Introduction to Superstrings

Joseph L. Birman Helmut Faissner Jeffrey W. Lynn

Springer

Michio Kaku

Introduction to Superstrings

With 48 Illustrations

Springer-Verlag
New York Berlin Heidelberg
London Paris Tokyo Hong Kong

Michio Kaku
Department of Physics
City College of the City University of New York
New York, NY 10031, USA

Printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data
Kaku, Michio.
Introduction to superstrings/Michio Kaku.
p. cm.—(Graduate texts in contemporary physics)
Bibliography: p.
1. Superstring theories. I. Title. II. Title: Introduction to superstrings. III. Series.
QC794.6.S85K35 1988
530.1'4—dc19 87-36080

Typeset by Asco Trade Typesetting Ltd., Hong Kong.

9 8 7 6 5 4 3 2 1 (Corrected Second Printing, 1990)

ISBN-13: 978-1-4684-0321-3 e-ISBN-13: 978-1-4684-0319-0
DOI: 10.1007/978-1-4684-0319-0

Contents

Part II. Second Quantization and the Search for Geometry

Part III. Phenomenology and Model Building

Part I

First Quantization and Path Integrals

CHAPTER 1

Path Integrals and Point Particles

§1.1. Why Strings?

One of the greatest scientific challenges of our time is the struggle to unite the two fundamental theories of modern physics, quantum field theory and general relativity, into one theoretical framework. Remarkably, these two theories together embody the sum total of all human knowledge concerning the most fundamental forces of nature. Quantum field theory, for example, has had phenomenal success in explaining the physics of the microcosm, down to distances less than 10^{-15} cm. General relativity, on the other hand, is unrivaled in explaining the large-scale behavior of the cosmos, providing a fascinating and compelling description of the origin of the universe itself. The astonishing success of these two theories is that together they can explain the behavior of matter and energy over a staggering 40 orders of magnitude, from the subnuclear to the cosmic domain.

The great mystery of the past five decades, however, has been the total incompatibility of these two theories. It's as if nature had two minds, each working independently of the other in its own particular domain, operating in total isolation of the other. Why should nature, at its deepest and most fundamental level, require two totally distinct frameworks, with two sets of mathematics, two sets of assumptions, and two sets of physical principles?

Ideally, we would want a unified field theory to unite these two fundamental theories:

$$\left.\begin{matrix}\text{Quantum field theory}\\ \text{General relativity}\end{matrix}\right\} \text{Unified field theory}$$

However, the history of attempts over the past decades to unite these two theories has been dismal. They have inevitably been riddled with infinities or

have violated some of the cherished principles of physics, such as causality. The powerful techniques of renormalization theory developed in quantum field theory over the past decades have failed to eliminate the infinities of quantum gravity. Apparently, a fundamental piece of the jigsaw puzzle is still missing.

Although quantum field theory and general relativity seem totally incompatible, the past two decades of intense theoretical research have made it increasingly clear that the secret to this mystery most likely lies in the power of *gauge symmetry*. One of the most remarkable features of nature is that its basic laws have great unity and symmetry when expressed in terms of group theory. Unification through gauge symmetry, apparently, is one of the great lessons of physics. In particular, the use of local symmetries in Yang–Mills theories has had enormous success in banishing the infinities of quantum field theory and in unifying the laws of elementary particle physics into an elegant and comprehensive framework. Nature, it seems, does not simply incorporate symmetry into physical laws for aesthetic reasons. Nature *demands* symmetry.

The problem has been, however, that even the powerful gauge symmetries of Yang–Mills theory and the general covariance of Einstein's equations are insufficient to yield a finite quantum theory of gravity.

At present, the most promising hope for a truly unified and finite description of these two fundamental theories is the superstring theory [1–12]. Superstrings possess by far the largest set of gauge symmetries ever found in physics, perhaps even large enough to eliminate all divergences of quantum gravity. Not only does the superstring's symmetry include that of Einstein's theory of general relativity and the Yang–Mills theory, it also includes supergravity and the Grand Unified Theories (GUTs) [13] as subsets.

Roughly speaking the way in which superstring theory solves the riddle of infinities can be visualized as in Fig. 1.1, where we calculate the scattering of two point particles by summing over an infinite set of Feynman diagrams with loops. These diagrams, in general, have singularities that correspond to "pinching" one of the internal lines until the topology of the graph is altered. By contrast, in Fig. 1.2 we have the single-loop contribution to the scattering

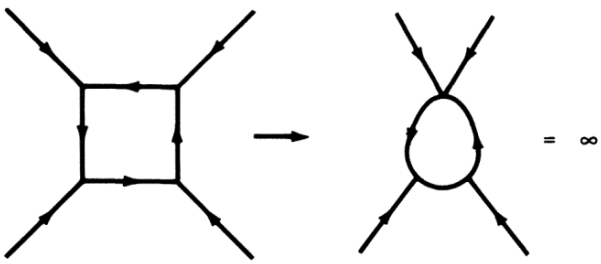

Figure 1.1. Single-loop Feynman diagram for four-particle scattering. The ultraviolet divergence of this diagram corresponds to the pinching of one internal leg, i.e., when one internal line shrinks to a point.

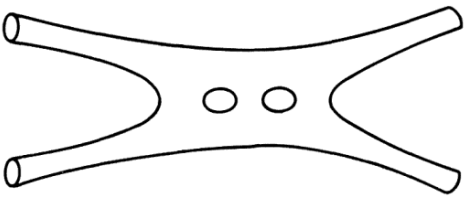

Figure 1.2. Two-loop Feynman diagram for closed string scattering. The diagram is ultraviolet finite because it cannot be pinched as in the point particle case. From topological arguments alone, we can see that string theory is less divergent than point particle theory. Infrared divergences, however, may still exist.

of two closed string states. Notice that we cannot "pinch" one of the internal lines as in the point particle case. Thus, we naively expect that the superstring theory is less divergent or even finite because of the symmetries that forbid this topological deformation.

Any theory that can simultaneously eliminate the infinities of the S-matrix and incorporate quantum mechanics, the general theory of relativity, GUT theory, and supergravity obviously possesses mathematics of breathtaking beauty and complexity. In fact, even the mathematicians have been startled at the mathematics emerging from the superstring theory, which links together some of the most dissimilar, far-ranging fields of mathematics, such as Kac–Moody algebras, Riemann surfaces and Teichmüller spaces, modular groups, and even Monster group theory.

The great irony of string theory, however, is that the theory itself is not unified. To someone learning the theory for the first time, it is often a frustrating collection of folklore, rules of thumb, and intuition. At times, there seems to be no rhyme or reason for many of the conventions of the model. For a theory that makes the claim of providing a unifying framework for all physical laws, it is the supreme irony that the theory itself appears so disunited! The secrets of the model, at its most fundamental level, are still being pried loose.

Usually, when we write down a quantum theory, we start with the geometry or symmetry of the theory and then write down the action. From the action, in turn, we derive all the predictions of the model, including the unitary S-matrix. Thus, a *second quantized* action is the proper way in which to formulate a quantum field theory. The fundamental reason why superstring theory seems, at times, to be a loose collection of apparently random conventions is that it is usually formulated as a first quantized theory. Because of this, we must appeal to intuition and folklore in order to construct all the Feynman diagrams for a unitary theory.

Unfortunately, the second quantized action and the geometry of the superstring are some of the last features of the model to be developed. In fact, as seen from this perspective, the model has been developing *backward* for the past 20 years, beginning with the accidental discovery of its quantum theory in 1968!

By contrast, when Einstein first discovered general relativity, he started with physical principles, such as the equivalence principle, and formulated it in the language of general covariance. Once the geometry was established, he then wrote down the action as the unique solution to the problem. Later, classical solutions to the equations were discovered in terms of curved manifolds, which provided the first successful theoretical models for the large-scale behavior of the universe. Finally, the last step in the evolution of general relativity is the development of a quantum theory of gravity. The crucial steps in the historical evolution of general relativity can thus be represented as

$$\text{Geometry} \rightarrow \text{Action} \rightarrow \text{Classical theory} \rightarrow \text{Quantum theory}$$

Furthermore, both general relativity and Yang–Mills theory are mature theories: they both can be formulated from first principles, which stresses the geometry and the physical assumptions underlying the theory. Superstring theory is just beginning to reach that stage of development.

Remarkably, Yang–Mills theory and gravity theory are the *unique* solution to two simple geometric statements:

(1) Global Symmetry
The free theory must propagate pure ghost-free spin-one and spin-two fields transforming as irreducible representations of SU(N) and the Lorentz group.
(2) Local Symmetry
The theory must be locally SU(N) and generally covariant.

What is remarkable is that the coupled Yang–Mills–gravity action is the unique solution of these two simple principles:

$$L = -\tfrac{1}{4}\sqrt{-g}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2\kappa^2}\sqrt{-g}R_{\mu\nu}g^{\mu\nu} \tag{1.1.1}$$

(The first principle contains the real physics of the theory. It cannot be included as a subset of the second principle. There is an infinite number of generally covariant and SU(N) symmetric invariants, so we need the first principle to input the physics and select the irreducible representations of the basic fields. By "pure" fields, we mean ghost-free fields that have at most two derivatives, which rules out R^2 and F^4 higher derivative theories.)

The question remains: *What is the counterpart to these two simple principles for superstring theory*? Much work still has to be done to formulate a truly geometric theory of strings, but the most promising candidate is presented in Chapter 8, where I discuss the geometric formulation of string field theory.

The plan of this book, of course, must reflect the fact that the theory has been evolving backward. For pedagogical reasons, we will mostly follow the historical development of the theory. Thus, Part I of this book, which introduces the first quantized theory, will at times appear to be a loose collection of conventions without any guiding principle. That is why we have chosen, in Part I, to emphasize the *path integral or functional* approach to string theory.

Only with Feynman path integrals do we have a formalism in which we can derive the other formalisms, such as the harmonic oscillator formalism. Although the path integral formulation of a first quantized theory is still woefully inadequate compared to a genuine second quantized theory, it is the most convenient formalism in which to tie together the loose ends of the first quantized theory.

In Part II of the book we will discuss the field theory itself, from which we can derive all the results of the theory from one action. However, once again we have followed historical order and presented the field theory backward. We begin with the broken theory, and then present a candidate for the geometric one.

Finally, in Part III we present the "phenomenology" of strings. Although it may be presumptuous to do phenomenology starting at 10^{19} GeV, it is important to establish the kinds of predictions that the theory makes.

However, to really appreciate the successes and possible defects of the superstring theory, we must first try to understand the historical problems that have plagued physicists for the past five decades. Let us now turn to a quick review of the development of gauge theories in order to appreciate the difficulty of constructing a finite theory of gravity. We will also briefly sketch the historical development of the superstring theory.

§1.2. Historical Review of Gauge Theory

In the 1960s, elementary particle physics seemed hopelessly mired in confusion. The weak, electromagnetic, strong, and gravitational forces were each studied separately, largely in isolation of the others. Moreover, investigations into each force had reached a fundamental roadblock:

(1) The weak interactions. Theoretical models of the weak interactions had progressed embarrassingly little beyond the Fermi theory first proposed three decades earlier in the 1930s:

$$L_{\text{Fermi}} \sim \bar{\psi}_p \Gamma^A \psi_n \bar{\psi}_e \Gamma_A \psi_\nu \tag{1.2.1}$$

where the Γ^A represents various combinations of Dirac matrices. The next major step, a theory of W bosons, was plagued with the problem of infinities. Furthermore, no one knew the underlying symmetry among the leptons, or whether there was any.

(2) The strong interactions. In contrast to weak interactions, the Yukawa meson theory provided a renormalizable theory of the strong interactions:

$$L_{\text{Yukawa}} \sim g\bar{\psi}\psi\phi \tag{1.2.2}$$

However, the Yukawa theory could not explain the avalanche of "elementary" particles that were being discovered in particle accelerators. J. Robert Oppenheimer even suggested that the Nobel Prize in physics should go

to the physicist who *didn't* discover a particle that year. Furthermore, the quark model, which seemed to fit the data much better than it had any right to, was plagued with the fact that quarks were never seen experimentally.

(3) The gravitational force. Gravity research was totally uncoupled from research in the other interactions. Classical relativists continued to find more and more classical solutions in isolation from particle research. Attempts to canonically quantize the theory were frustrated by the presence of the tremendous redundancy of the theory. There was also the discouraging realization that even if the theory could be successfully quantized, it would still be nonrenormalizable.

This bleak landscape changed dramatically in the early 1970s with the coming of the gauge revolution. One of the great achievements of the past 15 years has been the development of a fully renormalizable theory of spin-1 gauge particles in which, for the first time, physicists could actually calculate realistic S-matrix elements. Thus, it took fully 100 years to advance beyond the original gauge theory first proposed by Maxwell in the 1860s! (See the Appendix for an elementary introduction to gauge theories and group theory.)

Apparently the key to eliminating the divergences of relativistic quantum mechanics is to go to larger and more sophisticated gauge groups. Symmetry, instead of being a purely aesthetic feature of a particular model, now becomes its most important feature.

For example, Maxwell's equations, which provided the first unification of the electric force with the magnetic force, has a gauge group given by U(1). The unification of the weak and electromagnetic forces into the electro-weak force requires $SU(2) \otimes U(1)$. The forces that bind the quarks together into the hadrons, or quantum chromodynamics (QCD), are based on SU(3). All of elementary particle physics, in fact, is compatible with the minimal theory of $SU(3) \otimes SU(2) \otimes U(1)$.

Although the verdict is still not in on the GUTs, which are supposed to unite the electroweak force with the strong force, once again the unifying theme is gauge symmetry, with such proposals as SU(5), O(10), etc. symmetry.

