How far are we from the quantum theory of gravity?

IOP PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **72** (2009) 126002 (42pp) doi:10.1088/0034-4885/72/12/126002

How far are we from the quantum theory of gravity?

R P Woodard

Department of Physics, University of Florida, Gainesville, FL 32611, USA

E-mail: woodard@phys.ufl.edu

Received 16 July 2009, in final form 30 September 2009
Published 23 November 2009
Online at stacks.iop.org/RoPP/72/126002

Abstract
I give a pedagogical explanation of what it is about quantization that makes general relativity go from being a nearly perfect classical theory to a very problematic quantum one. I also explain why some quantization of gravity is unavoidable, why quantum field theories have divergences, why the divergences of quantum general relativity are worse than those of the other forces, what physicists think this means and what they might do with a consistent theory of quantum gravity if they had one. Finally, I discuss the quantum gravitational data that have recently become available from cosmology.

This article was invited by Beverly K Berger.

Contents

1. Introduction

Gravity was the first of the fundamental forces to be recognized and it will be the last to be understood. Its early recognition came because it is both long range and universal; gravity acts over macroscopic distances, and no one has ever found a way to screen it. Mankind's problems comprehending quantum gravity are the subject of this paper, but they derive from two basic facts:

1. Humans are not good at guessing fundamental principles without guidance from nature and
2. Gravity is such a weak interaction that nature does not provide much guidance in the quantum regime of microscopic sources.

We know there is something wrong with perturbative quantum general relativity because one cannot consistently absorb the divergences it produces without introducing new degrees of freedom that would make the universe blow up [1–7]. Obtaining that result was the work of great physicists over many decades, but their achievement will remain incomplete until we can identify the problem and fix it.

I wish I could present a solution to quantum gravity or at least a program that could reasonably be expected to provide a solution. Because that is not possible I shall endeavor to

instead give a clear explanation of the problem. This entails answering a number of questions:

- What is the distinction between classical physics and quantum physics that makes general relativity give such a wonderful classical theory of gravity and such a problematic quantum one?
- Why do we have to quantize gravity?
- Why do quantum field theories have divergences?
- Why are the divergences of quantum general relativity worse than those of the other forces?
- How bad is the problem?
- What are the main schools of thought about quantizing gravity and why do they disagree?
- What would we do with the theory of quantum gravity if we had it?

I will also comment on the quantum gravitational data that have recently become available.

The tale I have to tell is of necessity a complex one, requiring many digressive explanations. However, there is no need for the exposition to transcend the knowledge one expects of any physics graduate student. Because every one of the basic issues behind the problem of quantum gravity has a counterpart in either electrodynamics or introductory quantum mechanics, I shall use those subjects as paradigms. This is not condescension; even experts can benefit from occasionally viewing a tough problem in a general way, without becoming lost in technical details.

I shall work in four spacetime dimensions, with coordinate points labeled thus, $x^\mu = (ct, \vec{x})$. A symbol with an arrow over it denotes a 3-vector, and I employ the usual notations for scalar and vector products. Differentiation with respect to time is denoted with a dot; the gradient operator is $\vec{\nabla}$. To be clear, and for future reference, Maxwell's equations in MKS units are

$$\epsilon_0 \vec{\nabla} \cdot \vec{E} = \rho, \qquad \vec{\nabla} \cdot \vec{B} = 0, \tag{1}$$

$$\frac{1}{\mu_0} \vec{\nabla} \times \vec{B} - \epsilon_0 \dot{\vec{E}} = \vec{J}, \qquad \vec{\nabla} \times \vec{E} + \dot{\vec{B}} = 0. \tag{2}$$

Here the various fields are electric, $\vec{E}(t, \vec{x})$; magnetic, $\vec{B}(t, \vec{x})$; the charge density, $\rho(t, \vec{x})$ and the current density, $\vec{J}(t, \vec{x})$. The two constants are the electric permittivity of free space, ϵ_0, and the magnetic permeability of free space, μ_0. I denote spatial Fourier transforms with a tilde

$$\tilde{f}(t, \vec{k}) = \int \mathrm{d}^3x \, \mathrm{e}^{-\mathrm{i}\vec{k}\cdot\vec{x}} f(t, \vec{x}) \Longleftrightarrow f(t, \vec{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3} \mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{x}} \tilde{f}(t, \vec{k}). \tag{3}$$

Finally, I use the term 'classical' to mean 'not quantum', irrespective of relativity. So it is perfectly valid to speak of 'classical general relativity'. The adjective for suspending relativity is 'nonrelativistic'.

2. Perturbative quantum general relativity

The central problem of quantum gravity is that the computational techniques used with great success for the other forces do not give consistent results when applied to quantum general relativity. To understand the problem I will have to explain a little bit about general relativity, what it means to quantize a theory, and the only technique we so far have for computing things in quantum field theory. Then I describe renormalization from the perspective of polarization in electrodynamics, and I explain why the only way of consistently renormalizing perturbative quantum general relativity would make the universe virulently unstable. The section closes with a brief discussion of fixes that do not work.

2.1. General relativity

One defines a physical theory by specifying three things:

- The dynamical variable;
- How the dynamical variable affects the rest of physics and
- How the rest of physics affects the dynamical variable.

For example, the fundamental dynamical variables of electromagnetism are the scalar potential $\Phi(t, \vec{x})$ and the vector potential $\vec{A}(t, \vec{x})$, whose derivatives give the electric and magnetic fields,

$$\vec{E} = -\vec{\nabla}\Phi - \dot{\vec{A}} \qquad \text{and} \qquad \vec{B} = \vec{\nabla} \times \vec{A}. \tag{4}$$

They affect a particle of charge q which has position $\vec{x}$ and velocity $\vec{v}$ through the Lorentz force,

$$\vec{F} = q\vec{E}(t, \vec{x}) + q\vec{v} \times \vec{B}(t, \vec{x}). \tag{5}$$

And the various charges and currents affect them through the Maxwell equations (1) and (2), which have the general structure,

$$\partial^2(\Phi, \vec{A}) = (\rho, \vec{J}). \tag{6}$$

My notation is that $\partial^2(\Phi, \vec{A})$ stands for the derivative with respect to any two coordinates of any combination of Φ and $\vec{A}$. For the purposes of this paper no greater specificity is required than to note:

1. The electrodynamic field equations are linear in the potentials;
2. The electrodynamic field equations involve second derivatives of the potentials and
3. The electrodynamic field equations are sourced by the charge density $\rho(t, \vec{x})$ and the current density $\vec{J}(t, \vec{x})$.

In these terms, the dynamical variable of general relativity is known as the *metric field*, $g_{\mu\nu}(t, \vec{x})$. It is a 4×4, symmetric matrix whose components are functions of space and time. It affects the rest of physics by controlling the physical distances and times between points. Recall that in special relativity the frame independent concept of the distance between nearby points x^μ and $x^\mu + \mathrm{d}x^\mu$ is given by the 'invariant interval',

$$(\mathrm{d}s^2)_{\substack{\text{special} \\ \text{relativity}}} = -c^2 \, \mathrm{d}t^2 + \mathrm{d}\vec{x} \cdot \mathrm{d}\vec{x}. \tag{7}$$

The invariant interval of general relativity is

$$(\mathrm{d}s^2)_{\substack{\text{general}\\ \text{relativity}}} = \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} g_{\mu\nu}(t,\vec{x})\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu. \tag{8}$$

The fact that the metric defines true distances and times affects how derivatives and integrals are constructed in a way that is analogous to the minimal coupling rule of electrodynamics. The details need not concern us.

The rest of physics affects the metric through the Einstein equations which have the general form

$$[A(g)\partial^2 g + B(g)\partial g\partial g + \Lambda g]_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}. \tag{9}$$

My notation is that $\partial^2 g$ stands for the derivative of any component of the metric tensor with respect to any two coordinates. In the same sense, $\partial g\partial g$ stands for any product of first derivatives of the metric. The parameter Λ is known as the *cosmological constant*. It is often set to zero in general relativity but retaining it will facilitate our eventual discussion of renormalization, and nature seems to have chosen a small nonzero value for it [8, 9]. At the level I propose to work one need only note:

1. The gravitational field equations are not linear in the metric;
2. The gravitational field equations involve up to second derivatives of the metric and
3. The gravitational field equations are sourced by the stress–energy tensor $T_{\mu\nu}(t,\vec{x})$.

The stress–energy tensor $T_{\mu\nu}$ has the units of energy per volume (which is the same as force per area or stress) and its various components have the meanings,

T_{00}: energy density, (10)

T_{i0} : momentum density in the ith direction, (11)

T_{0j}: energy flux in the jth direction, (12)

T_{ij} : ith component of momentum flux in the jth direction. (13)

(The 'flux' of any quantity in the jth direction is the amount of that quantity which passes through a unit area perpendicular to the jth direction, per unit time.) The stress–energy tensor is composed of the other fields in physics, and it also depends on the metric. For example, if we set the metric equal to the value it has in special relativity (that is, $g_{00} = -1$, $g_{0i} = 0$ and $g_{ij} = \delta_{ij}$), the energy density contributed by electromagnetism is

$$[T_{00}(t,\vec{x})]_{\substack{\text{flat}\\ \text{space}}} = \tfrac{1}{2}\epsilon_0[\vec{E}\cdot\vec{E} + c^2\vec{B}\cdot\vec{B}]. \tag{14}$$

Of course gravity cannot be absent if there are electrodynamic (or any other) fields present so the actual expression for T_{00} depends upon the metric in a manner that need not concern us. It is worth commenting that the nonlinearity of the gravitational field equations, and the fact that their sources depend upon the metric they help to determine, means that no general solution of the Einstein equations is known. This is why physicists resort to the perturbative approximation technique I will describe in section 2.3.

It turns out that all fundamental force fields have a triune nature: one part is completely arbitrary, another part is totally determined by the sources of the force and the final part consists of independent degrees of freedom. For example, the completely arbitrary part of electromagnetism is the ability to change the scalar potential and vector potentials by a gauge transformation,

$$\Phi(t,\vec{x}) \longrightarrow \Phi(t,\vec{x}) + \dot{\theta}(t,\vec{x}), \tag{15}$$

$$\vec{A}(t,\vec{x}) \longrightarrow \vec{A}(t,\vec{x}) - \vec{\nabla}\theta(t,\vec{x}), \tag{16}$$

where $\theta(t,\vec{x})$ is an arbitrary function of space and time. One can easily check that the transformation equations (15) and (16) makes no change in the electric and magnetic fields (4). One can also show that it makes Maxwell's equations (1) and (2) consistent with current conservation.

The other two parts of any force field are illustrated by solving (1) and (2) for the magnetic field, assuming the current density is some known function,

$$\vec{B}(t,\vec{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{x}}\int_0^t \mathrm{d}t'\,\frac{\sin[ck(t-t')]}{ck\epsilon_0}\mathrm{i}\vec{k}\times\tilde{\vec{J}}(t',\vec{k}) + \int\frac{\mathrm{d}^3k}{(2\pi)^3}\,\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{x}}\left\{\tilde{\vec{B}}_0(\vec{k})\cos(ckt) - \frac{\mathrm{i}}{ck}\vec{k}\times\tilde{\vec{E}}_0(\vec{k})\sin(ckt)\right\}. \tag{17}$$

The first line of (17) gives the part of the magnetic field which depends upon its source, which is the current density. This part of the field contains important physics, but it is not an independent degree of freedom because knowing the current density $\vec{J}(t,\vec{x})$ fixes it completely. The technical name for such a field is 'constrained'. The 'dynamical' part of $\vec{B}(t,\vec{x})$ comes on the final line of (17) which contains the independent, purely electromagnetic degrees of freedom that would be present even if the current density was zero for all time. These extra degrees of freedom consist of pure electromagnetic radiation that corresponds to the 'photons' of quantum electrodynamics.

Gravity has the same general structure as electromagnetism and all the other fundamental force fields: part of the metric can be changed arbitrarily by a symmetry transformation known as *general coordinate invariance*; another part of the metric is determined by the stress–energy tensor and a third part contains independent degrees of freedom, the pure gravitational radiation which comprises the gravitons of quantum gravity. Because the Einstein equations (9) are not linear, the breakup of $g_{\mu\nu}(t,\vec{x})$ into these three constituents is vastly more complex than for electromagnetism, but that is another detail which need not concern us here, although it is a major headache for people who work in gravity. A fact of great importance for the discussion of renormalization is that general coordinate invariance implies the left-hand side of the Einstein equation (9) is the unique combination of the metric and no more than two of its derivatives which is consistent with the conservation of stress energy.

To understand quantum field theory it is important to be clear about the distinction between the constrained and

dynamical parts of a force field. It is the constrained part of electrodynamics which holds the hydrogen atom together. For example, if $\vec{q}(t)$ and $\vec{p}(t)$ are the position and momentum of the electron then the Hamiltonian for nonrelativistic hydrogen is

$$H = \frac{\|\vec{p}\|^2}{2m} - e\Phi(\vec{q}) \qquad \text{where } \Phi(\vec{q}) = \frac{e}{4\pi\epsilon_0\|\vec{q}\|}. \tag{18}$$

Note that the scalar potential $\Phi(\vec{q})$ is a completely determined function of the electron's position $\vec{q}(t)$. If we were doing quantum mechanics it would be a quantum operator, but only because the electron's position is a quantum operator; it would not possess any independent quantum degrees of freedom of its own. Of course the nonrelativistic Hamiltonian (18) is only an approximation and the full, relativistic system does incorporate quantized photon degrees of freedom. These degrees of freedom are needed to explain the Lamb shift of about one part in 10^6 in the frequencies of light emitted from decays of the $^2S_{\frac{1}{2}}$ and $^2P_{\frac{1}{2}}$ states [10]. However, it is worth emphasizing that the basic structure of the atom is determined by the constrained part of the electrodynamic potentials and one has to work quite hard to see even the first evidence for quantized photons from it. For example, the Lamb shift was only detected by stimulating a transition *between* the two levels.

Now consider the vastly weaker gravitational force. We have so far not been able to directly detect gravitational radiation, much less the gravitational radiation from a quantum transition, or the even subtler shift due to quantized gravitons. The gravitational effects which hold the solar system together derive from the constrained part of metric. There is only indirect evidence that gravitational radiation exists [11], and there is no evidence at all for its quantization. Which brings me to one of the major points of this paper: some quantization of gravity is inevitable because part of the metric depends upon the other fields whose quantum nature has been well established. It turns out that the first divergences of quantum gravity are due to quantum effects from these other fields [2–4]. The quantum effects of gravitons—if there is gravitational radiation, and if it is quantized—do cause problems [6, 7], but these difficulties occur at higher order in the approximation scheme described in section 2.3. So the central problem of quantum general relativity has nothing to do with gravitons.

2.2. *Quantum mechanics*

It will be noted that I have written field equations such as (1) and (2), and even solved them (17), without stating whether the system is classical or quantum. There is a good reason for this: it does not matter! The operator equations of motion in the Heisenberg Picture of quantum mechanics are the same as those for the corresponding classical theory. Further, 'solving' these equations means precisely the same thing: one expresses the dynamical variables at any time in terms of the initial values of the dynamical variables. Those initial values are the fundamental degrees of freedom of the system, and the only difference between classical physics and quantum physics is what they represent. In classical physics the initial values are just numbers and each of them can take any value, whereas in quantum physics they are noncommuting operators which must obey the uncertainty principle.

