Quantum mechanics of the gravitational field

Claudio Teitelboim
Center for Theoretical Physics, University of Texas, Austin, Texas, 78712
and Institute for Advanced Study, Princeton, New Jersey, 08540
(Received 25 January 1982)

An approach to the quantum theory of gravitation is developed by analogy with the quantum mechanics of the simplest generally covariant system—the relativistic point particle. The central object in the formalism is the transition amplitude from one three-geometry to another which is given by a path integral. In that path integral one sums over all possible histories which connect two three-geometries separated by a given local proper time and then integrates over all possible proper-time separations. The choice of the range of integration for the proper time fixes the boundary conditions for the transition amplitude. If only positive proper times are allowed, the resulting amplitude is causal. A perturbation theory is developed in which the expansion parameter is the signature which takes the value minus one when the field histories (spacetimes) have hyperbolic signature and plus one for the Euclidean case. The "free" theory corresponds to zero signature and may be viewed as the result of replacing the Lorentz group as a symmetry group of the tangent spaces by one of its contractions, namely that one where the speed of light approaches zero. It is argued that besides the processes in which the universe starts or finishes at a singularity, there are also processes with a nonzero amplitude in which the universe starts and finishes in the same regular configuration without ever going through a singularity. These latter processes may be pictured as a loop in the configuration space of the gravitational field. The work remains formal throughout in that no definite meaning is given to the functional integrals considered.

I. INTRODUCTION

An important lesson learned from the development of quantum field theory is that, remarkably enough, it is worthwhile to develop formal schemes ignoring the fact that one's equations many times possess only catastrophic solutions, or what is the same, they possess no solutions at all. Thus one works in practice with formal generalizations to a continuous number of variables of expressions valid for a finite number of them, assuming that basic linear algebra properties and rules for differentiation and integration still hold in the case at hand. It then turns out, sometimes, that one can formulate in this way an approximation scheme to solve the (formal) equations of the theory and—turning the process upside down—use the approximation method to *define* what is meant by one's equations in the first place.

The procedure described above has been carried out successfully so far for theories such as quantum electrodynamics and, more generally, Yang-Mills gauge fields which have been shown to be renormalizable. However, it would appear fair to say that, up to now, straightforward extension of the methods used for the Yang-Mills theories to the quantum theory of the gravitational field, described classically by Einstein's equations, has not been successful.

A natural way to proceed in the face of this impasse is to develop a formal approach to the quantum theory of gravitation which is sufficiently different from previous ones as to suggest either an approximation method of its own or another procedure which could be used in turn to give a precise meaning to the formal equations that one started with. Clearly, only after the equations have been given a precise meaning can one speak of a theory instead of a formalism. However, even before that step is taken (if ever) there is value in a different formal point of view as it still contributes to one's understanding of the problem.

The purpose of this article is to discuss an approach to the quantum theory of gravitation that may be regarded as a blend of the superspace ideas of Wheeler[1] and the path-integral methods originally developed by Feynman.[2] The central object in the formalism is the transition amplitude from one three-dimensional geometry to another which is given by a path integral. In that path integral one sums over all possible histories which connect two given three-geometries and which are separated by a given local proper time (in a sense made precise below) and then integrates over all possible proper-time separations. The choice of the range of integration in the proper-time variable fixes the boundary conditions for the transition amplitude. In particular, if only positive proper times are allowed, the resulting amplitude is causal, that is, it corresponds to the Feynman propagator in superspace.

Within the proper-time formalism there is a natural and apparently unique choice for a perturbation theory. The parameter in which one expands is the signature σ which takes the value minus one when the classical solutions of the theory are Riemannian spacetimes with Minkowskian signature $(-,+,+,+)$ and plus one in the Euclidean case $(+,+,+,+)$. What deserves to be called the "free theory" in this approach is obtained by setting $\sigma=0$, and here the corresponding field histories possess a geometrical structure which is halfway between Euclidean and Minkowskian signatures. Such geometry is that of a manifold where the tangent spaces have by symmetry group a contraction of the Poincaré group which is obtained by letting the speed of light go to zero and which is therefore just the opposite of the ordinary nonrelativistic limit. Consequently, in the free theory the light cones are shrunk to lines and there is no propagation in spacetime.

Although the zero-signature limit looks quite singular when looked at from the viewpoint of spacetime and would therefore seem an unsuitable starting point for perturbation theory, it appears to be regular when viewed from the configuration space of the gravitational field (superspace). In particular, wave propagation in superspace remains perfectly well behaved, and simplifies considerably, by setting $\sigma=0$, and the Feynman propagator remains well defined.

The scope of this article will remain formal in that, for example, the possibility of actually implementing the small-signature perturbation theory as a well-defined scheme, effectively free of infinities, will not be discussed; and hence the functional integrals involved will remain as expressions without a definite meaning.

A short account of these results has appeared in Ref. 3.

II. GENERAL COVARIANCE VERSUS INTERNAL GAUGE SYMMETRY

There are many similarities between a gauge theory of the Yang-Mills type and a generally covariant theory like general relativity. In both cases the action is invariant under a set of transformations, the parameters of which are functions of space and time. As a consequence, in both cases there are constraints among the canonical variables of the theory.

There is, however, one central difference between internal gauge symmetry and general covariance, a difference which at first sight would seem to be a matter of detail, but which on more careful analysis turns out to be a central one. The difference in question resides in the boundary conditions which the symmetry transformations of the action must satisfy. In the Yang-Mills case the action contained in a spacetime region is invariant under gauge transformations even when those transformations are not restricted on the region's boundary. In the case of general coordinate transformations, on the other hand, the action is invariant *only when the transformation is restricted to map the boundary onto itself.* This difference may be taken as *the* distinction between internal and spacetime gauge symmetries. In fact, if one is given an action principle, one should be able to recognize just by inspection of the necessary boundary conditions for invariance, and without any knowledge of differential geometry or fiber-bundle theory, which kind of theory one is dealing with. This is quite satisfactory as one should expect all criteria and distinctions of physical relevance to be supplied by the action integral itself.

The amount of freedom for the gauge transformations on the boundary has important implications for quantum mechanics in that it determines the kind of gauge conditions which is permissible to impose on the system at hand without eliminating physically permissible histories from the path integral. Thus for an internal gauge symmetry (Yang-Mills), one is free to impose conditions which restrict more strongly the behavior of the dynamical variables on the boundary than in the case of a generally covariant theory. This is so because in the latter case there is less gauge freedom available on the boundary.

The most efficient general, and foolproof, way of analyzing the quantum mechanics of a gauge system appears to be through the Hamiltonian form of the path integral,[4] and we shall follow that procedure in the study of the quantum mechanics of the gravitational field. However, instead of addressing immediately that problem, we will analyze first the simplest example of a generally covariant system, namely, a relativistic spinless particle in Minkowski space. This has the advantage that many of the key features of the more complicated systems already appear in this physically well-understood problem. In particular, one of our central results, namely the fact that *canonical gauges are not permissible for generally covariant systems,* and its implications, will be established first in this simple case.

III. RELATIVISTIC POINT PARTICLE ("GENERAL RELATIVITY IN ZERO SPATIAL DIMENSIONS")

A. Lagrangian

The action for a relativistic particle in (flat) configuration space is

$$S[x(\tau)]=\int_{\tau_1}^{\tau_2} L\,d\tau \tag{3.1}$$

with

$$L=-m\left[\frac{dx^{\mu}}{d\tau}\frac{dx_{\mu}}{d\tau}\right]^{1/2}=-m(-x^2)^{1/2} \tag{3.2}$$

and

$$x(\tau_1)=x_1,\;\; x(\tau_2)=x_2\;. \tag{3.3}$$

This action is invariant under reparametrizations:

$$x(\tau)\rightarrow\bar{x}(\tau)=x(f(\tau))\;, \tag{3.4}$$

provided the transformation becomes the identity at the boundary:

$$f(\tau_1)=\tau_1,\;\; f(\tau_2)=\tau_2\;. \tag{3.5}$$

Note that since the boundary has dimension zero this statement is equivalent to mapping the boundary onto itself.

The infinitesimal version of (3.4) is obtained by setting

$$f(\tau)=\tau+\epsilon(\tau)\;, \tag{3.6}$$

and the corresponding change in the action is

$$\delta S=(-\dot{x}^2)^{1/2}\epsilon\,\Big|_{\tau_1}^{\tau_2}\;, \tag{3.7}$$

which vanishes for arbitrary $\dot{x}(\tau_2)$, $\dot{x}(\tau_1)$ (the velocities are not restricted at the end points) if and only if

$$\epsilon(\tau_1)=\epsilon(\tau_2)=0\;, \tag{3.8}$$

which is the infinitesimal version of (3.5).

Now, one might attempt doing quantum mechanics directly in terms of the action (3.1) by defining the propagator as a sum over all possible histories of the exponential of $(i/\hbar)$ times the action (3.1). This meets, however, several difficulties. First of all, there is the practical problem of evaluating a functional integral which is not Gaussian. Second, it is not clear how to treat spacelike histories for which the integrand in (3.1) becomes imaginary. Third, it is not clear either how to eliminate the multiple counting of histories associated with (3.4), i.e., fixing the gauge, by imposing a condition on the $x^{\mu}(\tau)$ (see below for more on this). Fourth, there is no indication of how to distinguish the Feynman propagator—which is the desired answer—from other Green's functions of the Klein-Gordon equation. All these difficulties are overcome by going to the Hamiltonian form of the path integral which we will now discuss.

B. Hamiltonian

The Hamiltonian form of the action (3.1) is obtained in the standard manner. The momentum conjugate to $x^{\mu}(\tau)$ is

$$p_{\mu}=\frac{\partial L}{\partial\dot{x}^{\mu}}=m(-x^2)^{-1/2}\dot{x}_{\mu}\;, \tag{3.9}$$

which implies the constraint

$$\mathscr{H}_0=p_{\mu}p^{\mu}+m^2=0\;. \tag{3.10}$$

Furthermore the canonical Hamiltonian

$$p_{\mu}\dot{x}^{\mu}-L$$

vanishes identically on account of the homogeneity in the velocities of (3.2), which is in turn a consequence of reparametrization invariance.

Consequently, the equations of motion for the system are obtained by extremizing the Hamiltonian action,

$$S[x(\tau),p(\tau);N(\tau)]=\int_{\tau_1}^{\tau_2}(p_{\mu}\dot{x}^{\mu}-N\mathscr{H}_0)d\tau\;, \tag{3.11}$$

under variations of $x(\tau)$, $p(\tau)$, $N(\tau)$, subject again only to the conditions

$$x(\tau_1)=x_1,\;\; x(\tau_2)=x_2\;. \tag{3.12}$$

Thus, neither the momenta p_μ nor the Lagrange multiplier N are restricted at the end points. The reason is that after (3.12) is imposed there would be in general no solutions for the equations of motion compatible with restricting the p's or N at the end points in addition to the x's.