Although the gauge revolution is perhaps one of the most important developments in decades, it is still not enough. There is a growing realization that the Yang–Mills theory by itself cannot push our understanding of the physical universe beyond the present level. Not only do the GUTs fail to explain important physical phenomena, but also there is the crucially important problem of formulating a quantum theory of gravity.

Grand Unified Theories, first of all, cannot be the final word on the unification of all forces. There are several features of GUTs that are still unresolved:

(1) GUTs cannot resolve the problem of why there are three nearly exact copies or families of elementary particles. We still cannot answer Rabi's question, "Who ordered the muon?"

(2) GUTs still have 20 or so arbitrary parameters. They cannot, for example, calculate the masses of the quarks, or the various Yukawa couplings. A truly unified field theory should have at most one arbitrary parameter.
(3) GUTs have difficulty solving the hierarchy problem. Unless we appeal to supersymmetry, it is hard to keep the physics of incredibly massive particles from mixing with everyday energies and destroying the hierarchy.
(4) The unification of particle forces occurs around 10^{-28} cm which is very close to the Planck length of 10^{-33} cm, where we expect gravitational effects to become dominant. Yet GUTs say nothing whatsoever about gravitation.
(5) So far, proton decay has not been conclusively observed, which already rules out minimal SU(5). There is, therefore, still no compelling experimental reason for introducing the theory.
(6) It is difficult to believe that no new interactions will be found between present-day energies and the unification scale. The "desert" may very well bloom with new interactions yet unknown.

The most perplexing and the most challenging of these problems, from a foundational point of view, has been to find a way of quantizing Einstein's theory of general relativity. Although Yang–Mills theories have had spectacular successes in unifying the known laws of particle physics, the laws of gravity are curiously different at a fundamental level. Clearly, Yang–Mills theory and conventional gauge theory are incapable of dealing with this problem. Thus, GUTs are faced with formidable experimental and theoretical problems when pushed to their limits.

General relativity is also plagued with similar difficulties when pushed to its limits:

(1) Classically, it has been established that Einstein's equations necessarily exhibit pointlike singularities, where we expect the laws of general relativity to collapse. Quantum corrections must dominate over the classical theory in this domain.
(2) The action is not bounded from below, because it is linear in the curvature tensor. Thus, it may not be stable quantum mechanically.
(3) General relativity is not renormalizable. Computer calculations, for example, have now conclusively shown that there is a nonzero counterterm in Einstein's theory at the two-loop level.

Naive attempts to quantize Einstein's theory of gravitation have met with disappointing failure. One of the first to point out that general relativity would be incompatible with quantum mechanics was Heisenberg, who noted that the presence of a dimensional coupling constant would ruin the usual renormalization program.

If we set

$$\frac{h}{2\pi} = 1; \qquad c = 1 \tag{1.2.3}$$

there still remains a dimensional constant even in the Newtonian theory of gravity, the gravitational constant G:

$$F = G\frac{m_1 m_2}{r^2} \tag{1.2.4}$$

which has dimensions of centimeters squared. When we power expand the metric tensor $g_{\mu\nu}$ around flat space with the metric $\eta_{\mu\nu} = (-+++)$, we introduce the coupling constant κ, which has dimensions of centimeters:

$$g_{\mu\nu} = \eta_{\mu\nu} + \kappa h_{\mu\nu} \tag{1.2.5}$$

Therefore:

$$G \sim \kappa^2 \tag{1.2.6}$$

In this system of units, where the only unit is the centimeter, this coupling constant κ becomes the Planck length, 10^{-33} cm or 10^{19} GeV, which is far beyond the reach of experimentation!

Renormalization theory, however, is founded on the fundamental premise that we can eliminate all divergences with an infinite redefinition of certain constants. Having a dimensional coupling constant means that this complicated reshuffling and resumming of graphs is impossible. Unlike standard renormalizable theories, in quantum gravity we cannot add diagrams that have different powers of the coupling constant. *This means that general relativity cannot be a renormalizable theory.* The amplitude for graviton–graviton scattering, for example, is now a power expansion in a dimensional parameter (see Fig. 1.3):

$$A = \sum_{n=2}^{\infty} \kappa^n A_n \tag{1.2.7}$$

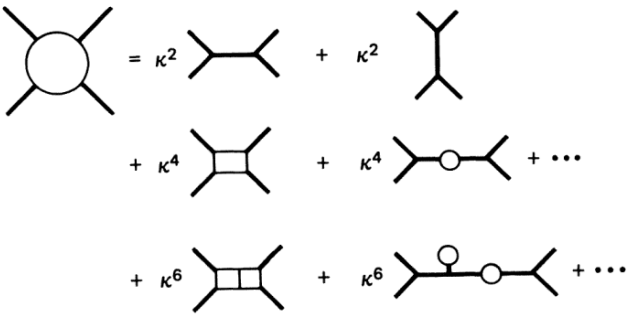

Figure 1.3. Scattering amplitude for graviton–graviton scattering. Because the coupling constant has dimensions, graphs of different order cannot be added to renormalize the theory. Thus, theories containing quantum gravity must be either divergent or completely finite order-by-order. Pure quantum gravity has been shown on computer to diverge at the two-loop level. Counterterms have also been found for quantum gravity coupled to lower-spin particles. Thus, superstring theory is the only candidate for a finite theory.

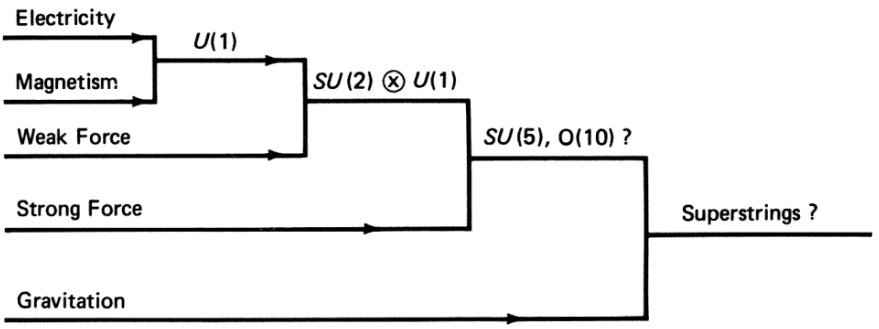

Figure 1.4. Chart showing how gauge theories based on Lie groups have united the fundamental forces of nature. Maxwell's theory, based on U(1), unites electricity and magnetism. The Weinberg–Salam model, based on SU(2) ⊗ U(1), unites the weak force with the electromagnetic force. GUT theories (based on SU(5), O(10), or larger groups) are the best candidate to unite the strong force with the electro-weak force. Superstring theory is the only candidate for a gauge theory that can unite gravity with the rest of the particle forces.

where we are no longer able to shuffle graphs with different values of n to cancel the infinities, which is the heart of renormalization theory. Thus, renormalization theory breaks down.

Because general relativity is hopelessly outside the domain of conventional renormalization theory, one must reconsider Dirac's fundamental objection. It was Dirac who said that the success of quantum mechanics was based on approximation schemes where each correction term was increasingly small. But renormalization theory is flawed because it maximally violates this principle and manipulates infinite quantities and discards them at the end.

One solution might be to construct a theory of gravity that is finite to every order in the coupling constant, with no need for renormalization at all. For a while, one bright hope was supergravity [14, 15], based on the local gauge group Osp(N/4) (see Appendix), which was the first nontrivial extension of Einstein's equations in 60 years. The hope was that this gauge group would offer us a large enough set of Ward–Takahashi identities to cancel a large class of divergent diagrams. The larger the gauge group, the more likely troublesome infinites would cancel (see Fig. 1.4):

Theory	Gauge group
Electromagnetism	U(1)
Electro-weak	SU(2) ⊗ U(1)
Strong	SU(3)
GUT(?)	SU(5), O(10)
Gravity (?)	GL(4), O(3, 1)
Supergravity (?)	Osp(N/4)

The basic strategy being pursued was

$$\text{Gauge symmetry} \rightarrow \text{Ward} - \text{Takahashi identities}$$

$$\rightarrow \text{Cancellation of graphs} \rightarrow \text{Renormalizable theory}$$

For example, even Einstein's theory of gravity can be shown to be trivially finite at the first loop level. There exists a remarkable identity, called the Gauss–Bonnet identity, which immediately shows that all one-loop graphs in general relativity (which would take a computer to write down) sum to zero. In fact, the super-Gauss–Bonnet identities eliminate many of the divergences of supergravity, but probably not enough to make the theory finite.

The largest and most promising of the supergravities, the O(8) supergravity, is probably divergent. Unfortunately, it is possible to write down locally supersymmetric counterterms at the seventh loop level. It is highly unlikely that the coefficients of this and probably an infinite number of other counterterms can all vanish without appealing to an even higher symmetry. This is discouraging, because it means that the gauge group of the largest supergravity theory, Osp(8/4), is still too small to eliminate the divergences of general relativity.

Furthermore, the O(8) gauge group is too small to accommodate the minimal $SU(3) \otimes SU(2) \otimes U(1)$ of particle physics. If we go to higher groups beyond O(8), we find that we must incorporate higher and higher spins into the theory. However, an interacting spin-3 theory is probably not consistent, making one suspect that O(8) is the limit to supergravity theories.

In conclusion, supergravity must be ruled out for two fundamental reasons:

(1) It is probably not a finite theory because the gauge group is not large enough to eliminate all possible supersymmetric counterterms. There is a possible counterterm at the seventh loop level.
(2) Its gauge group O(8) is not large enough to accommodate the minimal symmetry of particle physics, namely $SU(3) \otimes SU(2) \otimes U(1)$.

Physicists, faced with these and other stumbling blocks over the years, have concluded that perhaps one or more of our cherished assumptions about our universe must be abandoned. Because general relativity and quantum mechanics can be derived from a small set of postulates, one or more of these postulates must be wrong. The key must be to drop one of our commonsense assumptions about nature on which we have constructed general relativity and quantum mechanics. Over the years, several proposals have been made to drop some of our commonsense notions about the universe:

(1) Continuity
This approach assumes that space–time must be granular. The size of these grains would provide a natural cutoff for the Feynman integrals, allowing us to have a finite S-matrix. Integrals like

$$\int_{\varepsilon}^{\infty} d^4x \tag{1.2.8}$$

would then diverge as ε^{-n}, but we would never take the limit as ε goes to zero. Lattice gravity theories are of this type. In Regge calculus [16], for example, we latticize Riemannian space with discrete four-simplexes and replace the curvature tensor by the angular deficit calculated when moving in a circle around a simplex:

$$-\frac{1}{2\kappa^2}\sqrt{-g}R \to \text{angular deficit}$$

(In flat space, there is no angular deficit when walking around a closed path, and the action collapses.) Usually, in lattice theories, we take the limit as the lattice length goes to zero. Here, however, we keep it fixed at a small number [17]. At present, however, there is no experimental evidence to support the idea that space–time is granular. Although we can never rule out this approach, it seems to run counter to the natural progression of particle physics, which has been to postulate larger and more elegant groups.

(2) Causality

This approach allows small violations in causality. Theories that incorporate the Lee–Wick mechanism [18] are actually renormalizable, but permit small deviations from causality. These theories make the Feynman diagrams converge by adding a fictitious Pauli–Villars field of mass M that changes the ultraviolet behavior of the propagator. Usually, the Feynman propagator converges as p^{-2} in the ultraviolet limit. However, by adding a fictitious particle, we can make the propagator converge even faster, like p^{-4}:

$$\frac{1}{p^2+m^2}-\frac{1}{p^2+M^2}\to\frac{1}{p^4} \tag{1.2.9}$$

Notice that the Pauli–Villars field is a ghost because of the -1 that appears in the propagator. (This means that the theory will be riddled with negative probabilities.) Usually, we let the mass of the Pauli–Villars field tend to infinity. However, here we keep it finite, letting the pole go onto the unphysical sheet. Investigations of the structure of the resulting Feynman diagrams show, however, that causality is violated; that is, you can meet your parents before you are born.

(3) Unitarity

We can replace Einstein's theory, which is based on the curvature tensor, with a conformal theory based on the Weyl tensor:

$$\sqrt{-g}R_{\mu\nu}g^{\mu\nu}\to\sqrt{-g}C^2_{\mu\nu\rho\sigma} \tag{1.2.10}$$

where the Weyl tensor is defined as

$$C_{\mu\nu\rho\sigma}=R_{\mu\nu\rho\sigma}+g_{\mu[\sigma}R_{\rho]\nu}+g_{\nu[\rho}R_{\sigma]\mu}+\tfrac{1}{3}Rg_{\mu[\rho}g_{\sigma]\nu} \tag{1.2.11}$$

where the brackets represent antisymmetrization. The conformal tensor possesses a larger symmetry group than the curvature tensor, that is,

invariance under local conformal transformations:

$$\begin{cases} g_{\mu\nu} \to e^{\sigma} g_{\mu\nu} \\ C^{\mu}_{\nu\rho\sigma} \to C^{\mu}_{\nu\rho\sigma} \end{cases} \tag{1.2.12}$$

The Weyl theory converges because the propagators go as p^{-4}; that is, it is a higher derivative theory. However, there is a "unitary ghost" that also appears with a -1 in the propagator, for the same reasons cited above. The most optimistic scenario would be to have these unitary ghosts "confined" by a mechanism similar to quark confinement [19, 20].

(4) Locality
Over the years, there have also been proposals to abandon some of the important postulates of quantum mechanics, such as locality. After all, there is no guarantee that the laws of quantum mechanics should hold down to distances of 10^{-33} cm. However, there have always been problems whenever physicists tried to deviate from the laws of quantum mechanics, such as causality. At present, there is no successful alternative to quantum mechanics.

(5) Point Particles
Finally, there is the approach of superstrings, which abandons the concept of idealized point particles, first introduced 2000 years ago by the Greeks.