An example which has great significance for our discussion is the simple harmonic oscillator. The dynamical variable is the position as a function of time, $q(t)$. For the moment we imagine this system to exist in isolation, so it has no effect on the rest of the universe. Nor does the rest of the universe affect it; its dynamics are controlled by its Lagrangian and Euler–Lagrange equations,

$$L = \frac{1}{2}m\dot{q}^2 - \frac{1}{2}m\omega^2 q^2 \Longrightarrow \frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}}$$
$$= -m[\ddot{q}(t) + \omega^2 q(t)] = 0. \tag{19}$$

The general solution in the sense of fundamental theory is

$$q(t) = q_0\cos(\omega t) + \frac{\dot{q}_0}{\omega}\sin(\omega t), \tag{20}$$

where $q_0 = q(0)$ and $\dot{q}_0 = \dot{q}(0)$.

Now break up the trigonometric functions using the Euler relation

$$\cos(\omega t) = \frac{1}{2}[\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}],$$
$$\sin(\omega t) = \frac{\mathrm{i}}{2}[-\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}]. \tag{21}$$

The solution can be expressed as a sum of positive and negative frequencies,

$$q(t) = \frac{1}{2}\,\mathrm{e}^{-\mathrm{i}\omega t}\left[q_0 + \frac{\mathrm{i}\dot{q}_0}{\omega}\right] + \frac{1}{2}\,\mathrm{e}^{\mathrm{i}\omega t}\left[q_0 - \frac{\mathrm{i}\dot{q}_0}{\omega}\right]. \tag{22}$$

As noted above, this same solution applies both for the classical theory and for the quantum one. In the latter we can recognize that the operator coefficient of $\mathrm{e}^{-\mathrm{i}\omega t}$ must lower the energy by $\hbar\omega$, whereas the operator coefficient of $\mathrm{e}^{\mathrm{i}\omega t}$ must raise it by the same amount. The canonical momentum tells us how q_0 and $\dot{q}_0$ commute,

$$p = \frac{\partial L}{\partial \dot{q}} = m\dot{q} \Longrightarrow [q_0, \dot{q}_0] = \frac{\mathrm{i}\hbar}{m}. \tag{23}$$

We can use this to canonically normalize the raising and lowering operators,

$$a \equiv \sqrt{\frac{m\omega}{2\hbar}}\left(q_0 + \frac{\mathrm{i}\dot{q}_0}{\omega}\right) \Longrightarrow [a, a^\dagger] = 1. \tag{24}$$

And the final result for the position operator takes the form

$$q(t) = a\varepsilon(t) + a^\dagger\varepsilon^*(t), \tag{25}$$

where the 'mode coordinates' are,

$$a \equiv \sqrt{\frac{m\omega}{2\hbar}}\left(q_0 + \frac{\mathrm{i}\dot{q}_0}{\omega}\right) \qquad \text{and}$$
$$a^\dagger \equiv \sqrt{\frac{m\omega}{2\hbar}}\left(q_0 - \frac{\mathrm{i}\dot{q}_0}{\omega}\right), \tag{26}$$

and the 'mode functions' are

$$\varepsilon(t) \equiv \sqrt{\frac{\hbar}{2m\omega}}\,\mathrm{e}^{-\mathrm{i}\omega t} \qquad \text{and} \qquad \varepsilon^*(t) \equiv \sqrt{\frac{\hbar}{2m\omega}}\,\mathrm{e}^{\mathrm{i}\omega t}. \tag{27}$$

The harmonic oscillator is so important because of an amazing fact: the spatial Fourier components of every quantum field degenerate to independent harmonic oscillators in the limit that interactions vanish. For example, if we turn off the current density in (17) then the spatial Fourier transform of the magnetic field vector $\tilde{\vec{B}}(t,\vec{k})$ behaves as a pair of independent harmonic oscillators for each wave vector $\vec{k}$. The resulting free field mode sum can be rendered thus

$$\vec{B}(t,\vec{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3} \sum_{\lambda=\pm} \{\mathrm{i}\vec{k}\times\vec{\varepsilon}(t,\vec{x};\vec{k},\lambda)a(\vec{k},\lambda) - \mathrm{i}\vec{k}\times\vec{\varepsilon}^*(t,\vec{x};\vec{k},\lambda)a^\dagger(\vec{k},\lambda)\}. \tag{28}$$

In this expression the canonically normalized mode coordinates are

$$a(\vec{k},\lambda) \equiv \mathrm{i}\sqrt{\frac{\epsilon_0}{4\hbar ck}}[\hat{\theta}-\mathrm{i}\lambda\hat{\phi}]\cdot\left[c\hat{r}\times\tilde{\vec{B}}_0(\vec{k}) - \tilde{\vec{E}}_0(\vec{k})\right], \tag{29}$$

$$a^\dagger(\vec{k},\lambda) \equiv -\mathrm{i}\sqrt{\frac{\epsilon_0}{4\hbar ck}}[\hat{\theta}+\mathrm{i}\lambda\hat{\phi}]\cdot[c\hat{r}\times\tilde{\vec{B}}_0(-\vec{k}) - \tilde{\vec{E}}_0(-\vec{k})]. \tag{30}$$

The associated mode functions are

$$\vec{\varepsilon}(t,\vec{x};\vec{k},\lambda) \equiv \sqrt{\frac{\hbar}{4\epsilon_0 ck}}[\hat{\theta}+\mathrm{i}\lambda\hat{\phi}]\mathrm{e}^{-\mathrm{i}kct+\mathrm{i}\vec{k}\cdot\vec{x}}. \tag{31}$$

And it should be noted that I have expressed the wave number $\vec{k}$ in spherical coordinates with the usual spherical unit vectors,

$$\hat{r} \equiv (\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)), \tag{32}$$

$$\hat{\theta} \equiv (\cos(\theta)\cos(\phi), \cos(\theta)\sin(\phi), -\sin(\theta)), \tag{33}$$

$$\hat{\phi} \equiv (-\sin(\phi), \cos(\phi), 0). \tag{34}$$

In quantum electrodynamics acting $a^\dagger(\vec{k},\lambda)$ on a state adds a photon with energy $E = \hbar kc$, 3-momentum $\vec{p} = \hbar\vec{k}$ and polarization λ. ($\lambda = +1$ stands for left-handed, circular polarization and $\lambda = -1$ is right-handed.) Acting $a(\vec{k},\lambda)$ would remove a photon with the same quantum numbers. It should be noted that the 'particles' of fundamental theory are always the Fourier modes of quantum fields. For example, the electrons and positrons of quantum electrodynamics are represented by terms in the free field mode sum for the electron field,

$$\Psi_i(t,\vec{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3} \sum_{s=\pm\frac{1}{2}} \{\varepsilon_i(t,\vec{x};\vec{k},s)b(\vec{k},s) + \bar{\varepsilon}_i(t,\vec{x};\vec{k},s)c^\dagger(\vec{k},s)\}. \tag{35}$$

Acting $c^\dagger(\vec{k},s)$ on a state adds a positron of energy $E = \sqrt{(mc^2)^2+(\hbar ck)^2}$, 3-momentum $\vec{p} = \hbar\vec{k}$ and z-component spin (in its rest frame) $s = \pm\frac{1}{2}$. Acting $b(\vec{k},s)$ on a state removes an electron with the same 4-momentum and spin. As always, the mode coordinates are simply linear combinations of the initial values of the dynamical variable, which are the true degrees of freedom of the system

$$b(\vec{k},s) \equiv \sqrt{\frac{c}{2E}}\sum_{i=1}^{4} u_i^*(\vec{k},s)\tilde{\Psi}_i(0,\vec{k}),$$

$$c^\dagger(\vec{k},s) \equiv \sqrt{\frac{c}{2E}}\sum_{i=1}^{4} v_i^*(\vec{k},s)\tilde{\Psi}_i(0,\vec{k}). \tag{36}$$

Here the spinor wave functions are

$$u(\vec{k},s) \equiv \sqrt{\frac{\hbar c}{2(E+mc^2)}}\begin{pmatrix}\left[\frac{E+mc^2}{\hbar c}I - \vec{k}\cdot\vec{\sigma}\right]\xi(s)\\ \left[\frac{E+mc^2}{\hbar c}I + \vec{k}\cdot\vec{\sigma}\right]\xi(s)\end{pmatrix}, \tag{37}$$

$$v(\vec{k},s) \equiv \sqrt{\frac{\hbar c}{2(E+mc^2)}}\begin{pmatrix}\left[\frac{E+mc^2}{\hbar c}I - \vec{k}\cdot\vec{\sigma}\right]\eta(s)\\ -\left[\frac{E+mc^2}{\hbar c}I + \vec{k}\cdot\vec{\sigma}\right]\eta(s)\end{pmatrix}, \tag{38}$$

and the various 2-component quantities are familiar from nonrelativistic quantum mechanics,

$$I \equiv \begin{pmatrix}1&0\\0&1\end{pmatrix}, \qquad \sigma^1 = \begin{pmatrix}0&1\\1&0\end{pmatrix},$$
$$\sigma^2 = \begin{pmatrix}0&-\mathrm{i}\\ \mathrm{i}&0\end{pmatrix}, \qquad \sigma^3 = \begin{pmatrix}1&0\\0&-1\end{pmatrix}, \tag{39}$$

$$\xi\left(+\frac{1}{2}\right) \equiv \begin{pmatrix}1\\0\end{pmatrix}, \qquad \xi\left(-\frac{1}{2}\right) \equiv \begin{pmatrix}0\\1\end{pmatrix},$$
$$\eta\left(+\frac{1}{2}\right) \equiv \begin{pmatrix}0\\1\end{pmatrix}, \qquad \eta\left(-\frac{1}{2}\right) \equiv \begin{pmatrix}-1\\0\end{pmatrix}. \tag{40}$$

The mode functions in (35) are

$$\varepsilon_i(t,\vec{x};\vec{k},s) \equiv \sqrt{\frac{\hbar^2c}{2E}}\,\mathrm{e}^{-\mathrm{i}Et/\hbar+\mathrm{i}\vec{k}\cdot\vec{x}}u_i(\vec{k},s),$$
$$\bar{\varepsilon}_i \equiv \sqrt{\frac{\hbar^2c}{2E}}\,\mathrm{e}^{\mathrm{i}Et/\hbar-\mathrm{i}\vec{k}\cdot\vec{x}}v_i(\vec{k},s). \tag{41}$$

At this point I must discuss a little about quantum states. Operators such as q_0 and $\dot{q}_0$ have the potential for being anything; it is the state wave function which describes how they are distributed. These wave functions are time independent in the Heisenberg picture of quantum mechanics; it is the operators which evolve in time. In the position representation we can write the wave function as $\psi(x)$ and the two fundamental operators act by multiplication and differentiation,

$$q_0\psi(x) = x\psi(x), \tag{42}$$

$$\dot{q}_0\psi(x) = -\frac{\mathrm{i}\hbar}{m}\psi'(x). \tag{43}$$

The inner product between any two states is defined by integration,

$$\langle\phi|\psi\rangle \equiv \int_{-\infty}^{\infty}\mathrm{d}x\,\phi^*(x)\psi(x). \tag{44}$$

And a very important property of any state is *normalization*,

$$\langle\psi|\psi\rangle \equiv \int_{-\infty}^{\infty}\mathrm{d}x\,\psi^*(x)\psi(x) = 1. \tag{45}$$

(This is what puts the 'quantum' in quantum mechanics.) The expectation value of any operator $\mathcal{O}(q_0,\dot{q}_0)$ in the state $|\psi\rangle$ is

$$\langle\psi|\mathcal{O}(q_0,\dot{q}_0)|\psi\rangle = \int_{-\infty}^{\infty}\mathrm{d}x\,\psi^*(x)\mathcal{O}\left(x,-\frac{\mathrm{i}\hbar}{m}\frac{\partial}{\partial x}\right)\psi(x). \tag{46}$$

Of course expression (46) suffices for time dependent operators such as $q(t)$ because they can be expressed in terms of q_0 and $\dot{q}_0$.

The notions of state wave functions, and inner products involving them, all have straightforward generalizations to quantum field theory (if one is good with functional calculus!) however, they are very seldom used. The reason for this is that quantum field theories possess an infinite number of mode coordinates, one or more for every wave vector $\vec{k}$. Only a finite number of these modes can be excited because it costs energy to excite a mode and there is only a limited amount of free energy available. Hence there are so many more ground state modes than excited ones that most quantum field theoretic effects derive from the vast number of modes in their ground states.

The Uncertainty Principle provides a powerful, intuitive way of using classical physics to understand the effects of modes which are in their ground states. When the expectation values of q_0 and $\dot{q}_0$ are zero we can express the uncertainty principle as follows

$$\langle\psi|q_0^2|\psi\rangle \cdot \langle\psi|\dot{q}_0^2|\psi\rangle \geqslant \frac{\hbar^2}{4m^2}. \tag{47}$$

Equality is achieved for a *minimum uncertainty* state. For such a state we can think of the Hamiltonian as a function of q_0 alone,

$$H \equiv \frac{1}{2}m\dot{q}_0^2 + \frac{1}{2}m\omega^2 q_0^2 \longrightarrow \frac{\hbar^2}{8mq_0^2} + \frac{1}{2}m\omega^2 q_0^2 \equiv E(q_0). \tag{48}$$

The term $\hbar^2/8mq_0^2$ in $E(q_0)$ is known as *uncertainty pressure*. It reflects the physical import of the uncertainty principle, which is that concentrating q_0 more tightly about its mean value of zero makes $\dot{q}_0$ proportionately less concentrated. With uncertainty pressure we can understand many quantum effects classically. For example, the minimum energy is

$$\frac{\partial E}{\partial q_0} = -\frac{\hbar^2}{4mq_0^3} + m\omega^2 q_0 \Longrightarrow q_{\rm min} = \sqrt{\frac{\hbar}{2m\omega}} \qquad \text{and}$$
$$E_{\rm min} = \frac{1}{2}\hbar\omega. \tag{49}$$

One does not always get the factors of two right with this level of analysis but it is a powerful technique for understanding complex things in a simple way, and I will apply it to explain where the divergences of quantum gravity arise and why they are worse than those associated with the other forces.

2.3. Perturbation theory

The alert reader will have noted that the exact solutions of the previous subsections were all obtained for noninteracting theories. There is a good reason for that: not a single interacting field theory has been solved in four spacetime dimensions. Note the essential distinction between 'exactly solving a theory', which means expressing the dynamical variable at any time in terms of arbitrary initial value data, and obtaining an 'exact solution to the field equations,' which means solving the equations of motion for one particular choice of initial values. The various 'exact solutions' of Einstein's equations involve setting almost all the degrees of freedom to zero. This is fine for classical physics, but it is not permitted in quantum mechanics. For example, in free quantum electrodynamics the mode coordinates $a(\vec{k},\lambda)$ and $a^\dagger(\vec{k},\lambda)$ of expression (28) are not commuting variables,

$$[a(\vec{k},\lambda), a^\dagger(\vec{k}',\lambda')] = \delta_{\lambda\lambda'}(2\pi)^3\delta^3(\vec{k}-\vec{k}'). \tag{50}$$

A classical picture is possible, but one must imagine that each mode experiences the 0-point motion we found above for the harmonic oscillator.

Because there is no hope of exactly solving the field equations for arbitrary initial value data, computing anything in quantum field theory requires the use of approximation techniques. The standard one is known as *perturbation theory* and a good way of motivating it is by observing that, even though our expression (17) is correct, it does not give the magnetic field because the current density $\vec{J}(t,\vec{x})$ is affected by the electrodynamic potentials. One can see this in quantum electrodynamics for which the charge density and the current density are formed from the electron field $\Psi_i(t,\vec{x})$,

$$\rho(t,\vec{x}) = \frac{e}{\hbar}\Psi^\dagger(t,\vec{x})\Psi(t,\vec{x}),$$
$$\vec{J}(t,\vec{x}) = \frac{ec}{\hbar}\Psi^\dagger(t,\vec{x})\begin{pmatrix}-\vec{\sigma} & 0\\ 0 & \vec{\sigma}\end{pmatrix}\Psi(t,\vec{x}). \tag{51}$$

And the Dirac equation for Ψ involves the potentials Φ and $\vec{A}$,

$$\left[\frac{\partial}{\partial ct} - \frac{ie\Phi}{\hbar c}\right]\Psi + \begin{pmatrix}-\vec{\sigma} & 0\\ 0 & \vec{\sigma}\end{pmatrix}\cdot\left(\vec{\nabla} + \frac{ie\vec{A}}{\hbar}\right)\Psi$$
$$+ \frac{imc}{\hbar}\begin{pmatrix}0 & I\\ I & 0\end{pmatrix}\Psi = 0. \tag{52}$$

One can see what must happen, even without working out the details:

- 0-point motions of Ψ modes engender Φ and $\vec{A}$;
- These electromagnetic fields change Ψ;
- Which changes Φ and $\vec{A}$;
- Which changes Ψ, and so on.

