper limit on the proton lifetime is at most a factor of three to ten higher than the empirical lower limit, the argument in favor of building an improved detector seems compelling.

To conclude, the discovery of proton decay would undoubtedly constitute a landmark in the history of physics. It would provide the last, missing piece of gauge unification and would shed light on how such a unification may be extended to include gravity in the context of a deeper theory.

Acknowledgments: I would like to thank Kaladi S. Babu and Frank Wilczek for a most enjoyable collaboration on topics covered in these lectures. Discussions with Kaladi S. Babu in updating the results of the previous study have been most helpful. Correspondence with Michael Dine and Edward Witten on issues pertaining to the questions of unification and proton decay have been beneficial. I am grateful to K. Turznyski for communicating the details of his results on renormalization effects. I would like to thank Antonio Masiero, Goran Senjanovic and Alexie Smirnov for inviting me to lecture at the summer school at the ICTP, Trieste and also for their kind hospitality. The research presented here is supported in part by DOE grant no. DE-FG02-96ER-41015 and by the Department of Energy under contract number DE–AC03–76SF00515.

APPENDIX: A Natural Doublet-Triplet Splitting Mechanism in SO(10)

In supersymmetric SO(10), a natural doublet–triplet splitting can be achieved by coupling the adjoint Higgs $\mathbf{45_H}$ to a $\mathbf{10_H}$ and a $\mathbf{10'_H}$, with $\mathbf{45_H}$ acquiring a unification–scale VEV in the B–L direction [92, 93]: $\langle\mathbf{45_H}\rangle = (a,a,a,0,0)\times\tau_2$ with $a \sim M_U$. As discussed in Section 5, to generate CKM mixing for fermions we require $(\mathbf{16_H})_d$ to acquire a VEV of the electroweak scale. To ensure accurate gauge coupling unification, the effective low energy theory should not contain split multiplets beyond those of MSSM. Thus the MSSM Higgs doublets must be linear combinations of the $\mathrm{SU(2)}_L$ doublets in $\mathbf{10_H}$ and $\mathbf{16_H}$. A simple set of superpotential terms that ensures this and incorporates doublet-triplet splitting is [14]:

$$W_H \;=\; \lambda\,\mathbf{10_H}\,\mathbf{45_H}\,\mathbf{10'_H} \;+\; M_{10}\,\mathbf{10'_H}^{2} \;+\; \lambda'\,\overline{\mathbf{16}}_H\,\overline{\mathbf{16}}_H\,\mathbf{10}_H \;+\; M_{16}\,\mathbf{16}_H\overline{\mathbf{16}}_H\,. \tag{A1}$$

A complete superpotential for $\mathbf{45_H}$, $\mathbf{16_H}$, $\overline{\mathbf{16}}_\mathbf{H}$, $\mathbf{10}_H$, $\mathbf{10'}_H$ and possibly other fields, which ensure that (a) $\mathbf{45_H}$, $\mathbf{16_H}$ and $\overline{\mathbf{16}}_\mathbf{H}$ acquire unification scale VEVs with $\langle\mathbf{45_H}\rangle$ being along the $(B\text{-}L)$ direction; (b) that exactly two Higgs doublets (H_u, H_d) remain light, with H_d being a linear combination of $(\mathbf{10_H})_d$ and $(\mathbf{16_H})_d$; and (c) there are no unwanted pseudoGoldstone bosons, can be constructed. With $\langle\mathbf{45_H}\rangle$ in the B–L direction, it does not contribute to the Higgs doublet mass matrix, so one pair of Higgs doublet remains light, while all triplets acquire unification scale masses. The light MSSM Higgs doublets are [14]

$$H_u \;=\; \mathbf{10}_u\,, \quad H_d \;=\; \cos\gamma\,\mathbf{10}_d \;+\; \sin\gamma\,\mathbf{16}_d\,, \tag{A2}$$

with $\tan\gamma \equiv \lambda'\langle\overline{\mathbf{16}}_\mathbf{H}\rangle/M_{16}$. Consequently, $\langle\mathbf{10}\rangle_d = (\cos\gamma)\,v_d$, $\langle\mathbf{16}_d\rangle = (\sin\gamma)\,v_d$, with $\langle H_d\rangle = v_d$ and $\langle\mathbf{16}_d\rangle$ and $\langle\mathbf{10}_d\rangle$ denoting the electroweak VEVs of those multiplets. Note that H_u is purely in $\mathbf{10}_H$ and that $\langle\mathbf{10}_d\rangle^2 + \langle\mathbf{16}_d\rangle^2 = v_d^2$. This mechanism of doublet-triplet (DT) splitting is the simplest for the minimal Higgs systems. It has the advantage that it meets the requirements of both D-T splitting and CKM-mixing. In turn, it has three special consequences:

(i) It modifies the familiar SO(10)-relation $\tan\beta \equiv v_u/v_d = m_t/m_b \approx 60$ to [28]:

$$\tan\beta/\cos\gamma \approx m_t/m_b \approx 60\ . \tag{A3}$$

[28] It is worth noting that the simple relationship between $\cos\gamma$ and $\tan\beta$—i.e. $\cos\gamma \approx \tan\beta/(m_t/m_b)$—would be modified if the superpotential contains an additional term like

As a result, even low to moderate values of $\tan\beta \approx 3$ to 10 (say) are perfectly allowed in SO(10) (corresponding to $\cos\gamma \approx 1/20$ to 1/6).

(ii) The most important consequence of the DT-splitting mechanism outlined above is this: In contrast to SU(5), for which the strengths of the standard $d = 5$ operators are proportional to $(M_{H_c})^{-1}$ (where $M_{H_C} \sim few \times 10^{16}$ GeV (see Eq. (44)), for the SO(10)-model, they become proportional to $M_{\rm eff}^{-1}$, where $M_{\rm eff} = (\lambda a)^2/M_{10'} \sim M_X^2/M_{10'}$. As noted in Ref. [14], $M_{10'}$ can be naturally smaller (due to flavor symmetries) than M_X and thus $M_{\rm eff}$ correspondingly larger than M_X by even one to three orders of magnitude. Now the proton decay amplitudes for SO(10) in fact possess an intrinsic enhancement compared to those for SU(5), owing primarily due to differences in their Yukawa couplings for the up sector (see Appendix C in Ref. [14]). As a result, these larger values of $M_{\rm eff} \sim (10^{18} - 10^{19})$ GeV are in fact needed for the SO(10)-model to be compatible with the observed limit on the proton lifetime. At the same time, being bounded above by considerations of threshold effects (see below), they allow optimism as regards future observation of proton decay.

(iii) $M_{\rm eff}$ gets bounded above by considerations of coupling unification and GUT-scale threshold effects as follows. Let us recall that in the absence of unification-scale threshold and Planck-scale effects, the MSSM value of $\alpha_3(m_Z)$ in the $\overline{\rm MS}$ scheme, obtained by assuming gauge coupling unification, is given by $\alpha_3(m_Z)^\circ_{\rm MSSM} = 0.125 - 0.13$ [7]. This is about 5 to 8% *higher* than the observed value: $\alpha_3(m_Z) = 0.118 \pm 0.003$ [13]. Now, assuming coupling unification, the net (observed) value of α_3, for the case of MSSM embedded in SU(5) or SO(10), is given by:

$$\alpha_3(m_Z)_{\rm net} = \alpha_3(m_Z)^\circ_{\rm MSSM} + \Delta\alpha_3(m_Z)^{\rm MSSM}_{\rm DT} + \Delta'_3 \qquad \text{(A4)}$$

where $\Delta\alpha_3(m_Z)_{\rm DT}$ and Δ'_3 represent GUT-scale threshold corrections respectively due to doublet-triplet splitting and the splittings in the other multiplets (like the gauge and the Higgs multiplets), all of which are evaluated at m_Z. Now, owing to mixing between $\mathbf{10}_d$ and $\mathbf{16}_d$ [see Eq. (A2)], one finds that $\Delta\alpha_3(m_Z)_{\rm DT}$ is given by $[\alpha_3(m_Z)^2/2\pi](9/7)\ln(M_{\rm eff}\cos\gamma/M_X)$ [14].

As mentioned above, constraint from proton lifetime sets a lower limit

$\lambda''\mathbf{16}_H \cdot \mathbf{16}_H \cdot \mathbf{10}'_H$, which would induce a mixing between the doublets in $\mathbf{10}'_d$, $\mathbf{16}_d$ and $\mathbf{10}_d$. That in turn will mean that the upper limit on $M_{\rm eff}\cos\gamma$ following from considerations of threshold corrections (see below) will not be strictly proportional to $\tan\beta$. I thank Kaladi Babu for making this observation.

on M_{eff} given by $M_{\text{eff}} > (1-6) \times 10^{18}$GeV. Thus, even for small $\tan\beta \approx 2$ (i.e. $\cos\gamma \approx \tan(\beta/60) \approx 1/30$), $\Delta\alpha_3(m_Z)_{\text{DT}}$ is positive; and it increases logarithmically with M_{eff}. Since $\alpha_3(m_Z)^{\circ}_{\text{MSSM}}$ is higher than $\alpha_3(m_Z)_{\text{obs}}$, and as we saw, $\Delta\alpha_3(m_Z)_{\text{DT}}$ is positive, it follows that the corrections due to *other* multiplets denoted by $\delta_3' = \Delta_3'/\alpha_3(m_Z)$ should be appropriately negative so that $\alpha_3(m_Z)_{\text{net}}$ would agree with the observed value.

In order that coupling unification may be regarded as a natural prediction of SUSY unification, as opposed to being a mere coincidence, it is important that the magnitude of the net other threshold corrections, denoted by δ_3', be negative but not any more than about 8 to 10% in magnitude (i.e. $-\delta_3' \leq (8-10)\%$). It was shown in Ref. [14] that the contributions from the gauge and the minimal set of Higgs multiplets (i.e. $\mathbf{45}_H, \mathbf{16}_H, \overline{\mathbf{16}}_H$ and $\mathbf{10}_H$) leads to threshold correction, denoted by δ_3', which has in fact a negative sign and quite naturally a magnitude of 4 to 8%, as needed to account for the observed coupling unification. The correction to $\alpha_3(m_Z)$ due to Planck scale physics through the effective operator $F_{\mu\nu}F^{\mu\nu}\mathbf{45}_H/M$ does not alter the estimate of δ_3' because it vanishes due to antisymmetry in the SO(10)- contraction.

Imposing that δ_3' (evaluated at m_Z)be negative and not any more than about 10-11% in magnitude in turn provides a restriction on how big the correction due to doublet-triplet splitting—i.e. $\Delta\alpha_3(m_Z)_{D\bar{T}}$—can be. That in turn sets an upper limit on $M_{\text{eff}}\cos\gamma$, and thereby on M_{eff} for a given $\tan\beta$. For instance, for MSSM, with $\tan\beta = (2,3,8)$, one obtains (see Ref. [14]): $M_{\text{eff}} \leq (4, 2.66, 1) \times 10^{18}$ GeV. Thus, conservatively, taking $\tan\beta \geq 3$, one obtains:

$$M_{\text{eff}} \lesssim 2.7 \times 10^{18}\text{GeV (MSSM)} \quad (\tan\beta \geq 3)\ . \qquad \text{(A5)}$$

Limit on M_{eff} For The case of ESSM

Next consider the restriction on M_{eff} that would arise for the case of the extended supersymmetric standard model (ESSM), which introduces an extra pair of vector-like families $(16+\bar{1}6)$ of SO(10)) at the TeV scale [21](see also footnote 16). In this case, α_{unif} is raised to 0.25 to 0.3, compared to 0.04 in MSSM. Owing to increased two-loop effects the scale of unification M_X is raised to $(1/2-2) \times 10^{17}$ GeV, while $\alpha_3(m_Z)^{\circ}_{\text{ESSM}}$ is lowered to about 0.112-0.118 [21, 66].

With raised M_X, the product $M_{\text{eff}}\cos\gamma \approx M_{\text{eff}}(\tan\beta)/60$ can be higher by almost a factor of five compared to that for MSSM, without altering $\Delta\alpha_3(m_Z)_{\text{DT}}$. Furthermore, since $\alpha_3(m_Z)^{\circ}_{\text{MSSM}}$ is typically lower than the observed value of $\alpha_3(m_Z)$ (contrast this with the case of ESSM), for ESSM, M_{eff} can be higher than that for MSSM by as much as a factor of 2 to 3, without requiring an enhancement of δ_3'. The net result is that for ESSM embedded in SO(10), $\tan\beta$ can span a wide range from 3 to even 30 (say) and simultaneously the upper limit on M_{eff} can vary over the range (60 to 6)$\times 10^{18} GeV$, satisfying

$$M_{\text{eff}} \lesssim (6 \times 10^{18}\text{GeV})(30/\tan\beta) \text{ (ESSM)}, \tag{A6}$$

with the unification-scale threshold corrections from "other" sources denoted by $\delta_3' = \Delta_3'/\alpha_3(m_Z)$ being negative, but no more than about 5% in magnitude. As noted above, such values of δ_3' emerge quite naturally for the minimal Higgs system. Thus, one important consequence of ESSM is that by allowing for larger values of M_{eff} (compared to MSSM), without entailing larger values of δ_3', it can be perfectly compatible with the limit on proton lifetime for almost *central values* of the parameters pertaining to the SUSY spectrum and the relevant matrix elements (see Eq. (40)). Further, larger values of $\tan\beta$ (10 to 30, say) can be compatible with proton lifetime only for the case of ESSM, but not for MSSM. These features are discussed in the text, and also exhibited in Table 2.

• Since we are interested in exhibiting expected proton lifetime near the upper end, we are not showing entries in Table 2 corresponding to values of the parameters for the SUSY spectrum and the matrix element [see Eq. (40), for which the curly bracket {C} appearing in Eqs. (47), (49), (52)] would be less than one (see however Table 1). In this context, we have chosen here "nearly central", "intermediate" and "nearly extreme" values of the parameters such that the said curly bracket is given by 2, 8 and 32 respectively, instead of its extreme upper-end value of 64. For instance, the curly bracket would be 2 if β_H = (0.0117) GeV^3, $m_{\tilde{q}} \approx 1.2$ TeV and $m_{\tilde{W}}/m_{\tilde{q}} \approx (1/7.2)$, while it would be 8 if $\beta_H = 0.010$ GeV^3, $m_{\tilde{q}} \approx 1.44$ TeV and $m_{\tilde{W}}/m_{\tilde{q}} \approx 1/10$; and it would be 32 if, for example, $\beta_H = 0.007$ GeV^3, $m_{\tilde{q}} \approx \sqrt{2}(1.2$ TeV) and $m_{\tilde{W}}/m_{\tilde{q}} \approx 1/12$.