The superstring theory, because it abandons only the assumption that the fundamental constituents of matter must be point particles, does the least amount of damage to cherished physical principles and continues the tradition of increasing the complexity and sophistication of the gauge group. Superstring theory does not violate any of the laws of quantum mechanics, yet manages to eliminate most, if not all, of the divergences of the Feynman diagrams. The symmetry group of the superstring model, the largest ever encountered in the history of physics, is probably large enough to make the theory finite to all orders. Once again, it is symmetry, and not the breakdown of quantum mechanics, that is the fundamental key to rendering a theory finite.

In Fig. 1.5 we see diagrammatically the evolution of various theories of gravity. First, there was Newton's theory of action at a distance, where gravitational interactions travel faster than the speed of light. Einstein replaced this with the classical interpretation of curved manifolds. Quantum gravity, in turn, makes quantum corrections to Einstein's theory by adding in loops. Finally, the superstring theory makes further corrections to the point particle quantum theory by summing over all possible topological configurations of interacting strings.

Superstring theory, however, is quite unlike its predecessors in its historical development. Unlike other physical theories, superstring theory has perhaps one of the strangest histories in science, with more twists and turns than a roller coaster.

First, two young physicists, Veneziano and Suzuki [21, 22], independently

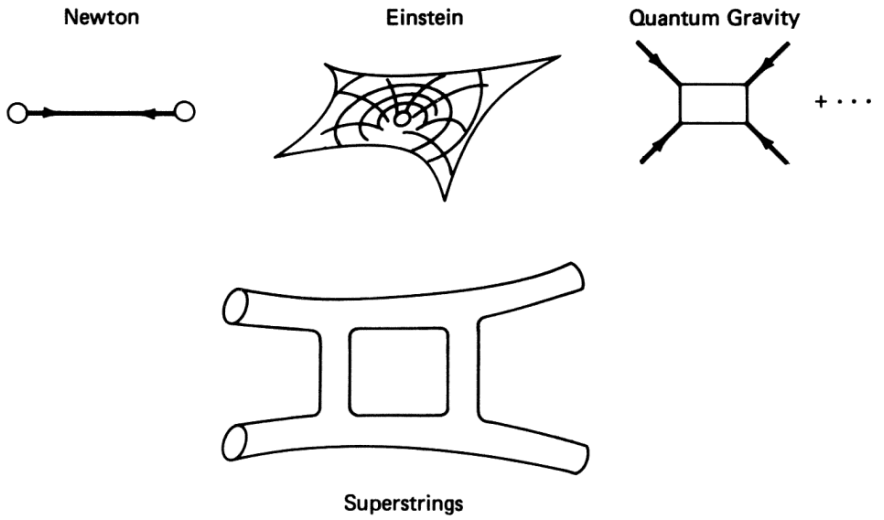

Figure 1.5. Steps in the evolution of the theory of gravitation. Each step in this chart builds on the successes of the previous step. Newton thought gravity was a force that acted instantly over a distance. Einstein proposed that gravitation was caused by the curvature of space–time. The naive merger of general relativity and quantum mechanics produces a divergent theory, quantum gravity, which assumes that gravitation is caused by the exchange of particle-like gravitons. Superstring theory proposes that gravitation is caused by the exchange of closed strings.

discovered its quantum theory when they were thumbing through a math book and accidentally noted that the Euler Beta function satisfied all the postulates of the S-matrix for hadronic interactions (except unitarity). Neveu, Schwarz, and Ramond [23–25] quickly generalized the theory to include spinning particles. To solve the problem of unitarity, Kikkawa, Sakita, and Virasoro [26] proposed that the Euler Beta function be treated as the Born term to a perturbation series. Finally, Kaku, Yu, Lovelace, and Alessandrini [27–33] completed the quantum theory by calculating bosonic multiloop diagrams. The theory, however, was still formulated entirely in terms of on-shell S-matrix amplitudes.

Next, Nambu and Goto [34–35] realized that lurking behind these scattering amplitudes was a classical relativistic string. In one sweep, they revolutionized the entire theory by revealing the unifying, classical picture behind the theory. The relationship between the classical theory and the quantum theory was quickly made by Goldstone, Goddard, Rebbi, and Thorn [36] and further developed by Mandelstam [37]. The theory, however, was still formulated as a first quantized theory, so that the measure, the vertices, the counting of graphs, etc. all had to be postulated ad hoc and not deduced from first principles.

The action (in a particular gauge) was finally written down by Kaku and

Kikkawa [38]. At last, the model could be derived from one action strictly in terms of physical variables, although the action did not have any symmetries left. However, when it was discovered that the theory was defined only in 10 and 26 dimensions, the model quickly died. Furthermore, the rapid development of QCD as a theory of hadronic interactions seemingly put the last nail in the coffin of the superstring.

For 10 years, the model languished because no one could believe that a 10- or 26-dimensional theory had any relevance to 4-dimensional physics. When Scherk and Schwarz [39] made the outrageous (for its time) suggestion that the dual model was actually a theory of all known interactions, no one took the idea very seriously. The idea fell like a lead balloon.

Finally, the discovery in 1984 by Green and Schwarz [40] that the superstring theory is anomaly-free and probably finite to all orders in perturbation theory has revived the theory. The $E_8 \otimes E_8$ "heterotic string" of Gross, Harvey, Martinec, and Rohm [41] at present seems to be the best candidate for unifying gravity with physically reasonable models of particle interactions.

One of the active areas of research now is to complete the evolution of the theory, to discover why all the "miracles" occur in the model. There has been a flurry of activity in the direction of writing down the covariant action using methods discovered in the intervening 10 years, such as BRST. However, there is now a growing realization that the covariant BRST formalism itself is a gauge-fixed formalism, much like the light cone formalism. Recently, however, there has been work on a truly *geometric field theory* where all the features of the theory can be deduced from simple physical principles. This is explained in Chapter 8. This would truly complete the evolution of the theory, which has been progressing backward for the past 20 years

$$\text{Quantum theory} \rightarrow \text{Classical theory} \rightarrow \text{Action} \rightarrow \text{Geometry}$$

Let us summarize some of the promising positive features of the superstring model:

(1) The gauge group includes $E_8 \otimes E_8$, which is much larger than the minimal group $SU(3) \otimes SU(2) \otimes U(1)$. There is plenty of room for phenomenology in this theory.
(2) The theory has no anomalies. These small but important defects in a quantum field theory place enormous restrictions on what kinds of theories are self-consistent. The symmetries of the superstring theory, by a series of "miracles," can cancel all of its potential anomalies.
(3) Powerful arguments from the theory of Riemann surfaces indicate that the theory is finite to all orders in perturbation theory (although a rigorous proof is still lacking).
(4) There is very little freedom to play with. Superstring models are notoriously difficult to tinker with without destroying their miraculous properties. Thus, we do not have the problem of 20 or more arbitrary coupling constants.

(5) The theory includes GUTs, super-Yang–Mills, supergravity, and Kaluza–Klein theories as subsets. Thus, many of the features of the phenomenology developed for these theories carry over into the string theory.

Superstring theory, crudely speaking, unites the various forces and particles in the same way that a violin string provides a unifying description of the musical tones. By themselves, the notes *A*, *B*, *C*, etc. are not fundamental. However, the violin string is fundamental; one physical object can explain the varieties of musical notes and even the harmonies one can construct from them. In much the same way, the superstring provides a unifying description of elementary particles and forces. In fact, the "music" created by the superstring is the forces and particles of nature.

Although superstring theory, because of its fabulously large set of symmetries, has "miraculous" cancellations of anomalies and divergences, we must also present a balanced picture and point out its shortcomings. To be fair we must also list the potential problems of the theory that have been pointed out by critics of the model:

(1) It is impossible experimentally to reach the tremendous energies found at the Planck scale. Therefore, the theory is in some sense untestable. A theory that is untestable is not an acceptable physical theory.
(2) Not one shred of experimental evidence has been found to confirm the existence of supersymmetry, let alone superstrings.
(3) It is presumptuous to assume that there will be no surprises in the "desert" between 100 and 10^{19} GeV. New, totally unexpected phenomena have always cropped up when we have pushed the energy scale of our accelerators. Superstring theory, however, makes predictions over the next 17 orders of magnitude, which is unheard of in the history of science.
(4) The theory does not explain why the cosmological constant is zero. Any theory that claims to be a "theory of everything" must surely explain the puzzle of a vanishing cosmological constant, but it is not clear how superstrings solves this problem.
(5) The theory has an embarrassment of riches. There are apparently *thousands* of ways to break down the theory to low energies. Which is the correct vacuum? Although the superstring theory can produce the minimal theory of $SU(3) \otimes SU(2) \otimes U(1)$, it also predicts many other interactions that have not yet been seen.
(6) No one really knows how to break a 10-dimensional theory down to 4 dimensions.

Of these six objections to the model, the most fundamental is the last, the inability to calculate dimensional breaking. The reason for this is simple: to every order in perturbation theory, the dimension of space–time is stable. Thus, in order to have the theory spontaneously curl up into 4- and 6-dimensional universes, we must appeal to nonperturbative, dynamical effects, which are notoriously difficult to calculate. This is why the search for the geometry underlying the theory is so important. The geometric formulation

of the model may give us the key insight into the model that will allow us to make nonperturbative calculations and make definite predictions with the theory.

Thus, the criticism that the model cannot be tested at the Planck length is actually slightly deceptive. *The superstring theory, if it could be successfully broken dynamically, should be able to make predictions down to the level of everyday energies.* For example, it should be able to predict the masses of the quarks. Therefore, we do not have to wait for several centuries until we have accelerators that can reach the Planck length.

Thus, the fundamental problem facing superstrings is not necessarily an experimental one. It is mainly theoretical. The outstanding problem of the theory is to calculate dynamical symmetry breaking, so that its predictions can be compared with experimental data at ordinary energies.

A fundamental theory at Planck energies is also a fundamental theory at ordinary energies. Thus, the main stumbling block to the development of the theory is an understanding of its nonperturbative behavior. And the key to this understanding probably lies in a second quantized, geometric formulation of the model.

In Part I of this book, however, we will follow historical precedent and present the first quantized formulation of the model. As we will stress throughout this book, the first quantized theory seems to be a loose collection of random facts. As a consequence, we have emphasized the path integral formulation (first written down for the Veneziano model by Hsue, Sakita, and Virasoro [42, 43]) as the most powerful method of formulating the first quantized theory. Although the path integral approach cannot reveal the underlying geometric formulation of the model, it provides the most comprehensive formulation of the first quantized theory.

We will now turn to the functional formulation [44] of point particle theory, which can be incorporated almost directly into the string theory.

§1.3. Path Integrals and Point Particles

Let us begin our discussion by analyzing the simplest of all possible systems, the classical nonrelativistic point particle. Surprisingly, much of the analysis of this simple dynamical system carries over directly to the superstring theory. The language we will use is the formalism of path integrals, which is so versatile that it can accommodate both first quantized point particles and second quantized gauge fields with equal ease.

As in classical mechanics, the starting point is the Lagrangian for a point particle:

$$L = \tfrac{1}{2}m\dot{x}_i^2 - V(x) \tag{1.3.1}$$

where the particle is moving in an external potential. The real physics is contained in the statement that the action S must be minimized. The equations

of motion can be derived by minimizing the action:

$$S = \int L(x_i, \dot{x}_i, t)\, dt$$
$$\delta S = 0 \tag{1.3.2}$$

To calculate the equations of motion, let us make a small variation in the path of the particle given by

$$\delta x_i; \qquad \delta \dot{x}_i \tag{1.3.3}$$

Under this small variation, the action varies as follows:

$$\int dt \left\{ \frac{\delta L}{\delta x_i} \delta x_i + \frac{\delta L}{\delta \dot{x}_i} \delta \dot{x}_i \right\} = 0 \tag{1.3.4}$$

Integrating by parts, we arrive at the Euler–Lagrange equations:

$$\frac{\delta L}{\delta x_i} - \frac{d}{dt} \frac{\delta L}{\delta \dot{x}_i} = 0 \tag{1.3.5}$$

For our point particle, the equations of motion become

$$m \frac{d^2 x_i}{dt^2} = -\frac{\partial V(x)}{\partial x_i} \tag{1.3.6}$$

which correspond to the usual classical Newtonian equations of motion.

In addition to the Lagrangian formulation of classical mechanics, there is also the Hamiltonian form. Instead of introducing the position and the velocities as fundamental objects, we now introduce the position and the momentum:

$$p_i = \frac{\delta L}{\delta \dot{x}_i} \tag{1.3.7}$$

With this definition of the conjugate variable, we have

$$H = p_i \dot{x}_i - L$$
$$H(p_i, x_i) = \frac{p_i^2}{2m} + V(x) \tag{1.3.8}$$

Finally, the Poisson brackets between the momenta and the coordinates are given by

$$[p_i, x_j]_{\mathrm{PB}} = -\delta_{ij} \tag{1.3.9}$$

A celebrated theorem in classical mechanics states that the equations of motion of Newton and the action principle method can be shown to be identical. Beginning with the action principle, we can derive Newton's laws of motion, and vice versa.

Equations of motion $\leftrightarrow$ Action principle

This equivalence, however, breaks down at the quantum level. Quantum mechanically, there is a fundamental difference between the two, with the equations of motion being only an approximation to the actual quantum behavior of matter. *Thus, the action principle is the only acceptable framework for quantum mechanics.*

Let us now reformulate the principles of quantum mechanics in terms of Feynman path integrals [44]:

(1) The probability $P(a, b)$ of a particle moving from point a to point b is the square of the absolute value of a complex number, the transition function $K(a, b)$:

$$P(a, b) = |K(a, b)|^2 \tag{1.3.10}$$

(2) The transition function is given by the sum of a certain phase factor, which is a function of the action S, taken over all possible paths from a to b:

$$K(a, b) = \sum_{\text{paths}} k e^{i2\pi S/h} \tag{1.3.11}$$

where the constant k can be fixed by

$$K(a, c) = \sum_{\text{paths}} K(a, b)K(b, c) \tag{1.3.12}$$

and the intermediate sum is taken over paths that go through all possible intermediate points b.