The progression is endless, and we could just have easily begun it with 0-point motions of the electromagnetic fields affecting Ψ! The process could be made to terminate if an external force fixed Ψ or if we made some very special choice of initial values. But there are no external forces in fundamental theory, and quantum mechanics demands that we consider generic initial value data.

The cycle of action and reaction I have sketched precludes obtaining the exact solution, but a good approximate solution can be found if each cycle of action and reaction induces a smaller subsequent cycle. So we start with noninteracting, charged matter and electrodynamics, which defines the 0th order. In quantum electrodynamics the 0th order would be the free field mode sums (35) and the analogue of (28) for the vector and scalar potentials. The 1st order perturbation for the electrodynamic potentials comes from solving Maxwell equations (1) and (2) with the charge density and current

density formed from the fixed, 0th order electron field (35). The 1st order perturbation for the electron field Ψ^1 comes solving the Dirac equation (52) with the terms involving the electron charge e evaluated at 0th order,

$$\frac{\partial\Psi^1}{\partial ct} + \begin{pmatrix} -\vec{\sigma} & 0 \\ 0 & \vec{\sigma} \end{pmatrix} \cdot \vec{\nabla}\Psi^1 + \frac{imc}{\hbar} \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \Psi^1 = \frac{ie\Phi^0\Psi^0}{\hbar c} - \begin{pmatrix} -\vec{\sigma} & 0 \\ 0 & \vec{\sigma} \end{pmatrix} \cdot \frac{ie\vec{A}^0}{\hbar}\Psi^0. \quad (53)$$

The second order perturbation is obtained by computing what the first order fields do, and so on. If the effect of each cycle is reduced by a small enough factor then we do not need to carry the process out for many cycles before the theoretical prediction is as accurate as any experiment that can be performed. This is how perturbation theory works.

It should be clear from the sketch I have given that each order of quantum electrodynamic perturbation theory comes with an extra power of the electron charge e. It should also be clear that the effects of different lower order modes add in perturbation theory. In a fixed volume V one sums over the Fourier wave vectors $\vec{k}$ using the famous density of states formula,

$$\sum_{\vec{k}} = V \int \frac{\mathrm{d}^3k}{(2\pi)^3}. \quad (54)$$

Because we typically want densities, the factor of V drops out. Because there are as many positron modes as electron modes, with the same kinematic properties and opposite charge, it turns out that the odd powers of e cancel out and the net effect comes from even powers. It also happens that one always gets a certain number of 2π's from (54), and some factors of ϵ_0, $\hbar$ and c in the dimensionless combination,

$$\frac{e^2}{4\pi^2\epsilon_0\hbar c} \equiv \frac{\alpha}{\pi} \simeq \frac{1}{430}. \quad (55)$$

This is the expansion parameter of quantum electrodynamics and its smallness is what makes perturbation theory effective.

I have glossed over quite a number of details in describing how perturbation theory works for quantum electrodynamics, and I am not going to be any more thorough for quantum general relativity. However, the coupling constant is important because it controls how reliable perturbation theory ought to be. One can see from the Einstein equation (9) that the source terms involve G/c^4, so one more power of this must obviously appear at each successive order in perturbation theory. Each higher order in perturbation theory also contributes a factor of $1/\hbar c$, just as in quantum electrodynamics (and all other quantum field theories). The resulting combination of fundamental constants has the dimension of an inverse energy squared,

$$\frac{G}{\hbar c^5} \simeq \left(\frac{1}{2.0 \times 10^9\,\mathrm{J}}\right)^2 \simeq \left(\frac{1}{1.2 \times 10^{19}\,\mathrm{GeV}}\right)^2. \quad (56)$$

Therefore, quantum gravitational perturbation theory must amount to an expansion in the dimensionless parameter $GE^2/\hbar c^5$, where E is some energy in whatever process is under study.

An issue which sometimes confuses people is that the series approximation perturbation theory generates is only asymptotic, rather than convergent. Suppose we have function $S(x)$ and we obtain an N-term series approximation $S_N(x)$ in terms of some standard set of functions $f_n(x)$ which are organized so that, for the x-range of interest,

$$f_0(x) > f_1(x) > f_2(x) > f_3(x) > \cdots. \quad (57)$$

The typical case is that $f_n(x) = x^n$ but I want to allow for fractional powers or powers times logarithms. The series approximation $S_N(x)$ takes the form

$$S_N(x) = \sum_{n=0}^{N-1} s_n f_n(x), \quad (58)$$

where the s_n are numbers. We say that the series is *convergent* if taking N to infinity recovers the original function $S(x)$,

$$\text{Convergent} : \Longrightarrow \lim_{N\to\infty} [S(x) - S_N(x)] = 0. \quad (59)$$

We say the series is *asymptotic* at $x = x_0$ if the difference between $S(x)$ and $S_N(x)$ goes to zero faster than $f_{N-1}(x)$ as x approaches x_0, for any fixed N,

$$\text{Asymptotic} : \Longrightarrow \lim_{x\to x_0} \left[\frac{S(x) - S_N(x)}{f_{N-1}(x)}\right] = 0. \quad (60)$$

It is entirely possible for a series to possess both properties but it is a fact that perturbative expansions of quantum field theory are only asymptotic (for zero coupling constant), not convergent.

The difference between a convergent series and an asymptotic series is beautifully illustrated by a special function known as the 'exponential integral' [12],

$$\mathrm{E}_1(x) \equiv \int_x^\infty \mathrm{d}t \frac{\mathrm{e}^{-t}}{t}. \quad (61)$$

A convergent series for small x can be obtained by extracting the integral down to $x = 1$ as a constant, and then adding zero in the form of $-\ln(x) + \ln(x)$,

$$\mathrm{E}_1(x) = \int_1^\infty \mathrm{d}t \frac{\mathrm{e}^{-t}}{t} + \int_0^1 \mathrm{d}t \left[\frac{\mathrm{e}^{-t} - 1}{t}\right] - \ln(x) - \int_0^x \mathrm{d}t \left[\frac{\mathrm{e}^{-t} - 1}{t}\right]. \quad (62)$$

The first two integrals of (62) just give minus the Euler–Mascheroni constant $\gamma \approx 0.577$. The final integral of (62) can be evaluated by expanding $(\mathrm{e}^{-t} - 1)/t$ in powers of t and then integrating termwise. The resulting expansion is

$$\mathrm{E}_1(x) = -\gamma - \ln(x) - \sum_{n=1}^\infty \frac{(-1)^n x^n}{n \cdot n!}. \quad (63)$$

Of course this series is also asymptotic for $x \to 0$.

One can get an asymptotic expansion of $\mathrm{E}_1(x)$ for large x by successive partial integrations,

$$\mathrm{E}_1(x) = \sum_{n=0}^{N-1} \frac{(-1)^n n!\,\mathrm{e}^{-x}}{x^{n+1}} + (-1)^N N! \int_x^\infty \mathrm{d}t \frac{\mathrm{e}^{-t}}{t^{N+1}}. \quad (64)$$

The asymptotic series (for $x \rightarrow \infty$) is the first term of (64), the second term is the remainder. Its magnitude can be bounded by replacing the $1/t^{N+1}$ with $1/x^{N+1}$,

$$R_N \equiv N! \int_x^\infty dt \frac{e^{-t}}{t^{N+1}} < \frac{N!}{x^{N+1}} \int_x^\infty dt\, e^{-t} = \frac{N!\, e^{-x}}{x^{N+1}}. \quad (65)$$

Hence the magnitude of the difference between $E_1(x)$ and $S_N(x)$ is

$$\left| E_1(x) - \sum_{n=0}^{N-1} \frac{(-1)^n n!\, e^{-x}}{x^{n+1}} \right| < \frac{N! e^{-x}}{x^{N+1}}. \quad (66)$$

This proves the series is asymptotic as $x \rightarrow \infty$. To see that it is not convergent, note that the sum of R_N and R_{N+1} can be evaluated exactly,

$$R_N + R_{N+1} = -N! \int_x^\infty dt \frac{\partial}{\partial t} \left[\frac{e^{-t}}{t^{N+1}} \right] = \frac{N!\, e^{-x}}{x^{N+1}}. \quad (67)$$

Increasing N by one changes the right-hand side of (67) by a factor of $(N+1)/x$. As long as this is less than one, accuracy is increased, but for any fixed value of x there must eventually be an N past which accuracy deteriorates. The best asymptotic series approximation for $E_1(x)$ is therefore obtained by carrying the expansion out to about $N \sim 1/x$ and no further.

The fact that nonconvergent asymptotic series expansions cannot be made arbitrarily accurate bothers the mathematically inclined. It does preclude defining a quantum field theory by its perturbative expansion, however, whether or not there is any practical problem depends upon the value of the coupling constant and the state of the experimental art. If the coupling constant is so small that no experiment can measure the deviation between reality and the best asymptotic series result then there is no operational problem. That is the case for quantum electrodynamics at conventional energy scales. A process which involves $2N$ vertices will acquire a factor of $(\alpha/2\pi)^N$ from its coupling constants and the factors of 2 and π from momentum integrations. There will also be a multiplicity factor of about $(2N-1)!!$, so one expects the series to begin diverging at about $N \sim \pi/\alpha \sim 430$. At that point the accuracy is so great one can only estimate it using Stirling's approximation,

$$\left(\frac{\alpha}{2\pi}\right)^N \times (2N-1)!! \longrightarrow \sqrt{2}\left(\frac{\alpha N}{e\pi}\right)^N \sim 10^{-187}. \quad (68)$$

No experiment can approach that accuracy; the current state of the art is sensitive to about the *fourth* order in α, not the *four hundredth*!

On the other hand, the coupling constant of the strong interaction is large enough, at low energies, that the asymptotic expansion is practically worthless. So how about quantum general relativity? We have seen that perturbation theory generates an expansion in powers of $GE^2/\hbar c^5$. For the highest proton energy we shall be able to reach at the LHC this number would be about

$$\frac{GE^2}{\hbar c^5} \sim \left(\frac{7 \times 10^3\,\text{GeV}}{1.2 \times 10^{19}\,\text{GeV}}\right)^2 \sim 3.4 \times 10^{-31}. \quad (69)$$

This means the perturbative series for quantum general relativity should be wonderfully more accurate than that of quantum electrodynamics, for which analogous factor is $\alpha/2\pi \sim 1.2 \times 10^{-3}$. In fact, it should be so good that some special circumstance would be needed to make quantum gravitational effects observable at all. I will have more to say about this later but let us for now note that the asymptotic nature of perturbative results is not high on the list of problems for quantum gravity.

2.4. Renormalization

It turns out that the perturbative corrections from any four-dimensional quantum field theory diverge when they are expressed in terms of the parameters such as e and m which appear in the field equations. The reason for this divergence is very simple: all the modes contribute a little bit, and there are so many modes that one gets a divergence from the sum (54) over them. For all the other quantum field theories these divergences can be absorbed by regarding the parameters of the field equations to depend in a divergent way upon physically measured quantities such as the electron charge and mass. When that is done correctly the perturbative corrections really are small, if the coupling constant is, and they agree wonderfully with experiment.

The procedure for using the parameters of the field equations to absorb divergences is known as *renormalization*. I will describe how it works by showing that the vacuum polarization of quantum electrodynamics is completely analogous to the phenomenon of classical polarization in a medium. Then I will use the fact that gravity couples to stress energy, rather than charge, to show that renormalizing the divergences of quantum gravity involves adding new sorts of terms to the gravitational field equations. These new terms would remove all the divergences [5], but we will see in the next subsection that they would also make the universe blow up. That is the central problem of perturbative quantum general relativity.

Let us consider the phenomenon of polarization in a static, classical medium. The actual charge density of the medium consists of an enormous number of positive and negative point charges q_α located at equilibrium positions $\vec{X}_\alpha$,

$$\rho(\vec{x})|_{\substack{\text{undisturbed} \\ \text{medium}}} = \sum_\alpha q_\alpha \delta^3(\vec{x} - \vec{X}_\alpha). \quad (70)$$

The medium as a whole is electrically neutral; it is only microscopically that one can see its vast collection of positive and negative charges. If we apply a static electric field $\vec{E}(\vec{x})$, the charges shift to new equilibrium positions $\vec{X}_\alpha \rightarrow \vec{X}_\alpha + \Delta\vec{x}_\alpha$. We can expand the density of each charge around its equilibrium value,

$$q_\alpha \delta^3(\vec{x} - \vec{X}_\alpha - \Delta\vec{x}_\alpha) = q_\alpha \delta^3(\vec{x} - \vec{X}_\alpha) - \vec{\nabla} \cdot [q_\alpha \Delta\vec{x}_\alpha \delta^3(\vec{x} - \vec{X}_\alpha)] + \cdots. \quad (71)$$

Hence we can write the charge density of the disturbed medium as its undisturbed value plus a series of corrections,

$$\rho(\vec{x})|_{\substack{\text{disturbed}\\\text{medium}}} = \rho(\vec{x})|_{\substack{\text{undisturbed}\\\text{medium}}} - \vec{\nabla}\cdot\left[\sum_{\alpha} q_\alpha \Delta\vec{x}_\alpha \delta^3(\vec{x}-\vec{X}_\alpha)\right] + \cdots . \tag{72}$$

We have already noted that the undisturbed charge density appears to be zero on macroscopic scales because the positive and negative charges cancel one another. However, the sum in the square bracketed term on the right-hand side of (72) tends to give a coherent effect because the positive charges move with the applied electric field and the negative charges move the other way. The higher terms in the expansion tend to be small because the charges do not move much, so the result is

$$\rho(\vec{x})|_{\substack{\text{disturbed}\\\text{medium}}} \simeq -\vec{\nabla}\cdot\vec{P}(\vec{x}), \tag{73}$$

where the *polarization* is

$$\vec{P}(\vec{x}) \equiv \sum_{\alpha} q_\alpha \Delta\vec{x}_\alpha \delta^3(\vec{x}-\vec{X}_\alpha). \tag{74}$$

Now suppose we add a small number of 'free' charges to the vast collection of 'bound' ones in the medium. The Gauss law equation reads,

$$\epsilon_0 \vec{\nabla}\cdot\vec{E} \simeq \rho_{\text{free}} - \vec{\nabla}\cdot\vec{P}. \tag{75}$$

The smart way to solve this equation is to combine $\vec{P}$ with $\epsilon_0\vec{E}$ to form the electric displacement $\vec{D}$,

$$\vec{\nabla}\cdot(\epsilon_0\vec{E}+\vec{P}) \equiv \vec{\nabla}\cdot\vec{D} = \rho_{\text{free}}. \tag{76}$$

For the case of a linear, isotropic medium the polarization is proportional to the electric field,

$$\vec{P}(\vec{x}) = \Delta\epsilon \times \vec{E}(\vec{x}). \tag{77}$$

In that case we can subsume the effect of the medium into a change in the electric permittivity,

$$\epsilon\vec{\nabla}\cdot\vec{E} = \rho_{\text{free}} \qquad \text{where } \epsilon \equiv \epsilon_0 + \Delta\epsilon. \tag{78}$$

Let us now switch from classical physics with a medium to quantum electrodynamics in empty space. Of course the space can never really be empty because it is pervaded by the electron field $\Psi_i(t,\vec{x})$, which gives the charge density of quantum electrodynamics,

$$\rho(t,\vec{x}) = \frac{e}{\hbar}\,\Psi_i^*(t,\vec{x})\Psi_i(t,\vec{x}). \tag{79}$$

As we saw in expression (35), the electron field consists of a an infinite collection of the operators which create and destroy charged particles. We also saw that it is possible to think about this sum of operators classically provided one imagines each mode to be executing 0-point motion. When an electric field is applied these 0-point motions change, and that produces an observable, coherent effect, just as it does for the classical medium and for the same reasons.