Now, since the original Lagrangian action (3.1) was invariant under gauge transformations involving one arbitrary function of time $\epsilon(\tau)$, one would expect that invariance to show itself somehow in the Hamiltonian formalism. This means that the action functional (3.11) should be invariant under a transformation of its arguments x, p, and N, which depends on a function parameter $\epsilon(\tau)$.

The way in which the canonical variables x and p must change is the standard one from the theory of constrained Hamiltonian systems,[5] namely one writes, in terms of the Poisson brackets in x and p,

$$\delta x(\tau)=\epsilon(\tau)[x,\mathscr{H}_0] \tag{3.13a}$$

and

$$\delta p(\tau)=\epsilon(\tau)[p,\mathscr{H}_0] , \tag{3.13b}$$

where $\mathscr{H}_0$ is the constraint (3.10). To these equations one must add a transformation law for N which is determined by the requirement that the action be invariant.[6,7] In the present simple case that transformation turns out to be just

$$\delta N(\tau)=\dot{\epsilon}(\tau) . \tag{3.14}$$

In effect, if one evaluates the change of the action (3.10) under (3.12)–(3.14) one finds

$$\delta S=\epsilon\left[p\frac{\partial\mathscr{H}_0}{\partial p}-\mathscr{H}_0\right]\Bigg|_1^2 . \tag{3.15}$$

Now, the constraint (3.9) is quadratic in the momenta. Consequently, the expression

$$p\frac{\partial\mathscr{H}_0}{\partial p}-\mathscr{H}_0 \tag{3.16}$$

is different from zero (it equals in this case p^2-m^2), and in order for the action to be invariant for arbitrary values of p_μ at the end points (the p's are not restricted at τ_1,τ_2) we must have

$$\epsilon(\tau_1)=0=\epsilon(\tau_2) , \tag{3.17}$$

in agreement with (3.8).

Thus we learn that general covariance manifests itself in the Hamiltonian formalism in the appearance of constraint generators *which are not linear and homogeneous in the momenta.* (In all cases of interest found so far, the constraints in question turn out to be quadratic in p.)

That quadratic constraints are the distinguishing mark of generally covariant systems is confirmed by observing that, for example, in Yang-Mills theory the gauge constraints are linear functions of the p's and therefore expression (3.16) is identically zero in that case. As a consequence the Yang-Mills Hamiltonian action is invariant under the analog of (3.13) and (3.14) even when (3.17) does not hold. A similar feature will be found below in gravitation theory where the constraints which generate mappings of the boundary onto itself are linear, whereas those which generate deformations of the boundary in a direction normal to itself turn out to be quadratic.

C. Gauge fixation

In order to pass to quantum mechanics one must eliminate the multiple counting of histories inherent in (3.13) and (3.14) by imposing an additional restriction (gauge condition) on the allowed histories of the system. That condition must be such as to make it possible to deform any given history into another one where the gauge condition holds, by an iteration of transformations of the type (3.13) and (3.14), *subject to the boundary conditions (3.17).* If this is not possible, the gauge condition is not a permissible one since it eliminates physically relevant modes from the system. Furthermore, the gauge must be fixed completely. That is, if a transformation of the type (3.13) and (3.14) subject to (3.17) preserves the gauge condition, then it must be the identity $\epsilon(\tau)=0$ for all τ.

One usually fixes the gauge by imposing additional conditions on the canonical variables of the theory of the form

$$\tau=\phi(x,p) \tag{3.18}$$

for all times (including end points). A condition of this form is customarily called a canonical gauge. However, for the gauge freedom associated with a quadratic constraint such as (3.10) *a canonical gauge is not permissible.* This is due to the fact that, on account of (3.17), there is no gauge freedom available at the end points and consequently (3.18) implies a restriction on physically relevant modes. This is an important difference between generally covariant systems and ordinary gauge symmetries of the Yang-Mills type where canonical gauges are allowed because there is full gauge freedom at the end points.

It is important to stress here that the impossibility of using a canonical gauge in the case of a quadratic constraint generator is not a matter of practical difficulty but an obstacle of principle.

If canonical gauges are not allowed, how should one fix the gauge? The answer is provided by observing that an admissible gauge condition should be such that if it is itself subject to a gauge transformation it should yield a *second*-order linear differential operator acting on $\epsilon(\tau)$ and that operator should have a unique inverse subject to the *two* boundary conditions (3.17). The existence of the inverse ensures that any history can be deformed into another obeying the gauge condition and its uniqueness guarantees that the gauge has been fixed completely.

Since the canonical variables x and p transform with ϵ undifferentiated, a condition involving only x and p would require at least one second derivative $\ddot{x}$ or $\ddot{p}$ to appear. Admitting second derivatives would seriously complicate the action principle as they would be effectively incorporated into the Lagrangian when adding the gauge condition with a Lagrange multiplier to the action. Although it is conceivable that a consistent way of dealing with such second derivatives may exist, it is not an alternative that one would like to explore unless forced to do so, and fortunately that is not the case. In fact, one sees from (3.14) that the transformation law for the Lagrange multiplier N involves already the first derivative of $\epsilon(\tau)$. Hence a gauge condition involving the first derivative of N as well as N, x, and p undifferentiated will do. The simplest of all those conditions is just

$$\dot{N}=0\,, \tag{3.19}$$

which is the prototype of a noncanonical gauge.

Indeed, given an arbitrary history where $N=N(\tau)$ one may deform it into another satisfying (3.19) by a transformation (3.13) and (3.14) with $\epsilon(\tau)$ given by

$$\ddot{\epsilon}(\tau)=-\dot{N}(\tau)\,,$$

an equation which possesses one and only one solution under condition (3.17) (there is no difference in this case between infinitesimal and finite transformations since there is only one parameter involved).

A wide class of permissible gauge conditions consists of those of the form

$$\dot{N}=\chi(p,x,N)\,, \tag{3.20}$$

where χ is any given function.

D. Path integral

We will now study a path-integral expression for the transition amplitude from x_1 to x_2 for the free particle with action (3.11). That action has the full gauge freedom (3.13) and (3.14) of the theory; so in order to evaluate the path integral it is necessary to choose a gauge condition and we shall take the simplest one, namely (3.19). It may be shown that the result obtained is independent of the particular gauge condition chosen.

When writing down a path integral one has to integrate over all the variables which are varied in the action in order to obtain the classical equations of motion; this is so because the classical theory is obtained by means of a stationary phase approximation to the path integral. In the present case these variables are $x(\tau)$, $p(\tau)$, and $N(\tau)$ at all instants of time in the interval $[\tau_1;\tau_2]$, subject to the end-point conditions (3.12) on $x(\tau)$, while $p(\tau)$ and $N(\tau)$ are not fixed at either τ_1 or τ_2. However, after fixing the gauge by (3.19), only one independent variable of integration N—which may be thought of as being $N(\tau_1)$— is left from the infinitely many $N(\tau)$.

If we first perform the integration over $p(\tau)$ and $x(\tau)$ we obtain a functional integral of the form

$$\int \tilde{K}_0[x_2,x_1;N,\tau_2,\tau_1]=\exp\left[i\int_{\tau_1}^{\tau_2}(p\dot{x}-N\mathscr{H}_0)d\tau\right]\prod_{\tau}\frac{dp(\tau)dx(\tau)}{2\pi}\,, \tag{3.21}$$

where N is a fixed number. The right-hand side of (3.21) is just the path-integral expression of the matrix element

$$\left\langle x_2\left|e^{-iN(\tau_2-\tau_1)\mathscr{H}_0}\right|x_1\right\rangle. \tag{3.22}$$

Next, the amplitude of interest is obtained by integrating (3.21) over all possible values of N. This last integration over N is not a functional one but a plain ordinary integral over one variable.

Although technically simple, the integration over N involves two conceptually important aspects. The first concerns the measure and the second the range of integration.

The measure in the path integral is determined by the requirement that the answer should be independent of the particular gauge condition chosen. Thus, one should require, in particular, that the specific form of the function χ in (3.20) should be immaterial. Application to the simple case at hand of the general theorem proven in Ref. 7 yields that the measure is simply that indicated in (3.21) together with an integration over

$$T_{21}=N(\tau_2-\tau_1)\ , \tag{3.23}$$

instead of over N itself. The meaning of T_{21} given by (3.23) will be explained below.

The analysis concerning the range of integration proceeds as follows. Consider, to begin with, a trajectory which solves the equations of motion. Then we have, in particular,

$$\frac{dx}{d\tau}=N[x,\mathscr{H}_0]=2Np\ . \tag{3.24}$$

Next, observe that although arbitrary the parameter τ must be a good one, i.e., it must be such that two points on the trajectory have the same τ only if they coincide (this is the reason for introducing τ in the first place). This means that N must be different from zero for all times. Furthermore, N is assumed to be a continuous function of τ. It follows that it must be either always positive or always negative. Therefore we see that the trajectories which obey (3.24) may be divided in two disjoint classes, those with $N>0$ and those with $N<0$. At this point one makes a crucial assumption, namely, that the physical amplitude is obtained by integrating *over just one of these two classes.* Which class is actually selected is in the last instance a matter of convention, since both differ by an interchange of past and future. Normally one takes $N>0$.

We will give below some arguments to justify integrating over one class only, but the best argument is perhaps the result obtained. So, we proceed to write the answer which, on account of (3.21)–(3.23), reads

$$K_0(x_2,x_1)=\int_0^\infty \widetilde{K}_0(x_2,x_1;T_{21})dT_{21}\ , \tag{3.25}$$

with $\widetilde{K}$ given by (3.21) and (3.22), and which in the free case (3.10) becomes

$$\widetilde{K}_0(x_2,x_1;T_{21})=e^{-im^2T_{21}}\left\langle x_2\left|e^{-ip^2T_{21}}\right|x_1\right\rangle . \tag{3.26}$$

Equation (3.25) is recognized as the integral representation for the Feynman propagator. In order to damp the oscillations for large T_{21} and thus render (3.25) well defined, it is necessary to introduce the prescription

$$\mathscr{H}_0\to\mathscr{H}_0-i\eta\ , \tag{3.27}$$

with $\eta>0$. Notice that as a consequence of integrating over T_{21} given (3.23) rather than on N itself, there is no mention in the final result (3.25) of the arbitrary end-point parameters τ_1,τ_2.