The second principle says that a particle "sniffs out" all possible paths from point a to point b, no matter how complicated the paths may be. We calculate this phase factor for each of this infinite number of paths. Then the transition factor for the path between a and b is calculated by summing over all possible phase factors (see Fig. 1.6).

Remarkably, *the essence of quantum mechanics is captured in these two principles.* All the profoundly important implications of quantum mechanics, which represent a startling departure from classical mechanics, can be derived from these two innocent-sounding principles! In particular, these two principles summarize the essence of the quantum interpretation of the double-slit experiment, which, in turn, summarizes the essence of quantum mechanics itself.

It is apparent at this point that the results of classical mechanics can be reproduced from our two assumptions in a certain approximation. Notice that, for values of S that are large compared to Planck's constant, the phase factor fluctuates rapidly, canceling out these contributions:

$$\delta S \gg \frac{h}{2\pi}: \sum_{\text{paths}} e^{i2\pi S/h} \to 0 \tag{1.3.13}$$

Thus, the only contributions to the path integral that survive are those for which the deviations in the action from the classical path are on the order of

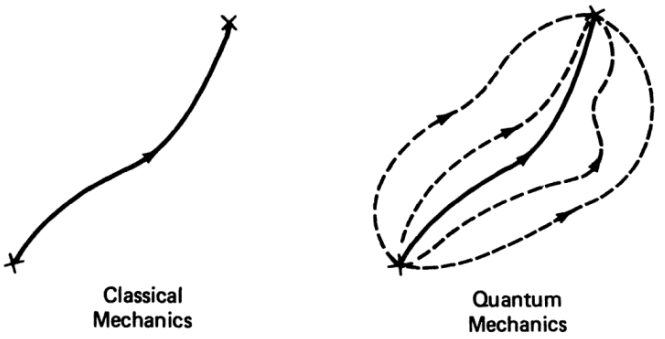

Figure 1.6. The essential difference between classical mechanics and quantum mechanics. Classical mechanics assumes that a particle executes just one path between two points based either on the equations of motion or on the minimization of the action. By constrast, quantum mechanics sums the contributions of probability functions (based on an action) for all possible paths between two points. Although the classical path is the one most favored, in principle all possible paths contribute to the path integral. Thus, the action principle is more fundamental than the equations of motion at the quantum level.

Planck's constant:

$$\delta S \sim \frac{h}{2\pi} \tag{1.3.14}$$

We see that the Euler–Lagrange equations of motion are reproduced only in a certain classical limit, that is, when Planck's constant goes to zero. Therefore the size of Planck's constant ultimately determines the probability that a particle will execute trajectories that are forbidden classically. We see the origin of Heisenberg's uncertainty principle embodied in these two principles.

Now let us try to reformulate this more precisely in terms of path integrals. The second principle now reads

$$K(a, b) = \int_a^b Dx\, e^{i2\pi S/h} \tag{1.3.15}$$

where

$$K(a, c) = \int K(a, b)K(b, c)\, Dx_b \tag{1.3.16}$$

and

$$\sum_{\text{paths}} \to \int Dx = \lim_{N\to\infty} \int \prod_{i=1}^{3} \prod_{n=1}^{N} dx_{i,n} \tag{1.3.17}$$

where the index n labels N intermediate points that divide the interval between the initial and the final coordinate. We will take the limit when N approaches infinity.

It is absolutely essential to understand that the integration Dx is not the ordinary integration over x. In fact, it is the product of all possible integrations over all intermediate points $x_{i,n}$ between points a and b. This crucial difference between ordinary integration and functional integration goes to the heart of the path integral formalism.

This infinite series of integrations, in turn, is equivalent to summing over all possible paths between a and b. Thus, we will have to be careful to include normalization factors when performing an integration over an infinite number of intermediate points.

If we take the simple case where $L = \frac{1}{2}m\dot{x}_i^2$, all functional integrations can actually be performed exactly. The integral in question is a Gaussian, which is fortunately one of the small number of functional integrals that can actually be performed. One of the great embarrassments of the method of path integrals is that one of the few integrals that can actually be performed is

$$\int_{-\infty}^{\infty} dx\, x^{2n} e^{-r^2x^2} = \frac{\Gamma(n+\frac{1}{2})}{r^{2n+1}} \tag{1.3.18}$$

We will be using this formula throughout the entire book.

Let us now break up the path into an infinite number of intermediate points, $x_{i,n}$. (Notice that the functional expression integrates over all possible values of the intermediate point $x_{i,n}$, so we cannot expect that $x_{i,n}$ and $x_{i,n+1}$ are close to each other even for small time separations.) Let us write

$$\begin{gathered} dt \to \varepsilon \\ \tfrac{1}{2}m\dot{x}_i^2\, dt \to \tfrac{1}{2}m(x_n - x_{n+1})_i^2 \varepsilon^{-1} \end{gathered} \tag{1.3.19}$$

In order to perform the functional integral over an infinite number of intermediate points, we will repeatedly use the following Gaussian integration:

$$\begin{aligned} &\int_{-\infty}^{\infty} dx_2 \exp[-a(x_1 - x_2)^2 - a(x_2 - x_3)^2] \\ &\quad = \sqrt{\frac{\pi}{2a}} \exp[-\tfrac{1}{2}a(x_1 - x_3)^2] \end{aligned} \tag{1.3.20}$$

One of the crucial points to observe here is that the integration over a Gaussian in one of the intermediate points yields another Gaussian with that intermediate point removed. This is the fundamental reason why we can perform the functional integration over an infinite number of intermediate points.

Finally, the path integral that we wish to perform is given by

$$\begin{aligned} K(a,b) = \lim_{\varepsilon \to 0} &\int\int \cdots \int dx_1\, dx_2 \cdots dx_{N-1} \\ &\times \left(\frac{2\pi i\varepsilon}{m}\right)^{-(1/2)N} \exp\left\{\frac{im}{2\varepsilon} \sum_{n=1}^{N} (x_n - x_{n-1})^2\right\} \end{aligned} \tag{1.3.21}$$

(where we have suppressed the vector index i). Using the previous relation (1.3.20), the final result is equal to

$$K(a, b) = \left| \frac{m}{2\pi(t_b - t_a)} \right|^{1/2} \exp \frac{\frac{1}{2}im(x_b - x_a)^2}{t_b - t_a} \tag{1.3.22}$$

The transition probability function K has some very interesting properties. For example, it solves the wave equation:

$$\frac{-1}{2m} \frac{\partial^2}{\partial x_a^2} K(a, b) = i \frac{\partial}{\partial t_a} K(a, b) \tag{1.3.23}$$

when t_a is greater than t_b.

Later, we will generalize these expressions for the case of freely propagating strings, and we will find that these expressions for the Green's functions carry over with only small, but important, changes.

To show the relationship between the Hamiltonian and Lagrangian formalisms in the path integral approach , it is helpful to insert a complete set of intermediate states when we divide up the path from a to b. Let us treat the variable x as an operator $\hat{x}$ acting on a set of eigenstates:

$$\hat{x}|x\rangle = x|x\rangle \tag{1.3.24}$$

The $|x\rangle$ represents an eigenstate of the position operator, treating $\hat{x}$ as an operator whose eigenvalue is equal to the number x. Then completeness over eigenstates for coordinates and for momenta can be represented as

$$\begin{aligned} 1 &= \int |x\rangle \, dx \, \langle x| \\ 1 &= \int |p\rangle \, dp \, \langle p| \end{aligned} \tag{1.3.25}$$

We normalize our states as follows:

$$\begin{aligned} \langle x|y\rangle &= \delta(x - y) \\ \langle p|x\rangle &= \frac{e^{ipx}}{\sqrt{2\pi}} \end{aligned} \tag{1.3.26}$$

(Because of the infinite number of normalization constants that constantly appear in the path integral formalism, we will often delete them for the sake of clarity in this book. We do not lose any generality, because we can, of course, reinsert them into the path integral if we desire.)

With these eigenstates, we can now rewrite the expression for the Green's function for going from point x_1 to x_N:

$$K(1, N) = \langle x_1, t_1 | x_N, t_N \rangle \tag{1.3.27}$$

In order to derive the previous expression (1.3.22) for the transition amplitude, let us insert a complete set of intermediate states at every intermediate point

between x_1 and x_N:

$$\langle x_1, t_1 | x_N, t_N \rangle = \langle x_1, t_1 | x_2, t_2 \rangle \int dx_2 \langle x_2, t_2 | \int dx_2 \cdots | x_{N-1}, t_{N-1} \rangle \int dx_{N-1} \langle x_{N-1}, t_{N-1} | x_N, t_N \rangle \quad (1.3.28)$$

Now let us examine each infinitesimal propagator in terms of the Hamiltonian, which we write as a function of the coordinates and derivatives:

$$H = H(x, \partial_x) \quad (1.3.29)$$

Then the transition for an infinitesimal interval is given by

$$\begin{aligned} \langle x_1, t_1 | x_2, t_2 \rangle &= \langle x_1 | e^{-iH(x,\partial_x)\delta t} | x_2 \rangle \\ &= e^{-iH(x,\partial_x)\delta t} \langle x_1 | x_2 \rangle \\ &= e^{-iH(x,\partial_x)\delta t} \langle x_1 | p \rangle \int dp \langle p | x_2 \rangle \\ &= e^{-iH(x,p)\delta t} \int \frac{dp}{2\pi} e^{ip(x_2 - x_1)} \\ &= e^{-iH(x,p)\delta t} \int \frac{dp}{2\pi} e^{ip\dot{x}\delta t} \end{aligned} \quad (1.3.30)$$

It is very important to notice that path integrals have made it possible to make the transition from classical to quantum commutators. The Hamiltonian can be expressed either as a function of derivatives with respect to the position or as a function of the canonical momenta because of the identity:

$$\partial_x e^{ipx} = ipe^{ipx} \quad (1.3.31)$$

This allows us to make the important identification:

$$\begin{cases} H(x, p) \leftrightarrow H(x, \partial_x) \\ p \leftrightarrow -i \dfrac{\delta}{\delta x} \end{cases} \quad (1.3.32)$$

In the functional formalism, the important correspondence between momenta and partial derivatives arises because of this identity.

Putting everything together, we can now write the complete transition amplitude as

$$\langle x_1, t_1 | x_N, t_N \rangle = \int_{x_1}^{x_N} Dp\, Dx \exp i \int_{t_1}^{t_N} [p\dot{x} - H(p, x)]\, dt \quad (1.3.33)$$

where

$$H = \frac{p_i^2}{2m} + V(x) \quad (1.3.34)$$

(As usual, we have dropped all the intermediate normalizations, which are just factors of 2π.) Notice that the functional integral, which was once only a function of the coordinates, is now a function of both the momenta and the coordinates.

In order to retrieve the original Lagrangian, we can perform the p integration exactly, because it is a simple Gaussian integral, and we arrive at

$$\langle x_1, t_1 | x_N, t_N \rangle = \int_{x_1}^{x_N} Dx \exp i \int_{t_1}^{t_N} [\tfrac{1}{2} m \dot{x}_i^2 - V(x)]\, dt \tag{1.3.35}$$

We have thus made the transition between the Lagrangian and the Hamiltonian formalism using functional methods. We can use either:

$$L = \tfrac{1}{2} m \dot{x}_i^2 - V(x) \leftrightarrow H = \frac{p_i^2}{2m} + V(x) \tag{1.3.36}$$

Functionally, the only difference between these two expressions is whether we integrate over the coordinates or a combination of the coordinates and the momenta. The transition probability can be represented as

$$\begin{aligned} K(a, b) &= \int_{x_a}^{x_b} Dx \exp i \int_{t_a}^{t_b} dt [\tfrac{1}{2} m \dot{x}_i^2 - V(x)] \\ &= \int_{x_a}^{x_b} Dx\, Dp \exp i \int_{t_a}^{t_b} dt \left[p \dot{x}_i - \frac{p_i^2}{2m} - V(x) \right] \end{aligned} \tag{1.3.37}$$

§1.4. Relativistic Point Particles

So far, our discussion has been limited to nonrelativistic particles, where all degrees of freedom are physical. However, nontrivial complications occur when we generalize our previous discussion to the case of relativistic particles. In particular, the (-1) appearing in the Lorentz metric will, in general, cause nonphysical states to propagate in the theory. These nonphysical "ghost" states, which have negative probability, must be eliminated carefully to ensure a sensible causal theory free of negative norm states.

For the relativistic case, let us assume that the location of a point particle is given by a four-vector:

$$x_\mu(\tau) \tag{1.4.1}$$

where parametrization τ does *not* necessarily refer to the time. The action is particularly simple, being proportional to the four-dimensional path length:

$$S = -m \int ds = -m(\text{length}) \tag{1.4.2}$$

The path length ds can be written in terms of the coordinates:

$$ds = \sqrt{-\dot{x}_\mu^2}\, d\tau \tag{1.4.3}$$

where the dot refers to differentiation with respect to the parameter τ. This

of second-order derivatives in the variable $x_\mu(\tau)$ and the field e. The "non-linear" Lagrangian (1.4.3) is expressed only in terms of $x_\mu(\tau)$. It can be derived from the second-order form by functionally integrating over the e field. And finally, the "Hamiltonian" form contains both $x_\mu(\tau)$ and the canonical conjugate $p_\mu(\tau)$ (it is first-order in derivatives):

$$\begin{aligned} \text{1st-order (Hamiltonian) form:} \quad & L = p_\mu \dot{x}^\mu - \tfrac{1}{2}e(p_\mu^2 + m^2) \\ \text{2nd-order form:} \quad & L = \tfrac{1}{2}(e^{-1}\dot{x}_\mu^2 - em^2) \\ \text{Non-linear form:} \quad & L = -m\sqrt{-\dot{x}_\mu^2} \end{aligned} \tag{1.4.16}$$

All three are invariant under reparametrization. Each of them has its own distinct advantages and disadvantages. This exercise in writing the action of the free relativistic particle in three different ways is an important one because it will carry over directly into the string formalism. Expressed in terms of path integrals, the point particle theory and the string theory are remarkably similar.