From the free field expansion (35) one can see that the spacetime dependence of 0-point motion is characterized by mode functions (41). The part of interest to us is the oscillatory factor on the electron creation operator $b^\dagger(\vec{k},s)$ in $\Psi^*(t,\vec{x})$,

$$e^{iEt/\hbar - i\vec{k}\cdot\vec{x}} \qquad \text{where } E = \sqrt{(mc^2)^2 + (\hbar ck)^2}. \tag{80}$$

We might cancel the spatial phase by combining this with a positron creation operator of opposite momentum and spin in $\Psi(t,\vec{x})$,

$$e^{iEt/\hbar - i\vec{k}\cdot\vec{x}} b^\dagger(\vec{k},s) \times e^{iEt/\hbar + i\vec{k}\cdot\vec{x}} c^\dagger(-\vec{k},-s), \tag{81}$$

but nothing can be done about the temporal phase factor. This temporal phase factor means that effects from this 'virtual' electron–positron pair cannot remain coherent longer than about $\Delta t \sim \hbar/E$. This is a very short time; the longest lived mode is the one with $\vec{k}=0$,

$$(\Delta t)_{\vec{k}=0} \sim \frac{\hbar}{mc^2} \sim \frac{10^{-34}\,\text{J s}}{(9\times10^{-31}\,\text{kg})(3\times10^8\,\text{m s}^{-1})^2} \sim 10^{-22}\,\text{s}. \tag{82}$$

Let us compute the polarization induced by a virtual pair of wave number $\vec{k}$. As we have seen, quantum physics tells us they effectively exist for a time $\Delta t \sim \hbar/E$, but the rest of the analysis is completely classical. The equation of motion for a charge e acted upon by an electric field $\vec{E}$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{m\vec{v}}{\sqrt{1-v^2/c^2}}\right] = e\vec{E}. \tag{83}$$

Over an interval as short as (82) we can regard the energy of the charge as constant,

$$\frac{m}{\sqrt{1-v^2/c^2}} \simeq \frac{E}{c^2}. \tag{84}$$

We can also forget about the variation in $\vec{E}(\vec{x})$ as the particle moves, so the induced deviation in time $\Delta t = \hbar/E$ is

$$\Delta\vec{x} \simeq \frac{ec^2\Delta t^2}{2E}\vec{E} = \frac{e\hbar^2c^2}{2E^3}\vec{E}. \tag{85}$$

It remains only to add the electron and positron contributions, and then sum over modes to find the total induced polarization,

$$\vec{P}(\vec{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\left[\mathrm{e}\times\left(\frac{\mathrm{e}\hbar^2c^2}{2E^3}\right)\times\vec{E}(\vec{x}) - \mathrm{e}\times\left(\frac{-\mathrm{e}\hbar^2c^2}{2E^3}\right)\times\vec{E}(\vec{x})\right]. \tag{86}$$

Expression (86) is proportional to the electric field, just like a linear, isotropic medium, so we can identify the change in the permittivity as

$$\Delta\epsilon = e^2\hbar^2c^2\int\frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{[(mc^2)^2+(\hbar ck)^2]^{\frac{3}{2}}} \tag{87}$$

$$= \frac{e^2}{2\pi^2\hbar c}\int_0^\infty \mathrm{d}(\hbar ck)\frac{(\hbar ck)^2}{[(mc^2)^2+(\hbar ck)^2]^{\frac{3}{2}}}. \tag{88}$$

Our expression for $\Delta\epsilon$ diverges logarithmically! On the other hand, this effect is ubiquitous because it comes from vacuum fluctuations, without any medium being present. That means we will never observe $\Delta\epsilon$ independently of ϵ_0, only their sum[1],

$$\epsilon \equiv \epsilon_0 + \Delta\epsilon. \tag{89}$$

It is this sum which must be finite, not $\Delta\epsilon$ or ϵ_0 separately. So what particle theorists do is to adjust the parameter in the field equation ϵ_0 to be the conventional value of about $8.85 \times 10^{-12}\,\mathrm{F\,m^{-1}}$ *minus* $\Delta\epsilon$. That is how renormalization works.

Some people find it disconcerting to have a parameter from the field equations changed when it appears in physical predictions. However, much of this sense of wrongness derives from insufficient experience with nonlinear systems. In a linear system, such as electrodynamics becomes when its sources are held fixed, there is indeed a simple relation between parameters in the equations and physical predictions. For example, a stationary point charge q induces a Coulomb field,

$$\rho(t,\vec{x}) = q\delta^3(\vec{x}) \Longrightarrow \vec{E}(t,\vec{x}) = \frac{q\vec{x}}{4\pi\epsilon_0\|\vec{x}\|^3}. \tag{90}$$

The parameters q and ϵ_0 that enter the classical field equations are the same ones which appear in the observed long range field. But even electrodynamics becomes nonlinear when one permits its sources to respond to electromagnetic fields, and this generally causes the observed quantities to differ from their cognates in the equations. For example, the conduction electrons in a metal behave, for many purposes, as if they are free to move inside the metal, but with a charge and mass different from their values in empty space.

Renormalization is the hallmark of nonlinear systems, even classical ones. We encountered it in quantum electrodynamics because that theory is nonlinear, not because it is quantum mechanical. Similar effects occur in nonlinear classical systems. For example, the combined mass of the Earth–Moon system is a little less than the sum of their masses in isolation, owing to their gravitational interaction energy,

$$\begin{aligned} -\frac{GM_{\rm E}M_{\rm M}}{R_{\rm EM}c^2} &\sim -\frac{(7\times10^{-11}\,\mathrm{N\,m^{-2}})(6\times10^{24}\,\mathrm{kg})(7\times10^{22}\,\mathrm{kg})}{(4\times10^{8}\,\mathrm{m})(3\times10^{8}\,\mathrm{m\,s^{-1}})^2} \\ &\sim -8\times10^{9}\,\mathrm{kg}. \end{aligned} \tag{91}$$

Nor is the Earth's mass equal to the sum of the masses of its constituents. If we imagine it to be a uniform sphere the actual mass is less by about

$$\begin{aligned} -\frac{3GM_{\rm E}^2}{5R_{\rm E}c^2} &\sim -\frac{3(7\times10^{-11}\,\mathrm{N\,m^{-2}})(6\times10^{24}\,\mathrm{kg})^2}{5(6\times10^{6}\,\mathrm{m})(3\times10^{8}\,\mathrm{m\,s^{-1}})^2} \\ &\sim -3\times10^{15}\,\mathrm{kg}. \end{aligned} \tag{92}$$

This 'gravitational renormalization' effect obviously becomes stronger the more compact the mass is. Arnowitt, Deser and Misner have shown the renormalization is actually infinite for a point mass [13].

Some people are resigned to renormalization in principle, but disturbed by the fact that quantum field theoretic renormalizations involves divergent quantities. This bothered even the physicists who devised renormalization! They eventually accepted it for two reasons:

- As I have emphasized, renormalization is inevitable in nonlinear systems, so we would need to choose ϵ_0 to make the observed quantity $\epsilon_0+\Delta\epsilon$ agree with experiment, even if $\Delta\epsilon$ had been finite and
- Once this is done, along with the analogous things for the electron mass and charge, all other quantum electrodynamic corrections are quite small and in wonderful agreement with experiment.

To see this last point let us return to vacuum polarization and do a better job of accounting for the spatial variation while still (incorrectly) ignoring temporal variation. The fundamental field equation is Gauss's law,

$$\epsilon_0\vec{\nabla}\cdot\vec{E}(t,\vec{x}) = \frac{\mathrm{e}}{\hbar}\Psi_i^*(t,\vec{x})\Psi_i(t,\vec{x}). \tag{93}$$

If the electric field is not constant in space we need to compute the vacuum polarization of each wave vector $\vec{p}$. Taking the spatial Fourier transform of the source and using Parseval's theorem gives

$$\int \mathrm{d}^3x\,\mathrm{e}^{-\mathrm{i}\vec{p}\cdot\vec{x}}\frac{\mathrm{e}}{\hbar},$$

$$\Psi_i^*(t,\vec{x})\Psi_i(t,\vec{x}) = \frac{\mathrm{e}}{\hbar}\int\frac{\mathrm{d}^3k}{(2\pi)^3}\tilde{\Psi}_i^*(t,\vec{k})\tilde{\Psi}(t,\vec{p}-\vec{k}). \tag{94}$$

From the free field mode sum (35) it is apparent that there are two energies involved, not the single one of expression (80),

$$E(\vec{k}) \equiv \sqrt{(mc^2)^2+(\hbar c)^2\|\vec{k}\|^2} \qquad \text{and}$$

$$E(\vec{p}-\vec{k}) \equiv \sqrt{(mc^2)^2+(\hbar c)^2\|\vec{p}-\vec{k}\|^2}. \tag{95}$$

In place of expression (81) we should expect the two operators to contribute as follows:

$$\tilde{\Psi}^*(t,\vec{k}) \longrightarrow \exp\left[\frac{\mathrm{i}E(\vec{k})t}{\hbar}\right]\times b^\dagger(\vec{k},s), \tag{96}$$

$$\tilde{\Psi}(t,\vec{p}-\vec{k}) \longrightarrow \exp\left[\frac{\mathrm{i}E(\vec{p}-\vec{k})t}{\hbar}\right]\times c^\dagger(\vec{p}-\vec{k},-s). \tag{97}$$

Now recall that the induced displacement we computed in (85) involved the inverse third power of 'the' energy. As we saw, there are really two energies and it turns out that it is their sum which appears in the full quantum field theoretic result. There is also a factor of $\frac{8}{3}$, so we can write the induced polarization as

$$\tilde{\vec{P}}(\vec{p}) = \frac{8}{3}e^2\hbar^2c^2\int\frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{[E(\vec{k})+E(\vec{p}-\vec{k})]^3}\times\tilde{\vec{E}}(\vec{p}). \tag{98}$$

We can write the position space result in terms of a momentum dependent shift in the permittivity,

$$\vec{P}(\vec{x}) = \int\frac{\mathrm{d}^3p}{(2\pi)^3}\,\mathrm{e}^{\mathrm{i}\vec{p}\cdot\vec{x}}\Delta\epsilon(\vec{p})\int\mathrm{d}^3x'\,\mathrm{e}^{-\mathrm{i}\vec{p}\cdot\vec{x}'}\vec{E}(\vec{x}'), \tag{99}$$

[1] Readers familiar with quantum field theory will recognize that ϵ is proportional to Z_2, the fermion field strength renormalization.

where the momentum dependent permittivity shift is

$$\Delta\epsilon(\vec{p}) = \frac{8e^2}{3\hbar c} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\hbar^3 c^3}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3}. \quad (100)$$

One way of understanding (100) is by expanding the energy denominator around $\vec{p} = 0$,

$$E(\vec{k}) + E(\vec{p}-\vec{k}) = 2E(\vec{k}) \left[1 - \frac{\hbar^2 c^2 \vec{k}\cdot\vec{p}}{2E^2(\vec{k})} + \frac{\hbar^2 c^2 p^2}{4E^2(\vec{k})} - \frac{\hbar^4 c^4 (\vec{k}\cdot\vec{p})^2}{4E^4(\vec{k})} + \cdots \right] \quad (101)$$

Substituting just the first term of (101) into (100) gives ($\frac{1}{3}\times$) the logarithmically divergent expression (88) we called $\Delta\epsilon$. The contributions from all the higher terms of expansion (101) are finite. Using some mathematical methods that were developed for quantum field theory they can be evaluated to give

$$\Delta\epsilon_{\rm fin}(\vec{p}) = -\frac{e^2}{2\pi^2\hbar c} \int_0^1 \mathrm{d}\tau\, \tau(1-\tau) \ln\left[1 + \tau(1-\tau)\frac{\hbar^2 p^2}{m^2 c^2} \right]. \quad (102)$$

We can therefore think of the total permittivity as

$$\epsilon(\vec{p}) = \epsilon_0 + \Delta\epsilon(0) + \Delta\epsilon_{\rm fin}(\vec{p}) = \epsilon_{\rm meas} + \Delta\epsilon_{\rm fin}(\vec{p}), \quad (103)$$

where $\epsilon_{\rm meas} \simeq 8.85 \times 10^{-12}\,\mathrm{F\,m^{-1}}$ is the measured value of the electric permittivity.

Expressions (102) and (103) illustrate why particle theorists are so pleased with renormalization. First, all the divergences are gone, as promised. Second, the quantum corrections from $\Delta\epsilon_{\rm fin}(\vec{p})$ are small, both because of the initial factor of e^2 and also because the ratio $\hbar^2 p^2/m^2c^2$ is minuscule for most applications. If we express the wave number in terms of a wavelength, $p = 2\pi/\lambda$ then the ratio is

$$\left(\frac{\hbar p}{mc}\right)^2 \simeq \left(\frac{2\times 10^{-12}\,\mathrm{m}}{\lambda}\right)^2. \quad (104)$$

Even at the Bohr radius of about $\lambda \sim 5\times 10^{-11}\,\mathrm{m}$ the ratio (104) would be only about 10^{-3}. Finally, and most important of all, the small effects from $\Delta\epsilon_{\rm fin}(\vec{p})$ have been verified in many experiments.

One of the strangest effects from $\Delta\epsilon_{\rm fin}(\vec{p})$ is that the electrodynamic force grows stronger at short distances. During the 1990s electrons and positrons were brought to within about $\lambda \sim 10^{-18}\,\mathrm{m}$ at colliders such as the SLC at Stanford University and LEP at the European Nuclear Research Center (CERN). At these separations the fractional change in permittivity is

$$\frac{\Delta\epsilon_{\rm fin}(\vec{p})}{\epsilon_{\rm meas}} \simeq -\frac{2\alpha}{3\pi} \times \ln\left(\frac{2\times 10^{-12}\,\mathrm{m}}{10^{-18}\,\mathrm{m}}\right) \simeq -0.02. \quad (105)$$

Because the electric force goes like $1/\epsilon$, this 2% reduction in the permittivity implies a 2% increase in the force, and that is just what was seen. The phenomenon is known as *running of the coupling constants*. We can understand it very simply from the fact that increasing the wave number p increases the energy of the virtual electron–positron pair, which decreases the time they can exist and hence the degree to which they can polarize the vacuum. So the charge screening at high p must be less than at low p.

Because the force fields of the strong interaction attract one another, it turns out that the strong interaction becomes weaker at large p. Discovering this in 1973 was what won the 2004 Nobel Prize for Politzer [14], Wilczek and Gross [15]. Our belief that we know the correct theory of the strong interaction is mostly based upon pushing to very small separations, at which point perturbative predictions from this theory become reliable. The curious fact that strong interactions are much stronger than electromagnetism at low energies, and grow weaker at high energies, whereas electromagnetism gets stronger, is one thing which makes particle physicists suspect both interactions are part of a grand unified theory whose unity becomes manifest at very high energy.