Let us now return to the reasons for restricting the integration in (3.25) to $T_{21}>0$. First of all one may argue that since the classes with $N>0$ and $N<0$ cannot be deformed continuously into each other (given that $N=0$ is excluded), the relative weight of contributions from both classes to the path integral remains undetermined and thus a choice must be made. Now, in order to motivate the choice, consider for definiteness the class with $N>0$. Equation (3.24) implies in that case that

$$\frac{dx^0}{d\tau}>0 \quad \text{if } p^0>0 \tag{3.28a}$$

and

$$\frac{dx^0}{d\tau}<0 \quad \text{if } p^0<0\ , \tag{3.28b}$$

so we see that the $N>0$ class contains trajectories with both orientations of x^0 relative to τ, provided we allow for positive and negative energies. Furthermore, according to the inequalities in (3.28) when $N>0$ a particle with positive energy "travels forward in time" whereas one of negative energy travels backward in time. Thus, from a physical point of view, all trajectories of interest (particles and antiparticles) are already included in that one class. Finally, one may rephrase the above argument by saying that restricting the integration to positive values of T_{21} amounts to inserting a factor $\Theta(T_{21})$ in the right-hand side of (3.21) which converts that equation into the usual expression for the *retarded,* i.e., causal, Green's function (in the time T_{21} not in x^0) $\widetilde{K}_0[x_2,x_1;T_{21}]$ of a nonrelativistic particle with space coordinates x^μ, obeying

$$\widetilde{K}_0[x_2,x_1;T_{21}]=0 \quad \text{if } T_{21}<0\ . \tag{3.29}$$

The reader may add to the above arguments his own favorite one for selecting the Feynman propagator over other possible choices. However, it is an important point here that by selecting the appropriate integration range for T_{21} one finds the correct propagator *directly from its fundamental definition as a path integral* without ever invoking the concept of a Green's function for the Klein-Gordon equation (other choices for the relative weight of the $N>0$ and $N<0$ classes yield other invariant functions for the Klein-Gordon equation, not all of them Green's functions). This is of importance since in the extension on general relativity discussed below, the analog of a Green's function in a functional configuration space is not an easy concept to deal with, whereas the analog of equation (3.25) may be written down in a simple manner.

Incidentally, it is of interest to notice that the arguments based on (3.24) rely only on the half of the classical equations of motion which is obtained by extremizing the action with respect to the momenta. Now, since the action is Gaussian in p one will find that (3.24) will be effectively implemented upon functional integration over p_μ. Therefore, the reasoning in question holds for every trajectory in the configuration (x^μ) space and not just for extremal ones. Last, a word about the meaning of T_{21}. The total proper time between τ_2 and τ_1 for a particle obeying equations (3.10), (3.19), and (3.24) is

$$\int_{\tau_1}^{\tau_2}\left[-\left[\frac{dx}{d\tau}\right]^2\right]^{1/2} d\tau=2Nm(\tau_2-\tau_1)$$

$$=2mT_{21}\ , \qquad (3.30)$$

for this reason T_{21} is referred to as "the proper time."

There is in fact a useful physical image to describe Eq. (3.25), which is as follows. One says that $\tilde{K}_0$ is the amplitude for propagation from 1 to 2 during the proper time interval T_{21}. Next one adds that the proper time employed in propagation is not observable and that one is interested in the total amplitude to arrive at x_2 from x_1 irrespective of how long it takes in propertime for the propagation to occur. Since in such a situation the individual amplitudes must be added, one integrates over all possible values of T_{21}. Only positive T_{21} are allowed for causality reasons ["retardation in proper time," Eq. (3.29)]. This way of speaking will be employed below.

E. Perturbation theory

Suppose that the particle interacts with an external scalar (for simplicity) potential $\mathscr{U}(x)$. That is, one adds a position-dependent contribution to the square of the mass, thus replacing $\mathscr{H}_0$ in (3.11) by

$$\mathscr{H}=\mathscr{H}_0+\mathscr{U}(x)\ . \qquad (3.31)$$

One is then interested in the deviation in the propagation amplitudes from their free value due to the presence of $\mathscr{U}$. The corresponding perturbation expansion for $\tilde{K}$ is then simply the familiar one,[8]

$$\tilde{K}=\tilde{K}_0+\tilde{K}^{(1)}+\tilde{K}^{(2)}+\cdots\ , \qquad (3.32)$$

with

$$V=N\mathscr{U} \qquad (3.33)$$

playing the role of the nonrelativistic potential V. Thus one has

$$\tilde{K}^{(1)}[3,1;N,\tau_3,\tau_1]=-i\int_{\tau_1}^{\tau_3}d\tau_2\int dx_2\tilde{K}_0[3,2;T_{32}]N\mathscr{U}(2)\tilde{K}_0[2,1;T_{21}]\ , \qquad (3.34a)$$

$$\tilde{K}^{(2)}[4,1;N,\tau_4,\tau_1]=(-i)^2\int_{\tau_1}^{\tau_4}d\tau_3\int_{\tau_1}^{\tau_4}d\tau_2\int dx_3dx_2\tilde{K}_0[4,3;T_{43}]N\mathscr{U}(3)$$
$$\times\tilde{K}_0[3,2;T_{32}]N\mathscr{U}(2)\tilde{K}_0[2,1;T_{21}]\ , \qquad (3.34b)$$

etc.

Now, the integrand on the right-hand side of (3.34a) depends in a rather complicated manner on N, τ_3, and τ_1 and not just through the product $N(\tau_3-\tau_1)$. It is therefore not possible to define an amplitude $\tilde{K}^{(1)}[3,1;T_{31}]$ through (3.34a). However, after integrating (3.34a) over $dN(\tau_3-\tau_1)$ one does obtain an answer independent of τ_3 and τ_1 as should be the case. This assertion may be checked by changing the integration variables from τ_2 and N to $T_{32}=N(\tau_3-\tau_2)$ and $T_{21}=N(\tau_2-\tau_1)$. One then finds that if τ_2 runs between τ_1 and τ_3 and N from zero to infinity, then T_{32} and T_{21} run from zero to infinity; furthermore,

$$dT_{32}dT_{21}=N\,dN\,d\tau_2(\tau_3-\tau_1)\ . \qquad (3.35)$$

Therefore one finds

$$K^{(1)}=\int\tilde{K}^{(1)}[3,1;N,\tau_3,\tau_1]dN(\tau_3-\tau_1)$$
$$=-i\int dT_{32}dT_{21}\tilde{K}_0[3,2;T_{32}]\mathscr{U}(2)$$
$$\times\tilde{K}_0[2,1;T_{21}]\ , \qquad (3.36)$$

which yields [since $\mathscr{U}(2)$ is independent of T_{32} and T_{31}]

$$K^{(1)}[3,1]=-i\int dx_2 K_0[3,2]\mathscr{U}(2)K_0[2,1] \;. \tag{3.37}$$

Similar changes of variables lead to

$$K^{(2)}[4,1]=(-i)^2\int dx_2 dx_3 K_0[4,3]\mathscr{U}(3)K_0[3,2]\mathscr{U}(2)K_0[2,1] \;, \tag{3.38}$$

etc. In this way one obtains the standard perturbation expansion for the propagator from the proper-time representation (3.25).

F. Extension to several particles

The previous discussion of perturbation theory can be extended to systems of several particles with spacetime coordinates $x_1,x_2,\ldots,x_p$ each of which interacts with an external potential. In such a case the action reads

$$S=\int_{\tau_i}^{\tau_f}\sum_{n=1}^{p}(p_n\dot{x}_n-N^n\mathscr{H}_n)d\tau \;, \tag{3.39}$$

where

$$\mathscr{H}_n=\mathscr{H}_n^{(0)}+\mathscr{U}_n \;. \tag{3.40}$$

Now, in order to have a theory which is invariant under independent reparametrizations of the n world lines, the functions $\mathscr{H}_n$ must be first class in Dirac's terminology, which means that the Poisson brackets of any two $\mathscr{H}$'s must be a linear combination of the $\mathscr{H}$'s themselves. It is easy to see that this implies that the nth $\mathscr{U}$ must depend only on the coordinates of the nth particle, which means that the particles do not couple to each other. Therefore, the propagator for the p-particle system is the product of one-particle propagators, each of which has a perturbation series of the type previously described.

It is, however, instructive to rederive the results that could be obtained by multiplying the expansion of the one-particle propagators directly from the action (3.39) as a whole. The reason for this seemingly futile exercise is that in the case of the gravitational field the analog of $\mathscr{U}_n$ *can* depend on coordinates other than x_n (this is due to the presence of other constraints besides the $\mathscr{H}_n$ which close the algebra—a feature which has no analog here) and therefore the propagator does not factorize. Hence in the gravitational case one has no alternative but to deal with the action as a whole. With this in mind we will not use the fact that $\mathscr{U}_n$ depends only on x_n until the very end of the discussion.

We may obtain the perturbation theory by noticing simply that

$$V=\sum_{n=1}^{p}N^n\mathscr{U}_n \tag{3.41}$$

plays in this case the role of the potential (3.33). Therfore, the first-order correction to $\tilde{K}$ is obtained by replacing $N\mathscr{U}$ in (3.34a) by (3.41) and letting the various subscripts refer collectively to the coordinates of all the particles. Thus, for example, dx_1 means now a product over all the particles of the spacetime volume elements associated with each one.

The $\tilde{K}$ so obtained will depend on the end-point coordinates x_f, x_i of all the particles, on the lapse functions N^n of all the particles, and on τ_f and τ_i. To obtain the physical amplitude one must in turn integrate this $\tilde{K}$ over $dN^1(\tau_f-\tau_i)dN^2(\tau_f-\tau_i)\cdots$ (one factor for each particle).

Now, in the case when only one particle was present it was possible to perform the integration over $d\tau$, and $dN(\tau_f-\tau_i)$ by means of the change of variables (3.35) and thus obtain an answer which was expressed solely in terms of the physical unperturbed propagator K_0 (Eq. 3.37). In the case of several particles *this last step cannot be taken,* namely, it is not possible to obtain an answer similar to (3.37) where $K^{(1)}$ is expressed solely in terms of the potentials $\mathscr{U}_n$ and the unperturbed propagator K_0 of the total system.

Technically this is due to the fact that if there are p particles, it is necessary to perform $2p$ proper-time integrations to convert the two propagators $\tilde{K}_0$ in (3.34a) to K_0's. There are, however, only $(p+1)$ time integrals available [p from $dN^1(\tau_f-\tau_i)\cdots dN^p(\tau_f-\tau_i)$ and one from $d\tau_1$]. If $p=1$ this just balances, if $p>1$ it does not.