† All the entries for the standard $d = 5$ operators correspond to taking an intermediate value of $\epsilon' \approx (1 \text{ to } 1.4) \times 10^{-4}$ (as opposed to the extreme values of 2×10^{-4} and zero for cases I and II, see Eq. (34)) and an intermediate phase-dependent factor such that the uncertainty factor in the square bracket appearing in Eqs. (47) and (49) is given by 5, instead of its extreme values of $2 \times 4 = 8$ and $2.5 \times 4 = 10$, respectively.

†† For the new operators, the factor [8-1/64] appearing in Eq. (52) is

taken to be 6, and K^{-2}, defined in Section 6.1, is taken to be 25, which are quite plausible, in so far as we wish to obtain reasonable values for proton lifetime at the upper end.

• The standard $d = 5$ operators for both MSSM and ESSM are evaluated by taking the upper limit on M_{eff} (defined in the text) that is allowed by the requirement of natural coupling unification. This requirement restricts threshold corrections and thereby sets an upper limit on M_{eff}, for a given $\tan\beta$ (see Section 6 and Appendix).

$*$ For all cases, the standard and the new $d = 5$ operators must be combined to obtain the net amplitude. For the three cases of ESSM marked with an asterisk, and other similar cases which arise for low $\tan\beta \approx 3$ to 6 (say), the standard $d = 5$ operators by themselves would lead to proton lifetimes typically exceeding (0.25-4) $\times$ 10^{34} years. For these cases, however, the contribution from the new $d = 5$ operators would dominate, which quite naturally lead to lifetimes in the range of $(10^{33} - 10^{34})$ years (see last column).

• As shown above, the case of MSSM embedded in SO(10) is tightly constrained to the point of being disfavored by present empirical lower limit on proton lifetime Eq. (42) [see discussion following Eq. (48)].

• Including contributions from the standard and the new operators, the case of ESSM, embedded in either G(224) or SO(10), is, however, fully consistent with present limits on proton lifetime for a wide range of parameters; at the same time it provides optimism that proton decay will be discovered in the near future, with a lifetime $\leq 10^{34}$ years.

• The lower limits on proton lifetime are not exhibited. In the presence of the new operators, these can typically be as low as about 10^{29} years (even for the case of ESSM embedded in SO(10)). Such limits and even higher are of course long excluded by experiments.

References

[1] The gauge symmetry $SU(2)_L \times U(1)$ was proposed by S. L. Glashow, Nucl. Phys. **22**, 57a (1961). The idea of achieving electroweak unification through a spontaneously broken $SU(2)_L \times U(1)_Y$ gauge symmetry was proposed by S. Weinberg, Phys. Rev. Lett. **19**, 1269 (1967) and A. Salam, in Elementary Particle Theory Nobel Symposium, ed. by N. Svartholm (Almqvuist, Stockholm, 1968), p. 367.

[2] The notion of global SU(3)-color symmetry was introduced by O. W. Greenberg, Phys. Rev. Lett. **13**, 598 (1964), and independently by M. Han and Y. Nambu, Phys. Rev. **139B**, 1006 (1965). That of generating (a) a fundamental "superstrong" force through an octet of gluons associates with SU(3)-color local symmetry, and (b) an *additional* fundamental "strong" force through the exchange of $(\rho, \omega, \phi, K^*)$ mesons, was introduced by Han and Nambu in the paper referenced above. In this attempt SU(3)-color was broken explicitly by electromagnetism. Up until 1972–73, there was, however, no clear idea on the origin of the fundamental strong interactions. Two considerations, pointing to the same conclusion provided a *clear choice* in this regard. The first came from initial attempts at a unification of quarks and leptons and of their three basic forces. It was realized that the *only way* to achieve such a unification is to assume that the fundamental strong force of quarks is generated *entirely* through the SU(3)-color local symmetry that commutes with flavor; the effective electroweak and strong interactions should then be generated by the *combined gauge symmetry* $SU(2)_L \times U(1)_Y \times SU(3)^c$, with ($\rho$, ω, ϕ and K^*) being composite [J. C. Pati and A. Salam; Proc. 15th High Energy Conference, Batavia, reported by J. D. Bjorken, Vol. 2, p. 301 (1972); Phys. Rev. **D8**, 1240 (1973)]. Compelling motivation for such an origin of the strong interaction came subsequently through the discovery of asymptotic freedom of non-abelian gauge theories which explained the scaling phenomena, observed at SLAC [D. Gross and F. Wilczek, Phys. Rev. Lett. **30**, 1343 (1973); H. D. Politzer, Phys. Rev. Lett. **30**, 1346 (1973)]. Some advantages of this framework were emphasized by H. Fritzsch, M. Gell-Mann and H. Leutwyler, Phys. Lett. **47B**, 365 (1973).

[3] J. C. Pati and Abdus Salam; Ref. [2]; Phys. Rev. **8**, 1240 (1973); J. C. Pati and Abdus Salam, Phys. Rev. Lett. **31**, 661 (1973); Phys. Rev.

D10, 275 (1974).

[4] H. Georgi and S. L. Glashow, Phys. Rev. Lett. **32**, 438 (1974).

[5] H. Georgi, H. Quinn and S. Weinberg, Phys. Rev. Lett. **33**, 451 (1974).

[6] H. Georgi, in Particles and Fields, Ed. by C. Carlson (AIP, NY, 1975), p. 575; H. Fritzsch and P. Minkowski, Ann. Phys. **93**, 193 (1975).

[7] The essential features pertaining to coupling unification in SUSY GUTS were noted by S. Dimopoulos, S. Raby and F. Wilczek, Phys. Rev. **D 24**, 1681 (1981); W. Marciano and G. Senjanovic, Phys. Rev. **D 25**, 3092 (1982) and M. Einhorn and D. R. T. Jones, Nucl. Phys. **B 196**, 475 (1982). For work in recent years, see P. Langacker and M. Luo, Phys. Rev. **D 44**, 817 (1991); U. Amaldi, W. de Boer and H. Furtenau, Phys. Rev. Lett. **B 260**, 131 (1991); F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, Nuov. Cim. **A 104** 1817 (1991).

[8] Y. A. Golfand and E. S. Likhtman, JETP Lett. **13**, 323 (1971). J. Wess and B. Zumino, Nucl. Phys. **B70**, 139 (1974); D. Volkov and V. P. Akulov, JETP Lett. **16**, 438 (1972).

[9] M. Green and J. H. Schwarz, Phys. Lett. **149B**, 117 (1984); D. J. Gross, J. A. Harvey, E. Martinec and R. Rohm, Phys. Rev. Lett. **54**, 502 (1985); P. Candelas, G. T. Horowitz, A. Strominger and E. Witten, Nucl. Phys. **B 258**, 46 (1985). For introductions and reviews, see: M. B. Green, J. H. Schwarz and E. Witten, "Superstring Theory" Vols. 1 and 2 (Cambridge University Press); J. Polchinski, "String Theory " vols.1 and 2 (Cambridge University Press).

[10] For a few pioneering papers on string-duality and M-theory, relevant to gauge-coupling unification, see E. Witten, Nucl. Phys. **B 443**, 85 (1995) and P. Horava and E. Witten, Nucl. Phys. **B 460**, 506 (1996). For reviews, see e.g. J. Polchinski, hep-th/9511157; and A. Sen, hep-th/9802051, and references therein.

[11] SuperKamiokande Collaboration, Y. Fukuda et. al., Phys. Rev. Lett. **81**, 1562 (1998).

[12] J. C. Pati, "*Implications of the SuperKamiokande Result on the Nature of New Physics*", in Neutrino 98, Takayama, Japan, June 98, hep-ph/9807315; Nucl. Phys. B (Proc. Suppl.) **77**, 299 (1999).

[13] Caso *et al.*, Particle Data Group, Review of Particle Physics, The European Physics Journal C, **3**, 1 (1998).

[14] K. S. Babu, J. C. Pati and F. Wilczek, "*Fermion Masses, Neutrino Oscillations and Proton Decay in the Light of the SuperKamiokande*" hep-ph/981538V3; Nucl. Phys. B (to appear).

[15] A. D. Sakharov, Pisma Zh. Eksp. Teor. Fiz. **5**, 32 (1967).

[16] V. Kuzmin, Va. Rubakov and M. Shaposhnikov, Phys. Lett **BM155**, 36 (1985).

[17] M. Fukugita and T. Yanagida, Phys. Lett. **B 174**, 45 (1986); M. A. Luty, Phys. Rev. **D 45**, 455 (1992); W. Buchmuller and M. Plumacher, hep-ph/9608308.

[18] J.C. Pati, "*With Grand Unification Signals in, Can Proton Decay Be Far Behind?*", hep-ph/0106082 (June, 2001), Talk presented at the International Summer School in Erice, Italy (September 2000), To appear in the Proceedings. This talk provides an update, relative to Ref. [14], on theoretical expectations for proton decay, and the present paper provides a further update that uses improved values of the matrix element and renormalizaton effects.

[19] N. Sakai and T. Yanagida, Nucl. Phys. **B 197**, 533 (1982); S. Weinberg, Phys. Rev. **D 26**, 287 (1982).

[20] K. S. Babu, J. C. Pati and F. Wilczek, "*Suggested New Modes in Supersymmetric Proton Decay*", Phys. Lett. **B 423**, 337 (1998).

[21] K. S. Babu and J. C. Pati, "*The Problems of Unification – Mismatch and Low α_3: A Solution with Light Vector-Like Matter*", hep-ph/9606215, Phys. Lett. **B 384**, 140 (1996).

[22] R. N. Mohapatra and J. C. Pati, Phys. Rev. **D 11**, 566, 2558 (1975); G. Senjanovic and R. N. Mohapatra, Phys. Rev. **D 12**, 1502 (1975).

[23] F. Gürsey, P. Ramond and P. Sikivie, Phys. Lett. **B 60**, 177 (1976).

[24] For recent reviews see e.g. P. Langacker and N. Polonsky, Phys. Rev. **D 47**, 4028 (1993) and references therein.

[25] See e.g. Refs. [24] and [7].

[26] P. Ginsparg, Phys. Lett. **B 197**, 139 (1987); V. S. Kaplunovsky, Nucl. Phys. **B 307**, 145 (1988); Erratum: *ibid.* **B 382**, 436 (1992).

[27] E. Witten, hep-th/9602070.

[28] For a recent discussion, see K. Dienes, Phys. Rep. **287**, 447 (1997), hep-th/9602045 and references therein; J. C. Pati, "*With Neutrino Masses Revealed, Proton Decay is the Missing Link*", hep-ph/9811442; Proc. Salam Memorial Meeting (1998), World Scientific; Int'l Journal of Modern Physics A, vol. 14, 2949 (1999).

[29] See e.g. D. Lewellen, Nucl. Phys. **B 337**, 61 (1990); A. Font, L. Ibanez and F. Quevedo, Nucl. Phys. **B 345**, 389 (1990); S. Chaudhari, G. Hockney and J. Lykken, Nucl. Phys. **B 456**, 89 (1995) and hep-th/9510241; G. Aldazabal, A. Font, L. Ibanez and A. Uranga, Nucl. Phys. **B 452**, 3 (1995); *ibid.* **B 465**, 34 (1996); D. Finnell, Phys. Rev. **D 53**, 5781 (1996); A.A. Maslikov, I. Naumov and G.G. Volkov, Int. J. Mod. Phys. **A 11**, 1117 (1996); J. Erler, hep-th/9602032 and G. Cleaver, hep-th/9604183; and Z. Kakushadze and S.H. Tye, hep-th/9605221, and hep-th/9609027; Z. Kakushadze *et al.*, hep-ph/9705202.

[30] I. Antoniadis, G. Leontaris and J. Rizos, Phys. Lett **B245**, 161 (1990); G. Leontaris, Phys. Lett. **B 372**, 212 (1996); G. Leontaris, hep-ph/9601337.

[31] A. Faraggi, Phys. Lett. **B 278**, 131 (1992); Phys. Lett. **B 274**, 47 (1992); Nucl. Phys. **B 403**, 101 (1993); A. Faraggi and E. Halyo, Nucl. Phys. **B 416**, 63 (1994).

[32] A. Faraggi and J.C. Pati, "*A Family Universal Anomalous U(1) in String Models as the Origin of Supersymmetry Breaking and Squark-degeneracy*", hep-ph/9712516v3, Nucl. Phys. **B 256**, 526 (1998). For a review and other references see A. Faraggi, hep-ph/9707311.

[33] K.S. Babu and J.C. Pati, "A Resolution of Supersymmetric Flavor-changing CP Problems Through String Flavor symmetries", UMD-PP0067, to appear.

[34] J.C. Pati, "*The Essential Role of String Derived Symmetries in Ensuring Proton Stability and Light Neutrino Masses*", hep-ph/9607446, Phys. Lett. **B 388**, 532 (1996).

[35] J.C. Pati, (Ref. [28]).

[36] N. Arkani-Hamed, S. Dimopoulos and G. Dvali, Phys. Lett. **B429**, 263 (1998); I. Antoniadis, N. Arkani-Hamed, S. Dimopoulos and G. Dvali, Phys. Lett. **B436**, 357 (1998); K. Dienes, E. Dudas, T. Gherghetta, Phys. Lett. **B436**, 55 (1998). For a recent comprehensive review of this scenario and other references, see N. Arkani-Hamed, S. Dimopoulos and G. Dvali, Physics Today, February 2002 (pages 35-40).

[37] See e.g., K.R. Dienes and J. March-Russell, hep-th/9604112; K.R. Dienes, hep-ph/9606467.

[38] M. Gell-Mann, P. Ramond and R. Slansky, in: *Supergravity*, eds. F. van Nieuwenhuizen and D. Freedman (Amsterdam, North Holland, 1979) p. 315; T. Yanagida, in: *Workshop on the Unified Theory and Baryon Number in the Universe*, eds. O. Sawada and A. Sugamoto (KEK, Tsukuba) 95 (1979); R. N. Mohapatra and G. Senjanovic, Phys. Rev. Lett. **44**, 912 (1980).