§1.5. First and Second Quantization

In this section, we will quantize the classical point particle and then show the relationship to the more conventional second quantized formalism of field theory. The first quantization program, we shall see, is rather clumsy compared to the second quantized formalism that most physicists are familiar with, but historically the string theory evolved as a first quantized theory. The great advantage of the second quantized formalism is that the entire theory can be derived from a single action, whereas the first quantized theory requires many additional assumptions.

The transition from the classical to the quantum system is intimately linked with the question of eliminating redundant infinites. As we said before, the path integral is formally ill-defined because we are summing over an infinite number of copies of the same thing. The trick is to single out just one copy.

There are at least three basic ways in which the first quantized point particle may be quantized: the Coulomb gauge, the Gupta–Bleuler formalism, and the BRST formalism.

Coulomb Quantization

Here, we choose the gauge

$$x_0 = t = \tau \tag{1.5.1}$$

In other words, we set the time component of the x variable equal to the real time t, which now parametrizes the evolution of the string. In this gauge, the

action reduces to

$$L = -m \int \sqrt{1 - v_i^2}\, dt \tag{1.5.2}$$

In the limit of velocities small compared to the velocity of light, we have

$$L \sim \tfrac{1}{2} m \dot{x}_i^2 \tag{1.5.3}$$

as before, so that the functional integral is modified to

$$\int Dx_\mu\, \delta(x_0 - t) e^{iS} = \int Dx_i \exp i \int \tfrac{1}{2} m \dot{x}_i^2\, dt \tag{1.5.4}$$

For the case of the string, this simple example will lay the basis for the light cone quantization. The advantage of the Coulomb gauge is that all ghosts have been explicitly removed from the theory, so we are dealing only with physical quantities. The other advantage is that the zeroth component of the position vector is now explicitly defined to be the time variable. The parametrization of the point particle is now given in terms of the physical time.

The disadvantage of the Coulomb formalism, however, is that manifest Lorentz symmetry is broken and we have to check explicitly that the quantized Lorentz generators close correctly. Athough this is trivial for the point particle, surprising features will emerge for the quantum string, fixing the dimension of space–time to be 26.

Gupta–Bleuler Quantization

This approach tries to maintain Lorentz invariance. This means, of course, that particular care must be taken to prevent the negative norm states from spoiling the physical properties of the S-matrix. The Gupta–Bleuler method keeps the action totally relativistic, but imposes the constraint (1.4.8) on state vectors:

$$[p_\mu^2 + m^2] |\phi\rangle = 0 \tag{1.5.5}$$

(Notice that the above equation is a ghost-killing constraint, because we can use it to eliminate p_0.) This formalism allows us to keep the commutators fully relativistic:

$$[p_\mu, x_\nu] = -i\eta_{\mu\nu} \tag{1.5.6}$$

where we choose $\eta_{\mu\nu} = (-+++\cdots)$. Notice that this gauge constraint naturally generalizes to the Klein–Gordon equation:

$$[\Box - m^2]\phi(x) = 0 \tag{1.5.7}$$

The Gupta–Bleuler formalism is an important one because most of the calculations in string theory have been carried out in this formalism.

BRST Quantization

The advantage of the BRST formalism [45, 46] is that it is manifestly Lorentz-invariant. But instead of regaining unitarity by applying the gauge constraints on the Hilbert space, which may be quite difficult in practice, the BRST formulation uses the Faddeev–Popov ghosts to cancel the negative metric particles. Thus, although the Green's functions are not unitary because of the propagation of negative metric states and ghosts, the final S-matrix is unitary because all the unwanted particles cancel among each other. Thus, the BRST formalism manages to incorporate the best features of both formalisms, i.e. the manifest Lorentz invariance of the Gupta–Bleuler formalism and the unitarity of the Coulomb or light cone formalism. In order to study the BRST formalism, however, we must first understand Faddeev–Popov quantization.

§1.6. Faddeev–Popov Quantization

Before we discuss the BRST method, it is essential to make a digression and review the formalism developed by Faddeev and Popov [47]. As we said earlier, the path integral measure Dx_μ is ill-defined because it possesses a gauge degree of freedom, so we are integrating over an infinite number of copies of the same thing. Naively, one might insert the gauge constraint directly into the path integral. If the constraint is given by some function F of the fields being set to zero:

$$F(x_\mu) = 0 \tag{1.6.1}$$

then we insert this delta functional directly into the path integral:

$$Z = \int Dx \prod_x \delta[F(x_\mu)] e^{iS} \tag{1.6.2}$$

However, this naive approach is actually incorrect because the delta functional contributes a *nontrivial measure* to the functional integral.

The key to the Faddeev–Popov method is to insert the number 1 into the functional, which obviously has the correct measure. For our purposes, the most convenient formulation of the number 1 is given by

$$1 = \Delta_{\text{FP}} \int D\varepsilon \delta[F(x_\mu^\varepsilon)] \tag{1.6.3}$$

where ε is the parametrization of the gauge symmetry of the coordinate, in (1.4.6), x_μ^ε is the variation of the field with respect to this symmetry, and the Faddeev–Popov determinant Δ_{FP} is defined by the previous equation.

Notice that the integral appearing in the previous equation is an integration over all possible parametrizations of the field. Since we are integrating out over all parametrizations, then, by construction, the Faddeev–Popov determinant is gauge independent of any particular parametrization:

$$\Delta_{\text{FP}}(x) = \Delta_{\text{FP}}(x^\varepsilon) \tag{1.6.4}$$

Let us now insert the number 1 into the functional integral and make a gauge transformation to reabsorb the ε dependence in x:

$$Z = \int Dx\, \Delta_{\mathrm{FP}}(x) \int D\varepsilon\, \delta[F(x^{\varepsilon})] e^{iS}$$
$$= \int Dx\, \Delta_{\mathrm{FP}}(x) \int D\varepsilon\, \delta[F(x)] e^{iS} \tag{1.6.5}$$

Notice here that x^{ε} was gauge rotated back into the original variable x. Since all other parts of the functional integral were already gauge independent, we now have

$$Z = \left[\int D\varepsilon\right] \int Dx\, \Delta_{\mathrm{FP}} \delta[F(x)] e^{iS} \tag{1.6.6}$$

We can now extract out the integral over the gauge parameter, which measures the infinite volume of the group space:

$$\text{volume} = \int D\varepsilon \tag{1.6.7}$$

and obtain a new expression for the functional which no longer has this infinite redundancy:

$$Z = \int Dx\, \Delta_{\mathrm{FP}} \delta[F(x)] e^{iS} \tag{1.6.8}$$

Notice that a naive quantization of the path integral would simply insert the F constraint and would omit the Faddeev–Popov determinant, which is a new feature that makes the measure come out correctly.

Now let us calculate the Faddeev–Popov determinant, which carries all the information concerning the ghosts of the theory. The trick is to change variables from ε to F. We can do this because both ε and F have the same number of degrees of freedom. Thus, the Jacobian can be calculated:

$$\det\left[\frac{\delta F}{\delta \varepsilon}\right] D\varepsilon = DF \tag{1.6.9}$$

We can therefore write

$$\Delta_{\mathrm{FP}} = \left\{\int D\varepsilon\, \delta(F)\right\}^{-1}$$
$$= \left\{\int DF \det\left[\frac{\delta \varepsilon}{\delta F}\right] \delta(F)\right\}^{-1}$$
$$= \left\{\det\left[\frac{\delta \varepsilon}{\delta F}\right]_{F=0}\right\}^{-1}$$
$$= \det\left[\frac{\delta F}{\delta \varepsilon}\right]_{F=0} \tag{1.6.10}$$

Thus, the Faddeev–Popov factor can be expressed as a simple determinant of the variation of the gauge constraint. It is more convenient to introduce this factor directly into the action by exponentiating it. We use the following trick:

$$\Delta_{\mathrm{FP}} = \int D\theta\, D\bar{\theta}\, e^{iS_{\mathrm{gh}}} \tag{1.6.11}$$

where the new ghost contribution to the action is given by

$$S_{\mathrm{gh}} = \int d\tau\, \bar{\theta}\left[\frac{\delta F}{\delta \varepsilon}\right]_{F=0} \theta \tag{1.6.12}$$

where the θ variables are anticommuting c-numbers called *Grassmann numbers* (see Appendix). Normally, when performing functional integrations, we expect to find the determinant of the inverse of a matrix. With functional integration over Grassmann numbers, the determinant occurs in the *numerator*, not the denominator. Grassmann numbers have the strange property that

$$\theta_i \theta_j = -\theta_j \theta_i \tag{1.6.13}$$

In particular, this means

$$\theta^2 = 0 \tag{1.6.14}$$

Normally, this would mean that θ vanishes. However, this is not the case for a Grassmann number. Thus, we also have the strange identity

$$e^{\theta} = 1 + \theta \tag{1.6.15}$$

This identity makes the integration over exponentials of Grassmann-valued fields in the functional integral rather easy, because they are simply polynomials. More identities on Grassmann numbers are presented in the Appendix, where we show that

$$\int \prod_{i=1}^{N} d\theta_i\, d\bar{\theta}_i \exp\left[\sum_{i,j=1}^{N} \bar{\theta}_i A_{ij} \theta_j\right] = \det(A_{ij}) \tag{1.6.16}$$

This identity verifies that integration over Grassmann variables yields determinant factors in the numerator, not the denominator, so that we can express the Faddeev–Popov determinant in (1.6.11) as a Grassmann integral.

Now that we have developed the apparatus of Faddeev–Popov quantization, let us return to the BRST approach, where we wish to impose the gauge condition

$$e = 1 \tag{1.6.17}$$

(we omit some subtleties with respect to this gauge). In this gauge, we should be able to recover the usual covariant Feynman propagator. To show this, notice that our action (1.4.14) becomes

$$L = \tfrac{1}{2}(\dot{x}_\mu^2 - m^2) \tag{1.6.18}$$

be more conveniently represented as

$$A_N = \sum_{\text{topologies}} \left\langle \exp i \sum_{i=1}^{N} k_\mu x_i^\mu \right\rangle \tag{1.7.4}$$

Thus, we associate a factor e^{ikx} for each external particle coming from the Fourier transform term. This path integral formula for the scattering amplitude is important because it will carry over almost exactly into the string formalism.

Notice how clumsy this description is. We must fix the set of all topologically allowed configurations and their weights *by hand*. Furthermore, unitarity of the S-matrix is not at all obvious.

In a second quantized description, however, we introduce a field $\psi(x)$ and quantization relations between the fields themselves, not between the coordinates:

$$\text{Second quantization:} \quad [\pi(x), \psi(y)]_{x_0=y_0} = -i\delta^{(3)}(x_i - y_i) \tag{1.7.5}$$

The advantage of the second quantized approach is that the interacting Hamiltonian can be written explicitly, without having to introduce sums over topologies. Showing that the Hamiltonian is Hermitian is sufficient to fix the weights of all diagrams and to demonstrate the unitarity of the S-matrix.

In summary, the pros and cons of first and second quantizations are as follows:

First Quantization
(1) Interactions must be added in by hand, order by order in the coupling constant.
(2) Unitarity of the final S-matrix is not obvious. This must be explicitly checked order by order.
(3) The formalism is necessarily a perturbative one, since the expansion in topologies is intimately tied to the expansion in terms of the coupling constant.
(4) It is difficult to describe the theory off-shell.

Second Quantization
(1) The interactions are explicit in the action itself.
(2) Unitarity is guaranteed if the Hamiltonian is Hermitian.
(3) The theory can be formally written nonperturbatively as well as perturbatively.
(4) The theory is necessarily off-shell.

The transition from the first to the second quantized theory can also be performed most easily in the path integral formalism in the Coulomb gauge. Earlier, we showed that the Green's function for a propagating free point particle can be explicitly evaluated:

$$K(a, b) = \left\{ \frac{m}{2\pi i(t_b - t_a)} \right\}^{1/2} \exp i \frac{m(x_b - x_a)^2}{2(t_b - t_a)} \tag{1.7.6}$$

This Green's function can also be written in a second quantized fashion. Let us start with the Hamiltonian:

$$H = -\frac{1}{2m}\nabla^2 \tag{1.7.7}$$

The Green's function satisfies

$$(i\partial_t - H)K(x, t; x', t') = \delta^{(3)}(x - x')\delta(t - t') \tag{1.7.8}$$

Solving for this Green's function, we find

$$K(a, b) = [i\partial_t - H]^{-1}_{x_a, t_a; x_b, t_b} \tag{1.7.9}$$

where we are treating the inverse Green's function as if it were a discrete matrix in (x, t) space, and we have dropped trivial normalization factors. This allows us to write the integral in second quantized language. To demonstrate this, we will use the following identities throughout this book:

$$\int \prod_{i=1}^{N} dx_i \exp\left\{ \sum_{i,j=1}^{N} - x_i A_{ij} x_j + \sum_{i=1}^{N} J_i x_i \right\}$$
$$= \frac{\pi^{(1/2)N}}{\det |A_{ij}|} \exp\left\{ \tfrac{1}{4} \sum_{i,j=1}^{N} J_i (A^{-1})_{ij} J_j \right\} \tag{1.7.10}$$

(This integral can easily be derived using our earlier formula for the Gaussian integral (1.3.18). We simply diagonalize the A matrix by making a change of variables in x. Thus, the quadratic term in the integral becomes a function of the eigenvalues of the A matrix. Because all the modes have now decoupled, the Gaussian integral can be performed exactly by completing the square. Finally, we make another similarity transformation to convert the eigenvalues of A back into the A matrix itself.)