Physically the reason is that the elementary process involved is not the scattering of the p particles as a whole but rather the scattering of each individual particle by its own potential. To see this more clearly let us consider the case of two particles with the coordinates x,y and lapse functions N_x,N_y, respectively. Then the analog of (3.34a) reads

$$\tilde{K}^{(1)}=-i\int_{\tau_1}^{\tau_f}d\tau_1\int dx_1dy_1\tilde{K}_0[f,1;N_x(\tau_f-\tau_1),N_y(\tau_f-\tau_1)]\times[N_x\mathscr{U}_x(1)+N_y\mathscr{U}_y(1)]\tilde{K}_0[1,i;N_x(\tau_1-\tau_i),N_y(\tau_1-\tau_i)]\ . \tag{3.42}$$

Now, the free propagator $\tilde{K}_0$ factorizes as a product of two single-particle propagators. Hence we can carry out the integrations over $d\tau_1$, $dN_x(\tau_f-\tau_i)$ and $dN_y(\tau_f-\tau_i)$ in the first and second term in (3.42) just as in the one-particle case [i.e., using (3.35) for each particle]. This yields

$$\tilde{K}^{(1)}=(-i)\int dy_1dx_1\int K_0[x_f,x_1]\mathscr{U}_x(1)\tilde{K}_0[x_1,x_i]\tilde{K}_0[y_f,y_1,N_y(\tau_f-\tau_1)]\times\tilde{K}_0[y_1,y_i,N_y(\tau_1-\tau_i)]dN_y(\tau_f-\tau_i)+(x\leftrightarrow y)\ . \tag{3.43}$$

The amplitude (3.43) is the sum of two terms corresponding to mutually exclusive alternatives. In each of these alternatives, particles x and y play an asymmetric role and adding up the contributions from both alternatives restores the symmetry. If the potentials $\mathscr{U}_x$ and $\mathscr{U}_y$ depended on the coordinates of both particles, this is as far as one could go, and there would appear to be no concise physical image to describe each of these processes. However, if $\mathscr{U}_x$ depends only on particle x and $\mathscr{U}_y$ depends only on particle y, one can go one step further. In that case the integration over y_1 can be performed on the first term in (3.43), yielding

$$\int\tilde{K}_0[y_f,y_1,N_y(\tau_f-\tau_1)]\tilde{K}_0[y_1,y_i,N_y(\tau_f-\tau_i)]dy_1=\tilde{K}_0[y_f,y_i,N_y(\tau_f-\tau_i)]\ , \tag{3.44}$$

an expression which can consequently be integrated on $N_y(\tau_f-\tau_i)$ to yield just $K_0[y_f,y_i]$. The answer then reads

$$K^{(1)}=(-i)\int dx_1K_0[x_f,x_1]\mathscr{U}_x(1)K_0[x_1,x_i]K_0[y_f,y_i]+(x\leftrightarrow y)\ , \tag{3.45}$$

which is what could be obtained much more easily from the fact that if $U_n=U_n(x_n)$ the full propagator factorizes.

Equation (3.45) describes the first-order amplitude as a sum of two alternatives corresponding to one of the particles being scattered once while the other proceeds freely and vice versa. What is to be noted here is that Eq. (3.43) as it stands contains all this information essentially from the fact that the sum over all particles in (3.41) makes the amplitude itself a sum over exclusive alternatives. Each of these alternatives is an elementary process, and for this reason the amplitude cannot be expressed as a function of the totak K_0 but must involve the one-particle K_0's (Eq. 3.45). If the potentials $\mathscr{U}_n$ could couple different x_n's (as it happens in the gravitational case), one could not even reach stage (3.45) but (3.43) as it stands with its $\tilde{K}_0$'s would be the final expression.

IV. GRAVITATIONAL FIELD

We will discuss in what follows the development of a formalism for the quantization of the gravitational field along lines similar to those explained above for the point particle.

The analogy between both systems will be heavily relied upon. Many technical proofs will be omitted in order not to make this account unduly extended while still trying to convey the essential points as clearly as possible.

For the same reasons of simplicity it will be assumed that the three-dimensional spacelike sections of spacetime are topologically compact ("closed space" in the language of cosmology) since in this case the analogies with the point particle are closest and the departures from ordinary field theory are sharpest. The open, asymptotically flat, case can also be dealt with in the present approach provided due consideration is given to the role in the action of integrals over a remote two-dimensional surface.[9] Those integrals act as generators of Poincaré transformations at spacelike infinity and must be included because in the asymptotically flat situation the propagation amplitude depends not only on two three-geometries (as discussed below) but also on the "location" of those three geometries at infinity. This additional dependence must be taken into account both in writing expressions for the propagation amplitude and in interpreting them. We plan to describe the special features of the open case elsewhere.

Couplings to matter will not be considered; their inclusion does not introduce conceptually new features.

A. Classical action principle

The Hamiltonian action principle for Einstein's equations[10,11] is given by the statement that a functional of the form

$$S=\int (\pi^{ij}\dot{g}_{ij}-N\mathscr{H}-N^i\mathscr{H}_i)d^3x\,d\tau \qquad (4.1)$$

must be extremized with respect to variations of g_{ij}, π^{ij}, N, and N^i.

The integral in (4.1) is extended over the region of spacetime included between two generic spacelike surfaces $\tau=\tau_1$, $\tau=\tau_2$. The g_{ij} are the components of the metric tensor on the three-dimensional spacelike surfaces τ=constant and π^{ij} are their canonically conjugate momenta. The functions N and N^i are Lagrange multipliers which describe the relative position of two neighboring spacelike surfaces.

The only restriction on the variations of the functions involved in (4.1) is that the spatial metric tensor g_{ij} must be fixed up to a change of spatial coordinates both at $\tau=\tau_1$ and $\tau=\tau_2$.

The constraints $\mathscr{H}$ and $\mathscr{H}_i$ are given by

$$\mathscr{H}=(2\kappa)G_{ijkl}\pi^{ij}\pi^{kl}+(2\kappa)^{-1}g(\sigma R+2\Lambda)\ , \qquad (4.2)$$

$$\mathscr{H}_i=-2\nabla_j\pi^j{}_i \qquad (4.3)$$

with

$$G_{ijkl}=\tfrac{1}{2}(g_{ik}g_{jl}+g_{il}g_{jk}-g_{ij}g_{kl})\ , \qquad (4.4)$$

and where g, R, and ∇ are, respectively, the determinant, curvature, and covariant derivative corresponding to g_{ij}, Λ is the cosmological constant, and $\kappa=8\pi G/c^3$. The quantity σ is the signature of spacetime[6,11–13]; it takes the value -1 for the hyperbolic case and $+1$ for the Euclidean one.

The action (4.1) is invariant under a gauge transformation which is the generalization of (3.13) and (3.14). It takes the form[6,7]

$$\delta g_{ij}(x,\tau)=[g_{ij}(x,\tau),H(\epsilon^\mu)]\ , \qquad (4.5a)$$

$$\delta\pi^{ij}(x,\tau)=[\pi^{ij}(x,\tau),H(\epsilon^\mu)]\ , \qquad (4.5b)$$

$$\delta N^\rho(x'',\tau)=\dot{\epsilon}^\rho(x'',\tau)+\int \kappa_{\mu\nu}{}^\rho(x,x';x'')N^\nu(x')\times\epsilon^\mu(x'')dx'dx''\ . \qquad (4.6)$$

Here $H[\epsilon^\mu]$ denotes a linear combination over all points of space of the constraints (4.2) with arbitrary coefficient functions $\epsilon^\mu\equiv(\epsilon,\epsilon^i)$:

$$H[\epsilon^\mu]=\int (\epsilon\mathscr{H}+\epsilon^i\mathscr{H}_i)d^3x\ , \qquad (4.7)$$

and the functions κ are the structure coefficients appearing in the Poisson bracket algebra[6,13] of the constraints $\mathscr{H}_\mu\equiv(\mathscr{H},\mathscr{H}_i)$:

$$[\mathscr{H}_\mu(x),\mathscr{H}_\nu(x')]=\int \kappa_{\mu\nu}{}^\rho(x,x';x'')\times\mathscr{H}_\rho(x'')d^3x''\ . \qquad (4.8)$$

An example of equations (4.8)—actually, the most interesting one—when written explicitly reads

$$[\mathscr{H}(x),\mathscr{H}(x')]=-\sigma[g(x)g(x')]^{1/2}[\mathscr{H}^i(x)+\mathscr{H}^i(x')]\delta_{,i}(x,x')\ . \qquad (4.9)$$

This equation differs from the one given in Ref. 13 because the constraint $\mathscr{H}$ used here differs from the $\mathscr{H}_\perp$ used there by a factor: $\mathscr{H}=g^{1/2}\mathscr{H}_\perp$. The present choice of weight has been preferred because it simplifies considerably the measure for the path integral discussed below.

The transformation (4.5) and (4.6) leaves (4.1) invariant provided the function ϵ is restricted to vanish at the end points:

$$\epsilon(x,\tau_1)=\epsilon(x,\tau_2)=0 \quad \text{for all } x\ . \qquad (4.10)$$

Equation (4.10) is the analog of (3.17); it follows from the quadratic dependence of $\mathscr{H}$ given by (4.2) on the momenta π^{ij}. No restriction is imposed on the functions ϵ^i at the endpoints because the constraints $\mathscr{H}_i$ given by (4.3) are linear and homogeneous in π^{ij}.

Geometrically, the reason for the necessity of (4.10) is that the action of $\mathscr{H}$ deforms a hypersurface perpendicularly to itself and therefore if ϵ were not zero on the boundary, the region of spacetime over which the integral (4.1) is extended would be changed by the transformation. This reasoning does not apply to $\mathscr{H}_i$ which maps each hypersurface, and in particular the boundary, onto itself.

B. Analogy with point particle

It follows from the above discussion that (4.2) plays in gravitation theory a role similar to that of (3.10) [or, more precisely, (3.31) (see below)] for a particle. The constraint $\mathscr{H}_i$, on the other hand, enters into the formalism much as the generator of a symmetry of the Yang-Mills type would.

The analog of x^μ for the point particle is the three-dimensional metric g_{ij} up to a change of coordinates, or in other words, the three-geometry. This is so because in the action principle based on (4.1) one needs to fix the three-geometry at the end points. Therefore the basic object of interest in quantum gravity is the amplitude $K[2,1]$ for propagation from one three-geometry to another.

In order to specify a three-geometry, one needs, roughly speaking, three functions of three variables (six metric coefficients g_{ij} minus three changes of coordinates). In the classical theory two of those functions may be thought of as describing the two independent degrees of freedom of the gravitational field at a given instant of time. The third function corresponds then to specifying intrinsically the "instant of time" by giving the equation of a hypersurface within spacetime. In the analogy with the relativistic particle, the first two functions correspond to the three spatial coordinates x^i while the third corresponds to giving x^0. By giving x^i and x^0, i.e., x^μ, one therefore picks a point on the world line without ever mentioning a particular parametrization $x(\tau)$ of the world line. This is what is meant by the term "intrinsically" above.