[39] S. Weinberg, I.I. Rabi Festschrift (1977); F. Wilczek and A. Zee, Phys. Lett. **70 B**, 418 (1977); H. Fritzsch, Phys. Lett. **70 B**, 436 (1977).

[40] H. Georgi and C. Jarlskog, Phys. Lett. **B 86**, 297 (1979).

[41] For a related but different SO(10) model see C. Albright, K.S. Babu and S.M. Barr, Phys. Rev. Lett. **81**, 1167 (1998).

[42] See e.g., R. Gupta and T. Bhattacharya, Nucl. Phys. Proc. Suppl. **53**, 292 (1997); and Nucl. Phys. Proc. Suppl. **63**, 45 (1998).

[43] See e.g. V. Barger, M. Berger and P. Ohman, Phys. Rev. **D47**, 1093 (1993); M. Carena, S. Pokorski and C. Wagner, Nucl. Phys. **B406**, 59 (1993); P. Langacker and N. Polonsky, Phys. Rev. **D49**, 1454 (1994); D.M. Pierce, J. Bagger, K. Matchev and R. Zhang, Nucl. Phys. **B491**, 3 (1997); K. Babu and C. Kolda, hep-ph/9811308.

[44] S. Mikheyev and A. Smirnov, Nuov. Cim. **9C**, 17 (1986); L. Wolfenstein, Phys. Rev. **D17**, 2369 (1978).

[45] SuperK Collaboration; Data on solar neutrino studies, presented by M. Vagins at WHEPP-7 Conference, Allahabad, India (January 6, 2002).

[46] S. Dimopoulos, S. Raby and F. Wilczek, Phys. Lett. **B112**, 133 (1982).

[47] J. Ellis, D.V. Nanopoulos and S. Rudaz, Nucl. Phys. **B 202**, 43 (1982).

[48] P. Nath, A.H. Chemseddine and R. Arnowitt, Phys. Rev. **D 32**, 2348 (1985); P. Nath and R. Arnowitt, hep-ph/9708469.

[49] J. Hisano, H. Murayama and T. Yanagida, Nucl. Phys. **B 402**, 46 (1993). For a recent estimate of the lifetime for the $d = 6$ gauge boson mediated $e^+\pi^0$-mode, see J. Hisano, hep-ph/0004266.

[50] K.S. Babu and S.M. Barr, Phys. Rev. **D 50**, 3529 (1994); **D 51**, 2463 (1995).

[51] For a recent work, comparing the results of lattice and chiral lagrangian-calculations for the $p \to \pi^0, p \to \pi^+$ and $p \to K^0$ modes, see N. Tatsui *et al.* (JLQCD collaboration), hep-lat/9809151.

[52] S. Aoki *et al.*, JLQCD collaboration, hep-latt/9911026; Phys. Rev. **D 62**, 014506 (2000).

[53] K. Turznyski, hep-ph/0110282, V2.

[54] J. Arafune and T. Nihei, Prog. Theor. Phys. **93**, 665 (1995).

[55] R. Dermisek, A. Mafi and S. Raby, Phys. Rev. **D63**, 035001 (2001).

[56] J.L. Feng, K.T. Matchev and T. Moroi, Phys. Rev. **D 61**, 75005 (2000), hep-ph/9909334.

[57] H.N. Brown *et al.* [Muon g-2 collaboration], hep-ex/0102017.

[58] M. Knecht and A. Nyffeler, hep-ph/0111058; M. Knecht, A. Nyffeler, M. Perrottet and E. de Rafael, Phys. Rev. Lett. **88**, 071802 (2002); M. Hayakawa and T. Kinoshita, hep-ph/0112102.

[59] For the early and a sample of few recent works on supersymmetric contribution to $(g-2)_\mu$, see T.C. Yuan, R. Arnowitt, A.H. Chamseddine and P. Nath, Z. Phys. **C2b**, 407 (1984); D.A. Kosower, L.M. Krauss and N. Sakai, Phys. Lett. **B133**, 305 (1983); A. Czarnecki and W. Marciano, hep-ph/001021222; J.L. Feng and K.T. Matchev, hep-ph/0102146; L.L. Everett, G.L. Kane, S. Rigolin and L. Wang, hep-ph/0102145; and J. Ellis, D.V. Nanopoulos and K. Olive, Phys. Lett. **B508**, 65 (2001). A

recent review and a list of other relevant references may be found in U. Chattopadhyay, A. Corsetti and P. Nath, hep-ph/0202275.

[60] J.C. Pati, Phys. Lett. **B228**, 228 (1989); K.S. Babu, J.C. Pati and H. Stremnitzer, Phys. Rev. **D 51**, 2451 (1995); K.S. Babu, J.C. Pati and X. Zhang, Phys. Rev. **D46**, 21990 (1992).

[61] K.S. Babu, J.C. Pati, "Muon g-2 Anomaly and Vector-Like Families" (To appear).

[62] SuperK Collaboration: Y. Hayato, Proc. ICHEP, Vancouver (1998); M. Earl, NNN2000 Workshop, Irvine, Calif (Feb, 2000); Y. Totsuka (private comm. May, 2001); M. Vagins, Report on SuperK Results presented at WHEPP-7 meeting, Allahabad, India (January 6, 2002).

[63] For a few recent papers showing restriction on $\tan\beta$, that follows from the limit on Higgs mass, together with certain assumptions about the MSSM parameters and/or $(g\text{-}2)_\mu$ - constraint, see e.g. R. Arnowitt, B. Dutta, B. Hu and Y. Santoso, hep-ph/0102344; J. Ellis, G. Ganis, D.V. Nanopoulos and K. Olive, hep-ph/0009355, and J. Ellis *et al.* (Ref. [59]).

[64] H. Murayama and A. Pierce, hep-ph/0108104.

[65] V. Lucas and S. Raby, Phys. Rev. **D55**, 6986 (1997); R. Darmisek, A. Mafi and S. Raby, hep-ph/0007213, V2.

[66] C. Kolda and J. March-Russell, Phys. Rev. **D 55** 4252 (1997); R. Hempfing, Phys. Lett **351**, 206 (1995); M. Bastero-Gil and B. Brahmachari, Nuc. Phys. **B575**, 35, (2000).

[67] I. Antoniadis, J. Ellis, J. S. Hagelin and D. V. Nanopoulos, Phys. Lett. **B194**, 231 (1987).

[68] N. Arkani-Hamed, S. Dimopoulos, G. Dvali and J. March-Russell, hep-ph/9811448.

[69] P. Candelas, G.T. Horowitz, A. Strominger and E. Witten, Nucl. Phys. **B258**, 46 (1985); E. Witten, Nucl. Phys. **B258**, 75 (1985);

[70] J.H. Kawai, D.C. Lewellen and S.H. Tye, Nucl. Phys. **B288**, 1 (1987); I. Antoniadis, C. Bachas and C. Kounnas, Nucl. Phys. **B289**, 87 (1987).

[71] I. Antoniadis, J. Ellis, J. Hangelin and D.V. Nanopoulos, Phys. Lett. **B231**, 65 (1989).

[72] A. Faraggi, hep-ph/0107094, Phys. Lett. **B520**, 337 (2001).

[73] L. Dixon, J.A. Harvey, C. Vafa and E. Witten, Nucl. Phys. **B261**, 6778 (1985).

[74] Y. Kawamura, Prog. Theor. Phys. **105**, 999 (2001); hep-ph/0012125.

[75] L.J. Hall and Y. Nomura, Phys. Rev. **D64**, 055003, hep-ph/0103125.

[76] G. Altarelli and F. Feruglio, Phys. Lett. **B511**, 257 (2001); hep-ph/0102301.

[77] M. Kakizaki and M. Yamaguchi, hep-ph/0104103.

[78] A. Hebecker and J. March-Russell, hep-ph/0107039; hep-ph/0204037.

[79] L.J. Hall and Y. Nomura, hep-ph/0111068.

[80] T. Asaka, W. Buchmuller and L. Covi, Phys. Lett. **B523**, 199 (2001); hep-ph/0108021.

[81] L.J. Hall, Y. Nomura, T. Okui and D. Smith, hep-ph/0108071.

[82] E. Witten, hep-ph/0201018.

[83] Z. Kakushadze, Phys. Rev. **D58**, 1010901 (1998); hep-th/9806044.

[84] G. Aldazabal, L. E. Ibanez and F. Quevedo, hep-th/9909172.

[85] C. Kokorelis, hep-th/0203187.

[86] G. I. Leontaris and J. Rizos, Phys. Lett. **B510**, 295 (2001); hep-ph/012255.

[87] G. Shiu and S. H. Tye, Phys. Rev. **D58**, 106007 (1998); hep-th/9805157.

[88] L. L. Everett, G. L. Kane, S. F. King, S. Rigolin and Lian-Tao Wang, hep-ph/0202100.

[89] A. Murayama and A. Toon, Phys. Lett. **B318**, 298 (1993).

[90] P. H. Frampton, R. N. Mohapatra and S. Suh, hep-ph/0104211.

[91] F. Paccetti Correia, M. G. Schmidt and Z. Tavartkiladze, hep-ph/0204080.

[92] S. Dimopoulos and F. Wilczek, Report No. NSF-ITP-82-07 (1981), in *The unity of fundamental interactions*, Proc. of the 19th Course of the International School on Subnuclear Physics, Erice, Italy, Erice, Italy, 1981, Plenum Press, New York (Ed. A. Zichichi); K.S. Babu and S.M. Barr, Phys. Rev. **D 48**, 5354 (1993).

[93] It has recently been pointed out by K. S. Babu and S. Barr (hep-ph/0201130) that one can achieve doublet-triplet splitting in SO(10), by having only a single 45_H with a $VEV \propto I_{3R}$; and this can be done in a manner that can eliminate the $d = 5$ proton decay operator. In this case, however, the group-theoretic correlation between the suppression of V_{cb} and the enhancement of $\theta^{\rm osc}_{\nu_\mu \nu_\tau}$, which becomes a compelling feature if $\langle 45_H \rangle \propto$ B–L (see discussion in Section 5), does not emerge.

Table 1: Proton lifetime, based on contributions from only the standard operators for the case of ESSM embedded in SO(10), with parameters being in the "median" range.

$\tan\beta = 3$ [S]=3 {C}=1/2 to 4	$\tan\beta = 3$ [S]=6 {C}=1/2 to 1	$\tan\beta = 5$ [S]=5.4 {C}=1 to 6	$\tan\beta = 5$ [S]=6 {C}=1 to 4
$\Gamma^{-1}(\bar{\nu}K^+)^{std}_{ESSM} \approx$ $(1.2 \text{ to } 10) \times 10^{33}$ yrs	$\Gamma^{-1}(\bar{\nu}K^+)^{std}_{ESSM} \approx$ $(2.5 \text{ to } 5) \times 10^{33}$ yrs	$\Gamma^{-1}(\bar{\nu}K^+)^{std}_{ESSM} \approx$ $(1.6 \text{ to } 10) \times 10^{33}$ yrs	$\Gamma^{-1}(\bar{\nu}K^+)^{std}_{ESSM} \approx$ $(1.8 \text{ to } 7.3) \times 10^{33}$ yrs

Table 2: Values of proton lifetime $(\Gamma^{-1}(p \to \bar{\nu} K^{+}))$for a wide range of parameters.

Parameters (spectrum/Matrix element)	MSSM → SO(10) **Std. d=5** Intermed. ϵ' & phase†		ESSM → SO(10) **Std. d=5** Intermed. ϵ' & phase†		$\left\{\begin{array}{c}\text{MSSM}\\ \text{or}\\ \text{ESSM}\end{array}\right\} \to$ G(224)/SO(10) **New d=5**††
	$\tan\beta$=3	$\tan\beta$=10	$\tan\beta$=5	$\tan\beta$=10	Independent of $\tan\beta$
Nearly "central" {C}=2	0.2×10^{32} yrs	$1.6{\times}10^{30}$ yrs	$0.25{\times}10^{34}$ yrs*	$0.7{\times}10^{33}$ yrs	0.50×10^{33} yrs††
Intermediate {C}=8	$0.7{\times}10^{32}$ yrs	$0.6{\times}10^{31}$ yrs	$1{\times}10^{34}$ yrs*	$2.8{\times}10^{33}$ yrs	$2{\times}10^{33}$ yrs††
Nearly Extreme {C}=32	$0.3{\times}10^{33}$ yrs	$2.6{\times}10^{31}$ yrs	$4{\times}10^{34}$ yrs*	$1.1{\times}10^{34}$ yrs	$8{\times}10^{33}$ yrs††

$*$**In this case, lifetime is given by the last column.**

Introduction to Noncommutative Field Theory

J.L.F. Barbón[*]

Theory Division, CERN, Switzerland

Lectures given at the
Summer School on Particle Physics
Trieste, 18 June - 6 July 2001

LNS0210004

[*] barbon@cern.ch

Abstract

These Lecture Notes give an intuitive introduction to noncommutative field theory with an emphasis on the physics ideas and methods. We pay special attention to those aspects of noncommutative field theory that represent genuine novelties from the physical point of view, such as the UV/IR mixing. We also include brief discussions of possible applications of these ideas to phenomenology as well as the connection to string theory.

Contents

1 Introduction

Noncommutative Field Theory (NCFT) is a field theory defined over a space-time endowed with a Noncommutative Geometry (NCG) (c.f. [1, 2, 3]).

Although the motivations for considering NCG are mostly mathematical, aspects of the formalism do show up in various physical situations and, in priciple, it is a relevant generalization of the standard framework of local quantum field theory. Indeed, the existence of a nonlocal, and yet tractable, generalization of quantum field theory is a highly non-trivial fact of great intrinsic interest. This is not only linked to interesting mathematics but it is also related to the non-locality present in string theory [4].

In this vein, the recent discovery of subtle quantum mechanical effects in NCFT, having to do with the interplay between locality and renormalization (c.f. [5]), has prompted a wide interest in NCFT as a toy model for the most widely studied nonlocal theory: string theory. Other potential applications of the formalism to the study of large-N limits of ordinary gauge theories (c.f. [2, 6, 7]), as well as the Quantum Hall Effect [8], only add to the interest of these ideas.