From this, we can also derive the following:

$$\int x_n x_m \prod_{i=1}^{N} dx_i \exp\left\{ \sum_{i=1}^{N} - x_i A_{ij} x_j + \sum_{i=1}^{N} J_i x_i \right\}$$
$$\sim \left[\frac{\delta}{\delta J_n} \frac{\delta}{\delta J_m} \exp\left\{ \tfrac{1}{4} J_i (A^{-1})_{ij} J_j \right\} \right]_{J=0} \det |A_{ij}|^{-1}$$
$$\sim (A^{-1})_{nm} (\det |A_{ij}|)^{-1} \tag{1.7.11}$$

These are some of the most important integrals in this book. Using these equations, we can now write the Green's function totally in terms of second quantized fields:

$$K(a, b) = \int \psi^*(x_a, t_a)\psi(x_b, t_b)\, D\psi^*\, D\psi \exp i \int dx\, dt\, L(\psi) \tag{1.7.12}$$

where

$$L(\psi) = \psi^*(i\partial_t - H)\psi \tag{1.7.13}$$

where we are again treating $K(a, b)$ as if it were a matrix in discretized (x, t) space.

In summary, we now have two complementary descriptions of the point particle. We can write the theory either in terms of the particle's coordinates x_i or in terms of its fields $\psi(x)$.

At the free level, both descriptions are totally equivalent, both in ease of description and also in mathematics. However, at the interacting level, distinct differences appear. For example, it is easy to write

$$L_I \sim \phi^3; \quad \sim \phi^4 \tag{1.7.14}$$

and we are guaranteed to get a unitary description of an interacting field. However, in the first quantized approach, the sum over topologies:

$$\sum_{\text{topologies}} \tag{1.7.15}$$

is a clumsy way in which to describe a unitary theory. We must check unitarity order by order in increasingly complicated diagrams. Furthermore, we are forced to adopt a totally perturbative description for the first quantized description. The sum over topologies in the first quantized path integral is a sum over perturbative Feynman diagrams, so the formulation is necessarily perturbative from the very beginning. That is the fundamental reason why we have divided this book into first quantization and second quantization.

§1.8. Harmonic Oscillators

One example that will illustrate the relationship between first and second quantizations is the harmonic oscillator problem. This example will prove helpful in introducing the harmonic oscillator representation, which we will use extensively for the string model. Let us begin with a point particle governed by the following Hamiltonian:

$$H = \frac{p^2}{2m} + \tfrac{1}{2}kx^2 \tag{1.8.1}$$

where k is the spring constant. Because the momenta and coordinates are conjugates, we can use the same arguments presented earlier in our discussion of path integrals to set

$$[p, x] = -i \tag{1.8.2}$$

We can now redefine our coordinates and momenta in terms of harmonic oscillators:

$$\begin{aligned} p &= (\tfrac{1}{2}m\omega)^{1/2}(a + a^\dagger) \\ x &= i(2m\omega)^{-1/2}(a - a^\dagger) \end{aligned} \tag{1.8.3}$$

where

$$k = m\omega^2 \tag{1.8.4}$$

In order to satisfy the canonical commutation relation (1.8.2), we must have

$$[a, a^\dagger] = 1 \tag{1.8.5}$$

If we insert this expression back into the Hamiltonian, we find

$$H = \tfrac{1}{2}\omega(aa^\dagger + a^\dagger a) \tag{1.8.6}$$

By extracting a c-number term, we can write this in normal ordered fashion:

$$H = \omega(a^\dagger a + h_0) \tag{1.8.7}$$

where h_0 is the zero point energy. We can now introduce the Hilbert space of harmonic oscillators. Let us define the vacuum as

$$a|0\rangle = 0 \tag{1.8.8}$$

Then an element of the Fock space of the harmonic oscillator Hamiltonian is given by

$$|n\rangle = \frac{(a^\dagger)^n}{\sqrt{n!}}|0\rangle \tag{1.8.9}$$

such that the states form an orthonormal basis:

$$\langle n|m\rangle = \delta_{nm} \tag{1.8.10}$$

The energy of the system is quantized and given by

$$E_n = (n + \tfrac{1}{2})\omega \tag{1.8.11}$$

So far, the system has been presented only in a first quantized formalism. We are quantizing only a single point particle at any time. We would now like to make the transition to the second quantized wave function by introducing

$$|\Phi\rangle = \sum_{n=0}^{\infty} \phi_n |n\rangle \tag{1.8.12}$$

where we power expand in the basis states of the harmonic oscillator. Thus, instead of describing a single excited state of a point particle, we are now introducing the wave function, which will be a superposition of an arbitrary number of excited states.

Let us make the important definition

$$\langle x|\Phi\rangle = \Phi(x) \tag{1.8.13}$$

This can be calculated explicitly. Notice that we now have two independent basis states, the harmonic oscillator basis $|n\rangle$ and the position eigenvectors $|x\rangle$. We must now calculate how to go back and forth between these two bases.

Let us first analyze the simplest matrix element:

$$\sigma_0(x) = \langle x|0\rangle \tag{1.8.14}$$

This matrix element satisfies the equation

$$\begin{aligned} 0 &= \langle x|a|0\rangle \\ &= \langle x|\frac{p - im\omega x}{\sqrt{2m\omega}}|0\rangle \\ &= (2m\omega)^{-1/2}\left(-\frac{i\partial}{\partial x} - im\omega x\right)\langle x|0\rangle \\ &= -i(2m\omega)^{-1/2}\left(\frac{\partial}{\partial x} + m\omega x\right)\sigma_0(x) \end{aligned} \tag{1.8.15}$$

This last equation can be solved exactly:

$$\sigma_0(x) = (m\omega/\pi)^{1/4}e^{-1/2\xi^2} \tag{1.8.16}$$

where

$$\xi = (m\omega)^{1/2}x \tag{1.8.17}$$

It is now a straightforward step to calculate all such matrix elements. Let

$$\begin{aligned} \sigma_n(x) &= \langle x|n\rangle \\ &= \langle x|(n!)^{-1/2}a^{\dagger n}|0\rangle \\ &= (n!)^{-1/2}(2m\omega)^{-n(1/2)}\langle x|[p + im\omega x]^n|0\rangle \\ &= (n!)^{-1/2}(2m\omega)^{-(1/2)n}\left(-i\frac{\partial}{\partial x} + im\omega x\right)^n\sigma_0(x) \end{aligned} \tag{1.8.18}$$

The solution is therefore

$$\sigma_n(x) = i^n(2^n n!)^{-1/2}(m\omega/\pi)^{1/4}\left(\xi - \frac{\partial}{\partial\xi}\right)^n e^{-(1/2)\xi^2} \tag{1.8.19}$$

In general, these are nothing but Hermite polynomials H_n. In terms of these polynomials, we can express the eigenstate $|x\rangle$ and $|n\rangle$ in terms of each other:

$$\begin{cases} |x\rangle = \sum_{n=1}^{\infty} |n\rangle\langle n|x\rangle = \sum_{n=1}^{\infty} |n\rangle\sigma_n(x) \\ |n\rangle = |x\rangle\int dx\langle x|n\rangle = \int dx\ \sigma_n(x)|x\rangle \end{cases} \tag{1.8.20}$$

Thus, using (1.8.12) and (1.8.20), we have the power expansion of the wave function in terms of a complete set of orthogonal polynomials, the Hermite polynomials:

$$\Phi(x) = \langle x|\Phi\rangle = \langle x|\sum_{n=1}^{\infty}\phi_n|n\rangle = \sum_{n=1}^{\infty}\phi_n H_n(\xi)e^{-(1/2)\xi^2} \tag{1.8.21}$$

Similarly, it is not difficult to calculate the Green's function for the propaga-

Under this change, the volume element of the integral changes as

$$\delta d^D x = d^D x \partial_\mu \delta x^\mu \tag{1.9.14}$$

Therefore, the variation of the action under this change is

$$\begin{aligned} \delta S &= \int d^D x [L \partial_\mu \delta x^\mu + \delta L] \\ \delta L &= \delta x^\mu \partial_\mu L + \frac{\delta L}{\delta \phi} \delta \phi + \frac{\delta L}{\delta \partial_\mu \phi} \delta \partial_\mu \phi \end{aligned} \tag{1.9.15}$$

Now, if we assume that the equations of motion are satisfied, we have

$$\delta S = \int d^D x \partial_\mu \left\{ \left(+ L \delta^\mu_\nu - \frac{\delta L}{\delta \partial_\mu \phi} \partial_\nu \phi \right) \delta x^\nu \right\} \tag{1.9.16}$$

If we now define the *energy–momentum tensor* as

$$T_{\mu\nu} = \frac{\delta L}{\delta \partial^\mu \phi} \partial_\nu \phi - \eta_{\mu\nu} L \tag{1.9.17}$$

then we have the equation

$$\delta S = \int d^D x \partial_\mu (T^{\mu\nu} \delta x_\nu) \tag{1.9.18}$$

So if the action is invariant under this change, then the energy–momentum tensor is conserved:

$$\partial_\mu T^{\mu\nu} = 0 \tag{1.9.19}$$

For example, for the scalar particle action, the energy–momentum tensor becomes

$$T_{\mu\nu} = \partial_\mu \phi \partial_\nu \phi - \eta_{\mu\nu} L \tag{1.9.20}$$

which is conserved if the equations of motion hold.

Lastly, it is instructive to investigate how the various quantization procedures treat the Yang–Mills field (see Appendix). Let us begin with the SU(N) invariant action:

$$L = -\tfrac{1}{4}[F^a_{\mu\nu}]^2 \tag{1.9.21}$$

where

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - f^{abc} A^b_\mu A^c_\nu \tag{1.9.22}$$

The action is invariant under

$$\delta A^a_\mu = \partial_\mu \Lambda^a - f^{abc} A^b_\mu \Lambda^c \tag{1.9.23}$$

where Λ^a is a gauge parameter.

The path integral method begins with the functional

$$Z = \int \prod_{\mu, x} dA_\mu(x) e^{i \int d^4x - (1/4) F^2_{\mu\nu}} \tag{1.9.24}$$

Now we consider the three methods of quantization:

Coulomb Quantization

The gauge invariance permits us to take the gauge:

$$\nabla_i A_i^a = 0$$

We can integrate over the A_0 component because it has no time derivatives, so the Coulomb formulation is explicitly ghost-free. (The price we pay for this, of course, is the loss of manifest Lorentz invariance, which must be checked by hand.) In this gauge, the action becomes

$$L = +\tfrac{1}{2}(\partial_0 A_i^a)^2 - \tfrac{1}{4}(F_{ij}^a)^2 + \cdots \tag{1.9.25}$$

where all fields are transverse. This is the canonical form for the Lagrangian.

Gupta–Bleuler Quantization

The advantage of the Gupta–Bleuler formulation is that we can keep manifest Lorentz symmetry without violating unitarity. For example, let us take the gauge

$$\partial_\mu A^{\mu a} = 0 \tag{1.9.26}$$

In this gauge, the propagator for massless vector particles becomes

$$\frac{\eta_{\mu\nu}}{p^2} \tag{1.9.27}$$

Notice that the propagator explicitly contains a ghost. The timelike excitation has a coefficient of -1 in the propagator, which represents a ghost. However, we are free to quantize in this covariant approach beause we will impose the ghost-killing constraint on the Hilbert space:

$$\langle \phi | \partial_\mu A^{\mu a} | \psi \rangle = 0 \tag{1.9.28}$$

This constraint allows us to solve for and hence eliminate the ghost modes. Thus, although the free propagator will allow ghosts to propagate, the Hilbert space is ghost-free, so the theory itself is both Lorentz invariant and ghost-free.

BRST Quantization

The BRST approach begins by calculating the Faddeev–Popov determinant (1.6.10). Let us calculate the determinant of the matrix:

$$\begin{aligned} M^{ab}(x, y) &= \frac{\delta(\partial_\mu A^{\mu a}(x))}{\delta \Lambda^b(y)} \\ &= \partial_\mu D^\mu \frac{\delta \Lambda^a(x)}{\delta \Lambda^b(y)} \\ &= \partial_\mu D^\mu (\delta^4(x - y)\delta^{ab}) \end{aligned} \tag{1.9.29}$$

As before, we can write the determinant of M^{ab} by exponentiating it into the action using (1.6.10):

$$L = -\tfrac{1}{4}F_{\mu\nu}^{a2} + \frac{1}{2\alpha}(\partial_\mu A^{\mu a})^2 + \bar{c}^a M^{ab} c^b \tag{1.9.30}$$

where the anticommuting Faddev–Popov ghost fields are represented by c and $\bar{c}$. This action is invariant under the following BRST transformation:

$$\text{BRST:} \quad \begin{cases} \delta A_\mu^a = (\nabla_\mu c)^a \varepsilon \\ \delta c^a = -\frac{1}{2} f^{abd} c^b c^d \varepsilon \\ \delta \bar{c}^a = \dfrac{1}{\alpha}(\partial_\mu A^{\mu,a})\varepsilon \end{cases} \tag{1.9.31}$$

Once again, it is important to notice that the BRST transformation is nilpotent. The BRST symmetry is not connected to the conservation of any observable quantity. From the previous invariance, we can extract out the generator of this transformation Q such that

$$Q^2 = 0 \tag{1.9.32}$$

The physical states of the theory then satisfy

$$Q|\text{phy}\rangle = 0 \tag{1.9.33}$$

§1.10. Summary

The great irony of string theory, which is supposed to provide a unifying framework for all known interactions, is that the theory itself is so disorganized. String theory is often frustrating to the beginner because it is full of folklore, conventions, and arbitrary rules of thumb. The fundamental reason for this is that string theory has historically evolved backward as a first quantized theory, rather than as a second quantized theory, where the entire theory is defined in terms of a fundamental action. The disadvantages of the first quantized approach are that

(1) The interactions of the theory must be introduced by hand. They cannot be derived from a single action.
(2) Unitarity is not obvious in this approach. The counting of graphs must be checked tediously.
(3) The formulation is perturbative, so that crucial nonperturbative calculations, such as dimensional breaking, are beyond its scope.
(4) The formulation is basically on-shell, rather than off-shell.