Now, in the particle case, the quantum theory was developed by keeping the four x^μ on the same footing. This was not only natural since in the action principle the four x^μ enter symmetrically, but it was also compulsory in the sense that breaking that symmetry by choosing a gauge such as $x^0=\tau$ was not permissible.

In the gravitational case, for the same reasons, the quantum formalism will be developed below dealing all the time with the three-geometry as a whole rather than assigning a privileged role to any conceivable splitting which would separate the two functions necessary to describe the "independent degrees of freedom" from the "intrinsic time." That the theory should be built in a manner which is independent of any such splitting would appear to be as important for the quantization of the gravitational field as maintaining Lorentz covariance is for the relativistic particle of Sec. III.

Several comments are pertinent in this context. First, it should be clear that treating the three-geometry as a whole is a different statement from maintaining the invariance under changes of spatial coordinates. For example, one may split the three-geometry into a conformally invariant part and a local scale factor in a coordinate-independent manner and therefore preserve the coordinate invariance while breaking the uniformity in the treatment of the three-geometry. Second, in the particle case, one may choose to treat the timelike coordinates x^0 in a different footing from the x^i in an intermediate calculation or to fix the ideas in a particular context, the final formulas will however always be symmetric in the four x^μ. Similarly, here one may take advantage of a particular splitting of the three-geometry, such as that into a conformal part and a local scale mentioned above without breaking the symmetry of the theory itself. We shall indeed use this possibility below.

At this point a related remark is called for concerning the role of general covariance in the quantum theory. In the classical theory one may perform changes of the four spacetime coordinates and the action remains invariant. However, in the final equations of the quantum theory there is only room for changes of the spatial coordinates, which still play a role in the formalism if the three-geometry is specified by means of the metric coefficients, but there is no place for reparametrizations of the time coordinate.

This is due, in the last instance, to the fact that one can only speak meaningfully about reparametrization of the time coordinate when one deals with histories. Now, individual histories are of course fundamental for the evaluation of the propagator, but after one has summed over all of them, there remain only their common end points as the arguments of the propagator, and it is meaningless to ask about invariance under reparametrization of the time coordinate of an end point. This is not a peculiarity of general relativity. It also makes no sense whatsoever to demand that for a point particle the propagator $K(x_2^\mu,x_1^\mu)$ should be invariant under τ-time reparametrizations $\tau\rightarrow f(\tau)$.

The analogy with the point particle is sharpened by a closer examination of the structure of the generator $\mathscr{H}$ which shows that the coefficient (4.4) of the quadratic term in (4.2) plays the role of a six-dimensional (contravariant in spite of the position of the indices) metric. It can be shown[14] that this metric has signature $(-,+,+,+,+,+)$. (Actually, our "supermetric" differs from the one used in Ref. 14 by a factor $g^{1/2}$. This improves its properties making it, for example, geodesically complete.) A typical "timelike" displacement is given by a local change of scale,

$$\delta g_{ij}(x)=\delta\lambda(x)g_{ij}(x) . \tag{4.11}$$

Indeed one has for the inverse G^{ijkl} of (4.4)

$$G^{ijkl}=\tfrac{1}{2}(g^{il}g^{kj}+g^{ij}g^{jl}-2g^{ij}g^{kl}) , \quad (4.12)$$

so that one has for the displacement of (4.11)

$$G^{ijkl}\delta g_{ij}\delta g_{kl}=-6(\delta\lambda)^2<0 . \quad (4.13)$$

We see from (4.11) and (4.13) that the local volume plays the role of the coordinate x^0 for a point particle while the conformally invariant three-geometry may then be taken as the analog of the x^i. This definition may be made more explicit by splitting the metric in the form

$$g_{ij}=g^{1/3}\tilde{g}_{ij} , \quad (4.14)$$

where the conformal metric $\tilde{g}_{ij}$ has a determinant equal to unity. Then the change (4.11) leaves $\tilde{g}_{ij}$ unchanged while changing $g^{1/3}$ by

$$\delta(g^{1/3})=g^{1/3}\delta\lambda , \quad (4.15)$$

so that

$$\delta(\ln g^{1/3})=\delta\lambda . \quad (4.16)$$

Thus a natural "timelike variable," i.e., the analog of x^0 in the particle case is

$$X^0(x)=\ln g^{1/3} . \quad (4.17)$$

This choice is confirmed by observing that the following Poisson bracket relation holds:

$$[X^0(x),\pi(x')]=1 , \quad (4.18)$$

where

$$\pi(x)=g_{ij}g_{kl}\pi^{ij}\pi^{kl} \quad (4.19)$$

plays in (4.2) the role that p^0 plays in (3.1).

From the preceding discussion one sees that an expanding universe corresponds to a particle traveling forward in time while a contracting one is the analog of a particle traveling backward in time.

One may take the analogy further and picture the dynamics of the universe as a scattering in configuration space. Therefore an initially expanding universe may be scattered "forward in time," i.e., continue its expansion, or may be scattered "backward in time," i.e., recontract. This idea is the basis for the perturbation theory considered below.

Now, in quantum electrodynamics one reinterprets physically a particle going back in time as an antiparticle, and the scattering described above corresponds to the annihilation of an electron-positron pair. In the present context that reinterpretation is not necessary and it appears to be more natural to stick to the view of a system able to travel forward and backward in time.

It is worthwhile remarking at this point that in electrodynamics the annihilation (or creation) of an electron-positron pair can occur only quantum mechanically. Here, however, the universe can be scattered backward in time even classically. This difference, which is of course of immense physical importance, is however not a fundamental one from the point of view of the formal structure of the theory. It follows from a different coupling to the external potential in both cases. One may indeed invent examples[15] where pair creation occurs also classically in particle theory.

As a final comment, we should point out that for a classical history there will in general be more than one three-dimensional section where $X^0(x)$ takes a given value. In fact, more generally, one will find a given three-geometry to repeat itself one or more times within a four-dimensional spacetime. The possibility of such a multiple occurrence has been used as an argument against building the intrinsic time from the three-geometry.[16] However, in the present formulation where trajectories are allowed to go backward as well as forward in intrinsic time, this causes no problems, just as the possible repetition of a given x^0 for a particular worldline causes no problem in relativistic particle theory.

C. Gauge fixation

The propagation amplitude $K(2,1)$ from three-geometry 1 to three-geometry 2 is to be obtained by summing the exponential of i times the action (4.1) with the three-geometry fixed at both end points. In order to do so, it is necessary to impose a gauge condition that eliminates the freedom (4.5) and (4.6). Just as for the point particle, it is not permissible to do so by a canonical condition of the form, say, $\tau=f[g_{ij},\pi^{ij}]$ since such a relation would not reduce to the identity on the boundary $\tau=\tau_1$ or τ_2, and would therefore conflict with [4.10].

The simplest permissible gauge fixation appears to be that given by

$$\dot{N}=0 , \quad (4.20a)$$

$$N^i=0 , \quad (4.20b)$$

where $\dot{N}=\partial N(x,\tau)/\partial\tau$. Equations (4.20) will be referred to as *the proper-time gauge.* They eliminate the freedom (4.5) and (4.6) except for time-independent changes of the spatial coordinates, i.e., (4.5) and (4.6) with $\epsilon=0$, $\dot{\epsilon}^i=0$. That residual invariance may be eliminated at a later stage by imposing, for example, a canonical gauge condition at, say, the initial time τ_1.

Geometrically, (4.20b) says that the line which joins two points with the same spatial coordinates lying on two neighboring surfaces is normal to the first surface. Equation (4.20a) on the other hand states that for a fixed value of the spatial coordinates, the ratio of the normal distance between two neighboring surfaces to the local volume of the first surface does not depend on the location of the first surface.

It is important to realize that (4.20a) does not restrict the dependence of N on the spatial coordinates; this is essential for the compatibility of (4.20a) and (4.10).

D. Path integral

Straightforward generalization of the procedure followed for the point particle would suggest setting $\dot{N}=0=N^i$ in the action and to write simply

$$K[g_{ij}(2),g_{ij}(1)]=\int_{N(x)=0}^{N(x)=\infty}\tilde{K}[2,1;N(x)(\tau_2-\tau_1)]\prod_x d[N(x)(\tau_2-\tau_1)] \tag{4.21}$$

for the gravitational field propagator in the metric representation, where

$$\tilde{K}[2,1;N(\tau_2-\tau_1)]=\int \exp[iS]\prod_{x,\tau}\frac{dg_{ij}(x,\tau)d\pi^{ij}(x,\tau)}{2\pi} . \tag{4.22}$$

In the integral (4.21) one would sum over all positive-definite g_{ij} which have fixed values $g_{ij}(2)$, $g_{ij}(1)$ at τ_2 and τ_1 and over all π^{ij}. [We ignore for the moment the neccessity of introducing extra gauge conditions at τ_1 in order to fully fix the spatial coordinates. It is quite all right to do so since those conditions only restrict the variables at the end points and do not interfere with the integrations in (4.21) and (4.22).]

Equations (4.21) and (4.22) are however not correct as they stand, the reason being that if one would select a different set of gauge conditions and then follow the same naive procedure of inserting them in the action thereby summing afterwards directly over the variables which remain free, one would obtain a different answer for the propagator. This problem stems from the non-Abelian nature (4.8) of the composition of surface deformations whose generators are the constraints (4.2) and (4.3) and does not arise for the point particle where the composition is commutative.

It turns out that in order to obtain an expression for the propagator which treats all permissible gauge conditions on the same footing one must use instead of the action (4.1) a modified effective action which involves two extra real anticommuting scalar "ghost" fields $C,\bar{C}$ and their conjugate moment $\bar{P},P$.

To avoid being led astray from the main line of reasoning, we will omit the procedure for obtaining the necessary modification of the action. It is an application of the general procedure given in Ref. 7, which in turn generalizes the work reported in Ref. 17. The final result is that in the gauge $\dot{N}=0$, $N^i=0$ the effective action takes the remarkably simple form

$$S_{\text{eff}}=\int_{\tau_1}^{\tau_2}(\dot{g}_{ij}\pi^{ij}+i\dot{C}\bar{P}+i\dot{\bar{C}}P-N\mathscr{H}^{\text{eff}})d^3x\,dt , \tag{4.23}$$

where

$$\mathscr{H}^{\text{eff}}=\mathscr{H}+\mathscr{H}^{\text{ghost}}$$

with

$$\mathscr{H}^{\text{ghost}}=i[(2\kappa)\bar{P}P+(2\kappa)^{-1}\sigma g g^{ij}\bar{C}_{,i}C_{,j}] . \tag{4.24}$$

Equations (4.23) and (4.24) show that the necessary introduction of the ghost scalar fermions does not alter the essential features of the action. It merely results in the addition to the Hamiltonian of a piece describing a minimal coupling [g_{ij} enters (4.24) undifferentiated] of the massless ghost fields to the gravitational fields. Furthermore the extra piece has the same form of the original Hamiltonian; namely it consists of the sum of two contributions: a local term quadratic in the momenta and a "potential" where spatial derivatives appear. The potential part drops out in the limit of zero signature $\sigma \to 0$, a feature that will be essential for the feasibility of the perturbation theory discussed in the next section. Other remarkable properties of the ghost Hamiltonian, such as the fact that its addition preserves the surface deformation algebra will not be discussed here.