Here we give a very basic introduction to NCFT, emphasizing the physical methods and motivations, at the price of being considerably sloppy on the mathematical niceties of the subject. Other reviews with a much more comprehensive scope exist. See for example [9]. Reviews with a more mathematical outlook are for example [10, 11].

In preparing these notes, no attemp has been made of giving a careful set of references. Rather complete sets of references can be found in the reviews just quoted. In the text, we will only refer explicitly to some works that are particularly relevant to the discussion.

1.1 Noncommutative Geometry

Intuitively, NCG is the generalization of standard geometry ideas, such as manifolds, metrics and fiber bundles, to spaces where the “coodinates” are operators rather than c-numbers. In particular, they do not commute, but satisfy some operator algebra

$$[\hat{x}^i, \hat{x}^j] = C^{ij}(\hat{x}). \tag{1}$$

It is useful to think of the operators $\hat{x}^k$ as “generators” of an algebra $\mathcal{A}$, in the sense that the general element of $\mathcal{A}$ can be thought of as a function of the basic variables, $f(\hat{x})$, satisfying certain constraints. In this case, the functions $C(\hat{x})$ acquire the interpretation of "structure functions", generalizations of the notion of structure constants for ordinary Lie algebras. The basic idea of the development of NCG is then the recovery of geometrical notions about the "base space" (the space parametrized by the "coordinates" $\hat{x}_k$) in terms of the algebra $\mathcal{A}$ of functions on that space, where this algebra is required to be associative but in general non-commutative.

1.1.1 Examples

Rather than developing these ideas in full generality, here we collect some simple examples that are motivated by the applications of the formalism to physics.

- The trivial example is $C_k^{ij} = 0$, a commutative algebra. Then $\mathcal{A}$ is the algebra of (say smooth) functions $C(M)$ on the base manifold parametrized by the c-numbers x^k.

- The next example in order of triviality is when the noncommutative algebra is a direct product of a commutative algebra and a *finite-dimensional* noncommutative algebra, such as some Lie algebra $\mathcal{G}$:

$$\mathcal{A} = C(M) \otimes \mathcal{G}. \tag{2}$$

  This is the case of ordinary gauge theory, where fields are just matrix-valued functions.

- Another simple, albeit somewhat exotic example is the "fuzzy sphere". If we define $\mathbf{S}^2$ as the solution in $\mathbf{R}^3$ of

$$x_1^2 + x_2^2 + x_3^2 = R^2, \tag{3}$$

  the obvious definition of the fuzzy sphere whould be in terms of three non-commuting operators $\hat{x}_1, \hat{x}_2, \hat{x}_3$ that satisfy

$$\hat{x}_1^2 + \hat{x}_2^2 + \hat{x}_3^2 = R^2\,\mathbf{1}, \tag{4}$$

  with $\mathbf{1}$ the unit operator of the algebra. An obvious choice is

$$\hat{x}_a = \frac{R}{\sqrt{j(j+1)}}\,J_a, \tag{5}$$

where J_a are $SU(2)$ angular momenta in the spin-j representation. Hence, this is the particular choice

$$C_{ab}(\hat{x}) = i\,\frac{R}{\sqrt{j(j+1)}}\,\sum_c \epsilon_{abc}\,\hat{x}_c \tag{6}$$

for the structure "constants". Notice that the resulting operator space respects the $SO(3)$ isometry that characterizes $\mathbf{S}^2$. The space is "discrete" in some sense, because the spectrum of eigenvalues of any position operator $\hat{x}_a$ has dimension $2j+1$. So, it looks like some kind of "lattice approximation" to $\mathbf{S}^2$. Strictly speaking, we cannot build a quantum field theory with an infinite number of degrees of freedom on such space. In the limit $j \to \infty$ at fixed R, the number of degrees of freedom does diverge, but then we recover the commutative algebra of functions on $\mathbf{S}^2$.

- In the previous examples, the noncommutative character of $\mathcal{A}$ was "finite-dimensional", which leads to somewhat trivial examples. The next step in complexity is to regard $\hat{x}_j$ as operators represented in some infinite-dimensional Hilbert space, with continuous spectrum, *i.e.* we want to regard their eigenvalues are parametrizing standard flat space $\mathbf{R}^d$. Then the simplest choice of structure constants is a simple central extension:

  $$[\hat{x}^j, \hat{x}^k] = i\,\theta^{jk}, \tag{7}$$

  with θ^{jk} an antisymmetric matrix of constants with length-dimension two. This defines noncommutative flat space or $\mathbf{R}^d_\theta$, and an obvious restriction to periodic angular coordinates defines the noncommutative torus $\mathbf{T}^d_\theta = \mathbf{R}^d_\theta/\mathbf{Z}^d$.

1.2 Examples from Physics

NCG may arise in physical systems when some *effective* position operator becomes non-commutative as a result of interactions.

$$\left[\hat{X}^\mu_{\text{eff}}, \hat{X}^\nu_{\text{eff}}\right] \neq 0. \tag{8}$$

This involves typically non-relativistic systems in first-quantization and the non-commutatitivity of the position operator may or may not vanish in the classical limit $\hbar \to 0$. We will illustrate this with two examples: electrons in a strong magnetic field and D-branes.

1.2.1 Electrons in a Strong Magnetic Field

Let us consider planar electrons in a strong uniform magnetic field B_{ij}, with Hamiltonian

$$H = \frac{1}{2m_e}\left(\vec{p} - e\vec{A}\right)^2, \tag{9}$$

where

$$A_i = -\frac{1}{2} B_{ij}\, x^j \tag{10}$$

in an appropriate gauge. Defining

$$z = \sqrt{\frac{e|B|}{2\hbar}}\,(x+iy) \tag{11}$$

and the operators

$$a = \partial_{\bar{z}} + \frac{z}{2}, \qquad a^\dagger = -\partial_z + \frac{\bar{z}}{2}, \tag{12}$$

one finds a harmonic oscillator system

$$[a,a] = [a^\dagger, a^\dagger] = 0, \qquad [a, a^\dagger] = 1, \tag{13}$$

and the Hamiltonian

$$H = \hbar\,\omega_c \left(a^\dagger\, a + \frac{1}{2}\right), \tag{14}$$

with spectrum $E_\ell = \hbar\omega_c(\ell + \frac{1}{2}), \ell \in \mathbf{Z}$, where $\omega_c = e|B|/m_e$ denotes the cyclotron (Larmor) frequency. Each energy (Landau) level has an infinite degeneracy; the ground states satisfy:

$$a\,\psi(z,\bar{z}) = \left(\partial_{\bar{z}} + \frac{z}{2}\right)\psi(z,\bar{z}) = 0. \tag{15}$$

A basis of the lowest Landau level (LLL) can be chosen as

$$\psi_m(z,\bar{z}) = \frac{z^m}{\sqrt{m!}}\, e^{|z|^2/2}. \tag{16}$$

We can concentrate on the LLL wave functions if the magnetic field is large enough, so that mixing with the higher Landau levels is suppressed by the high cyclotron frequency gap. The interesting feature of the LLL wave functions is that they are almost analytic. We can consider analytic functions $v_m(z)$ by stripping off the exponential term:

$$v_m(z) \equiv e^{|z|^2/2}\,\psi_m(z,\bar{z}). \tag{17}$$

If we further define a specific inner product on the LLL:

$$(v_n|v_m) \equiv \int d\mu(z,\bar{z})\, \bar{v}_n(\bar{z})\, v_m(z) = \langle\psi_n|\psi_m\rangle \tag{18}$$

with the non-holomorphic exponential term in the measure:

$$d\mu(z,\bar{z}) = e^{-|z|^2} dz d\bar{z}, \tag{19}$$

then we have, integrating by parts:

$$(f|\partial_z|g) = (f|\bar{z}|g), \tag{20}$$

so that, on the LLL:

$$(\partial_z)_{LLL} = (\bar{z})_{LLL}. \tag{21}$$

Hence, $[\partial_z, z] = 1$ implies

$$[\bar{z}, z]_{LLL} = 1 \tag{22}$$

or, back to the original variables

$$[\hat{x}, \hat{y}]_{LLL} = i\,\theta_B, \qquad \theta_B = \frac{\hbar}{e|B|}. \tag{23}$$

Thus, the motion of electrons in the lowest Landau level is effectively described by a noncommutative plane. NCG is relevant to the physics of the Quantum Hall Effect.

It is worth deriving this result in a more heuristic fashion, using a Lagrangian argument. The Lagrangian of the system is

$$L = \frac{1}{2} m_e \dot{\vec{x}}^2 - \frac{e}{2}\, B_{ij}\, x^i\, \dot{x}^j. \tag{24}$$

In a situation where the kinetic energy term is negligible $|m_e \dot{x}^i| \ll |B_{ij} x^j|$, we may approximate the dynamics by the degenerate Lagrangian

$$L \approx -\frac{e}{2} B_{ij} x^i\, \dot{x}^j. \tag{25}$$

The canonical momenta are proportional to the coordinates themselves:

$$\pi_j = \frac{dL}{d\dot{x}^j} = -e B_{jk} x^k. \tag{26}$$

Upon canonical quantization

$$[\hat{\pi}_j, \hat{x}^l] = -i\hbar\, \delta^l_j = -e\, B_{jk} [\hat{x}^k, \hat{x}^l], \tag{27}$$

and finally:

$$[\hat{x}^k, \hat{x}^l] = i\hbar \left(\frac{1}{e|B|}\right)^{kl}. \tag{28}$$

1.2.2 D-branes

Dp-branes are specific states of string theory that resemble non-relativistic solitons extended in p spatial dimensions [12]. For the case of D-particles, their low-energy dynamics is primarily characterized by the position collective coordinates. For a system of *distant* N D-particles, we have a collection N vectors of positions $\vec{x}_i, i = 1, \dots, N$. When the D-particles' separation is in the stringy domain, $|\vec{x}_i - \vec{x}_j| < \ell_s$, with ℓ_s the string length, new light degrees of freedom appear, corresponding to open strings stretched between neighboring D-particles. Therefore, the number of collective coordinates is enlarged to N^2 and we may assemble them into a hermitian matrix $\mathbf{X}_{ij}$.

Thus, the notion of positon becomes "fuzzy" at short distances. An operational definition of the i-th particle position is

$$\langle \mathbf{X} \rangle_i \equiv \langle i | \mathbf{X} | i \rangle = \mathbf{X}_{ii}. \tag{29}$$

With this definition, any non-diagonal matrix of collective coordinates assigns a nonvanishing dispersion to the possition of the i-th particle:

$$(\Delta \mathbf{X})_i^2 = \langle \mathbf{X}^2 \rangle_i - \langle \mathbf{X} \rangle_i^2 = \sum_{j \neq i} |\mathbf{X}_{ij}|^2 \geq 0. \tag{30}$$

Once the positions are promoted to a matrix, the statistical permutation group of N particles, S_N, is naturally promoted to $U(N)$, whose Weyl subgroup is precisely S_N.

In fact, for a one-dimensional system we just have a single "position matrix" and we can always agree to define the positions in terms of the eigenvalues of this matrix. Starting with two spatial dimensions we have more than one position matrix and it is not possible to diagonalize all of them in the same basis, unless they commute. In D-brane theory, this condition is selected dynamically by the minima of the static interaction potential of a system of D-particles:

$$V(\mathbf{X}) = -\frac{1}{g_s \ell_s} \sum_{a,b=1}^{d} \left[X^a, X^b \right]^2. \tag{31}$$

Thus, in this case the noncommutativity survives the classical limit of the theory. In fact, taking into account the "statistical symmetry" $U(N)$ we are just constructing a $U(N)$ gauge theory with Higgs fields in the adjoint representation, and interpreting the expectation values of these scalar fields as generalized position coordinates of the soliton. Thus, from the point of view of the earlier list of simple NCG examples, the D-branes represent the noncommutative algebra $C(M) \times U(N)$. Notice however that here M is only the world-volume of the D-brane, whereas the space transverse to the D-brane is constructed out of the matrix degrees of freedom, via the Higgs fields X^a in the adjoint of $U(N)$.

In certain situations, the interaction potential depends on a background field through a "dielectric coupling" [13]:

$$\delta V(\mathbf{X}) = if\,\epsilon^{abc}\,\mathrm{tr}\,X_a X_b X_c. \tag{32}$$

In this example, it depends on a single constant parameter f and we take $d = 3$. The equations of motion become

$$\Big[[X^a, X^b], X_b\Big] + if\,\epsilon^{abc}\,[X_b, X_c]. \tag{33}$$

Although commuting (diagonal) matrices are still a solution, we see that the fuzzy sphere (6) is a solution with

$$X_a = f\,J_a, \tag{34}$$

and J_a in the spin-j representation of $SU(2)$.

2 Noncommutative Field Theory

In constructing NCFT we go one step further. As in the D-brane example, the underlying NCG is taken as a passive "arena", or background choice, for the dynamics, but we formally generalize the noncommutativity to infinite matrices, i.e. operator algebras. In these lectures we concentrate on the simple example of $\mathbf{R}^d_\theta$. The nontrivial structure

$$[\hat{x}^j, \hat{x}^k] = i\,\theta^{jk} \tag{35}$$

can be interpreted by regarding $\hat{x}^k$ as phase-space variables represented on a Hilbert space $\mathcal{H}_\theta$. This Hilbert space has nothing to do with the standard Quantum Hilbert space $\mathcal{H}_\hbar$ that arises upon quantization. In fact, $\mathcal{H}_\theta$ is part of the specification of the classical field theory, i.e. the classical field configurations are functions $\phi(\hat{x}^k)$ on the algebra of operators $\mathcal{A}_\theta$ that are represented on $\mathcal{H}_\theta$.

It is clear that such a structure imposes a physical nonlocality on length scales of $\mathcal{O}(\sqrt{\theta})$. There is a minimal area unit of $\mathcal{O}(\theta)$ in the sense of the Heisenberg uncertainty relation:

$$\Delta x^j \, \Delta x^k \geq \frac{1}{2} |\theta^{jk}|. \tag{36}$$

Thus, we may hope that $\sqrt{\theta}$ is an interesting physical cutoff in quantum field theory, presumably with interesting applications to the quantum gravity realm. Meanwhile, if space-time satisfies (35) at short distances, the most characteristic hint at low energies would be the short-distance breakdown of Lorentz invariance, a very well-tested symmetry.