By contrast, the advantage of the second quantized approach is precisely that everything can be derived from a single off-shell action, where unitarity is manifest and nonperturbative calculations can, in principle, be performed.

Unfortunately, string theory evolved historically as a first quantized theory. Thus, string theory has been evolving backward, with the second quantized geometric theory still in its infancy. For pedagogical reasons, we have introduced string theory from a semihistorical point of view, beginning with the first quantized theory and later developing the second quantized theory and the underlying geometry. We hope that future accounts of string theory will reverse this sequence.

To reduce the level of arbitrariness in the first quantized theory as much as possible, in this chapter we have tried to lay the groundwork for string theory in the formalism of path integrals. This functional formalism has the great advantage that we can express the first and second quantized gauge theories with equal ease. We find, in fact, that large portions of the path integral formulation of point particles can be incorporated wholesale into string theory.

The path integral method postulates two fundamental principle that express the essence of quantum mechanics:

(1) The probability $P(a, b)$ of a particle going from point a to point b is given by the absolute value squared of a transition function $K(a, b)$:

$$P(a, b) = |K(a, b)|^2$$

(2) The transition function is given by the sum of a phase factor e^{iS}, where S is the action, taken over all possible paths from a to b:

$$K(a, b) = \sum_{\text{paths}} k e^{iS}$$

In the limit of continuous paths, we have

$$K(a, b) = \int_a^b Dx\, e^{iS}$$

where

$$Dx = \lim_{N \to \infty} \prod_{i=1}^{3} \prod_{n=1}^{N} dx_{i,n}$$

The action S of the first quantized point particle is given by the length of the path that the particle sweeps out in space-time. We can represent the Lagrangian for the point particle in three ways:

$$\begin{aligned} \text{1st-order (Hamiltonian) form:} \quad & L = p_\mu \dot{x}^\mu - \tfrac{1}{2} e(p_\mu^2 + m^2) \\ \text{2nd-order form:} \quad & L = \tfrac{1}{2}(e^{-1}\dot{x}_\mu^2 - em^2) \\ \text{Non-linear form:} \quad & L = -m\sqrt{-\dot{x}_\mu^2} \end{aligned} \tag{1.10.1}$$

Unfortunately, because all three forms of the action are parametrization-invariant, the path integral diverges. Thus, the quantization procedure must break this gauge symmetry and yield the correct measure in the functional.

These actions can be quantized in three ways, each with its own advantages

and disadvantages:

(1) Coulomb Quantization
By explicitly fixing the value of some of the fields, such as

$$x_0 = t = \tau$$

we can eliminate the troublesome negative metric states and the Lagrangian becomes $\frac{1}{2}mv_i^2$. The Coulomb quantization method is therefore manifestly ghost-free. However, the disadvantage of this method is that it is very awkward because manifest Lorentz symmetry is broken and must be checked at every level.

(2) Gupta–Bleuler Quantization
The advantage of the Gupta–Bleuler quantization method is that we have a manifestly covariant quantization program. Of course, negative metric ghosts are now allowed to circulate in the theory, but they are eventually eliminated by imposing the gauge constraints directly onto the Hilbert space:

$$[p_\mu^2 + m^2]|\phi\rangle = 0$$

Thus, the S-matrix is ultimately ghost-free. The disadvantage of this approach, however, is that the imposition of these gauge constraints, especially at the interacting level, is frequently quite difficult.

(3) BRST Quantization
This method of quantization keeps the good features of both approaches. The theory is manifestly covariant, but the S-matrix is still unitary because the addition of ghost fields in the theory cancels precisely against the negative metric states. The BRST method imposes the gauge $e = 1$ in the first-order form and then inserts the Faddeev–Popov term Δ_{FP} into the functional to get the correct measure. We can exponentiate this determinant into the action by using Grassman variables:

$$\Delta_{\text{FP}} = \det|\partial_\tau| = \int d\theta\, d\bar{\theta} e^{i\int d\tau\, \bar{\theta}\partial_\tau\theta}$$

The resulting gauge-fixed action has a residual symmetry, called the BRST symmetry, which is generated by Q, the BRST charge. (This new symmetry does not result in the elimination of more fields).

When we generalize these methods to the interacting case, the path integral formulation begins with the fundamental formula for the transition function for N-particle scattering:

$$\begin{aligned} A(k_1, k_2, \ldots, k_N) &= \sum_{\text{topologies}} g^n \int Dx\, \Delta_{\text{FP}} \\ &\quad \times \exp\left\{ i \int dt\, L(t) + i \sum_{i=1}^{N} k_\mu x_i^\mu \right\} \\ &= \sum_{\text{topologies}} \int Dx \langle e^{i\sum_{i=1}^{N} k_\mu x_i^\mu} \rangle \end{aligned} \tag{1.10.2}$$

CHAPTER 2

Nambu–Goto Strings

§2.1. Bosonic Strings

String theory, at first glance, seems divorced from the standard techniques developed over the past 40 years for second quantized field theories. This is because string theory was first historically discovered as a *first quantized theory*. This is the reason why string theory at times appears to be a random collection of arbitrary conventions. Although a second quantized field theory can be derived completely from a single action, a first quantized theory requires additional assumptions. In particular, the vertices, the choice of interactions, and the weights of these perturbation diagrams must be postulated by hand and checked to be unitary later.

Fortunately, the path integral formalism for the first quantized point particle has been generalized for the string by J. L. Gervais and B. Sakita, which enables us to write down the dynamics of interacting strings with remarkable ease.

In the previous chapter, we laid the crucial mathematical groundwork for a discussion of the first quantized point particle theory. Surprisingly, almost all of the main features of the Nambu–Goto string have some form of analogue in the first quantized point particle theory. Of course, entirely new features are found in the string theory, such as the existence of powerful symmetries on the world sheet, but the basic methods of quantization can be carried over directly from the point particle case studied in the previous chapter.

We saw that the usual formulation of second quantized field theory can be rewritten in first quantized form. Thus, the traditional covariant Feynman

propagator (1.6.19) can be written via (1.3.28), (1.3.30), (1.3.37) as

$$\begin{aligned}\Delta_F(x_1, x_2) &= \langle x_1 | \frac{1}{\Box - m^2} | x_2 \rangle \\ &= \langle x_1 | \int_0^\infty d\tau \, e^{-\tau(\Box - m^2)} | x_2 \rangle \\ &= \int_0^\infty d\tau \int_{x_1}^{x_2} Dx \, e^{-(1/2)\int_0^\tau d\bar{\tau}(\dot{x}_\mu^2 - m^2)}\end{aligned} \tag{2.1.1}$$

where we integrate over all possible trajectories of a particle located at $x_\mu(\tau)$ which start at x_1 and end at point x_2. The interactions, we saw, were introduced by hand into the theory by postulating a particular set of topologies over which this particle can roam. The scattering amplitude, for example, is

$$\begin{aligned}A(k_1, k_2, \ldots, k_N) &= \sum_{\text{topologies}} \int Dx \, e^{-\int L \, d\tau + i \sum_{i=1}^N k_i \cdot x_i} \Delta_{FP} \\ &= \sum_{\text{topologies}} \left\langle \prod_{i=1}^N e^{ik_i \cdot x^i} \right\rangle\end{aligned} \tag{2.1.2}$$

where we integrate over topologies that form the familiar Feynman diagrams for ϕ^3 or ϕ^4 theory.

It is important to notice that the resulting Feynman diagram is a graph, *not* a manifold. At the interaction point, the local topology is not $\mathbf{R}^n$, so it cannot be a manifold. There is no correlation between the internal lines and the interaction points. This means that we can introduce arbitrarily high spins at the interaction point of the first quantized relativistic point particle. Thus, the first quantized point particle theory has an infinite degree of arbitrariness, corresponding to the different spins and masses we can place at the interaction point. Furthermore, the ultraviolet singularities of each Feynman diagram correspond to the number of ways we can "pinch" the diagram by shrinking an internal line to zero, thus deforming the local topology.

This picture, however, totally changes with the string. Although the path integral formalism looks almost identical, there are profoundly important differences. In particular, the sum over histories becomes a sum over all possible tubes or sheets that one can draw between two different strings (see Fig. 2.1). This world sheet, in turn, is a genuine manifold, a Riemann surface, so the set of interactions consistent with the propagator is severely limited. *Thus, we expect to find a very small number of string theories, in contrast to the infinite number of point particle theories one can write.* Furthermore, the superstring theory does not suffer from ultraviolet divergences caused by shrinking one of the internal lines to zero. You cannot "pinch" the string world sheet to obtain an ultraviolet divergence. Thus, string theory is free of ultraviolet divergences from strictly topological arguments. (We must be careful to point out that this pinched diagram, however, can be reinterpreted as an *infrared* divergence representing the emission of massless, spin-zero par-

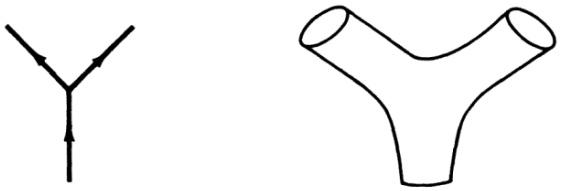

Figure 2.1. Vertex functions for point particles and strings. A large number of point particle theories are possible, based on different spins and isospins, because the Feynman diagrams are graphs. Only a few string theories are known, however, because the interactions are restricted to be manifolds, not graphs. Conformal symmetry, modular invariance, and supersymmetry place enormous restrictions on the manifolds one may use to construct superstring theories that have no counterpart in point particle theory.

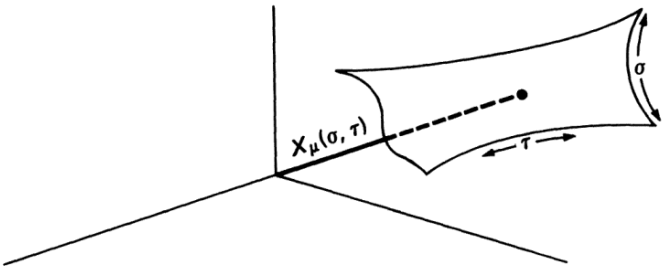

Figure 2.2. The two-dimensional world sheet swept out by a string. When a string, which is parametrized by σ, moves in space–time, it sweeps out a two-dimensional surface parametrized by σ and τ. The string variable $X_\mu(\sigma, \tau)$ is just a vector that extends from the origin to a point on this two-dimensional manifold.

ticle into the vacuum. Fortunately, supersymmetry eliminates these infrared divergences.)

In summary, although the path integral formalism can treat both the first quantized point particle and the first quantized string theory with relative ease, there are profoundly important physical differences between the two theories that arise from strictly topological arguments.

We begin our discussion of strings by first introducing the coordinate of a string vibrating in physical space–time. Let the points along the string be parametrized by the variable σ, and then let the string propagate in time. Let the vector

$$X_\mu(\sigma, \tau) \tag{2.1.3}$$

represent the space–time coordinates of this string (see Fig. 2.2) parametrized by two variables. When the string moves, it sweeps out a two-dimensional surface, which we call the "world sheet." We will parametrize this world sheet with two variables, σ and τ. The vectors that are tangent to the surface are

given by the derivatives of the coordinate:

$$\text{Tangent vectors} = \frac{\partial X_\mu}{\partial \tau}; \qquad \frac{\partial X_\mu}{\partial \sigma} \tag{2.1.4}$$

The contraction of two of these tangent vectors yields a metric:

$$g_{ab} = \partial_a X_\mu \partial_b X^\mu \tag{2.1.5}$$

where we have now replaced the two variables (τ, σ) with the set (a, b), where a, b can equal either 0 or 1. The infinitesimal area on this surface can be written simply as

$$d\,\text{Area} \sim \sqrt{\det |g_{ab}|}\, d\sigma\, d\tau \tag{2.1.6}$$

In analogy to the point particle case, where the action is the length swept out by the point, we now define our action to be the surface area of this world sheet. Our Lagrangian is therefore [1–4]:

$$L = \frac{1}{2\pi\alpha'} \sqrt{\dot{X}_\mu^2 X'^{\mu 2} - (\dot{X}_\mu X'^\mu)^2} \tag{2.1.7}$$

where the prime represents σ differentiation and the dot represents τ differentiation. The action is just the Lagrangian integrated over the world sheet, which is the total area of the two-dimensional surface:

$$S = \int d\sigma\, d\tau\, L(\sigma, \tau) \tag{2.1.8}$$

The Green's function for the propagation of a string from configuration X_a at "time" τ_a to configuration X_b at "time" τ_b, as well as the path integral over a surface that expresses the topology of several interacting strings, can be represented as

$$\begin{aligned} K(X_a, X_b) &= \int_{X_a}^{X_b} DX\; e^{-\int_{\tau_a}^{\tau_b} d\tau \int_0^\pi d\sigma\, L} \\ Z &= \sum_{\text{Topologies}} \int d\mu\, DX\; e^{-\text{area}} \end{aligned} \tag{2.1.9}$$

where $DX = \prod_{\mu,\sigma,\tau} dX_\mu(\sigma, \tau)$, $d\mu$ represents the measure of integeration over the location of the external legs, and where we have made a Wick rotation in the τ variable ($\tau \to -i\tau$) so the integral converges.