We are finally ready to consider how quantum matter affects the Einstein equation (9). For definiteness let us focus on the 0th order electromagnetic contribution to the energy density (14),

$$\frac{8\pi G}{c^4} \times \frac{1}{2}\epsilon_0 [\vec{E}(t,\vec{x})\cdot\vec{E}(t,\vec{x}) + c^2 \vec{B}(t,\vec{x})\cdot\vec{B}(t,\vec{x})]. \quad (106)$$

At lowest order this is a product of two free field mode sums, just like the charge density of quantum electrodynamics,

$$\frac{e}{\hbar}\Psi^*(t,\vec{x})\Psi(t,\vec{x}). \quad (107)$$

As might be expected, the terms they induce on the right-hand side of their respective field equations are very similar. For quantum electrodynamics in the static limit we got

$$\vec{\nabla}\cdot\vec{P}(\vec{x}) = \int \frac{\mathrm{d}^3p}{(2\pi)^3} \mathrm{e}^{\mathrm{i}\vec{p}\cdot\vec{x}} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\frac{8}{3}e^2(\hbar c\|\vec{p}\|)^2}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3} \times \tilde{\Phi}(\vec{p}). \quad (108)$$

For quantum gravity the analogous result takes the form

$$\mathcal{G}_{\mu\nu}(\vec{x}) = \int \frac{\mathrm{d}^3p}{(2\pi)^3} \mathrm{e}^{\mathrm{i}\vec{p}\cdot\vec{x}} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\frac{8\pi G}{c^4}\mathcal{E}^4}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3} \times \tilde{h}(\vec{p}), \quad (109)$$

where $h_{\mu\nu}(t,\vec{x})$ is the deviation of the metric from its quiescent value and I have suppressed its indices in (109) because more than one component can contribute to any one of the ten Einstein equations. The symbol $\mathcal{E}^4$ stands for any combination of four quantities built from $\vec{k}$ and $\vec{p}-\vec{k}$ and having the dimensions of energy^4. Possible values for $\mathcal{E}^4$ include,

$$(\hbar c\|\vec{p}\|)^4, \quad [\hbar^2 c^2 \vec{k}\cdot(\vec{p}-\vec{k})]^2, \quad E^2(\vec{k})E^2(\vec{p}-\vec{k}) \quad \text{and} \quad [E(\vec{k}) + E(\vec{p}-\vec{k})]^4. \quad (110)$$

If (109) represents a contribution from electrodynamics then the energy of a photon with wave vector $\vec{k}$ is $E(\vec{k}) = \hbar c\|\vec{k}\|$.

There are two outstanding differences between the quantum electrodynamic result (108) and its quantum gravitational analogue (109):

- The quantum gravitational integrand (109) contains four factors of energy in the numerator as opposed to only two in the quantum electrodynamic numerator (108); and

- All four of the energy factors $\mathcal{E}^4$ in the quantum gravitational integrand (109) can involve $\vec{k}$ whereas the two factors of $\hbar c\|\vec{p}\|$ in the quantum electrodynamic integrand (108) do not.

The first difference derives from the fact that stress energy is the source for gravity in general relativity, whereas the source for electromagnetism is charge. The second difference derives from the fact that the 0-point energy of a bosonic mode such as a photon is always positive (the 0-point energy of a fermionic meode such as an electron is always negative) whereas the same wave vector $\vec{k}$ contributes both positive and negative charges. That is why one gets no coherent effect from the undisturbed charges of a neutral medium; the lowest coherent effect in quantum electrodynamics comes from the *difference* of positive and negative charges subjected to an electric field, whereas quantum general relativity receives a coherent effect from the 0-point energy of each mode.

The two differences I have noted explain why the divergences of quantum general relativity are so much worse than those of quantum electrodynamics. To be quantitative, let us evaluate the $\vec{k}$ mode sum of (109) with the second possibility from the list (110) for the numerator factors $\mathcal{E}^4$. After some tedious but standard reductions we reach the form

$$\frac{8\pi G}{c^4}\int\frac{\mathrm{d}^3k}{(2\pi)^3}\frac{[\hbar^2c^2\vec{k}\cdot(\vec{p}-\vec{k})]^2}{[\hbar c\|\vec{k}\|+\hbar c\|\vec{p}-\vec{k}\|]^3}$$
$$=\frac{3G\hbar}{\pi c^3}\int_0^1\mathrm{d}x\,x(1-x)$$
$$\times\int_0^\infty\mathrm{d}k k^2\frac{[k^4+(\frac{1}{3}-\frac{10}{3}x+\frac{10}{3}x^2)k^2p^2+x^2(1-x)^2p^4]}{[k^2+x(1-x)p^2]^{\frac{3}{2}}}. \tag{111}$$

Expression (111) is a quarticly divergent integral, which is typical of the first order corrections in quantum general relativity. To make it well defined I will cut off the upper limit at $k=K$. This procedure is an example of a technique known as *regularization* which is employed to make mathematical sense of the divergences of quantum field theory before they are removed by renormalization. I should really have done it at the beginning of the computation, with a better technique, and used it consistently throughout (which is, rest assured, the standard practice) but this level of rigor is superfluous if one just wants an explanation of the problem without precise numbers.

Once regulated, expression (111) becomes well defined. We can evaluate it exactly but it suffices to give the expansion for large K,

$$\frac{3G\hbar}{\pi c^3}\int_0^1\mathrm{d}x\,x(1-x)$$
$$\times\int_0^K\mathrm{d}k k^2\frac{[k^4+(\frac{1}{3}-\frac{10}{3}x+\frac{10}{3}x^2)k^2p^2+x^2(1-x)^2p^4]}{[k^2+x(1-x)p^2]^{\frac{3}{2}}}$$
$$=\frac{3G\hbar}{\pi c^3}\left\{\frac{K^4}{4}-\frac{K^2p^2}{24}+\frac{61}{2^4\cdot7!!}p^4\ln\left(\frac{2K}{p}\right)\right.$$
$$\left.-\frac{190\,92\,1}{2^7\cdot(7!!)^2}p^4+O\left(\frac{p^6}{K^2}\right)\right\}. \tag{112}$$

All contributions to (109) have this same form. It is useful to break them up into a part that just has positive powers of p^2 but diverges when K goes to infinity and another part which can have logarithms of p or even inverse powers, but remains finite. For the logarithm term this breakup requires the introduction of a fixed length scale L,

$$\ln\left(\frac{2K}{p}\right)=\ln(2LK)-\ln(Lp)$$
$$=\frac{1}{2}\ln(L^2K^2)-\frac{1}{2}\ln(L^2p^2)+\ln(2). \tag{113}$$

We can therefore express the first order electromagnetic correction to the Einstein equations as the sum of divergent and finite parts having the form

$$\mathcal{G}^{\rm div}_{\mu\nu}(\vec{x})=\frac{G\hbar}{c^3}\int\frac{\mathrm{d}^3p}{(2\pi)^3}\mathrm{e}^{\mathrm{i}\vec{p}\cdot\vec{x}}\{AK^4+BK^2p^2$$
$$+C\ln(L^2K^2)\,p^4\}\times\tilde{h}(\vec{p}) \tag{114}$$
$$=\frac{G\hbar}{c^3}\{AK^4h(\vec{x})-BK^2\nabla^2h(\vec{x})$$
$$+C\ln(L^2K^2)\nabla^4h(\vec{x})\}, \tag{115}$$
$$\mathcal{G}^{\rm fin}_{\mu\nu}(\vec{x})=\frac{G\hbar}{c^3}\int\frac{\mathrm{d}^3p}{(2\pi)^3}\mathrm{e}^{\mathrm{i}\vec{p}\cdot\vec{x}}\{-C\ln(L^2p^2)p^4+Dp^4\}$$
$$\times\tilde{h}(\vec{p}). \tag{116}$$

Here A, B, C and D are pure numbers of order one and I am still suppressing the indices of the field $h_{\mu\nu}$. Had I considered 0-point contributions from a massive field, such as the electron, the resulting $\mathcal{G}^{\rm fin}_{\mu\nu}$ would have had a more complicated structure involving the mass, but the divergent part would have had the same form as (115).

One might think it is the most highly divergent parts of (115) which cause problems for quantum general relativity but that is not so. We can see this by taking the static, linearized limit of the Einstein equations (9), and including the effect we have just discussed from quantum 0-point motions,

$$[\nabla^2h+\Lambda h]_{\mu\nu}=\mathcal{G}_{\mu\nu}+\frac{8\pi G}{c^4}(T_{\mu\nu})_{\rm linear}. \tag{117}$$

(Because $\mathcal{G}_{\mu\nu}$ already includes the effect of matter 0-point motions one should think of the linearized stress–energy tensor as a classical source, just like $\rho_{\rm free}$ in our discussion (78) of electromagnetic polarization.) Now multiply (117) by $c^4/8\pi G$ and bring the $\mathcal{G}_{\mu\nu}$ terms to the left-hand side, as we do for electromagnetic polarization,

$$\left\{-\frac{C\hbar c\ln(L^2K^2)}{8\pi}\nabla^4h+\left[\frac{c^4}{8\pi G}+\frac{B\hbar cK^2}{8\pi}\right]\nabla^2h\right.$$
$$\left.+\left[\frac{c^4\Lambda}{8\pi G}-\frac{A\hbar cK^4}{8\pi}\right]h\right\}_{\mu\nu}+\frac{c^4}{8\pi G}\mathcal{G}^{\rm fin}_{\mu\nu}=\Big(T_{\mu\nu}\Big)_{\rm linear}. \tag{118}$$

We see that the quadratic and quartic divergences can be absorbed into renormalizations of Newton's constant and the

cosmological constant,

$$\frac{c^4}{8\pi G} + \frac{B\hbar c K^2}{8\pi} \equiv \left(\frac{c^4}{8\pi G}\right)_{\rm meas}, \tag{119}$$

$$\frac{c^4\Lambda}{8\pi G} - \frac{A\hbar c K^4}{8\pi} \equiv \left(\frac{c^4\Lambda}{8\pi G}\right)_{\rm meas}. \tag{120}$$

Just as neither ϵ_0 nor $\Delta\epsilon(0)$ is separately observable in electrodynamics, so it is only the combinations (119) and (120) which are observable in gravity.

Alas, there is no parameter in general relativity with which to absorb the logarithmic divergence! If only there were then the remaining, finite quantum gravitational effects from $\mathcal{G}^{\rm fin}_{\mu\nu}$ would be unobservably small. For example, the fractional change in Earth's surface gravity due to quantum gravitational effects would be about

$$\frac{G\hbar}{c^3 R_{\rm E}^2} \ln\left(\frac{L^2}{R_{\rm E}^2}\right) \sim 10^{-83} \times \ln\left(\frac{L^2}{R_{\rm E}^2}\right), \tag{121}$$

which is negligible even if L is chosen to be the Planck length of about 10^{-35} m. However, infinity is not small, and that is what one gets from the logarithmic divergence. One can only absorb it if new, 4th derivative terms are added to the gravitational field equations, but I will show in the next subsection that doing so would make the universe blow up. That is the fundamental obstacle to making sense of perturbative quantum general relativity.

Before concluding the discussion of renormalization I need to comment on three more issues. The first is that Einstein's equations are not linear in the field $h_{\mu\nu}$ and I have only considered quantum corrections which are linear in this field. Might there be additional divergences on nonlinear terms? There are indeed such divergences but the conservation of stress energy prescribes how the various powers of $h_{\mu\nu}$ can enter the field equations. With either zero derivatives or two derivatives of the full metric the results are unique,

$$\partial^0 \Longrightarrow \Lambda g_{\mu\nu}, \tag{122}$$

$$\partial^2 \Longrightarrow [A(g)\partial^2 g + B(g)\partial g \partial g]_{\mu\nu}. \tag{123}$$

These correspond to the two terms in the Einstein equations (9), and knowing their linear parts in $h_{\mu\nu}$ dictates all higher (and lower) powers as well.

The second point is that I have so far worked in the static limit. That is not even correct for classical electrodynamics! Real media cannot polarize infinitely rapidly, so dielectric response is frequency dependent. To compute the actual polarization one must first Fourier transform the electric field in time, as well as space, then multiply by the frequency and wave vector dependent permittivity $\epsilon(\omega, \vec{k})$, and only then transform back,

$$\vec{P}(t,\vec{x}) = \int \frac{{\rm d}\omega}{2\pi}\, {\rm e}^{-{\rm i}\omega t} \int \frac{{\rm d}^3k}{(2\pi)^3}\, {\rm e}^{{\rm i}\vec{k}\cdot\vec{x}} \times \epsilon(\omega,\vec{k}) \times \int {\rm d}t'\, {\rm e}^{{\rm i}\omega t'} \int {\rm d}^3x'\, {\rm e}^{-{\rm i}\vec{k}\cdot\vec{x}'}\, \vec{E}(t,\vec{x}'). \tag{124}$$

This obviously raises issues about causality! Those issues can all be resolved but doing so in quantum field theory involves a technique known as the 'Schwinger–Keldysh formalism' which even many particle theorists do not understand. (The original literature is [16–18]; for some nice reviews see [19].) They have evolved a series of tricks to avoid having to think about it and, although I do understand the technique, I employed such a trick to avoid a lengthy (and probably not very illuminating) digression to explain it. The trick was to study the static limit of no time dependence and then appeal to the fact that the Einstein equation is the unique combination of the metric and no more than two derivatives which is consistent with stress–energy conservation. So the renormalizations (119) and (120) which were found in the static limit of only spatial derivatives must apply as well for space and time derivatives in the full theory.

Finally, I should comment that stress–energy conservation allows two linearly independent combinations of four derivatives of the metric. Each of these terms has the general form

$$[A(g)\partial^4 g + B(g)\partial g\partial^3 g + C(g)\partial^2 g\partial^2 g + D(g)\partial g\partial g\partial^2 g + E(g)\partial g\partial g\partial g\partial g]_{\mu\nu}. \tag{125}$$

They are called the 'R^2 counterterm' and the 'C^2 counterterm' after the curvature scalars that comprise the Lagrangian densities from which they descend. The difference between them has to do with how the indices are arranged, which I have suppressed. As with (122) and (123), knowing just the linearized parts $\partial^4 h$ fixes all other powers.

2.5. The problem with higher derivatives

We have seen that the renormalization of perturbative quantum general relativity requires that the equations of motion be changed to include terms with up to four derivatives. Stelle has shown that if such terms are added to gravity (which we cannot any longer call general relativity) then the resulting quantum theory is perturbatively renormalizable [5]. However, it is also subject to a virulent instability that is totally inconsistent with the observed reality of a universe which is 13.7 billion years old.

This result is very old, and not specific to gravity; it was obtained in 1850 by the great Russian physicist Ostrogradsky [20]. Ostrogradsky's theorem is that there is a linear instability in the Hamiltonians associated with Lagrangians which depend upon more than one time derivative in such a way that the dependence cannot be eliminated by partial integration [20]. The result is so general that I can simplify the discussion by presenting it in the context of a single, one-dimensional point particle whose position as a function of time is $q(t)$.

In the usual case of $L = L(q,\dot{q})$, the Euler–Lagrange equation is

$$\frac{\partial L}{\partial q} - \frac{\rm d}{{\rm d}t}\frac{\partial L}{\partial \dot{q}} = 0. \tag{126}$$

The assumption that $\partial L/\partial\dot{q}$ depends upon $\dot{q}$ is known as *nondegeneracy*. If the Lagrangian is nondegenerate we can write (126) in the form Newton originally laid down for the laws of physics

$$\ddot{q} = \mathcal{F}(q,\dot{q}) \Longrightarrow q(t) = \mathcal{Q}(t, q_0, \dot{q}_0). \tag{127}$$

From this form it is apparent that solutions depend upon two pieces of initial value data: $q_0 = q(0)$ and $\dot{q}_0 = \dot{q}(0)$.