In terms of the effective action the correct version of (4.22) reads

$$\tilde{K}[2,1;N(x)(\tau_2-\tau_1)]=\int \exp[iS_{\text{eff}}]\prod_{x,\tau}\frac{dg_{ij}d\pi^{ij}dC\,d\bar{P}\,d\bar{C}\,dP}{2\pi}\,, \tag{4.25}$$

where

$$C(\tau_1)=\bar{C}(\tau_1)=0\,,\qquad C(\tau_2)=\bar{C}(\tau_2)=0\,, \tag{4.26}$$

and, in the metric representation,

$$g_{ij}(\tau_1)=g_{ij}(1)\,,\qquad g_{ij}(\tau_2)=g_{ij}(2)\,. \tag{4.27}$$

The gravitational momenta and the ghost momenta are summed over without restrictions at the end points.

On account of conditions (4.26) the ghost fields do not affect the classical equations of motion, since extremization of S_{eff} under (4.26) makes the ghosts vanish for all times and hence gives back the classical Einstein equations. However, for nonextremal trajectories the contribution of the host fields is nonzero and corresponds to introducing a nontrivial integration measure for g_{ij},π^{ij} in (4.22).

Writing the correct version (4.25) of (4.22) in terms of the ghost fields requires also a slight modification of (4.21). In fact the propagator is obtained by integration over N with a logarithmic measure:

$$K[g_{ij}(2),g_{ij}(1)]=\int_{N(x)=0}^{N(x)=\infty}\tilde{K}[2,1,N(\tau_2-\tau_1)]\prod_x d[\ln N(x)(\tau_2-\tau_1)]\,. \tag{4.28}$$

In the Abelian limit ($\sigma=0$) one may carry on explicitly the integration over the fermionic variables. That operation results in a factor $N(\tau_2-\tau_1)$ for each space point which cancels the N^{-1} in (4.28) and restores the trivial measure $dN(x)(\tau_2-\tau_1)$, just as in the particle case.

Equations (4.25) and (4.28) are the central result of this section. They provide an (formal) expression for the Feynman propagator of the gravitational field as a whole which does not rely on splitting off a background or a semiclassical approximation. The condition which selects the Feynman propagator is the restriction to integration over positive values of $N(x)$, or what is equivalent, the proper time retardation condition:

$$\tilde{K}[2,1;N(\tau_2-\tau_1)]=0\quad \text{for } N(\tau_2-\tau_1)<0\,. \tag{4.29}$$

In order to ensure convergence of the proper-time integral (4.28) at the upper end, one must also incorporate the (unique) prescription

$$\mathscr{H}^{\text{eff}}(x)\to\mathscr{H}^{\text{eff}}(x)-i\eta(x)\,, \tag{4.30}$$

where $\eta(x)$ is an infinitesimal positive-definite scalar density of weight two. If we write $\eta(x)=\eta g$ with η constant, (4.24) can be described as resulting from the addition of a small negative imaginary part to the cosmological constant.

C. The signature as a perturbation parameter

As discussed in the introduction, the need for a perturbation scheme appears to be not only a practical necessity but also a part of the actual definition of a quantum field theory.

A particular kind of perturbation theory suggests itself within the approach treated in this article. It corresponds to treating the signature σ as a perturbation parameter.

According to this idea one writes

$$\mathscr{H}^{\text{eff}}=\mathscr{H}_0+\mathscr{U} \tag{4.31}$$

with

$$\mathscr{H}_0=(2\kappa)(G_{ijkl}\pi^{ij}\pi^{kl}+i\bar{P}P)+(2\kappa)^{-1}g\Lambda\,, \tag{4.32}$$

and

$$\mathscr{U}=\sigma g(2\kappa)^{-1}(-R+ig^{ij}\bar{C}_{,i}C_{,j})\,. \tag{4.33}$$

Here $\mathscr{H}_0$ plays a role analogous to that of (3.10) whereas $\mathscr{U}$ appears as an "external potential" in configuration space. The splitting is mathematically remarkable in that $\mathscr{H}_0$ obeys a closed and, moreover, Abelian [Eq. (4.9) with $\sigma=0$] set of commutation rules and on these grounds is quite unique. In fact, dropping $\mathscr{H}$ from (4.31), a possibility first contemplated in Ref. 18, amounts to setting $\sigma=0$ which makes the right side of (4.9) vanish identically.

The field histories of the $\sigma=0$ theory possess a geometrical structure which lies halfway between Euclidean and Lorentzian curved spacetimes. Since its Hamiltonian generators obey a closed algebra, the theory is generally covariant (i.e., it involves four arbitrary functions of x and τ in the solution of the equations of motion). A Lagrangian formulation in which this four-dimensional invariance is manifest may be written in terms of a metric tensor density with zero determinant and a conformal factor.[19]

One may also view the $\sigma=0$ theory as a limiting case of general relativity in which the local light cones become narrower and narrower and eventually collapse to a line. From a group-theoretical point of view this limit corresponds to replacing the Lorentz group as a symmetry group of the tangent spaces by one of its contractions. That contraction is the opposite of the Galilean ("$c\rightarrow\infty$"), that is, it corresponds to letting $c\rightarrow 0$ and has been called[20] the "Carroll group". (' "A slow sort of country," said the Queen, "Now, here, you see, it takes all the running you can do to stay in the same place" '.[21])

Physically this state of affairs would not seem too unreasonable. In fact, it says that in the absence of interaction the light cones are closed and therefore there is no propagation, which implies that the different points of space are mutually disconnected. (This may also be seen by noticing that setting $\sigma=0$ eliminates spatial derivatives from $\mathscr{H}^{\text{eff}}$.) When the interaction is switched on ($\sigma\neq 0$), the light cones begin to open and physics as we usually understand it starts to take place. Thus, in this view propagation and interaction are two concepts inextricably linked to each other.

Now, the decoupling of different space points is precisely what happens in the early stages of cosmological evolution. In fact, one may show that already at the classical level the $\sigma=0$ Hamiltonian describes approximately the behavior of the universe near the cosmological singularity.[22] Hence the region of configuration space where three-space collapses plays the role of an asymptotic domain where the free theory ($\sigma=0$) takes over. This conclusion fits particularly well with the identification of X^0 given by (4.17) as a time variable, since that function approaches negative infinity near the singularity $g=0$.

The fact that the cosmological singularity corresponds to the "in" region in configuration space means that what comes out of the singularity must be prescribed as an initial condition and cannot be predicted by the theory. This state of affairs is in agreement with what one finds in the classical theory where a singularity is also something unpredictable. What is different, however, in the quantum case, is that the equations appear to be quite smooth ("free") in that limit. The role of the singularity as an "in" region for scattering has been previously discussed in the context of simplified cosmological models.[23]

Statements analogous to the previous ones for large volumes ($X^0\rightarrow+\infty$) are much less clear even at the qualitative level, and it is not known whether a simple Hamiltonian may reproduce the actual behavior of the theory in that limit. It may be necessary to introduce an *ad hoc* cutoff for large volumes so that the theory will behave freely in that region also (i.e., as if σ were zero). This could be a useful procedure even if the actual situation turns out to be that the "potential" becomes totally reflective in that limit, i.e., if all trajectories bent backward "in time." In that case as the cutoff went to infinity the barrier would become more and more reflective. This note is intended only as speculative at this point, since not enough work has been done on how to introduce such a cutoff into the theory.

F. Perturbation theory

The perturbation theory in the signature σ may be simply implemented in terms of the path integral for the propagator given by (4.28).

The analysis follows closely that of Sec. III F for a system of point particles. Here we take as the free, unperturbed propagator the one with signature zero, namely that obtained by setting $\sigma=0$ in (4.21)–(4.30), and write

$$\exp[iS_{\text{eff}}(\sigma)]=\exp[iS_{\text{eff}}(\sigma=0)]\exp\left[-i\int_{\tau_1}^{\tau_2}N(x)\mathscr{U}(x)d^3x\,d\tau\right], \tag{4.34}$$

with $\mathscr{U}$ given by (4.33).

In order to obtain $\tilde{K}[2,1;N(\tau_2-\tau_1)]$ we have to integrate (4.34) over coordinates and momenta, as indicated in (4.25). Now, the dependence on the momenta is all included in the first exponential factor in (4.34), and after the integrations over $\pi^{ij},\bar{P},P$ are performed, the problem takes the standard form discussed, for example, in Ref. 8. There one deals with a free propagator whose action is quadratic in the velocities perturbed by a potential $V(q)$ which is given in our case by

$$V[q;N]=\int N(x)\mathscr{U}(x)d^3x \ , \tag{4.35}$$

where q is shorthand for g_{ij}, C, and $\bar{C}$ at all space points x. At this stage $N(x)$ is regarded as a fixed parameter in the problem.

Having made this observation we can write immediately the perturbation expansion of (4.25). If we call $\tilde{K}_0[2,1;N(\tau_2-\tau_1)]$ the result of the integrations in (4.25) with $\sigma=0$, we have an expansion in powers of σ of the form

$$\tilde{K}=\tilde{K}_0+\tilde{K}_1+\tilde{K}_2+\cdots \ , \tag{4.36}$$

with the standard expressions

$$\tilde{K}_1[3,1;N,\tau_3,\tau_1]=-i\int_{\tau_1}^{\tau_3}d\tau_2\int Dq_2\tilde{K}_0[3,2;N(\tau_3-\tau_2)]V(q_2)\tilde{K}_0[2,1;N(\tau_2-\tau_1)] \ , \tag{4.37}$$

$$\begin{aligned}\tilde{K}_2[4,1;N,\tau_4,\tau_1]=(-i)^2\int_{\tau_1}^{\tau_4}d\tau_3\int_{\tau_1}^{\tau_4}d\tau_2\int Dq_3Dq_2\tilde{K}_0[4,3,N(\tau_4-\tau_3)]V(q_3)\\ \times\tilde{K}_0[3,2;N(\tau_3-\tau_2)]\\ \times V(q_2)\tilde{K}_0[2,1;N(\tau_2-\tau_1)] \ ,\end{aligned} \tag{4.38}$$

etc.