2.1 Elementary Construction of Classical NCFT

For simplicity, we begin with a single noncommutative plane with coordinates x, y satisfying

$$[\hat{x}, \hat{y}] = i\,\theta. \tag{37}$$

We consider the standard representation on "wave functions" on $L^2(\mathbf{R})$. The operator $\hat{x}$ is diagonal and represented multiplicatively, whereas $\hat{y}$ is the corresponding 'conjugated momentum':

$$\hat{x}\,\psi(x) = x\,\psi(x), \qquad \hat{y}\,\psi(x) = -i\theta\,\partial_x\,\psi(x). \tag{38}$$

We have then the standard operator identities:

$$e^{ip\hat{y}}\, f(\hat{x}) = f(\hat{x} - p\theta)\, e^{ip\hat{y}}, \tag{39}$$

so that $\hat{y}$ generates translations of $\hat{x}$ eigenvalues. Straightforward application of the Baker–Campbell–Hausdorff formula yields the plane-wave composition rule:

$$e^{ip_\mu \hat{x}^\mu}\, e^{iq_\mu \hat{x}^\mu} = e^{-\frac{i}{2} p\times q}\, e^{i(p+q)_\mu \hat{x}^\mu}, \tag{40}$$

where we have returned to a general $\theta^{\mu\nu}$ matrix and defined

$$p \times q \equiv p_\mu\, \theta^{\mu\nu}\, q_\nu. \tag{41}$$

A convenient way of manipulating the operator algebra is to map it to some deformed function algebra. This in turn allows a much more intuitive development of the physical set up for NCFT.

The basic idea is to work with the "components" of the operators in a conventionally chosen basis. This is the infinite-dimensional generalization of the standard choice of a basis in a finite-dimensional $U(N)$ Lie algebra:

$$A = \sum_{a=1}^{N^2} A^a \, T^a. \tag{42}$$

In this case, we say that the hermitian matrix A has "vector components" A^a in the basis of generators $\{T^a\}$. For a general operator $\hat{O}$ acting on $\mathcal{H}_\theta$ the "vector of components" in a given basis is in general a function of a continuous label $f_{\hat{O}}(x^\mu)$. This establishes a map from the operator algebra to the space of ordinary functions:

$$\hat{O} = \int d^d x \, f_{\hat{O}}(x^\mu) \, \hat{T}_{x^\mu}, \tag{43}$$

where $\hat{T}_{x^\mu}$ is a basis of the operator algebra. Associated to this choice of basis, there is a representation of the operator product "in components". This is a product in the space of component functions, the "star product", defined by the identity:

$$f_{\hat{O}\hat{O}'}(x) = f_{\hat{O}}(x) \star f_{\hat{O}'}(x). \tag{44}$$

For illustrative purposes, it is interesting to work out the star product in the finite-dimensional $U(N)$ Lie algebra. In a conventional basis of generators T^a we have

$$T^a \, T^b = \sum_c C_c^{ab} \, T^c \tag{45}$$

for some constants C_c^{ab}. Given two hermitian matrices $A = \sum_a A_a T^a, B = \sum_a B_a T^a$, the product can be written as

$$AB = \sum_{a,b} A_a \, B_b \, \sum_c C_c^{ab} \, T^c = \sum_c (AB)_c \, T^c. \tag{46}$$

So that the definition of "star product" is simply

$$A_c \star B_c \equiv (AB)_c = \sum_{a,b} C_c^{ab} \, A_a \, B_b. \tag{47}$$

A convenient choice for $\mathbf{R}^d_\theta$ is the so-called Weyl map, defined by choosing the operator basis as

$$\hat{T}_{x^\mu} = \int \frac{d^d k}{(2\pi)^d} \, e^{ik\cdot(\hat{x}-x)}, \tag{48}$$

with inverse

$$f_{\hat{O}}(x^\mu) = \int \frac{d^d k}{(2\pi)^d} \,\mathrm{Tr}\, e^{ik_\mu (x-\hat{x})^\mu}\, \hat{O}(\hat{x}^\mu). \tag{49}$$

The specific property of the Weyl map that makes it useful is that the plane-wave operator

$$\exp(ip\cdot\hat{x}) \tag{50}$$

is associated to the plane-wave function

$$\exp(ip\cdot x). \tag{51}$$

In particular, the composition law (44) holds for the star product of the component functions:

$$e^{ip\cdot x} \star e^{iq\cdot x} = e^{-\frac{i}{2}p\times q}\, e^{i(p+q)\cdot x}. \tag{52}$$

A general expression for arbitrary functions can be obtained by simple superposition of plane waves:

$$f(x) \star g(x) = f(x) \exp\left(\frac{i}{2}\overleftarrow{\partial}_\alpha\, \theta^{\alpha\beta}\, \overrightarrow{\partial}_\beta\right) g(x). \tag{53}$$

This associative, but noncommutative product is known as the Moyal product of functions. In this language, NCG amounts to a smooth deformation of the classical algebra of functions, i.e. we just change the composition rules, but not the elements of the algebra.

Since $\mathcal{A}_\theta$ can be viewed as a deformation of the ordinary algebra of functions on $\mathbf{R}^d$, we can construct NCFT by deforming action functionals in a straightforward way.

Therefore, a prescription to construct NCFT's is to exercise a "correspondence principle" in terms of the noncommutativity deformation parameter θ^{jk}: one just replaces ordinary products by Moyal products all over the place, i.e. for a scalar field:

$$S[\phi] = \int d^d x \left(\frac{1}{2}\partial_\mu\phi\star\partial^\mu\phi - \frac{1}{2}m^2\,\phi\star\phi - \frac{\lambda}{4!}\,\phi\star\phi\star\phi\star\phi\right). \tag{54}$$

An important property of any such action is the cyclic property of the Moyal product inside integrals:

$$\int d^d x\, f(x)\star g(x)\star h(x) = \int d^d x\, g(x)\star h(x)\star f(x), \tag{55}$$

provided one can neglect boundary terms at infinity. In particular, under the same conditions one can remove *one* Moyal product inside integrals:

$$\int d^d x\, f(x) \star g(x) = \int d^d x\, f(x)\, g(x). \tag{56}$$

As with any correspondence principle, the noncommutativity of products implies some ambiguities in translating actions. For example, for fields with indices, the interaction term

$$\int d^d x\ \phi_i \star \phi^i \star \phi_j \star \phi^j \tag{57}$$

is *not equivalent* to

$$\int d^d x\ \phi_i \star \phi_j \star \phi^i \star \phi^j. \tag{58}$$

2.2 Noncommutative Gauge Theories

The previous construction of a classical scalar field theory admits straightforward generalizations to other theories with polynomial interactions involving fermions and scalars with Yukawa-type couplings. Special features arise in the case of gauge fields.

Starting from an ordinary gauge theory based on a Lie group G, the naive correspondence principle yields

$$S_{\rm NCYM} = -\frac{1}{4g^2} \int d^d x\ {\rm tr}\ F_{\mu\nu} \star F^{\mu\nu}, \tag{59}$$

where

$$F_{\mu\nu} = \partial_\mu\, A_\nu - \partial_\nu\, A_\mu + i\, A_\mu \star A_\nu - i\, A_\nu \star A_\mu, \tag{60}$$

with infinitesimal gauge transformations acting as

$$\delta A_\mu = D_\mu\, \epsilon = \partial_\mu\, \epsilon + i\, A_\mu \star \epsilon - i\, \epsilon \star A_\mu. \tag{61}$$

Notice that taking the gauge fields valued on the standard Lie algebra of G is not in general consistent with the noncommutative deformation. Let us write $A(x) = \sum_a A_a(x) T^a$ for some basis of $\mathcal{G}$. Then the non-commutative character of the Moyal product implies that gauge transformations depend on the anticommutator $\{T^a, T^b\}$, together with the usual commutator terms $[T^a, T^b]$. In general, the anticommutator of two generators belongs to the Lie algebra only in the case of $U(N)$ in the fundamental representation. Thus, the discussions of NCYM theories are normally restricted to $U(N)$ groups. In principle, other options are possible (such as working with the univeral enveloping algebra that contains the products of all generators [14]) at the price of working with a theory whose degree of non-locality is considerably *larger* than that implied by the Moyal product.

Restricting the generators to the fundamental representation of $U(N)$ yields further constraints on the possible matter representations. These are restricted to the adjoint, Ψ, the fundamental, ψ, and the antifundamental, $\bar{\psi}$. If $g(x)$ is an $N \times N$ matrix-valued function satisfying

$$g(x) \star g(x)^{\dagger} = g(x)^{\dagger} \star g(x) = \mathbf{1}, \tag{62}$$

the finite gauge transformations are

$$A_{\mu} \to g \star (A_{\mu} - i\,\partial_{\mu}) \star g^{\dagger}, \qquad \Psi \to g \star \Psi \star g^{\dagger}, \qquad \psi \to g \star \psi, \qquad \bar{\psi} \to \bar{\psi} \star g^{\dagger}. \tag{63}$$

An important property that follows from these expressions is the non-existence of naive *local* gauge-invariant operators, i.e. $F^2 \to g \star F^2 \star g^{\dagger}$ under gauge transformations, but in order to cancel out $g(x)$ against $g(x)^{\dagger}$ we need to use the cyclic property of the trace. Since the "trace" for the Moyal product includes the ordinary integral, we conclude that standard local operators must be integrated over in order to remain gauge-invariant after the noncommutative deformation.

On can do slightly better and define quasi-local operators by using the so-called "open Wilson lines". Consider a Wilson line operator associated to the path γ_x with initial point x:

$$W(\gamma_x) = P_{\star} \exp\left(i \int_{\gamma} A\right), \tag{64}$$

with $P_{\star}$ denoting the instruction of path-ordering with respect to the Moyal product. Then, given any local operator $\mathcal{O}(x)$, formally constructed out of the field strength and covariant derivatives, the noncommutative Fourier transform

$$\tilde{\mathcal{O}}(k) = \int d^d x \;\mathrm{tr}\; \mathcal{O}(x) \star W(\gamma_x) \star e^{ikx} \tag{65}$$

is gauge-invariant provided the endpoint of the path γ_x lies at the point $x^{\mu} + k_{\nu}\theta^{\mu\nu}$ (c.f. for example [15]).

There are some interesting consequences of these algebraic restrictions when considering the rank-one NCYM theory, i.e. the one based on "$U(1)$

gauge fields". Notice that this theory contains non-linear interactions much like any other non-abelian gauge theory, such that the theory becomes free in the classical commutative limit $\theta \to 0$. Once the Yang–Mills coupling e is fixed, the restriction on the matter representations implies that the charge of the matter fields cannot be adjusted further. Thus, the $U(1)$ charge assignments of the Standard Model are not easily implemented in a noncommutative deformation (c.f. for example [16], and [17] for alternative constructions).

2.3 Perturbative Quantization

In carrying the quantization of the *classical* theory (54) we may proceed with a formal canonical quantization provided $\theta^{0i} = 0$. Otherwise, the infinite number of time derivatives in the action makes the canonical program rather ackward.

An alternative is to write down a formal path integral

$$Z[J] = \int d\mu[\phi]\; e^{iS[\phi]}\; e^{i\int d^dx\, J\star\phi} \tag{66}$$

with some specification of the integration measure. For the time being, we will restrict ourselves to the perturbative evaluation of $Z[J]$.

The crucial observation is that the free approximation is locally θ-independent:

$$S[\phi]_{\text{free}} = \frac{1}{2}\int d^dx\; \Big(\partial_\mu\phi\star\partial^\mu\phi - m^2\phi\star\phi\Big) = \frac{1}{2}\int d^dx\; \Big(\partial_\mu\phi\,\partial^\mu\phi - m^2\phi^2\Big)\,. \tag{67}$$

Therefore, in evaluating perturbation-theory integrals, we can consider the standard Gaussian θ-independent measure. This prescription gives a set of Feynman rules. We have standard propagators

$$\frac{i}{p^2 - m^2 + i\,0}, \tag{68}$$

and non-standard interaction terms. Upon Fourier transformation:

$$\int d^dx\; \phi(x)\star\ldots\star\phi(x) = \int d^dp\,(2\pi)^d\,\delta(\textstyle\sum p)\;\tilde\phi(p_1)\ldots\tilde\phi(p_n)\;W(p_1,\ldots,p_n), \tag{69}$$

where

$$W(p_1,\ldots,p_n) = \exp\left(-\frac{i}{2}\sum_{i<j} p_i\times p_j\right) \tag{70}$$

is the so-called Moyal phase. Thus, we are led to a simple Feynman rule for the interactions. We just need to "decorate" the standard Feynman vertex with the non-local Moyal phase:

$$-i\,\lambda_n \longrightarrow -i\,\lambda_n\,W(p_1,\dots,p_n). \tag{71}$$

Notice that the Moyal phase spoils the Bose symmetry of the vertex, the noncommutative vertex being only cyclically symmetric. This modifies the symmetry factors associated to the Feynman rules.

Since the vertices written as in (71) are only cyclically symmetric, they satisfy the same topological properties as planar vertices in 't Hooft's double line notation for gauge-theory Feynman rules [18]. Thus, diagrams in noncommutative field theories admit a similar topological classification by the genus of the surface on which they can be drawn.

Using simple topological arguments, plus momentum conservation at each vertex, one can prove a general result regarding the θ-dependence of the Feynman diagram integrands: the class of *planar* diagrams has a θ-dependence saturated by the external legs, i.e. the overall Moyal phase of the diagram with a given set of external legs equals the phase of a single-vertex diagram with the same external legs (c.f. [19]).

For *nonplanar* diagrams, the θ-dependence remains in non-trivial phases in the integrand. Nonplanar loop integrations are then sensitive to the Moyal phases.

2.3.1 Two Examples

Having noticed that the bosonic Feynman vertices are not Bose-symmetric in general, it is still useful in practice to symmetrize them in order to manipulate them in a standard fashion, without paying special attention to the different topological classes of diagrams. We can illustrate this with two examples.