The fact that solutions require two pieces of initial value data means that there must be two canonical coordinates, Q and P. They are traditionally taken to be

$$Q \equiv q \qquad \text{and} \qquad P \equiv \frac{\partial L}{\partial \dot{q}}. \tag{128}$$

The assumption of nondegeneracy is that we can invert the phase space transformation (128) to solve for $\dot{q}$ in terms of Q and P. That is, there exists a function $v(Q, P)$ such that

$$\left.\frac{\partial L}{\partial \dot{q}}\right|_{\substack{q = Q \\ \dot{q} = v}} = P. \tag{129}$$

For example, one finds $v(Q, P) = P/m$ for the harmonic oscillator (19).

The canonical Hamiltonian is obtained by Legendre transforming on $\dot{q}$,

$$H(Q, P) \equiv P\dot{q} - L, \tag{130}$$

$$= Pv(Q, P) - L(Q, v(Q, P)). \tag{131}$$

It is easy to check that the canonical evolution equations reproduce the inverse phase space transformation (129) and the Euler–Lagrange equation (126),

$$\dot{Q} \equiv \frac{\partial H}{\partial P} = v + P\frac{\partial v}{\partial P} - \frac{\partial L}{\partial \dot{q}}\frac{\partial v}{\partial P} = v, \tag{132}$$

$$\dot{P} \equiv -\frac{\partial H}{\partial Q} = -P\frac{\partial v}{\partial Q} + \frac{\partial L}{\partial q} + \frac{\partial L}{\partial \dot{q}}\frac{\partial v}{\partial Q} = \frac{\partial L}{\partial q}. \tag{133}$$

This is what we mean by the statement, 'the Hamiltonian generates time evolution'. When the Lagrangian has no explicit time dependence, H is also the associated conserved quantity. Hence it is 'the' energy by any usual standard, of course up to canonical transformation.

Now consider a system whose Lagrangian $L(q, \dot{q}, \ddot{q})$ depends nondegenerately upon $\ddot{q}$. The Euler–Lagrange equation is

$$\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}} + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial L}{\partial \ddot{q}} = 0. \tag{134}$$

Non-degeneracy implies that $\partial L/\partial \ddot{q}$ depends upon $\ddot{q}$, in which case we can cast (134) in a form radically different from Newton's,

$$q^{(4)} = \mathcal{F}(q, \dot{q}, \ddot{q}, q^{(3)}) \Longrightarrow q(t) = \mathcal{Q}(t, q_0, \dot{q}_0, \ddot{q}_0, q_0^{(3)}). \tag{135}$$

Because solutions now depend upon four pieces of initial value data there must be four canonical coordinates. Ostrogradsky's choices for these are

$$Q_1 \equiv q, \qquad P_1 \equiv \frac{\partial L}{\partial \dot{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \ddot{q}}, \tag{136}$$

$$Q_2 \equiv \dot{q}, \qquad P_2 \equiv \frac{\partial L}{\partial \ddot{q}}. \tag{137}$$

The assumption of nondegeneracy is that we can invert the phase space transformation (136) and (137) to solve for $\ddot{q}$ in terms of Q_1, Q_2 and P_2. That is, there exists a function $a(Q_1, Q_2, P_2)$ such that

$$\left.\frac{\partial L}{\partial \ddot{q}}\right|_{\substack{q = Q_1 \\ \dot{q} = Q_2 \\ \ddot{q} = a}} = P_2. \tag{138}$$

Note that one only needs the function $a(Q_1, Q_2, P_2)$ to depend upon *three* canonical coordinates—and not all four—because $L(q, \dot{q}, \ddot{q})$ only depends upon three configuration space coordinates. This simple fact has great consequence.

Ostrogradsky's Hamiltonian is obtained by Legendre transforming, just as in the first derivative case, but now on $\dot{q} = q^{(1)}$ and $\ddot{q} = q^{(2)}$,

$$H(Q_1, Q_2, P_1, P_2) \equiv \sum_{i=1}^{2} P_i q^{(i)} - L \tag{139}$$

$$= P_1 Q_2 + P_2 a(Q_1, Q_2, P_2)$$

$$-L(Q_1, Q_2, a(Q_1, Q_2, P_2)). \tag{140}$$

The time evolution equations are just those suggested by the notation,

$$\dot{Q}_i \equiv \frac{\partial H}{\partial P_i} \qquad \text{and} \qquad \dot{P}_i \equiv -\frac{\partial H}{\partial Q_i}. \tag{141}$$

Let us check that they generate time evolution. The evolution equation for Q_1,

$$\dot{Q}_1 = \frac{\partial H}{\partial P_1} = Q_2, \tag{142}$$

reproduces the phase space transformation $\dot{q} = Q_2$ in (137). The evolution equation for Q_2,

$$\dot{Q}_2 = \frac{\partial H}{\partial P_2} = a + P_2\frac{\partial a}{\partial P_2} - \frac{\partial L}{\partial \ddot{q}}\frac{\partial a}{\partial P_2} = a, \tag{143}$$

reproduces (138). The evolution equation for P_2,

$$\dot{P}_2 = -\frac{\partial H}{\partial Q_2} = -P_1 - P_2\frac{\partial a}{\partial Q_2} + \frac{\partial L}{\partial \dot{q}} + \frac{\partial L}{\partial \ddot{q}}\frac{\partial a}{\partial Q_2}$$

$$= -P_1 + \frac{\partial L}{\partial \dot{q}}, \tag{144}$$

reproduces the phase space transformation $P_1 = (\partial L/\partial \dot{q}) - (\mathrm{d}/\mathrm{d}t)(\partial L/\partial \ddot{q})$ (136). And the evolution equation for P_1,

$$\dot{P}_1 = -\frac{\partial H}{\partial Q_1} = -P_2\frac{\partial a}{\partial Q_1} + \frac{\partial L}{\partial q} + \frac{\partial L}{\partial \ddot{q}}\frac{\partial a}{\partial Q_1} = \frac{\partial L}{\partial q}, \tag{145}$$

reproduces the Euler–Lagrange equation (134). So Ostrogradsky's system really does generate time evolution. When the Lagrangian contains no explicit dependence upon time it is also the conserved Noether current. By any standard definition, it is therefore 'the' energy, again up to canonical transformation.

There is one, overwhelmingly bad thing about Ostrogradsky's Hamiltonian (140): *it is linear in the canonical momentum* P_1. Note the power and generality of the result. It applies to *every* Lagrangian $L(q, \dot{q}, \ddot{q})$ which depends nondegenerately upon $\ddot{q}$, independent of the details. The only assumption is nondegeneracy, and that simply means one cannot eliminate

$\ddot{q}$ by partial integration. This is why Newton was right to assume the laws of physics take the form (127) when expressed in terms of fundamental dynamical variables.

The Ostrogradskian instability is not a problem with the potential energy, in which the dynamical variable can reach arbitrarily negative energies by approaching a special value. It is instead an instability of the kinetic energy in which arbitrarily negative energies are associated with a special sort of time dependence. The problematic term in Ostrogradsky's Hamiltonian (140) is $P_1 Q_2$. One makes its large by adjusting the third time derivatives in P_1, which can be done while the dynamical variable $q(t)$ is still quite small.

At this point I need to debunk the misconception that unstable systems decay 'because the system wants to lower its energy'. Total energy is conserved in fundamental theory, so the decay of an excited atomic state into the ground state atom plus some photons leads to no change in energy. What drives the decay is entropy: quantum systems explore the space of classical configurations which have the same energy, and there are many more ways for an atom to decay as opposed to only one for it not to decay. For example, the decay photons can go off in any direction.

This insight about what drives decays means that the Ostrogradskian instability is *instantly* fatal for an interacting field theory. Recall that each Fourier mode of such a theory contributes its own, independent kinetic energy, and there are an infinite number of such Fourier modes. One's usual intuition about modes with large $\|\vec{k}\|$ is completely wrong for theories which possess the Ostrogradskian instability. We are used to thinking that these modes cannot be excited because doing so costs a large energy and there is only a finite free energy available. But that is only true when all modes carry positive energy. Exciting a negative energy mode *frees up energy*, which can then be used to excite positive energy modes. And exciting a negative energy mode with higher $\|\vec{k}\|$ frees up even more energy. Now use expression (54) to count up the number of modes per unit volume which have $\|\vec{k}\| < K$,

$$\int \frac{\mathrm{d}^3k}{(2\pi)^3}\theta(K - \|\vec{k}\|) = \frac{K^3}{6\pi^2}. \tag{146}$$

Of course there is no limit on K because the higher modes actually participate more strongly when the Ostrogradskian instability is present. So one can see that an interacting field theory with the Ostrogradskian instability decays instantly, no matter how weak the interaction is.

Adding even more higher derivatives to the Lagrangian just makes the problem worse. Suppose the Lagrangian depends on q and its first N derivatives, $L(q, \dot{q}, \ldots, q^{(N)})$. Ostrogradsky's choice for the canonical variables is

$$Q_i \equiv q^{(i-1)}, \qquad P_i \equiv \sum_{j=0}^{N-i} \left(-\frac{\mathrm{d}}{\mathrm{d}t}\right)^j \frac{\partial L}{\partial q^{(i+j)}}. \tag{147}$$

Provided the equation for P_N can be inverted to solve for $q^{(N)}$ in terms of P_N and the Q_i's, the Hamiltonian is linear in the first $N-1$ momenta,

$$H = \sum_{i=1}^{N-1} P_i Q_{i+1} + P_N q^{(N)}(\vec{Q}, P_N) - L(Q_1, \ldots, Q_N, q^{(N)}(\vec{Q}, P_N)). \tag{148}$$

One might think that a fully nonlocal Lagrangian must inherit the Ostrogradskian instability but this is not necessarily so. If the nonlocality is restricted to entire functions of the derivative operator—such as $\mathrm{e}^{T^2\partial_t^2}$—then the nonlocal Lagrangian can be viewed as the limit of a sequence of ever higher derivative Lagrangians, and it indeed possesses the Ostrogradskian instability. However, if the nonlocality involves poles or cuts which invalidate the expansion in powers of the derivative operator, then there need be no problem. This is precisely what happens when a local degree of freedom is integrated out. For example, consider the coupled oscillator system

$$L = \tfrac{1}{2}m(\dot{q}_1^2 + \dot{q}_2^2) - \tfrac{1}{2}m\omega^2(q_1^2 + 2gq_1q_2 + q_2^2). \tag{149}$$

For small g this is a stable system with frequencies $\omega_\pm^2 = (1 \pm g)\omega^2$. Integrating out the variable $q_2(t)$ results in a nonlocal Lagrangian

$$L \longrightarrow \frac{1}{2}m\dot{q}_1^2 - \frac{1}{2}m\omega^2 q_1^2 + \frac{1}{2}m\omega^2 q_1\left[\frac{g^2\omega^2}{(\mathrm{d}/\mathrm{d}t)^2 + \omega^2}\right]q_1. \tag{150}$$

This nonlocal system is obviously stable as well; it fails to inherit the Ostrogradskian instability because the pole at $\mathrm{d}/\mathrm{d}t = \pm\mathrm{i}\omega$ prevents one from representing it as the limit of a sequence of ever higher derivative Lagrangians.

A final point is that the problematic term $P_1 Q_2$ of the Ostrogradskian Hamiltonian (140) can have either sign. We have been concerned with the fact that it can be arbitrarily negative, but it can also be arbitrarily positive. This means that two things happen when you add a higher derivative term to a lower derivative theory:

- There can be changes in the original, lower derivative degrees freedom and
- The higher derivative term introduces new degrees of freedom which carry the opposite kinetic energy to the changed, lower derivative degrees of freedom.

The usual case is that the lower derivative theory has positive energy, so adding a higher derivative induces negative energy degrees of freedom. That is what happens with the C^2 counterterm; it adds a negative energy, spin two graviton which would make the universe decay instantly. However, it turns out that the R^2 counterterm adds a positive energy, spin zero particle which is harmless. This represents no violation of Ostrogradsky's theorem because the spin zero part of the metric in general relativity carries negative energy. It is better known as the Newtonian potential and its negative energy poses no problem for stability because this part of the metric is completely determined by the stress–energy tensor. The new spin zero degree of freedom induced by the R^2 counterterm is an independent, purely gravitational degree of freedom, just like the gravitons.

To summarize, although adding the R^2 counterterm to general relativity would be no problem, adding the C^2 counterterm would make the universe blow up. If only we did not need the C^2 counterterm! But careful analyses show that we do need it for scalar particles such as the Higgs [2],

for electromagnetism [3] and for the particles which carry the weak and strong interactions [4].

The situation for pure gravity is more complicated but no more satisfactory. Virtual gravitons induce the C^2 counterterm at first order in perturbation theory. However, treating this term perturbatively at first order means we can evaluate it at the zeroth order solution, and it is straightforward to show that doing this gives zero as long as the cosmological constant is zero and no matter is present [2]. (The same thing is true as well for the R^2 counterterm.) So pure gravity with zero cosmological constant is actually finite at first order, which is why I have emphasized that the basic problem of quantum gravity concerns quantum matter effects which must be present, whether or not there is gravitational radiation or it is quantized. That is not to say gravitons pose no problems. A heroic second order computation by Goroff and Sagnotti [6], verified by van de Ven [7], demonstrates that they induce a higher derivative counterterm as unacceptable as the C^2 counterterm, and which fails to vanish with the zeroth order field equations.

2.6. *The impact of primordial inflation on two fixes*

Inflation is defined as a period of accelerated cosmological expansion. We know this can happen because it is taking place right now [8, 9]. Guth has proposed that a very early phase of *primordial inflation* would explain why the current universe is so nearly homogeneous and isotropic on the largest scales, why it is so nearly spatially flat, and why it contains no exotic relics such as magnetic monopoles (which typically occur when all the forces are unified) and primordial black holes [21]. There is a lot of evidence in favor of this idea and none against it, although physicists are not yet ready to regard it as proven. I shall have a lot more to say about primordial inflation in section 5 but let me for now explore the consequences for quantum gravity of three tenets of primordial inflation:

1. Quantum gravitational fluctuations in the metric at the beginning of inflation were no larger than about one part in 10^6;
2. The universe has expanded by a factor of at least 10^{51} and
3. The structures of today's universe derived from 13.7 billion years of gravitational collapse into the tiny (one part in about 10^5) inhomogeneities provided by quantum fluctuations of the stress–energy tensor near the end of inflation.

Assumptions 1–3 can be used to rule out two proposals which are sometimes advanced for resolving the apparent inconsistency of perturbative quantum general relativity:

- Regard *all* components of the metric as classical and change the source of gravity from the quantum stress–energy tensor to its expectation value in some state or
- Regard space as discrete at some very small length scale.

Of course assumption #3 immediately falsifies the first proposal. If inflation is correct then the expectation value of the stress–energy tensor at the end of inflation cannot retain inhomogeneities of more than about one part in 10^{78}, otherwise they would have been so big at the beginning of inflation that gravitational collapse would have ensued[2]. But the measured strength of primordial perturbations in the cosmic microwave background and in the matter density is about one part in 10^5 [22, 23]. I remark in passing that, if inflation is correct, primordial perturbations are the first ever data from quantum gravity. One can see how having these data has shifted debates over quantum gravity from philosophy and aesthetics to the interpretation of hard evidence. Quantum gravity is coming of age.

One has to work a little harder to debunk discretization, and I need to be clear that I mean only the most naive form of discretization in which the number of spatial points in a fixed coordinate volume becomes finite and no new degrees of freedom are added in the course of time evolution. Some models which contain an infinite number of degrees of freedom are said to endow spacetime with a 'discrete structure' below some scale due either to changes in the field equations or to nonperturbative effects. There is nothing necessarily wrong with these models, but I choose to regard them as either modifications of general relativity or of the perturbative computational technique. What I wish to focus on here is models in which neither thing is done and one instead changes only the continuum nature of space.