It is important to emphasize that although the ghost fields vanish at the initial and final states they must be included when summing over intermediate states in (4.37) and (4.38). Therefore, in (4.37) we set $C(1)$, $\bar{C}(1)$, $C(3)$, and $\bar{C}(3)$ equal to zero but integrate over $C(2)$ and $\bar{C}(2)$. Thus the volume element Dq_2 is shorthand for

$$Dq_2=\prod_{\substack{\text{all space}\\ \text{points } x\\ i\le j}}\frac{dg_{ij}(x,2)d\bar{C}(x,2)dC(x,2)}{2\pi} \ . \tag{4.39}$$

Similarly, for (4.38) and the higher-order terms. Finally, the physical propagator is obtained to each order by integrating separately the various terms in the expansion (4.37), (4.38), etc., over $N(x)$ with the logarithmic measure (4.28). This completes the formal discussion of the perturbation expansion.

The picture which emerges from the perturbation theory is the following. In the absence of interaction ($\sigma=0$) the light cone is closed and different space points evolve uncoupled. Accordingly, the free propagator factorizes as

$$\tilde{K}_0[2,1;N(\tau_2-\tau_1)]=\prod_x\tilde{K}_0(x)[q_2(x),q_1(x);N(x)(\tau_2-\tau_1)] \ . \tag{4.40}$$

Furthermore, at any given point the ghosts fields and the metric are decoupled when $\sigma=0$. Therefore, at each point we have

$$\tilde{K}_0(x)[q_2(x),q_1(x);N(x)(\tau_2-\tau_1)]=\tilde{K}_0^{\text{ghost}}\tilde{K}_0^{\text{metric}} \ . \tag{4.41}$$

Here it is a simple matter to write $\tilde{K}_0^{\text{ghost}}$ in closed form. A short calculation yields

$$\begin{aligned}\tilde{K}_0^{\text{ghost}}[C_2,\bar{C}_2,C_1,\bar{C}_1;N(\tau_2-\tau_1)]=\left[\prod_x N(x)(\tau_2-\tau_1)\right]\\ \times\exp\left[\int[N(\tau_2-\tau_1)]^{-1}(\bar{C}_2-\bar{C}_1)(C_2-C_1)d^3x\right] .\end{aligned} \tag{4.42}$$

However, a closed expression for K_0^{metric} is not known at the moment of this writing. *(Note added in proof.* Such an expression has been found in Ref. 24.)

When the interaction is switched on ($\sigma \neq 0$), the light cone begins to open and neighboring space points couple through the potential $\mathscr{U}(x)$ which contains first derivatives of the ghosts and up to second derivatives of the metric.

Now we may imagine inserting expression (4.35) for $V(q)$ into the first-order perturbation (4.37). That amplitude becomes then an integral over all space which is analogous to the sum over all particles (Eq. (3.41) of Sec. 3 F. This means that the first-order transition amplitude is in itself a sum of infinitely many mutually exclusive alternatives. Each of those alternatives corresponds to selecting a point x and letting the points in its neighborhood be scattered by each other under the influence of the interaction potential $\mathscr{U}(x)$, while the points outside that neighborhood proceed freely, i.e., as if the local light cone were closed.

Similarly, the second-order transition amplitude which becomes a double integral over all space points upon insertion of (4.35) into (4.38) may be thought of as a sum of mutually exclusive alternatives. Each alternative corresponds to coupling among themselves the points within two different neighborhoods through the potential $\mathscr{U}$, while letting all other pairs of neighborhoods evolve freely.

F. External lines and loops

Although the propagation amplitude $K(2,1)$ is the basic element of the quantum theory, it does not have directly the meaning of a probability amplitude. In order to obtain probability amplitudes it is necessary to fold the propagator onto wave functionals which represent states.

Here we shall give a short account of how this is done. The basic strategy is again to deal with the intermediate amplitude $\widetilde{K}$ given by (4.25) and perform the proper-time integration [i.e., that over $d\ln N(x)$] at the very end.

Thus, one represents the state at the initial time τ by a functional $\Psi(q)$ which is then propagated to the final τ by means of $\widetilde{K}$. The resulting expression is then folded onto a final state functional $\Phi(q)$, which yields an expression which still depends upon $N(x)$ and the initial and final τ's. One then integrates over $N(x)$ and the dependence on the end-point τ's drops out automatically.

If this process is followed for the relativistic particle of Sec. III one recovers[25] directly from the proper-time approach the standard formulas of positron theory.[26]

In the case of positron theory, there are two physically important choices for the initial and final states Ψ,Φ. They correspond either to the situation in which the particle is free for large times or to that for which the particle starts at a spacetime point which is not at infinity and ends at the same point.

In the first situation the states $\Psi[x^\mu]$ and $\Phi[x^\mu]$ are solutions of the free wave equation $\mathscr{H}_0\Psi=0$, that is, of the Klein-Gordon equation. Such a state is represented by an external line in a Feynman diagram. In the second situation both states are taken to be of the form $\psi(x^\mu)=\delta(x-y)$. In a Feynman diagram such states are represented simply by a point (at y) and the whole process is pictured as a loop that starts and ends at y.

Both situations have direct analogs in the gravitational case which we discuss cursorily now.

G. Real universes (external lines)

By analogy with the particle case, we associate with a universe which emerges from the cosmological singularity a solution of the $\sigma=0$ equations,

$$\mathscr{H}_0^{\text{grav}}\Psi[g_{ij}]=0 \ . \tag{4.43a}$$

Here $\mathscr{H}_0^{\text{grav}}$ is the expression obtained by dropping the ghost part of the generator (4.32). The reason for omitting the ghost contributions in the operator appearing in (4.43a) is that the ghosts must be equal to zero at the end points in the path integral, so the dependence on those fields in the state is of the form $\delta[\overline{C}]\delta[C]=\overline{C}C$, rather than through a solution of $i\overline{P}P\psi=0$.

We will not be concerned here with the problem of defining properly the operator $\mathscr{H}_0$ in (4.43a). The factor ordering in that operator is nontrivial because the metric G_{ijkl} given by (4.12) is not flat. As a consequence the measure in (4.39) below may acquire some g_{ij} factors when defined properly. (See Ref. 24 for more on this.) This issue is also intimately connected with the need for integrating over positive-definite geometries only. A similar problem occurs if one uses spherical coordinates instead of Cartesian ones when describing the quantum mechanics of a point particle. However, the general features discussed in this report do not depend on this ordering problem.

To (4.45) one must add the condition

$$\mathscr{H}_i\Psi(g_{ij})=0 \tag{4.43b}$$

with $\mathscr{H}_i$ given by (4.3). Equation (4.43b) is the statement that the functional Ψ depends only on the three-geometry and not on the particular system of coordinates chosen to write the metric tensor g_{ij}. This equation must be imposed because it is the three-geometry rather than g_{ij} which is fixed at the end points in the path integral.

In order to derive a perturbation expression for the transition element from a state Ψ to a state Φ, both obeying (4.43a) and (4.43b) we first multiply (4.37) from the right by a $\Psi[g_{ij}(1)]$ and integrate over $dg_{ij}(1)$ at all points of space. This yields the integral

$$\int \widetilde{K}_0[2,1;N(\tau_2-\tau_1)]\Psi[g_{ij}(1)]\prod_{x,ij} dg_{ij}(x,1)\ , \tag{4.44}$$

where it is understood that $C_1=\bar{C}_1=0$. Next we recall that when $\sigma=0$ the metric and the ghosts decouple so that the propagator factorizes. Thus, if we imagine inserting (4.41) into (4.44), the ghost part of the propagator comes out of the integral whereas on account of (4.43b) the effect of K_0^{metric} upon integration in $g_{ij}(1)$ is simply to reproduce Ψ with argument $g_{ij}(2)$. Therefore, the expression (4.44) reduces to

$$\Psi(2)=\Psi[g_{ij}(2)]\widetilde{K}_0^{\text{ghost}}[C_2,\bar{C}_2,0,0;N(\tau_2-\tau_1)]\ , \tag{4.45}$$

with $\widetilde{K}_0^{\text{ghost}}$ given by (4.42).

The presence of the fermionic factor in (4.45) is a reflection of the fact that ghosts appear in intermediate times (τ_2 in this case) even though they are absent at the initial and final times. It is in terms of the functionals (4.45) that the transition elements are most simply written and we shall adopt them, rather than $\Psi[g_{ij}]$ as the basic description of an "external-line state."

To complete the transition element we multiply (4.44) from the left by $\Phi^*[g_{ij}(3)]$ and integrate over $g_{ij}(3)$. The same analysis leading to (4.45) shows then that the first-order effect on the transition amplitude to go from Ψ to Φ is obtained by integrating the familiar-looking expression

$$(-i)\int_{\tau_1}^{\tau_3}\Phi^*(2)V(2)\Psi(2)d\tau_2 \tag{4.46}$$

over q_2 (which includes the metric and the ghosts) and also over all positive values of the lapse $N(x)$ according to (4.37) and (4.28).

The integration of (4.46) over $g_{ij}(x)$, $C(x)$, $\bar{C}(x)$, and $N(x)$ cannot however be performed with that expression as it stands. In fact (4.46) when regarded as a functional of these variables is invariant under reparametrizations of the spatial coordinates x provided g_{ij} is changed as a covariant tensor, C and $\bar{C}$ as scalars, and N as a density of weight minus one. Indeed, the ghost part of (4.46) may be seen to be invariant by inspection [the infinite product in (4.42) may be included in the measure over N without destroying its formal reparametrization invariance], whereas the metric part is invariant by assumption (Eq. 4.42). The potential V given by (4.35) is also evidently reparametrization invariant.

Since (4.46) is reparametrization invariant, one must integrate over classes rather than over individual field configurations much in the same way as was done for the propagator in the first place, with the difference that one is concerned here with changes of the spatial coordinates at a fixed time only.

In order to integrate over classes one must impose a set of coordinate conditions,

$$\chi_i(x)[g_{ij},C,\bar{C},N]=0\ , \tag{4.47}$$

and include in (4.46) a factor[17]

$$\delta[\chi]\Delta[\chi]\ . \tag{4.48}$$

Here $\delta[\chi]$ represents a product of Dirac δ functions with argument (4.47) over all points x and indices i, and $\Delta[\chi]$ denotes the functional determinant of the matrix $\delta\chi_i(x)/\delta\xi^k(x')$, which yields the change

$$\delta\chi_i(x)=\int\frac{\delta\chi_i(x)}{\delta\xi^k(x')}\delta\xi^k(x') \tag{4.49}$$

induced by an infinitesimal reparametrization by the vector field $\delta\xi^k$ on the coordinate conditions (4.47).

The need for imposing (4.47) is a consequence of the fact that the gauge is not fully fixed by the conditions $\dot{N}=0$, $N^i=0$. As was previously explained these conditions still leave the freedom of performing time-independent changes of the spatial coordinates. That freedom may be eliminated by imposing coordinate conditions at any fixed time which is precisely what is accomplished by (4.48).