Consider first ϕ^3 theory. The vertex can be obtained directly by considering the Moyal product of two plane-waves

$$\phi(p_1)e^{ip_1 x} \star \phi(p_2)\,e^{ip_2 x} = \phi(p_1)\phi(p_2)\;e^{-\frac{i}{2}p_1\times p_2}\;e^{i(p_1+p_2)x}. \tag{72}$$

Since the momentum variables p_1, p_2 are integrated over in writing the interaction action, they are dummy variables can be switched over. So we can symmetrize the Moyal product above and write

$$\phi(p_1)\phi(p_2)\,\cos\left(\frac{p_1\times p_2}{2}\right)\;e^{i(p_1+p_2)x}. \tag{73}$$

Therefore, we can use the Feynman rule

$$\text{Vertex } = -i\,\lambda\,\cos(p_1\times p_2/2), \tag{74}$$

where Bose symmetry is restored. Consider now the one-loop contribution to the two-point function. It contains a factor of $\cos^2(p_1\times p_2/2)$ from the vertices. The two structures, planar and non-planar, arise upon writing:

$$\cos^2\left(\frac{p_1\times p_2}{2}\right)=\frac{1}{2}+\frac{1}{2}\cos\left(p_1\times p_2\right). \tag{75}$$

The first term, θ independent, yielding the planar part.

A second example of the same nature involves the Feynman rules of a $U(N)$ NCYM theory. Let us write for the plane-wave field:

$$A_\mu(x)=\sum_{a=1}^{N^2}A^a_\mu(p)\,T^a\,e^{ipx} \tag{76}$$

and reduce the commutator:

$$[A_\mu,A_\nu]_\star=\frac{1}{2}\sum_{a,b}\{T^a,T^b\}\,[A^a_\mu,A^b_\nu]_\star+\frac{1}{2}\sum_{a,b}[T^a,T^b]\,\{A^a_\mu,A^b_\nu\}_\star \tag{77}$$

Defining now the usual symmetric and antisymmetric tensor structures:

$$[T^a,T^b]=i\sum_c f^{abc}\,T^c,\qquad \{T^a,T^b\}=\sum_c d^{abc}\,T^c, \tag{78}$$

one obtains

$$\begin{aligned}[A_\mu,A_\nu]_\star &= \sum_c\Big(i\,d^{abc}\,T^c\sin\left(p_1\times p_2/2\right)+i\,f^{abc}\,T^c\cos\left(p_1\times p_2/2\right)\Big)\\ &\qquad\qquad\times A^a_\mu(p_1)\,A^b_\nu(p_2)\;e^{i(p_1+p_2)x}\end{aligned} \tag{79}$$

It follows that the Feynman rule for a $U(N)$ noncommutative gauge theory can be constructed from the Feynman rule of the ordinary $SU(N)$ theory by the substitution of the structure constants:

$$f^{abc}\longrightarrow f^{abc}\cos\left(\frac{p_a\times p_b}{2}\right)+d^{abc}\sin\left(\frac{p_a\times p_b}{2}\right). \tag{80}$$

where now the group indices a,b,c include also the diagonal $U(1)$ subgroup of $U(N)$. For example, the noncommutative rank-one theory, $U(1)$, has a three-point coupling of the photon given by

$$\begin{aligned}V_{\gamma\gamma\gamma} &= -2g\sin\left(\frac{p_1\times p_2}{2}\right)\,\left[(p_1-p_2)^{\mu_3}\,\eta^{\mu_1\mu_2}+(p_2-p_3)^{\mu_1}\,\eta^{\mu_2\mu_3}\right.\\ &\qquad\left.+(p_3-p_1)^{\mu_2}\,\eta^{\mu_1\mu_3}\right].\end{aligned} \tag{81}$$

2.3.2 Asymptotically Free Photons

As an example of the peculiar new features introduced by noncommutativity we make a heuristic discussion of a surprising fact: the rank-one noncommutative Yang–Mills theory (pure noncommutative photons) is asymptotically free (see for example [20]). According to the previous paragraph, the perturbative structure of this theory is rather similar to that of $SU(N)$ Yang–Mills theory in the limit $N \to 1$. The perhaps surprising fact is that a characteristic dynamical feature such as asymptotic freedom does survive in the limit.

Consider the ordinary $SU(N)$ Yang–Mills theory with Wilsonian cutoff Λ and bare coupling g_Λ (we now switch to Euclidean signature):

$$S = \frac{1}{4g_\Lambda^2} \int^\Lambda \mathrm{tr}\, |F|^2. \tag{82}$$

Integrating out quantum fluctuations in a momentum slice $|k| < |q| < \Lambda$, the operator $|F|^2$ is renormalized as

$$S_{\mathrm{eff}} = \frac{1}{4} \int^{|k|} \frac{1}{g^2(k)}\, \mathrm{tr}\, |F|^2 + \dots, \tag{83}$$

where the effective coupling is given, with logarithmic precission, by

$$\frac{1}{g^2(k)} \sim \frac{1}{g_\Lambda^2} + N \int_{|k|}^{\Lambda} \frac{d^4 q}{(p-q)^2 q^2} + \dots = \frac{1}{g_\Lambda^2} + \frac{\beta_0 N}{(4\pi)^2} \log\left(|k|^2/\Lambda^2\right) + \text{finite} \tag{84}$$

For $SU(N)$ gauge group, we have $\beta_0 = 22/3$, the usual one-loop beta function coefficient. Notice that the effective coupling corrected by the effect of quantum fluctuations grows towards the infrared, the behaviour that signals asymptotic freedom. Perturbation theory is then expected to break down at scales of order

$$\Lambda_{\mathrm{QCD}} \sim \Lambda \, \exp\left(-\frac{8\pi^2}{N\beta_0 g_\Lambda^2}\right). \tag{85}$$

For an ordinary $U(N)$ gauge theory, the same running takes place, except for the coupling of the global $U(1)$ subgroup, that remains decoupled. Separating this part through the identity

$$\mathrm{tr}\, F^2 = \frac{1}{N} \left(\mathrm{tr}\, F\right)^2 + \mathrm{tr}\, F^2_{SU(N)} \tag{86}$$

we end up with a one-loop corrected effective action:

$$\begin{aligned} S_{\text{eff}}^{U(N)} &= \frac{1}{4}\int^{|k|}\left(\frac{1}{g_\Lambda^2}+\frac{\beta_0 N}{(4\pi)^2}\log(|k|^2/\Lambda^2)\right)\text{tr }|F|^2 \\ &\quad -\frac{1}{N}\frac{\beta_0 N}{(4\pi)^2}\log(|k|^2/\Lambda^2)\,|\text{tr}\,F|^2, \end{aligned} \tag{87}$$

where the second term subtracts the running of the $U(1)$ coupling. It can be thought of as the contribution of the one-loop non-planar diagram to the two point function of the field strength.

We now consider the noncommutative theory with $\theta \neq 0$. The integrand has a factor of

$$\sin^2\left(\frac{k\times q}{2}\right)=\frac{1}{2}-\frac{1}{2}\sin\,(k\times q) \tag{88}$$

from (88). The planar diagram contribution is identical to the first term in (87), since θ-dependence only affects external legs. On the other hand, the nonplanar contribution has a surviving factor of

$$\sin\,(k\times q)$$

from the Feynman rules. This factor oscillates very fast for large values of the loop momentum $|q|$. Thus, the loop momentum integral in the nonplanar graph is effectively cut-off at

$$\Lambda_{\text{eff}}\sim\frac{1}{|\tilde{k}|^2}, \tag{89}$$

where we have defined

$$\tilde{k}^\mu \equiv k_\nu\,\theta^{\nu\mu}. \tag{90}$$

In other words, for $|k|^2\theta \gg 1$ the effective coupling runs only at the planar level, with

$$S(|k|^2\theta\gg 1)_{\text{eff}}\approx\frac{1}{4}\int^{|k|}\left(\frac{1}{g_\Lambda^2}+\frac{\beta_0 N}{(4\pi)^2}\log(|k|^2/\Lambda^2)\right)\ \text{tr}\,|F|^2. \tag{91}$$

This still makes sense for $N=1$, so we learn that the NC $U(1)$ theory is asymptotically free! The NC $U(N)$ theory has in fact the same beta function as the ordinary $SU(N)$ theory:

$$\beta(g^2)_{U(N)_\star}=\frac{dg_\Lambda^2}{d\,\log\,\Lambda}=\beta(g^2)_{SU(N)}=-\frac{11g^4N^2}{12\pi^2}. \tag{92}$$

In particular, this would suggest that the NC $U(1)$ theory becomes strongly coupled for

$$\Lambda_{\rm strong} \sim \Lambda \, \exp\left(-\frac{12\pi^2}{11 g_\Lambda^2} \right) . \tag{93}$$

On the other hand, perhaps we should expect some kind of threshold effect at the classical scale of noncommutativity $|k|^2\theta \sim 1$. In fact, this is the case. Recall that the effective ultraviolet cutoff of the nonplanar diagram was $\Lambda_{\rm eff} = 1/|\tilde{k}|^2$. So, for $|k|^2\theta \leq 1$ the logarithmic divergence in the nonplanar diagram gives a term proportional to

$$\log\left(\frac{|k|^2}{\Lambda_{\rm eff}^2} \right) = \log\left(|k|^2 \, |\tilde{k}|^2\right), \tag{94}$$

and we obtain

$$\begin{aligned} S(|k|^2\theta \leq 1)_{\rm eff} \quad \approx \quad & \frac{1}{4}\int^{|k|} \left(\frac{1}{g_\Lambda^2} + \frac{\beta_0 N}{(4\pi)^2} \log\left(|k|^2/\Lambda^2\right) \right) \, {\rm tr}\, |F|^2 \\ & - \frac{\beta_0}{(4\pi)^2} \log\left(|k|^2 |\tilde{k}|^2\right) |{\rm tr}\,(\partial A)|^2 \end{aligned} \tag{95}$$

In the second term we have written ∂A instead of F because the effective action is evaluated at quadratic order only, and in fact the gauge-invariant completion of (95) cannot be written entirely in terms of the field strength F (c.f. [21]). For us, the important point about (95) is that the second term grows at low energies and produces screening rather than the antiscreening that is characteristic of asymptotic freedom. Thus, we can combine these results and extract the effective coupling of the diagonal $U(1)$ degrees of freedom with running

$$\left(\frac{1}{g_{U(1)}^2} \right)_{|k|^2\theta \leq 1} \approx \frac{1}{g_\Lambda^2} - \frac{\beta_0}{(4\pi)^2} \log\left(|\tilde{k}|^2 \, \Lambda^2\right). \tag{96}$$

The result is that the effective $U(1)$ coupling grows towards the infrared, with the running induced by the planar contribution, as in an $SU(N)$ theory in the formal $N \to 1$ limit, up to energies of order $1/\sqrt{\theta}$. At this threshold, the screening effects start to dominate and the effective coupling grows back up. At energies of order $1/\Lambda\theta$ the effective coupling has again the ultraviolet value g_Λ and ceases to run. In principle, one can still have an infrared Landau pole in the pure $U(1)$ noncommutative theory provided $\Lambda_{\rm strong}\sqrt{\theta} > 1$.

The phenomenon just discussed is the first example of a "mild" UV/IR effect, since we see that, after removal of the UV cutoff Λ, the $\theta \to 0$ limit of the theory is no longer the ordinary free $U(1)$ Maxwell model.

2.4 Physical Interpretation of the Moyal Product

Consider a particle described by a noncommutative field $\phi(x)$, interacting with a fixed external potential $V(x)$ by a term

$$\int d^d x \, \left(V(x) \star \phi(x) - \phi(x) \star V(x)\right). \qquad (97)$$

For a plane wave configuration $\phi(x) \sim e^{ip\cdot x}$ we have

$$V(x) \star e^{ip\cdot x} - e^{ip\cdot x} \star V(x) = \left(V(x + p\cdot\theta/2) - V(x - p\cdot\theta/2)\right) \, e^{ip\cdot x}. \qquad (98)$$

Thus, the noncommutative interaction is exactly reproduced by that of a rigid dipole oriented along the vector

$$L^\mu = \theta^{\mu\nu} p_\nu, \qquad (99)$$

interacting ordinarily through the end-points, exactly like a rigid open string. This analogy is actually rather literal, as we will see in the next section.

Fields interacting in the "fundamental representation" as

$$\int d^d x \, V(x) \star \phi(x) \qquad (100)$$

behave as half-dipoles of length $L^\mu/2$ (c.f. [22]).

Therefore, the non-locality of the noncommutative theories constructed out of Moyal products amounts to reinterpreting the elementary excitations as extended rigid objects [23]. This leads to an interesting extension of the heuristic Heisenberg principle. The effecive size of a noncommutative particle grows linearly with the momentum at very high velocity:

$$L_{\text{eff}} = \max\left(\frac{1}{|p|}, |\theta \cdot p|\right). \qquad (101)$$

This type of relation is known to appear in string theory with the noncommutativity scale replaced by the Regge slope parameter α' (c.f. for example [24]). This is essentially the reason why NCFT is an interesting toy model of string dynamics; it combines some essential features of strings with a much simpler dynamics with finite particle degrees of freedom.

2.5 Connection to String Theory

The dipole picture implies that elementary quanta of NCFT are analogous to open strings. This analogy is actually the source of one of the most important recent developments in the subject.

Indeed, oriented open strings are naturally dipoles. The coupling of an electromagnetic $U(1)$ vector potential to an open string is given by a Wilson line coupling to the end-points of the string. Consider a string worldsheet with proper time τ and string coordinate σ, the endpoints given by $\sigma = 0$ and $\sigma = \pi$. The $U(1)$ coupling is then

$$S_{U(1)} = \int_{\sigma=0} A_\mu dx^\mu - \int_{\sigma=\pi} A_\mu dx^\mu = \int_{\partial\Sigma} A_\mu dx^\mu = \frac{1}{2}\int_\Sigma F_{\mu\nu}\, dx^\mu \wedge dx^\nu . \tag{102}$$

The complete sigma-model action for a string moving in a background metric $g_{\mu\nu}$ and background magnetic field B_{ij} is

$$S = \frac{1}{4\pi\alpha'}\int_\Sigma g_{\mu\nu}\, dx^\mu\, dx^\nu + \frac{1}{2}\int_\Sigma B_{\mu\nu}\, dx^\mu \wedge dx^\nu , \tag{103}$$

where $(2\pi\alpha')^{-1}$ is the tension of the string.