We only have direct evidence for the continuum nature of space down to about 10^{-19} m [24]. If space was discrete at some smaller scale ΔL the strength of quantum gravitational corrections would be roughly what one gets from taking the spatial momentum cutoff K to be $1/\Delta L$. That would remove the divergences but one also has to keep quantum gravitational corrections sufficiently small and, it turns out that discretization cannot accomplish this if one accepts primordial inflation.

First note that the scale of discreteness ΔL cannot be much smaller than about the Planck length of $[\hbar G/c^3]^{\frac{1}{2}} \sim 1.6 \times 10^{-35}$ m. This might seem surprising in view of the first order results (115),

$$\mathcal{G}^{\rm div}_{\mu\nu} = \frac{G\hbar}{c^3}\{AK^4 h(\vec{x}) - BK^2\nabla^2 h(\vec{x}) + C\ln(L^2K^2)\nabla^4 h(\vec{x})\}. \tag{151}$$

Just because $K \sim 1/\Delta L$ will stay finite does not preclude renormalization, so we can still absorb the K^4 contribution into a shift of the cosmological constant, and the K^2 contribution into a change in the Newton constant. That leaves only the $\ln(L^2K^2)$ correction, which would be minuscule at low scales, even if the argument of the logarithm is enormous.

However, one must consider higher order corrections. If one does not add 4th derivative terms to the classical field equations then the divergences of quantum gravity grow worse as the order of perturbation theory increases. For example, at

[2] The number $10^{-78} = (10^{-26})^3$ comes from assuming that the universe expanded by a factor of 10^{26} during primordial inflation (half the total expansion), that perturbations were initially less than order one, and that they obey the equation of state of nonrelativistic matter. For a relativistic equation of state the reduction would be even larger, $10^{-106} = (10^{-26})^4$. More details about the connection between the equation of state and the expansion factor can be found in section 5.2.

second order the divergences would take the form

$$\mathcal{G}^{\rm div}_{\mu\nu} = \left(\frac{G}{\hbar c^5}\right)^2 \{AK^6 + BK^4\nabla^2 + CK^2\nabla^4 + D\ln(L^2K^2)\nabla^6\}h(\vec{x}), \quad (152)$$

where A, B, C and D are numbers of order one. One can absorb the K^6 divergence in the cosmological constant and the K^4 divergence in the Newton constant, but the other two divergences must be regulated by the cutoff $K \sim 1/\Delta L$. The term proportional to $\nabla^4 h(\vec{x})$ could be written as the first order term times an extra factor of $GK^2/\hbar c^5$,

$$\left(\frac{G}{\hbar c^5}\right)^2 \times CK^2\nabla^4 h(\vec{x}) = \frac{GK^2}{\hbar c^5} \times \left\{\frac{G}{\hbar c^5} \times C\nabla^4 h(\vec{x})\right\}. \quad (153)$$

The term in brackets is roughly the same strength as the $\nabla^4 h$ term we got at first order, and very small under normal circumstances, but the initial factor of $GK^2/\hbar c^5$ would be huge if scale of discretization drops much below the Planck length. The third order correction would contain two factors of this huge number, etc. The only way to avoid eventually getting an unacceptably large quantum correction is to prevent the cutoff scale ΔL from dropping below the Planck Length.

Now consider assumption #1 from primordial inflation, namely that quantum gravitational effects were small at the beginning of inflation. This means that, in terms of the physical length measured at the beginning of inflation, the scale of discretization cannot have been below about 10^{-35} m. But assumption #2 says that the universe has expanded by a factor of at least 10^{51} since the beginning of inflation. That means the physical length between discrete points (the number of which cannot change with time) ought to be about 10^{16} m today! That is roughly the distance between stars in our part of the Milky Way galaxy, and of course utterly inconsistent with current, direct checks of continuum spacetime to about 35 orders of magnitude smaller. Turning the argument around, for the current physical length of discreteness to be less than 10^{-19} m, its value at the beginning of primordial inflation must have been 10^{-70} m, which is 35 orders of magnitude too small to explain why quantum gravitational effects are small during inflation.

Note that the same argument can be invoked for *any* early event during which we have reason to believe quantum gravitational fluctuations were small. Some events and the associated cosmological expansion factors are:

$$\text{Recombination} \Longrightarrow 10^3, \quad (154)$$

$$\text{Nucleosynthesis} \Longrightarrow 10^9, \quad (155)$$

$$\text{Quark gluon plasma} \Longrightarrow 10^{12}, \quad (156)$$

$$\text{Electroweak symmetry breaking} \Longrightarrow 10^{15}. \quad (157)$$

Primordial inflation provides a vastly stronger bound because it came much earlier, but the expansion since electroweak symmetry breaking (which may well have seen the formation of the asymmetry between matter and anti-matter) lacks only an order of magnitude to connect the Planck length to the current experimental bound on discreteness. We therefore conclude that, while spacetime may well be discrete at some, very small scale, this cannot explain what is suppressing quantum gravitational effects.

3. General reactions to the problem

We have seen that the problem of quantum gravity arises from a conflict between four physics principles:

- *Continuum field theories* possess an infinite number of modes;
- *Quantum mechanics* requires each mode to have a minimum amount of energy;
- *General relativity* stipulates that stress energy is the source of gravitation and
- *Perturbation theory* simply adds up the contribution from each mode at lowest order.

When a problem can be shown to derive from a well-defined set of propositions then one or more of these propositions must be wrong. In the previous section I argued that it cannot be the first two. Although spacetime may well be discrete at some level, the expansion of the universe implies that this discreteness must be at too small a scale to be useful for making sense of quantum gravity. And the lowest order divergences of quantum gravity derive from the quantum properties of matter, which have been too thoroughly checked to abandon. It follows that the problem must lie either with general relativity or with the use of perturbation theory. The great fault line which divides fundamental theorists is which of these two is held suspect.

3.1. Particle theorists versus relativists

The two major schools of thought on quantum gravity consist of those who approach the subject from the perspective of particle theory and those who approach it from the perspective of classical general relativity[3]. Particle theorists are much attached to perturbation theory, so they are willing to alter general relativity. This is what led to supergravity and superstring theory, and to study of on-shell finiteness and asymptotic safety. Relativists are equally attached to general relativity, so they are willing to ignore perturbative problems. This is what has led to loop quantum gravity.

Both views reflect the body of experience of those who hold them. The history of particle physics has involved aggressively using perturbation theory to derive predictions for proposed models of fundamental interactions, and then ruthlessly discarding any model which failed to agree with observation and experiment. Among particle theorists the 'crackpots' were those who became too attached to a particular model and either failed to check it using perturbation theory or else refused to abandon the model when perturbative checks indicated a problem.

It should also be mentioned that particle theorists are much attracted to the idea of unifying gravity with the other forces. This led to progress, first with Maxwell's unified theory of electricity and magnetism and, a century later, with the unification of the weak and electromagnetic forces for which Weinberg, Salam and Glashow shared the 1979 Nobel Prize.

[3] There are of course exceptions: string theorists who were trained as relativists and loop quantum gravity researchers who were trained in particle theory. I hope they will not take offense at the names I have chosen to characterize their disciplines.

And the fact that the electroweak and strong coupling constants become equal at about 10^{15} GeV, and that this energy is close to the Planck energy, seems (to particle theorists) to point to a fully unified theory at very high energies.

Relativists come to quantum gravity with a completely different historical perspective. For them general relativity is a model which has stood the test of time. Whenever people thought there was a problem which necessitated changing general relativity it turned out, upon closer examination of either the theory or the data, that general relativity was right and the proposed changes were wrong. Among relativists the people who made mistakes were those who tinkered with the model on the basis of incomplete data or anything less than a rigorous theoretical analysis. The first example was none other than Albert Einstein, who in 1917 introduced the cosmological constant Λ into his gravitational field equations (9) in order to prevent the universe from expanding in the simplest cosmological realization of general relativity. Of course Hubble actually quantified this expansion in 1929 [25], and Einstein could have predicted it had he just stuck with his original formulation of general relativity. He called this the greatest blunder of his life.

Relativists are also familiar with a vast collection of 'paradoxes' which purport to show that either special or general relativity is wrong, and which can only be debunked by carefully identifying false assumptions. So it seems very natural to a relativist to reject the result of an asymptotic series expansion, especially when divergences are present. They distrust the idea of considering something as 'small' when it is actually divergent, and they will not be satisfied that there is anything wrong with quantum general relativity until a rigorous proof is supplied which is not based on perturbation theory.

I am myself a particle theorist but I have friends in both camps and it is sometimes difficult to make them see any worth in the other side's views. Although I share the fondness of my particle colleagues for perturbation theory, let me reply to an objection they sometimes raise about doubting the validity of perturbative results for quantum general relativity. The objection takes the form of an exasperated question: perturbation theory is supposed to be valid when the corrections it generates are small and, whatever is the right fundamental theory, quantum gravitational corrections must be small at low energies because we have never observed a single one! How then can you people refuse to accept a perturbative treatment of quantum general relativity?

My answer is twofold:

1. There might be low energy quantum gravitational effects which masquerade as something else and
2. The correct asymptotic series expansion for quantum general relativity may involve terms which are not analytic in Newton's constant G, even though they are still very small at low energies.

I will have more to say about the first possibility in the next subsection. Concerning the second, suppose the actual series expansion consists not of just powers of $GE^2/\hbar c^5$ but also logarithms,

$$\sum_{\ell=0}^{\infty}\left(\frac{GE^2}{\hbar c^5}\right)^{\ell}\sum_{k=0}^{\ell}a_{k\ell}\ln\left(\frac{GE^2}{\hbar c^5}\right) = a_{00}+a_{11}\left(\frac{GE^2}{\hbar c^5}\right)\ln\left(\frac{GE^2}{\hbar c^5}\right)+a_{01}\ln\left(\frac{GE^2}{\hbar c^5}\right)+\cdots \tag{158}$$

Note first that trying to beat this into the form of an expansion in powers of G would result in uncontrollable logarithmic divergences, which is what we see in quantum general relativity. Note also that low energy quantum gravitational corrections would still be unobservably small, just not quite as small as without the logarithms. For example, at LHC energies the suppression factor would not be the figure of $\sim 3.4\times 10^{-31}$ we got in expression (69). We would instead get this number times its logarithm,

$$(3.4\times 10^{-31})\cdot\ln(3.4\times 10^{-31})\sim -2.4\times 10^{-29}. \tag{159}$$

The extra factor of about 100 represents an enormous enhancement, but the effect is still many orders of magnitude below observability.

Exotic terms occur in many familiar asymptotic expansions, and divergences are the typical signature of their appearance. Consider the logarithm of the grand canonical partition function for noninteracting, nonrelativistic bosons of mass m in a three-dimensional volume V:

$$\ln(\Xi) = Vn_Q\sum_{k=1}^{\infty}k^{-\frac{5}{2}}\exp(k\beta\mu). \tag{160}$$

Here $n_Q \equiv (mk_{\rm B}T/2\pi\hbar^2)^{\frac{3}{2}}$ is the quantum concentration, $\mu<0$ is the chemical potential and $\beta=(k_{\rm B}T)^{-1}$. Near Bose–Einstein condensation one has $0<-\beta\mu\ll 1$ so it should make sense to expand $\ln(\Xi)$ for small $\beta\mu$. Straightforward perturbation theory corresponds to the following expansion:

$$\ln(\Xi) = Vn_Q\sum_{k=1}^{\infty}k^{-\frac{5}{2}}\sum_{\ell=0}^{\infty}\frac{(k\beta\mu)^{\ell}}{\ell!} \rightarrow Vn_Q\sum_{\ell=0}^{\infty}\frac{(\beta\mu)^{\ell}}{\ell!}\sum_{k=1}^{\infty}k^{\ell-\frac{5}{2}}. \tag{161}$$

Although the $\ell=0$ and $\ell=1$ terms are finite, the sum over k diverges for $\ell\geqslant 2$.

The divergences we have encountered do not mean that higher corrections are large, just that they are not as small as $(\beta\mu)^2$. One sees this by expanding the second derivative around its integral approximation:

$$\frac{1}{Vn_Q}\frac{\partial^2\ln(\Xi)}{\partial(\beta\mu)^2} = \sum_{k=1}^{\infty}k^{-\frac{1}{2}}\exp(k\beta\mu), \tag{162}$$

$$= \int_0^{\infty}{\rm d}y\,y^{-\frac{1}{2}}\exp(y\beta\mu)+\sum_{k=1}^{\infty}\left[k^{-\frac{1}{2}}\exp(k\beta\mu)-\int_{k-1}^{k}{\rm d}y\,y^{-\frac{1}{2}}\exp(y\beta\mu)\right], \tag{163}$$

$$= \left(\frac{-\pi}{\beta\mu}\right)^{\frac{1}{2}}+\sum_{k=1}^{\infty}[k^{-\frac{1}{2}}-2k^{\frac{1}{2}}+2(k-1)^{\frac{1}{2}}]+O(\beta\mu). \tag{164}$$

Integration reveals the true asymptotic expansion:

$$\ln(\Xi) = Vn_Q\{\zeta(\tfrac{5}{2}) + \zeta(\tfrac{3}{2})\beta\mu + \tfrac{4}{3}\sqrt{\pi}(-\beta\mu)^{\frac{3}{2}} + O(\beta^2\mu^2)\}. \quad (165)$$

The oscillating series of ever-increasing divergences in the perturbative expansion (161) has resolved itself into a perfectly finite, fractional power. If only we had the analytical power to check for such behavior in quantum general relativity!

Another example of how perturbation theory can break down in gravity is the Vainshtein mechanism for classical, massive graviton models [26]. In that case conventional perturbation theory fails to apply nearby static, spherically symmetric sources, in spite of the fact that the curvature is small. It should also be noted that some quantum field theories receive exponentially suppressed, but sometimes numerically significant, corrections which are simply invisible to conventional perturbation theory [27].

3.2. *How we would use quantum gravity if we had it*

The unsatisfactory state of affairs in quantum general relativity is that the only computational tool we currently possess offers the choice between two hopelessly incorrect predictions:

- Either quantum gravitational effects are infinitely strong
- Or else the universe blows up instantly.

But there is obviously some quantum theory of gravity, because quantum matter gravitates, and one can discuss what it would tell us if we found it. To be specific, suppose we discovered a consistent and plausible quantum theory of gravity whose dimensionless coupling strength is that of quantum general relativity, $GE^2/\hbar c^5 \sim (E/10^{19}\,\mathrm{GeV})^2$. An irony of this subject is that, as soon as we manage to avoid predicting infinitely strong effects, we are almost inevitably left with no observable predictions at all because $GE^2/\hbar c^5$ is so small for any process which can be contrived in the laboratory!

One might think observable effects could be obtained by resorting to large masses, such as that of the Earth. This indeed gives measurable gravitational effects from the incoherent sum over many sources, but *quantum* gravitational effects derive from the energy of only a single mode. So the energy E appropriate for quantum gravitational corrections to the Earth's potential is not the Earth's enormous rest mass energy of $M_E c^2$ but rather the minuscule energy $E = \hbar c/R$ of the mode whose wave length is the radius R at which the potential is measured. Quantum gravitational corrections to the Earth's classical potential of $-GM_E/R$ take the form [28],

$$\Phi = -\frac{GM_E}{R}\left\{1 + \text{const} \times \frac{G\hbar}{R^2c^3} + \cdots\right\}. \quad (166)$$

At the surface of the Earth the fractional change would be about

$$\frac{G\hbar}{R_E^2 c^3} \sim 10^{-84}. \quad (167)$$

Even if we could measure gravity that accurately (and we cannot), this change is vastly smaller than the classical effect from the mass of the human taking the data!