Thus we may finally write the first-order effect on the transition amplitude as

$$-i\int\int_{\tau_1}^{\tau_3}\Phi^*(2)V(2)\Psi(2)\delta[\chi(2)]\Delta[\chi(2)]d\tau_2 Dq(2)D[\ln N]\ , \tag{4.50}$$

where

$$D[\ln N]=\prod_{x} d[\ln N] \, . \tag{4.51}$$

Similarly we obtain for the second-order contribution

$$(-i)^2 \int \Phi^*(3)V(3)\widetilde{K}_0[3,2]V(2)\Psi(2)d\tau_3 Dq(3)d\tau_2 Dq(2)\delta[\chi(2)]\Delta[\chi(2)]D[\ln N] \, . \tag{4.52}$$

The pattern for higher-order terms is clear. It should be noted that only one coordinate fixation factor $\delta[\chi]\Delta[\chi]$ is needed for every order in σ. The reason is that the integrand is invariant under only one reparametrization, common to all fields involved. Thus if the integral involves two separate fields $g_{ij}(2)$ and $g_{ij}(1)$, say, one can integrate directly in $g_{ij}(2)$ provided $g_{ij}(1)$ is held fixed. The very fact of keeping $g_{ij}(1)$ fixed breaks the gauge invariance at that level. However, when performing the last integration there will be no "external" field left and it is necessary to fix the gauge by means of χ. Another way of seeing this is noticing that the different vertices in the perturbation expansion correspond to scatterings taking place at different times, and one is allowed to impose the coordinate conditions (4.47) at only one (arbitrarily chosen) time. That time was conventionally chosen in (4.52) to be τ_2 but it could have been taken equally well to be τ_3.

We shall not be concerned in this paper, which deals only with the general formalism, with specific choices for the gauge conditions χ_i. There is however one relevant comment we wish to make. After the factors $\delta[\chi]\Delta[\chi]$ are included in the measure in (4.50), (4.52), etc., one is effectively integrating over all possible three-dimensional geometries (we include in this terminology the ghosts $C,\bar{C}$). Thus one would expect that if a procedure which avoided the introduction of spatial coordinates altogether could be devised (such as the use of Regge calculus on three-space) those equations would remain essentially valid, but with the complicated measure $\delta[\chi]\Delta[\chi]Dq$ replaced by an expression depending only on the invariant quantities involved in such a coordinate-free description. We hope to return to this question in the future, since it would appear that only in that way [(i.e., by dropping the spatial coordinates in (4.50) and (4.52)] can one attempt to deal satisfactorily with changes of topology and other related issues. For this reason we cannot help but feel that (4.50), (4.52), etc., with their need for (4.47), should be taken only as a provisional representation for the transition elements. These same comments apply to the loop amplitudes discussed below.

H. Virtual universes (loops)

The previous section dealt with states describing a universe emerging from a cosmological singularity [solutions of the $\sigma=0$ equations (4.43a)] which then proceed to be scattered by the potential (4.35) as the light cone begins to open ($\sigma\neq 0$). Those states will eventually go back to the cosmological singularity or proceed indefinitely to larger and larger volumes.

There are, however, other processes which are in principle conceivable in the present framework. Those processes are the analog of closed loops in positron theory. Thus one may conceive a three-dimensional space starting from a state of perfectly regular g_{ij} (and zero ghost field) being scattered by the potential V and coming back to the original regular geometry. The amplitude for those processes is obtained just by replacing the functional $\Psi[g_{ij}(1)]$, satisfying (4.43a) used in (4.44) by another of the form $\delta[g_{ij}(1)-g_{ij}]$. Similarly, $\Phi^*[g_{ij}(3)]$ is to be replaced by $\delta[g_{ij}(3)-g_{ij}]$. Therefore, to a first approximation the effect of the potential V on the amplitude for the universe to describe a loop starting and ending at g_{ij} is, from (4.38),

$$(-i)^2 \int_{\tau_1}^{\tau_4} \widetilde{K}_0[4,3]V(3)\widetilde{K}_0[3,2]V(2)\widetilde{K}_0[2,1]Dq(2)d\tau_2 Dq(3)d\tau_3 D[\ln N] \, , \tag{4.53}$$

where

$$g_{ij}(4)=g_{ij}(1)=g_{ij} \, , \tag{4.54}$$

$$C(4)=\bar{C}(4)=C(1)=\bar{C}(1)=0 \, . \tag{4.55}$$

Now, in order to find the total amplitude for all possible values of g_{ij} we have to integrate (4.53) over all possible values of the initial configuration g_{ij} and divide by two since the loop may be taken as starting either at $g_{ij}(4)$ or $g_{ij}(1)$. In order to perform this final integration we must introduce a coordinate fixation factor, so the final integration includes a measure

$$\delta[\chi(g_{ij})]\Delta[\chi(g_{ij})]\prod_{\substack{x\\ i\leq j}} dg_{ij}(x)\ . \tag{4.56}$$

As for other cases, after performing all the integrations τ_1 and τ_4 should drop out from the problem so that the answer is expressed only in terms of the Planck length and any cutoff which may eventually be introduced.

The processes in which the universe describes a closed loop offer interesting grounds for speculation and are among the most unconventional features which appear in the present approach to quantum gravity. They are in the last instance a consequence of the quadratic nature of the Hamiltonian constraint in all momenta and cannot therefore even be conceived of in ordinary field theory.

One may, however, take the view that only "real processes" are of interest in nature, i.e., those which correspond to states emerging from the cosmological singularity. That view may consistently be maintained only if there are no interactions between the loops and the real processes or if those interactions can consistently be ignored.

Now, the kind of interaction that we are talking about could be described by means of a diagram where a universe splits into two and one of the branches merges subsequently with a loop. Three possibilities are then open: (1) the splitting of a universe into two is not at all possible; (2) the splitting is possible but requires the introduction of a new propagator with one or more new coupling constants for the vertex in the splitting; (3) the splitting is possible and the vertex is completely determined by the structure of the theory itself.

Of these three possibilities the presumption is strong that (1) is not the correct one; in fact, there seems to be not only the place, but also need for splittings to occur— for example, to account for the final state in the black-hole evaporation process.[27] It is however not clear at the moment of writing whether one should believe in (2) or (3).

Clearly (3) is theoretically more attractive, but on the other hand analogy with the decay of point particles would suggest that the decay of three-dimensional space as a whole may occur through more than one fundamental process.

In any case the propagation studied in this paper would only be valid for the propagation between splittings (i.e., it would be the analog of the free propagator in particle theory), so that if splittings are compulsory one could not attempt to build a unitary theory based on it. It appears therefore more urgent to obtain more insight into the splitting question than to attempt discussing unitarity. We hope to return to this issue in the future.

ACKNOWLEDGMENTS

I would like to thank many colleagues for helpful discussions, especially S. Adler, P. Cordero, R. Dashen, F. Dyson, M. Henneaux, M. Pilati, T. Regge, and J. Wheeler. This work was supported in part by the U.S. National Science Foundation grants to the University of Texas and the Institute for Advanced Study. The author is a recipient of an Alfred P. Sloan Fellowship.

---

[1]See, for example, J. A. Wheeler, in *Analytical Methods in Mathematical Physics,* edited by R. Gilbert and R. Newton (Gordon and Breach, New York, 1970).
[2]R. P. Feynman, Rev. Mod. Phys. 20, 267 (1948).
[3]C. Teitelboim, Phys. Lett. 96B, 77 (1980).
[4]See, for example, L. D. Faddeev and A. A. Slavnov, *Gauge Fields, Introduction to Quantum Theory* (Benjamin/Cummings, Reading, Mass., 1980).
[5]P. A. M. Dirac, *Lectures in Quantum Mechanics* (Academic, New York, 1964); see also A. Hanson, T. Regge, and C. Teitelboim, *Constrained Hamiltonian Systems* (Accademia Nazionale dei Lincei, Rome, 1976).
[6]C. Teitelboim, Ph.D. thesis, Princeton University, 1973 (unpublished); Phys. Rev. Lett. 38, 1106 (1977).
[7]E. S. Fradkin and G. A. Vilkovisky, CERN Report No. TH-2332, 1977 (unpublished). See also I. A. Batalin and G. A. Vilkovisky, Phys. Lett. 69B, 310 (1977).
[8]R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).
[9]T. Regge and C. Teitelboim, Ann. Phys. (N.Y.) 88, 268 (1974).
[10]P. A. M. Dirac, Proc. R. Soc. London A246, 333 (1958); R. Arnowitt, S. Deser, and C. W. Misner, in *Gravitation, an Introduction to Current Research,* edited by L. Witten (Wiley, New York, 1962).
[11]C. Teitelboim, in *General Relativity and Gravitation* :

One Hundred Years After the Birth of Albert Einstein, edited by A. Held (Plenum, New York, 1980).

[12]C. Teitelboim, in *Proceedings of the Seventh International Colloquium and Integrative Conference on Group Theory and Mathematical Physics,* edited by W. Beiglbock, A. Bohm, and E. Takasugi (Springer, Heidelberg, 1979).

[13]C. Teitelboim, Ann. Phys. (N.Y.) 79, 592 (1973).

[14]B. S. DeWitt, Phys. Rev. 162, 1195 (1967).

[15]E. C. G. Stückelberg, Helv. Phys. Acta 15, 23 (1942); R. P. Feynman, Phys. Rev. 79, 939 (1948).

[16]K. Kuchar, J. Math. Phys. 11, 3322 (1970).

[17]L. D. Faddeev, Teor. Mat. Fiz. 1, 3 (1969), [Theor. Math. Phys. 1, 1 (1969)].

[18]C. J. Isham, Proc. R. Soc. London A351, 209 (1976).

[19]M. Henneaux, Bull. Soc. Math. Belg. 31, 47 (1979).

[20]J. M. Levy Leblond, Ann. Inst. Henri Poincaré A3, 1 (1965).

[21]L. Carroll, *Through the Looking Glass, and What Alice Found There* (MacMillan, London, 1871).

[22]See, for example, L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields,* 4th ed. (Pergamon, New York, 1975).

[23]C. W. Misner, in *Magic Without Magic: J. A. Wheeler,* edited by J. R. Klauder (Freeman, San Francisco, 1972).

[24]M. Henneaux, M. Pilati, and C. Teitelboim, Phys. Lett. 110B, 123 (1982).

[25]M. Henneaux and C. Teitelboim, Ann. Phys. (N.Y.) (to be published).

[26]R. P. Feynman, Phys. Rev. 76, 749 (1949).

[27]S. W. Hawking, Nature 248, 30 (1974).