Let us now suppose that B_{ij} is constant and moreover $|g_{ij}| \ll |\alpha' B_{ij}|$, so that we can approximate the action by

$$S \approx \frac{1}{2}\int B_{ij}\, dx^i \wedge dx^j = \frac{1}{2} B_{ij} \int_{\partial\Sigma} x^i \partial_\tau x^j . \tag{104}$$

Thus, we see that the endpoints of the open string behave like electrons in the LLL in this limit! The same arguments as in the electron case yield then

$$[x^j, x^k]_{\partial\Sigma} = i\,\theta^{jk} \tag{105}$$

with

$$\theta^{jk} = \left(\frac{1}{B}\right)^{jk} . \tag{106}$$

In order to obtain a NCFT of rigid dipoles we would like to project out all the massive (oscillatory) degrees of freedom of the open string theory, i.e. we would like to take the zero-slope limit $\alpha' \to 0$. But we just have learnt that at the same time we must keep $\theta \sim 1/B$ constant and also $|g_{ij}| \ll |\alpha' B_{ij}|$. A scaling limit that satisfies these constraints and produces NCFT interaction Lagrangians out of the open-string perturbative interactions is the so-called Seiberg–Witten limit [25]:

$$g_{ij} \sim (\alpha')^2 B_{ij} B^{ij} \longrightarrow 0 \tag{107}$$

at fixed $B_{ij} = (1/\theta)_{ij}$. Physically, what is being stated is very simple. In order to make the open string into a rigid dipole, we must take the nominal tension to infinity to decouple all oscillator modes (rigidity). Normally this produces the effective collapse of the open string to a pointlike object. However, if the magnetic field is kept large in the scaling limit, the Lorentz force tending to stretch the open string endpoints can compensate for this effect and one reaches a rigid open string of finite extent given by $L \sim \theta p$.

2.6 The UV/IR Mixing

The phenomenon of UV/IR mixing represents the most radical departure of NCFT from the standard behaviour of ordinary field theories. It occurs in perturbation theory, so that it can be studied with considerable detail, and represents the fact that the two deformation operations: the noncommutative deformation $\theta \neq 0$, and the quantum deformation $\hbar \neq 0$, do not commute [5].

The UV/IR mixing is a lack of Wilsonian decoupling between UV and IR scales, even in the presence of explicit masses. Technically, it comes about in a rather elementary fashion. Recall that nonplanar diagrams have improved convergence properties because of Moyal phases that depend on loop momenta. For example, two loop momenta q, q' tied by a Moyal phase

$$e^{-\frac{i}{2}q\times q'}$$

will introduce an effective cutoff in the diagram at the scale $\Lambda_{\text{eff}} \sim 1/\sqrt{\theta}$. On the other hand, a loop momentum q tied to an exteral momentum p will introduce

$$e^{-\frac{i}{2}q\times p},$$

which in turn gives an effective cutoff $\Lambda_{\text{eff}} \sim 1/|p\cdot\theta|$. Since the corresponding UV divergences are absent, they are not explicity subtracted in the renormalization procedure. However, since the effective cutoff is non-analytic in θ, these singularities in physical quantities show up in the $\theta \to 0$ limit. Alternatively, in Green's functions depending on external momenta, they show up in the limit $|\theta\cdot p| \to 0$. This may be interpreted as non-analytic behaviour in the $\theta \to 0$ limit at finite $|p|$, or as an infrared singularity at fixed θ.

Therefore, we see that in general the noncommutative *quantum* field theory is *not* a smooth deformation of the ordinary $\theta = 0$ theory, even if it was so in the classical approximation. We also learn that, at fixed non-zero

θ, the NCFT is IR singular as a result of divergences that originally had an UV interpretation, hence the name UV/IR mixing.

2.6.1 A Simple Example

In order to illustrate this important phenomenon, we consider the simplest setting in which it arises: the one-loop mass renormalization of the ϕ^4 model in four dimensions (in this section we work in Euclidean signature):

$$S = \int \left(\frac{1}{2}(\partial\phi)^2 + \frac{m^2}{2}\phi^2 + \frac{\lambda}{4!}\phi \star \phi \star \phi \star \phi \right). \tag{108}$$

In the ordinary ($\theta = 0$) model the leading mass renormalization comes from the normal-ordering diagram contribution to the self-energy:

$$\Sigma = \frac{\lambda}{2} \int \frac{d^4k}{(2\pi)^4} \frac{1}{k^2+m^2} \approx \frac{\lambda}{32\pi^2} \left(\Lambda^2 - m^2 \,\log(\Lambda^2/m^2) + \text{finite} \right), \tag{109}$$

in terms of the ultraviolet cutoff Λ. We find the standard quadratic renormalization together with a subleading logarithmic piece.

In the noncommutative theory, we have two contributions, planar and nonplanar. The planar diagram gives exactly the contribution (109), except for the different symmetry factor of the diagram, which is $1/3$ instead of $1/2$. On the other hand, the nonplanar diagram has a surviving Moyal phase that makes it finite:

$$\Sigma_{\text{NP}} = \frac{\lambda}{6} \int \frac{d^4k}{(2\pi)^4} \frac{e^{-ik\times p}}{k^2+m^2} = \frac{\lambda}{24\pi^2} \frac{m^2}{\sqrt{m^2\tilde{p}^2}} K_1 \left[\sqrt{m^2\tilde{p}^2} \right]. \tag{110}$$

In order to compare the planar and nonplanar parts, we introduce an ultraviolet cutoff via a Schwinger proper-time parametrization:

$$\left[\frac{1}{k^2+m^2} \right]_\Lambda = \int_0^\infty ds \, e^{-s(k^2+m^2)} \; e^{-1/\Lambda^2 s}. \tag{111}$$

We find

$$\Sigma_{\text{NP}} = \frac{\lambda}{96\pi^2} \left(\Lambda_{\text{eff}}^2 - m^2 \,\log\left(\Lambda_{\text{eff}}^2/m^2\right) + \dots \right), \tag{112}$$

where the effective cutoff is given by

$$\Lambda_{\text{eff}}^2 = \frac{1}{\tilde{p}^2 + 1/\Lambda^2}. \tag{113}$$

Notice that $\Lambda_{\rm eff} \approx \Lambda$ for $|p| \ll 1/\Lambda\theta$, whereas $\Lambda_{\rm eff} \approx 1/|\tilde{p}|$ for $|p| \gg 1/\Lambda\theta$. So, if we renormalize the theory at fixed p and fixed θ, by subtracting the planar divergence in the $\Lambda \to \infty$ limit:

$$m^2 \to M^2 = m^2 + \frac{\lambda}{48\pi^2}\left(\Lambda^2 - m^2 \log\left(\Lambda^2/m^2\right)\right) + \text{constant} \tag{114}$$

we have a quadratic 1PI effective action:

$$\Gamma_{\rm 1PI} = \int d^4p\, \phi(-p)\, \Gamma^{(2)}(p)\, \phi(p) + \dots \tag{115}$$

with

$$\Gamma^{(2)}(p) = p^2 + M^2 + \frac{\lambda}{96\pi^2\tilde{p}^2} - \frac{\lambda M^2}{96\pi^2} \log\left(1/M^2\tilde{p}^2\right) + \dots \tag{116}$$

Thus, as promised, the effective action has a singularity at $p = 0$ that can be interpreted either as an IR singularity at fixed θ or as a non-analiticity as a function of θ at fixed p.

We may wonder to what extent the leading IR-singular term

$$\Sigma_{\rm NP} \sim \frac{1}{|\tilde{p}|^2}$$

can be reliably calculated in perturbation theory. An indication is given by the following estimation. Higher-order perturbative corrections to the leading $1/\tilde{p}^2$ behaviour have the form

$$\frac{\lambda}{\tilde{p}^2}\left[\lambda \log\left(M^2\tilde{p}^2\right)\right]^n .$$

These corrections are significant only for momenta such that the term in brackets is of $\mathcal{O}(1)$. Thus, we see that perturbation theory will break down at nonperturbatively small momenta of order

$$|p|_{\rm breakdown} \sim \frac{1}{M\theta}\, e^{-C/\sqrt{\lambda}}. \tag{117}$$

For the present model, we can give a simple physical interpretation of the UV/IR mixing provided the noncommutativity is purely spatial, i.e. $\theta^{0i} = 0$. Notice that the just computed 1PI effective action implies a modified dispersion relation for the ϕ-quanta of the form:

$$p^2 + M^2 + \frac{\lambda}{96\pi^2\tilde{p}^2} = 0. \tag{118}$$

After Wick rotation back to $(-+++)$ signature one finds:

$$\omega = \sqrt{|\vec{p}|^2 + M^2 + \frac{c}{\theta^2 |\vec{p}|^2_\theta}} \tag{119}$$

where $c = \lambda/96\pi^2$ and $\vec{p}_\theta$ is the projection of the spatial momentum onto the plane of noncommutativity.

This expression shows dramatically the UV/IR mixing effects, since the entire energy spectrum below noncommutative momenta of order $\lambda^{1/4}/\sqrt{\theta}$ has been removed!

2.6.2 The Case of Gauge Theories

The UV/IR mixing in the case of gauge theories shows some specific features of interest [26]. Consider the polarization tensor of the NC $U(1)$ theory:

$$S^{(2)} = \frac{1}{2} \int A_\mu(k)\, \Pi^{\mu\nu}(k)\, A_\nu(-k). \tag{120}$$

In the ordinary (or planar) case, gauge invariance together with Lorentz invariance forbids a quadratic divergence in the polarization $\Pi_{\mu\nu} \sim \eta_{\mu\nu}\, \Lambda^2$. It would violate transversality. In fact

$$\Pi_{\mu\nu}(k) = \left(k_\mu\, k_\nu - \eta_{\mu\nu}\, k^2\right)\, \Pi(k), \tag{121}$$

where

$$\Pi(k) \sim \log\left(\frac{|k|^2}{\Lambda^2}\right) + \text{finite}. \tag{122}$$

The nonplanar contribution has the standard effective cutoff $\Lambda_{\text{eff}} = \min(\Lambda, 1/|\tilde{k}|)$. Because of gauge invariance at $\theta = 0$, we would expect that UV/IR phenomena would only appear at logarithmic level $\Pi(k) \sim \log(|k|^2|\tilde{k}|^2)$, and indeed we found such terms in the previous section in our discussion of asymptotic freedom.

However, the explicit breaking of Lorentz symmetry allows now for other kinematical structures with IR singularity stronger than logarithmic and still transverse. In particular, quadratic divergences do appear with the structure

$$\Pi^{\text{NP}}_{\mu\nu} = -g^2\, C\, \frac{\tilde{k}^\mu \tilde{k}^\nu}{\tilde{k}^2}\, \Lambda^2_{\text{eff}} = -g^2\, C\, \frac{\tilde{k}^\mu \tilde{k}^\nu}{\tilde{k}^4}. \tag{123}$$

Notice that transversality is ensured by $k^\mu \tilde{k}_\mu = k_\mu \theta^{\mu\nu} k_\nu = 0$. At one-loop, the constant C has been calculated to be

$$C = \frac{2N}{\pi^2}\,(2 + n_s - 2n_f), \tag{124}$$

where N is from the $U(N)$ gauge group, n_s is the number of complex scalars in the adjoint representation and n_f is the number of Majorana fermions also in the adjoint representation. Notice that $C = 0$ for supersymmetric or soflty broken supersymmetric spectra.

Thus, we learn that the strength of the UV/IR mixing responds to the naive power-counting rather than to the effective divergence structure of the $\theta = 0$ model. In particular, one finds unstable dispersion relations in NC $U(1)$

$$\omega(k) = \sqrt{\vec{k}^2 - \frac{g^2 C}{\theta^2 |\vec{k}_\theta|^2}} \tag{125}$$

with low-momentum tachyonic excitations as soon as $C > 0$.

2.6.3 Heuristic Explanation of the UV/IR Mixing

The dipole picture of NCFT that was developed before provides a simple heuristic explanation of the UV/IR mixing. Since a virtual loop of momentum p carries dipoles of transverse length $|\theta \cdot p|$, we understand that the loop corrections to the Green's functions will have strong θ-dependence down to arbitrarily low energies, unless these effects are cancelled by some mechanism (such as enough amount of supersymmetry).

Notice that, if an explicit UV cutoff is present, Λ, it sets the maximum possible momentum of the virtual dipoles circulating in the loop. This in turn means that significant θ-dependence only appears down to momenta of order $1/\Lambda\theta$.

Thus, we have the following general hierarchycal structure. NCFT with ultraviolet cutoff $\Lambda\sqrt{\theta} \gg 1$ has significant classical effects (tree level) associated to noncommutativity up to length scales of $\mathcal{O}(\sqrt{\theta})$. However, one-loop effects "transport" the effects of noncommutativity to the larger length scale of $\mathcal{O}(\Lambda\theta)$. This larger length scale is true dynamical scale of noncommutativity. Of course, this picture would be invalidated if perturbation theory would break down at some intermediate scale. For example, if we insist on removing the ultraviolet cutoff $\Lambda \to \infty$ at fixed θ, necessarily $\Lambda\theta \to \infty$ and perturbation theory is bound to break down before we reach the deep infrared domain.

2.6.4 UV/IR Mixing and Unitarity

There is an interesting interplay between the UV/IR mixing and the violation of unitarity in the case that the noncommutativity affects time. Instead of developing the general theory we will simply explain the basic phenomena by looking at a simple example. Let us consider a noncommutativity matrix of the skew-diagonal form $(\theta^{\mu\nu}) = i\,\text{diag}\,(\sigma_2\,\theta_e, \sigma_2\,\theta_m)$. That is, we have the noncommutativity relations:

$$[t, x] = i\theta_e, \qquad [y, z] = i\theta_m. \tag{126}$$

We return now to the ϕ^4 theory studied in the previous section and we consider the massless model for simplicity. The normal-ordering tadpole diagram has no interesting dynamical interpretation in the ordinary theory, simply inducing the quadratic renormalization of the mass parameter. However, this is no longer the case for the noncommutative theory, since the nonplanar tadpole diagram *does* have an interesting singularity structure when interpreted as a 1 → 1 scattering amplitude:

$$i\mathcal{M}(p \to p) = -i\frac{\lambda}{6}\int \frac{d^4q}{(2\pi)^4} e^{-i\tilde{p}\cdot q}\,\frac{i}{q^2 + i0} = -i\frac{\lambda}{24\pi^2}\frac{1}{-\tilde{p}^2 + i0}. \tag{127}$$

The striking fact about this explicit expression is that the imaginary part of the amplitude is a non-trivial distribution, i.e.

$$2\,\text{Im}\,\mathcal{M}(p) = \frac{\lambda}{12\pi}\,\delta(-\tilde{p}^2). \tag{128}$$

Therefore, if unitarity is to be satisfied, this imaginary part should be understadable in terms of a product of on-shell amplitudes corresponding to all the non-trivial cuttings of the diagram. Since the tadpole has no on-shell cuttings, it seems that we find a violation of unitarity [27].