The correspondence between the point particle path integral formalism that we carefully developed in the previous chapter and the string formalism is quite remarkable. We find that almost the entire point particle formalism can be imported into the string formalism:

$$\left\{ \begin{array}{c} x_\mu(\tau) \\ \text{length} \\ \prod_{\mu,\tau} dx_\mu(\tau) \end{array} \right\} \to \left\{ \begin{array}{c} X_\mu(\sigma, \tau) \\ \text{area} \\ \prod_{\mu,\sigma,\tau} dX_\mu(\sigma, \tau) \end{array} \right\}$$

Similarly, the path integral for the point particle and the string theory have surprising similarities. The N-point function for the N-string scattering amplitude can also be written as a Fourier transform, similar to point particle path integrals:

$$\left\{ \begin{array}{c} \displaystyle\int_{x_i}^{x_j} Dx\, e^{-\int_{\tau_1}^{\tau_2} L(x)\, d\tau} \\ \left\langle \displaystyle\prod_{i=1}^{N} e^{ik_i x^i} \right\rangle \end{array} \right\} \rightarrow \left\{ \begin{array}{c} \displaystyle\int_{X_i}^{X_j} DX\, e^{-\int_0^{\pi} d\sigma \int_{\tau_1}^{\tau_2} L(X)\, d\tau} \\ \left\langle \displaystyle\prod_{i=1}^{N} e^{ik_i X^i} \right\rangle \end{array} \right\}$$

Although there are remarkable similarities between point particle and string theories when expressed in the language of path integrals, the crucial difference between them emerges when we analyze the topologies over which the objects can move. For the point particle case, the topologies are *graphs*, as in Feynman graphs, whereas the topologies for string theories are *manifolds*:

$$\text{Graphs} \rightarrow \text{Manifolds}$$

One of the crucial reasons why there are so many point particle actions (and so few string actions) is the difference between graphs and manifolds. The nontrivial restrictions placed on manifolds severely restrict the number of consistent string theories.

As in the case of the point particle, the choice of parametrization was totally arbitrary. Thus, our action must be reparametrization invariant. To see this, let us make an arbitrary change of variables:

$$\begin{aligned} \tilde{\sigma} &= \tilde{\sigma}(\sigma, \tau) \\ \tilde{\tau} &= \tilde{\tau}(\sigma, \tau) \end{aligned} \tag{2.1.10}$$

Under this reparametrization, the string variable changes as

$$\delta X^\mu = X'^\mu \delta\sigma + \dot{X}^\mu \delta\tau \tag{2.1.11}$$

Because the area of a surface is independent of the parametrization, the action is manifestly reparametrization invariant, which is easily checked.

As before, let us now write down the canonical conjugates of the theory:

$$P_\mu = \frac{\delta L}{\delta \dot{X}^\mu} = \frac{1}{2\pi\alpha'} \frac{X'^2 \dot{X}_\mu - (\dot{X}_\nu X'^\nu) X'_\mu}{\sqrt{\det |\partial_a X^\nu \partial_b X_\nu|}}$$

As in the point particle case, these momenta are not all independent. In fact, we find two identities that are satisfied by the canonical momenta:

$$\text{Constraints:} \quad \left\{ \begin{array}{l} P_\mu^2 + \dfrac{1}{(2\pi\alpha')^2} X_\mu'^2 \equiv 0 \\ P_\mu X'^\mu \equiv 0 \end{array} \right. \tag{2.1.12}$$

Thus, the canonical momenta are constrained by these two conditions. If we calculate the Hamiltonian of the system, we find that it vanishes identically as in (1.4.9):

$$H = P_\mu \dot{X}^\mu - L \equiv 0 \tag{2.1.13}$$

Notice that the X decomposition is given strictly in terms of cosines. This is because, when we calculate the equations of motion of the string, we must integrate by parts and hence obtain unwanted surface terms at $\sigma = \pi$ and $\sigma = 0$. In order to eliminate these surface terms, we must impose

$$X'_\mu = 0 \tag{2.1.26}$$

at the boundary. This boundary condition eliminates all sine modes of the string.

We will sometimes find it convenient to take advantage of this form of the expansion. In particular, it means that

$$\begin{aligned} X_\mu(\sigma) &= X_\mu(-\sigma) \\ X'_\mu(\sigma) &= -X'_\mu(-\sigma) \end{aligned} \tag{2.1.27}$$

The same applies for the modes of the canonical conjugate P_μ. This, in turn, allows us to combine both constraints into one, using the properties of the string under reflection from σ into $-\sigma$. If we let the string parametrization length be π we can define:

$$L_f = \frac{1}{4\pi} \int_{-\pi}^{\pi} d\sigma \, f(\sigma) \left(\sqrt{2\alpha'} \pi P_\mu + \frac{X'_\mu}{\sqrt{2\alpha'}} \right)^2 \tag{2.1.28}$$

where $f(\sigma)$ is an arbitrary function defined from $-\pi$ to π. Notice that both constraints are now combined into one equation because of this reflection symmetry. Using (2.1.21), we can show that these generators form a closed algebra:

$$[L_f, L_g] = L_{f \times g} \tag{2.1.29}$$

where

$$f \times g = fg' - gf' \tag{2.1.30}$$

It is also possible to show that this algebra satisfies the Jacobi identities:

$$[L_{[f}, [L_g, L_{h]}] = 0 \tag{2.1.31}$$

where the brackets represent all possible cyclic symmetrizations. This algebra is called the *Virasoro algebra* [6], which will turn out to be one of the most powerful tools we have in constructing the string theory.

As in (1.4.11), we can elevate the constraints into the action with Lagrange multipliers $\lambda(\sigma, \tau)$ and $\rho(\sigma, \tau)$:

$$L = P_\mu \dot{X}^\mu + \pi\alpha' \lambda \left[P_\mu^2 + \frac{X'^2_\mu}{(2\pi\alpha')^2} \right] + \rho P_\mu X'^\mu \tag{2.1.32}$$

By functionally integrating out over these Lagrange multipliers, we arrive at the previous set of constraints. Not surprisingly, this new action has its own

reparametrization group parametrized by η and ε:

$$\begin{aligned}
\delta X_\mu &= 2\pi\alpha'\varepsilon P_\mu + \eta X'_\mu \\
\delta P_\mu &= \left[\frac{\varepsilon X'_\mu}{2\pi\alpha'} + \eta P_\mu\right]' \\
\delta\lambda &= -\dot{\varepsilon} + \lambda'\eta - \eta'\lambda + \rho'\varepsilon - \rho\varepsilon' \\
\delta\rho &= -\dot{\eta} + \lambda'\varepsilon - \lambda\varepsilon' + \rho'\eta - \eta'\rho
\end{aligned} \tag{2.1.33}$$

The advantage of this form for the action is that it is first-order and does not have the bothersome square roots of the original action. As in the point particle case, this indicates that there exists yet one more form for the action, expressed in terms of an auxiliary field. To find this third formulation of the action, let us introduce a new independent field

$$g_{ab}(\sigma, \tau) \tag{2.1.34}$$

which represents a metric on a two-dimensional surface. Unlike our previous discussion, this metric is now totally independent of the string variable. Let us write down the Polyakov form of the action [7] ($g = |\det g_{ab}|$):

$$L = -\frac{1}{4\pi\alpha'}\sqrt{g}g^{ab}\partial_a X_\mu \partial_b X^\mu \tag{2.1.35}$$

This is a generalization of the second-order point particle action (1.4.14). Notice that the Polyakov action resembles an action with scalar fields interacting with an external two-dimensional gravitational field. This action, too, possesses manifest reparametrization invariance:

$$\begin{aligned}
\delta X^\mu &= \varepsilon^a \partial_a X^\mu \\
\delta g^{ab} &= \varepsilon^c \partial_c g^{ab} - g^{ac}\partial_c \varepsilon^b - g^{bc}\partial_c \varepsilon^a \\
\delta\sqrt{g} &= \partial_a(\varepsilon^a\sqrt{g})
\end{aligned} \tag{2.1.36}$$

The action is also trivially invariant under Weyl rescaling:

$$\delta g^{ab} = \Lambda g^{ab} \tag{2.1.37}$$

The Polyakov action is entirely equivalent at the classical level to the earlier Nambu–Goto action. As in the Nambu–Goto formalism, we can derive the Virasoro algebra. By varying with respect to the metric tensor, we obtain the energy–momentum tensor, which we can set to zero:

$$T_{ab} = -4\pi\alpha'\frac{1}{\sqrt{g}}\frac{\delta L}{\delta g^{ab}} \tag{2.1.38}$$

Working this out explicitly, we find

$$T_{ab} = \partial_a X_\mu \partial_b X^\mu - \tfrac{1}{2} g_{ab} g^{cd}\partial_c X^\mu \partial_d X_\mu \tag{2.1.39}$$

The moments of the energy–momentum tensor will correspond to the Virasoro generators. Thus, we have another way of deriving the Virasoro generators from this new but equivalent formalism.

Notice that the metric field g_{ab} is *not* a propagating field. The metric tensor does not have any derivatives acting on it. Thus, we can eliminate it via its own equations of motion. This leads us to

$$\frac{\delta L}{\delta g_{ab}} = 0 \rightarrow g_{ab} = \frac{2\partial_a X_\mu \partial_b X^\mu}{g^{cd}\partial_c X_\nu \partial_d X^\nu} \tag{2.1.40}$$

Substituting this value of the metric tensor back into the action, we rederive the original Nambu–Goto action. Thus, at the classical level, the two actions are identical.

In summary, as in the point particle case, we now have three different ways in which to write down the action, all of which are equivalent classically. Each has its own particular advantages and disadvantages when we make the transition to the quantum system. These string equations are direct generalizations of the three point particle Lagrangians found in (1.4.16). As before, we have the second-order formalism, which is expressed in terms of the string variable X_μ as well as the metric tensor g_{ab}; the non-linear formalism, which is expressed entirely in terms of X_μ; and the Hamiltonian formalism, where we have X_μ and its canonical conjugate P_μ (or the pair $\partial_a X_\mu$ and $P^{a\mu}$):

$$\begin{aligned}
\text{1st-order (Hamiltonian) form: } L &= P_\mu \dot{X}^\mu + \pi\alpha'\lambda\left[P_\mu^2 + \frac{X_\mu'^2}{(2\pi\alpha')^2}\right] + \rho P_\mu X'^\mu \\
&\sim \sqrt{g} P^a_\mu g_{ab} P^{b\mu} + P^{a\mu}\partial_a X_\mu \sqrt{g}/\pi\alpha' \\
\text{2nd-order form: } L &= \frac{-1}{4\pi\alpha'}\sqrt{g} g^{ab}\partial_a X_\mu \partial_b X^\mu \\
\text{Non-linear form: } L &= \frac{1}{2\pi\alpha'}(\dot{X}_\mu^2 X_\nu'^2 - (\dot{X}_\mu X'^\mu)^2)^{1/2}
\end{aligned} \tag{2.1.41}$$

At first, we suspect that these actions are totally equivalent, so that we can choose one and drop the others. This is apparently not so, for two subtle reasons:

(1) Because we are dealing with a first quantized theory, we have to take the sum over all interacting topologies that are swept out by the string. For the Nambu–Goto string, the precise nature of these topologies is ambiguous and must be specified by hand. However, for the Polyakov form of the action, which contains an independent metric tensor, we can eliminate most of this ambiguity by specifying that we sum over all *conformally and modular inequivalent configurations.* (These terms will be defined later.) This will become a powerful constraint once we start to derive loops and will determine the functional measure uniquely. The measure and the topologies in the Nambu–Goto action, however, are not well defined. (We must point out, however, that this rule of integrating over inequivalent

surfaces does not automatically satisfy unitarity. This still must be checked by hand.)

(2) The gauge fixing of Weyl invariance for the Polykov action, although trivial classically, poses problems when we make the transition to quantum mechanics. An anomaly appears when we carefully begin the quantization process. In fact, this conformal anomaly will disappear only in 26 dimensions!

Let us now discuss the quantization of the string action. The strategy we will take in quantizing the free theory to obtain the physical Hilbert space will be first to extract the symmetry of the action, then the currents, and then the algebra formed by the generators of this symmetry. (For the string, the symmetry will be reparametrization invariance and the algebra will be the Virasoro algebra.) Then we must apply the constraints onto the Hilbert space, which eliminates the ghosts and creates a unitary theory. It is important to keep this strategy in mind as we begin the quantization of the string:

Action → Symmetry → Current → Algebra → Constraints → Unitarity

As in the point particle case, we can begin the quantization program in several ways. There are three formalisms in which to fix the gauge of the theory: (1) Gupta–Bleuler (conformal gauge), (2) light cone gauge, and (3) BRST formalism. The advantages and disadvantages of each are as follows:

(1) The Gupta–Bleuler is perhaps the simplest of the three formalisms. We allow ghosts to appear in the action, which permits us to maintain manifest Lorentz invariance. The price we must pay, however, is that we must impose ghost-killing constraints on the Hilbert space. Projection operators must be inserted in all propagators. For trees, this is trivial. For higher loops, however, this is exceedingly difficult.

(2) The advantage of the light cone gauge formalism is that it is explicitly ghost-free in the action as well as the Hilbert space. There are no complications when going to loops. However, the formalism is very awkward and Lorentz invariance must be checked at each step of the way.

(3) The BRST formalism combines the best features of the previous two formalisms. It is manifestly covariant, like the Gupta–Bleuler formalism, and it is unitary, like the light cone formalism, because the negative metric ghosts cancel against the Faddeev–Popov ghosts.

Let us now discuss each quantization scheme separately.

§2.2. Gupta–Bleuler Quantization

The Gupta–Bleuler formalism will maintain Lorentz invariance by imposing the Virasoro constraints on the state vectors of the theory:

$$\langle \phi | L_f | \psi \rangle = 0 \tag{2.2.1}$$