Despite this fact, one can still manipulate $\text{Im}\,\mathcal{M}$ in a purely formal fashion so that it looks like a contribution from the optical theorem. Take $\theta_m = 0$ and $\theta_e \neq 0$, and introduce

$$1 = \int d^4k\,\delta(p - k)$$

to obtain

$$2\,\text{Im}\,\mathcal{M}(p) = \frac{\lambda}{12\pi}\int d^4k\,\delta(k-p)\,\delta(-\tilde{p}^2) = \int \frac{d^3\vec{k}}{2(2\pi)^3|k_1|}\left(\frac{(2\pi)^2\lambda}{6\theta_e^2}\right)\delta(p-k). \tag{129}$$

This formula can be interpreted as the amplitude for the mixing of the ϕ quanta with particle states $|\chi\rangle$ with dispersion relation $|k_0| = |k_1|$. The $\phi - \chi$ coupling is given by

$$\lambda_{\phi\chi} = \sqrt{\frac{(2\pi)^2 \lambda}{6\theta_e^2}}. \tag{130}$$

Thus, it seems that we can save unitarity at the expense of enlarging the Hilbert space of asymptotic states, just like one can make the S-matrix of open-string theory unitary by introducing the closed-string states. In fact, while this is true at a formal level, it turns out that the added Hilbert space of 'closed-string' states $|\chi\rangle$ does not satisfy appropriate physical conditions. In particular these states come with a continuous spectrum, they are tachyonic and moreover have negative norm in general.

For example, just considering the more general case with $\theta_m \neq 0$ in our example above yields

$$2\,\mathrm{Im}\,\mathcal{M}(p) = \int \frac{d^3\vec{k}}{2(2\pi)^3\omega_\chi} \left(\frac{(2\pi)^2\lambda}{6\theta_e^2}\right) \delta(p-k), \tag{131}$$

where the frequency of the χ particles is:

$$\omega_\chi = \sqrt{|\vec{p}_e|^2 - \frac{\theta_m^2}{\theta_e^2}\,|\vec{p}_m|^2}. \tag{132}$$

This dispersion relation shows clearly that the χ particles have an unbounded-below spectrum of tachyonic excitations [28]. Thus, timelike noncommutative theories are generically inconsistent in perturbation theory, at least to the extent that UV/IR mixing is present.

2.7 Remarks on θ-Phenomenology

The most obvious application of NCFT is to entertain the possibility that the noncommutativity of spacetime might be real and could be detected experimentally. In such a situation the most notorious feature of the physics is the breakdown of Lorentz invariance. Even if $\theta^{0i} = 0$, the spatial noncommutativity $\theta^{ij} = \epsilon^{ijk}\theta_k$ determines a privileged direction *in vacuo* $\vec{\theta} = (\theta_k)$. Thus, collider experiments put a bound of order

$$|\theta| < (100 \text{ GeV})^{-2} \tag{133}$$

to begin with. In fact, it is not easy to be more specific since the Standard Model doesn't fit naturally into a NCFT with Lorentz violation (recall the problem of $U(1)$ charge assignments). For this reason, most of the phenomenological discussions of NCG effects have been carried out in the noncommutative generalization of the QED sector.

The bound (133) can be improved by application of some elementary constraints from atomic physics. Because of the dipole picture given before, the leading interaction of electrons with the field of the atomic nucleus has a dipole moment induced by the substitution

$$x^\mu \longrightarrow x^\mu - \frac{1}{2} p_\alpha \, \theta^{\alpha\mu},$$

so that the Coulomb potential has terms:

$$V_C(|\vec{x} - \tfrac{1}{2}\,\vec{p}\cdot\vec{\theta}|) = -\frac{\alpha_{em} Z}{\sqrt{(\vec{x} - \frac{1}{2}\,\tilde{p})^2}} \approx -\frac{\alpha_{em} Z}{|\vec{x}|} + \frac{1}{2}\frac{\alpha_{em} Z}{|\vec{x}|^3}\,\vec{\theta}\cdot\vec{L} + \mathcal{O}(\theta^2 p^4), \quad (134)$$

where $\vec{L} = \vec{x} \wedge \vec{p}$. Thus, this term induces a "noncommutative hyperfine splitting" [29]. From limits on the Lamb shift we can put a bound of order

$$|\theta| < (10 \text{ TeV})^{-2}. \quad (135)$$

Constraints from collider experiments are not actually much better than this, if evaluated at tree level. Dependence on the noncommutativity parameter in the vertices comes with two powers of momenta (derivatives) and thus it corresponds generally to dimension five or six effective operators. For example, a leading correction to the $e^+\gamma e^-$ vertex is given by the operator

$$\theta^{\alpha\beta}\, \partial_\alpha \, \bar{\psi}\, \gamma^\mu A_\mu \, \partial_\beta \, \psi. \quad (136)$$

Corrections from such operators are or relative order $\mathcal{O}(\theta E^2)$ for processes at typical energies of $\mathcal{O}(E)$. Thus, collider physics at $E \sim 100$ GeV, known to within a few percent errors, give bounds of order

$$|\theta| < \frac{1}{100\, E^2} \sim (\text{TeV})^{-2}. \quad (137)$$

When quantum corrections are considered, the situation changes dramatically. The UV/IR mixing arising at one-loop order implies that noncommutative effects show up at energies much below $1/\sqrt{\theta}$. In fact, noncommutative QED has tachyonic photon excitations induced at one-loop order and therefore it is incompatible, not only with experiments, but with simple observations of everyday life. This means that, in exploring applications of NCG to phenomenology in the context of weakly coupled NCFT, we must assume the existence of an UV cutoff beyond which the effects of UV/IR mixing dissappear.

Since UV/IR mixing affects dispersion relations, this means that the breakdown of Lorentz symmetry is not restricted to (nonrenormalizable) operators of high dimension, but rather creeps in the operators of dimension two and three at the one-loop level. Correspondingly, the violations of Lorentz symmetry that affect dispersion relations are the subject of fantastic constraints from both low and high energy physics (see for example [30]).

Consider, for example, the dispersion relation of photons corrected at one loop in the pure NC $U(1)$ theory. The leading terms in the polarization tensor at low momentum are (we neglect the logarithmic corrections that only renormalize the coupling):

$$\Pi_{\mu\nu} = (p_\mu \, p_\nu - p^2 \, \eta_{\mu\nu}) + \tilde{p}_\mu \, \tilde{p}_\nu \, \Pi_{nc}, \tag{138}$$

where

$$\Pi_{nc} = -\frac{C \, g^2}{|\tilde{p}|^4}. \tag{139}$$

Considering transverse photons with polarization $A_\mu \sim \tilde{p}_\mu$ we obtain a mass-shell condition

$$p^2 - \frac{C \, g^2 \, \tilde{p}^2}{|\tilde{p}|^4} = 0. \tag{140}$$

Since $C > 0$ we find tachyonic excitations at low momentum. Therefore, we must assume some UV cutoff that eliminates the UV/IR mixing due to very long dipoles in the virtual loop. One such cutoff is provided for example by a softly broken supersymmetric spectrum broken at scale Λ_s. Then, we have an effective cutoff for the nonplanar diagram given by

$$\Lambda^2_{\text{eff}} = \frac{1}{(-\tilde{p}^2 + 1/\Lambda_s^2)^2} \tag{141}$$

and a corrected dispersion relation for photons polarized as $A_\mu \sim \tilde{p}_\mu$ given by

$$\omega^2 = |\vec{p}|^2 - \frac{C g^2 |\vec{p}|^2 \theta^2}{(\theta^2 |\vec{p}|^2 + 1/\Lambda_s^2)^2}, \tag{142}$$

where we assume that the photon propagates parallel to the noncommutative directions. Expanding this dispersion relation around low momenta we see that it produces a correction to the speed of light for these photons given by

$$c_s = 1 - C\, g^2\, \theta^2\, \Lambda_s^4. \tag{143}$$

This means, in particular, that we must have $\Lambda_s \sqrt{\theta} \gg 1$ in order not to conflict with observations. So we actually have an inverted hierarchy in which the noncommutativity scale is forced to be much higher than the supersymmetry breaking scale. Even in this situation, a variety of phenomenological constraints put bounds of order

$$|c_s - 1| < 10^{-15} \tag{144}$$

or even stronger, depending on how model-independent we wish to be (see for example [31]). This translates into bounds on the hierarchy between Λ_s and θ that easily render the classical bounds irrelevant.

In any case, the lesson to be learned from these considerations is that noncommutative phenomenology is probably a premature exercise. The absence of natural models and the strong bounds to be put on θ at the level of perturbative dynamics are rather neat arguments against the prospects of such phenomenonlogical exercises.

Acknowledgments

I would like to thank the organizers of the Trieste Summer School for their kind hospitality at the ICTP while these lectures were delivered.

References

[1] A. Connes, *Noncommutative Geometry*, Academic Press (1994).

[2] A. González-Arroyo and C.P. Korthals Altes, Phys. Lett. **B131** (1983) 396.

[3] A. Connes and M. Rieffel, Contemp. Math. Oper. Alg. Math. Phys. **62**, AMS (1987) 237.

[4] A. Connes, M.R. Douglas and A. Schwarz, J. High Energy Phys. **9802** (1998) 003, `hep-th/9711162`. M.R. Douglas and C. Hull, J. High Energy Phys. **9802** (1998) 008, `hep-th/9711165`.

[5] S. Minwalla, M. Van Raamsdonk and N. Seiberg, J. High Energy Phys. **0002** (2000) 020, `hep-th/9912072`.

[6] Z. Guralnik and J. Troost, J. High Energy Phys. **0105** (2001) 022, `hep-th/0103168`.

[7] L. Alvarez-Gaumé and J.L.F. Barbón, Nucl. Phys. **B623** (2002) 165 `hep-th/0109176`.

[8] L. Susskind, `hep-th/0101029`. A.P. Polychronakos, J. High Energy Phys. **04** (2001) 011 `hep-th/0103013`.

[9] M.R. Douglas and N. Nekrasov, `hep-th/0106048`. A. Konechny and A. Schwarz, `hep-th/0012145`. R.J. Szabo, `hep-th/0109162`.

[10] A. Connes, `hep-th/0003006`.

[11] J. Gracia Bondía, J.C. Varilly and H. Figueroa, *Elements of Noncommutative Geometry.* Birkhauser, Boston (2001).

[12] J. Polchinski, `hep-th/9611050`.

[13] R.C. Myers, J. High Energy Phys. **9912** (1999) 022 `hep-th/9910053`.

[14] B. Jurco, S. Schraml, P. Schupp and J. Wess Eur. Phys. J. **C17** (2000) 521 `hep-th/0006246`.

[15] D.J. Gross, A. Hashimoto and N. Itzhaki, Adv. Theor. Math. Phys. **4** (2000) 893 `hep-th/0008075`.

[16] M. Chaichian, P. Presnajder, M.M. Sheikh-Jabbari, A. Tureanu. Phys. Lett. **B526** (2002) 132 `hep-th/0107037.`

[17] C.-S. Chu, V.V. Khoze and G. Travaglini, `hep-th/0112139.`

[18] G. 't Hooft, Nucl. Phys. **B75** (1974) 461.

[19] A. González-Arroyo and M. Okawa, Phys. Lett. **B120** (1983) 174; Phys. Rev. **D27** (1983) 2397. T. Filk, Phys. Lett. **B376** (1996) 53.

[20] C.P. Martín and D. Sánchez-Ruiz, Phys. Rev. Lett. **83** (1999) 476, `hep-th/9903077.` M.M Sheikh-Jabbari, J. High Energy Phys. **9906** (1999) 015, `hep-th/9903107.` T. Krajewski and R. Wulkenhaar, `hep-th/9903187.`

[21] M. van Raamsdonk, J. High Energy Phys. **11** (2001) 006 `hep-th/0110093.` A. Armoni and E. López, `hep-th/0110113.`

[22] L. Alvarez-Gaumé and J.L.F. Barbón, Int. J. Mod. Phys. **A16** (2001) 1123 `hep-th/0006209.`

[23] C.-S. Chu and P.-M. Ho, Nucl. Phys. **B550** (1999) 151, `hep-th/9812219.` M.M. Sheikh-Jabbari, Phys. Lett. **B455** (1999) 129, `hep-th/9901080.` D. Bigatti and L. Susskind, `hep-th/9908056.` Z. Yin, Phys. Lett. **B466** (1999) 234, `hep-th/9908152.`

[24] T. Yoneya, Prog. Theor. Phys. **103** (2000) 1081 `hep-th/0004074.`

[25] N. Seiberg and E. Witten, J. High Energy Phys. **9909** (1999) 032, `hep-th/9908142.`

[26] M. Hayakawa, Phys. Lett. **B478** (2000) 394 `hep-th/9912094, hep-th/9912167.` A. Matusis, L. Susskind and N. Toumbas, J. High Energy Phys. **12** (2000) 002 `hep-th/0002075.`

[27] J. Gomis and T. Mehen, Nucl. Phys. **B591** (2000) 265, `hep-th/0005129.`

[28] L. Alvarez-Gaumé, J.L.F. Barbón and R. Zwicky, J. High Energy Phys. **0105** (2001) 057 `hep-th/0103069.`

[29] M. Chaichian, M.M. Sheikh-Jabbari, A. Tureanu Phys. Rev. Lett. **86** (2001) 2716 `hep-th/0010175.`

[30] A. Anisimov, T. Banks, M. Dine and M. Graesser, `hep-ph/0106356.`

[31] S.R. Coleman and S.L. Glashow, Phys. Rev. **D59** (1999) 116008 `hep-ph/9812418.`