How to observe a very weak interaction is not an unprecedented problem in the long history of physics. There are two general approaches:

- Find a regime in which the interaction is not so weak or
- Exploit some unique property of the interaction that gives rise to effects for which there is no background from other interactions.

I will comment on both approaches.

The obvious way of overcoming suppression by the factor $GE^2/\hbar c^5$ is to scale up the energy E. We cannot build accelerators which approach interesting energy ranges but nature does this for us in four cases:

- The initial singularity which must occur, on very general grounds, either without primordial inflation [29] or with it [30];
- The final stages of black hole collapse;
- The final stages of black hole evaporation and
- The phase of primordial inflation.

The first three processes can access modes of arbitrarily high energy. The final one might reach as high as $E \sim 10^{13}\,\mathrm{GeV}$, at which quantum gravitational effects would be small but not negligible. So there are good reasons to expect significant quantum gravitational effects in all four cases; the issue is finding some signature of them that reaches us here and now. Ideas about how to do this for the initial singularity and for black hole collapse are still quite speculative, and we have not discovered any black holes near the end of their existence. However, there is by now a well-developed formalism for tracing quantum gravitational effects from primordial inflation to the current epoch. The simplest interpretation of the data from anisotropies in the cosmic microwave background [22] and from large scale structure surveys [23] is that the primordial perturbations in the gravitational potential of our universe arose from quantum matter fluctuations near the end of inflation. I will review the argumentation in section 5.

So much for the first approach; the other technique is to identify processes driven by some special feature of gravity that no other force possesses. So if we see the effect at all, no matter how weak it is, it must be from gravity. For example, particle physicists did not discern the weak nuclear force in the background of vastly stronger quantum electrodynamic processes but rather because it alone mediates decays such as $\mu^- \to \mathrm{e}^- \nu_\mu \bar{\nu}_\mathrm{e}$.

There are four special features of gravity which deserve comment:

- One of the gravitational parameters is the cosmological constant Λ;
- It determines the maximum speed at which signals can propagate;
- Gravitons have zero mass without being driven to zero amplitude by the expansion of the universe and
- The gravitational interaction energy is negative.

The third point is what my own current research concerns and it will make more sense if I postpone it to section 5. I will discuss the first two points briefly and the last one at greater length.

The cosmological constant Λ multiplies a term in the Einstein equations (9) without any derivatives. It influences the rate at which the overall expansion of the universe is accelerating; positive Λ tends to make the universe accelerate whereas negative Λ tends to make it decelerate. (The spatially homogeneous parts of certain matter fields also play a role.) The current universe is accelerating [8, 9], which is consistent with Λ having a small, positive value,

$$(\Lambda)_{\rm meas} \sim +10^{-52}\,{\rm m}^{-2}. \tag{168}$$

The measured value (168) of the cosmological constant is outlandish! From equation (120) one can see that first order quantum gravitational correction has the form $\Delta\Lambda \sim G\hbar K^4/c^3$, where K is the cutoff wave number. Of course one must cancel the divergence when K goes to infinity, but there will obviously be a finite remainder which takes the same form with some finite K. (As particle theorists used to remark during the period renormalization was being worked out, 'Just because something is infinite does not mean it is zero' [31].) The trouble is that the other scales in physics give values for $\Delta\Lambda$ which are vastly larger than (168),

$$\text{Planck Scale}\left(K^2 = \frac{c^3}{G\hbar}\right) \Longrightarrow \Delta\Lambda = \frac{c^3}{G\hbar}$$
$$\sim 10^{121} \times 10^{-52}\,{\rm m}^{-2}, \tag{169}$$

$$Z\text{ Boson Mass}\left(K = \frac{m_Z c}{\hbar}\right) \Longrightarrow \Delta\Lambda = \frac{G c m_Z^4}{\hbar^3}$$
$$\sim 10^{53} \times 10^{-52}\,{\rm m}^{-2}, \tag{170}$$

$$\text{Electron Mass}\left(K = \frac{m_e c}{\hbar}\right) \Longrightarrow \Delta\Lambda = \frac{G c m_e^4}{\hbar^3}$$
$$\sim 10^{32} \times 10^{-52}\,{\rm m}^{-2}. \tag{171}$$

Particle theorists refer to a mismatch of this type as a *hierarchy problem*, and the one associated with the cosmological constant is the worst in all of fundamental theory [32]. Of course Λ is a free parameter in the Einstein equations (9) and it has to take some value, so why not precisely the number which gives (168)? That could be, but many people suspect we are missing a key principle from quantum gravity [33].

Because the metric field $g_{\mu\nu}(t,\vec{x})$ determines physical lengths and times, it controls the maximum rate at which signals can propagate. A tiny quantum fluctuation in the metric could allow some photons of light from a distant galaxy to reach us a little sooner than others. Because we do not know cosmic distances very well the potentially observable effects would be a blurring of images, fluctuations in luminosity and a broadening of spectral lines [35]. Since the earliest days of quantum gravity such effects have been termed *smearing of the light-cone* [36]. Surprisingly, the lowest order contributions can be computed using first order perturbation theory and they give finite results [37]. These results are still unobservably small, but not by much, and they may well be detectable in future laser interferometers [38]. It could be that history repeats itself because the first quantum electrodynamic correction to the electron magnetic moment is finite, and was derived by Schwinger in 1948 [39] while physicists were still puzzling out how to fully absorb all the divergences.

The final special feature of gravity is its negative interaction energy. That raises a fascinating possibility in the context of computing the contribution to a particle's measured mass from the interaction with its own force fields. Every beginning physics student is taught that the electric potential from a collection of charges q_i at fixed positions $\vec{x}_i$ is

$$\Phi(\vec{x}) = \sum_i \frac{q_i}{4\pi\epsilon_0 \|\vec{x} - \vec{x}_i\|}. \tag{172}$$

They are also taught to compute the electrostatic interaction energy between these charges by summing $\frac{1}{2} q_i \Phi(\vec{x}_i)$, where $\Phi(\vec{x}_i)$ is the potential due to all the other charges,

$$E_{\rm EM} = \sum_i \sum_{j\neq i} \frac{q_i q_j}{8\pi\epsilon_0 \|\vec{x}_i - \vec{x}_j\|}. \tag{173}$$

The curious rule about omitting the charge's interaction with its own field (which is the strongest contribution!) derives from the fact that setting $\vec{x} = \vec{x}_i$ in (172) produces a divergent potential. More advanced students are instructed to regulate this divergence and then absorb it into the unobservable bare mass of the particle in such a way as to make the total self-energy of the particle agree with its measured mass. (cf chapter 16 of the text by Jackson [34].)

Arnowitt, Deser and Misner have worked out how this procedure changes, on the classical level, when the gravitational interaction is included [13]. It is simplest to model the particle as a stationary spherical shell of radius R (that is the regularization) charge e and bare mass m_0. In Newtonian gravity the shell's energy would be

$$E_R = m_0 c^2 + \frac{e^2}{8\pi\epsilon_0 R} - \frac{G m_0^2}{2R}. \tag{174}$$

It turns out that all the effects of general relativity are accounted for by replacing E_R/c^2 and m_0 with the full mass m_R,

$$m_R c^2 = m_0 c^2 + \frac{e^2}{8\pi\epsilon_0 R} - \frac{G m_R^2}{2R}$$
$$= \frac{Rc^4}{G}\left[-1 + \sqrt{1 + \frac{2G}{Rc^4}\left(m_0 c^2 + \frac{e^2}{8\pi\epsilon_0 R}\right)}\right]. \tag{175}$$

It should be noted that Arnowitt, Deser and Misner rigorously solved the constraint equations of general relativity and electrodynamics, and then used the asymptotic metric to compute the ADM mass. They also developed the simple model I am presenting [13].

The perturbative result is obtained by expanding the square root,

$$m_{\rm pert} c^2 = m_0 c^2 + \frac{e^2}{8\pi\epsilon_0 R} + \sum_{n=2}^{\infty} \frac{(2n-3)!!}{n!}\left(-\frac{G}{Rc^4}\right)^{n-1}$$
$$\times \left(m_0 c^2 + \frac{e^2}{8\pi\epsilon_0 R}\right)^n, \tag{176}$$

and shows an oscillating series of increasingly singular terms. The alternating signs derive from the fact that gravity is

attractive. The positive divergence of order e^2/R evokes a negative divergence or order Ge^4/R^3, which results in a positive divergence of order G^2e^6/R^5, and so on. The reason these terms are increasingly singular is that the gravitational response to an effect at one order is delayed to a higher order in perturbation theory.

The correct result is obtained by taking R to zero before expanding in the coupling constants e^2 and G,

$$\lim_{R\rightarrow 0} m_R c^2 = \left(\frac{e^2}{4\pi\epsilon_0 G}\right)^{\frac{1}{2}} c^2 = \sqrt{\alpha} M_{\rm Planck} c^2 \sim 10^{18}\,{\rm GeV}. \tag{177}$$

Like the expansion of $\ln(\Xi)$ in expression (165) it is finite but not analytic in the coupling constants e^2 and G. Unlike the expansion of $\ln(\Xi)$, it diverges for small G. This is because gravity has regulated the linear self-energy divergence which results for a nongravitating charged particle.

One can understand the process from the fact that gravity has a built-in tendency to oppose divergences. A charge shell does not want to contract in pure electromagnetism; the act of compressing it calls forth a huge energy density concentrated in the nearby electric field. Gravity, on the other hand, tends to make things collapse, especially large concentrations of energy density. The dynamical signature of this tendency is the large negative energy density concentrated in the Newtonian gravitational potential. In the limit of $R \rightarrow 0$ the two effects balance and a finite total mass results.

Expressed this way, there seems to be no reason why gravitational interactions should not cancel divergences in quantum field theory the same way they do in classical field theory. It is significant that the divergences of some quantum field theories—such as quantum electrodynamics—are weaker than the linear ones which ADM have shown that classical gravity controls [40]. So less cancellation is necessary and one might expect a smaller final mass, closer to the values of known charged particles. The frustrating thing is that the only computational tool we possess for quantum field theory is perturbation theory, and one cannot hope to see the cancellation perturbatively. In perturbation theory the gravitational response to an effect at any order must be delayed to a higher order. This is why the perturbative result (176) consists of an oscillating series of ever higher divergences. What is needed is an approximation technique in which the gravitational response is able to keep pace with what is going on in other sectors.

Note that any finite bare mass drops out of the exact result (177) in the limit $R \rightarrow 0$. This makes for an interesting contrast with the usual program of renormalization. Without gravity one would adjust the bare mass m_0 to be whatever divergent quantity is necessary to produce the measured mass $m_{\rm meas}$,

$$m_0 c^2 = m_{\rm meas} c^2 - \frac{e^2}{8\pi\epsilon_0 R}. \tag{178}$$

Of course the same procedure would work with gravity as well,

$$m_0 c^2 = m_{\rm meas} c^2 - \frac{e^2}{8\pi\epsilon_0 R} + \frac{G m_{\rm meas}^2}{2R}. \tag{179}$$

The difference with gravity is that we have an alternative: keep m_0 finite and let the dynamical cancellation of divergences produce a unique result for the measured mass. This would fulfill the old dream of deriving particle masses from their self-interactions. It would also mean that fundamental particle masses represent disguised quantum gravitational effects which would provide sensitive tests of the theory of quantum gravity if only we possessed the analytical tools to predict them.

4. Current approaches to quantum gravity

Many fine physicists have burned away their lives grappling with the problem of quantum gravity. There is not space in this paper to discuss all their efforts nor do I possess the expertise for it. I will here review five of the most popular approaches that are still being pursued. Three of them derive from the particle theory belief that general relativity must be changed; this common origin and methodology dictates that they should be presented consecutively. The remaining two approaches derive from the relativist's belief that quantum general relativity might be alright if studied nonperturbatively. I will cite review papers and books specific to each approach but I would also like to recommend the general review paper by Carlip [41].

4.1. Superstring theory

A central point to understanding string theory is that it cannot be formulated the way all other fundamental theories are, by giving the dynamical variables and the equations they obey. We do not know what the fundamental dynamical variables of string theory are, nor the equations they obey. What we have instead is a formalism for perturbatively computing what is the usual observable of a quantum field theory, the S-matrix[4]. The reasons for this are historical so I will summarize how string theory was developed.

String theory began in the late 1960s as an attempt to understand the strong interactions. Experiment had shown a series of resonances whose mass-squared m^2 increases approximately linearly as a function of angular momentum J, starting from a positive intercept,

$$m^2 = +m_0^2 + \Delta m^2 J. \tag{180}$$

It was obvious to everyone that the strong interactions could not be treated perturbatively (no one then suspected that the strong interaction would become weaker at high energies) so, instead of proposing quantum field theories, physicists tried to guess scattering amplitudes which incorporated such resonances. The first to succeed was Gabriele Veneziano [43], who proposed what would become known as the 4-particle, open string tree amplitude. Miguel Virasoro found the analogous amplitude for a closed string [44]. These 4-particle amplitudes were quickly generalized to give N-particle scattering for open [45] and for closed strings [46].

[4] Maldacena's AdS/CFT correspondence [42] is now widely accepted as providing a nonperturbative formulation of string theory (in terms of an ordinary quantum field theory) for the boundary conditions associated with the most symmetric solution of general relativity with a negative cosmological constant.

The early string amplitudes had linearly rising resonances (180) but they suffered from three problems:

- They did not include fermions;
- They contained resonances with the wrong sign to come from physical particles and
- They started from a negative intercept, rather than a positive one.

Particles with the wrong sign are called 'ghosts' and they can be viewed as making the theory decay instantly through a kinetic instability such as the Ostrogradskian instability I discussed in section 2.5. Particles with a negative mass-squared are called 'tachyons'. People sometimes make the mistake of thinking of them as particles that move faster than the speed of light but what they really are is instabilities of the potential energy. In a field theory such an instability would not be serious, it just means the field decays to some value with a lower potential energy, as happens when electroweak symmetry breaking takes place in the standard model. But string theory is not based upon a field theory so there is no principle to tell us how the perturbative background shifts—or even what the perturbative background is on the fundamental level. All we have is the S-matrix about a handful of backgrounds, and the appearance of a tachyon in the spectrum implies that this S-matrix cannot be correct.

It was my University of Florida colleague, Pierre Ramond who worked out how to incorporate free fermions in 1971 [47]. By this time it had been recognized that string theory scattering amplitudes can be written as integrals with respect to the coordinates of a string $X^\mu(\sigma)$, similar to the way that ordinary quantum field theory scattering amplitudes can be written as integrals with respect to the spacetime coordinates of a point particle x^μ. In neither case are the true dynamical degrees of freedom these coordinates; for quantum field theory the true dynamical degrees of freedom are the various fields, no one knows what they are for string theory.

Shortly after Ramond's work, Neveu and Schwarz added a new kind of string fermions to produce the amplitudes for interacting bosons [48]. Although these models had supersymmetry in the string coordinate space, the amplitudes themselves were not supersymmetric. In 1976 Gliozzi, Scherk and Olive showed how to get supersymmetric amplitudes by combining Ramond's formalism with that of Neveu and Schwarz, and then projecting out a certain sector which includes the problematic tachyon [49]. This was the birth of superstring theory.

Ramond's work is tremendously significant for quantum gravity because it represents the first appearance of a supersymmetry which connects fermions and bosons. There are many interesting things about supersymmetry but its importance for quantum gravity derives from a fact I mentioned in section 2.4: the 0-point energies of bosons and fermions with the same mass m and wave number $k = \|\vec{k}\|$ cancel,

$$\text{Bosons} \Longrightarrow +\tfrac{1}{2}\sqrt{m^2c^4 + (\hbar c k)^2}, \tag{181}$$

$$\text{Fermions} \Longrightarrow -\tfrac{1}{2}\sqrt{m^2c^4 + (\hbar c k)^2}. \tag{182}$$

Supersymmetry involves a tight relation between fermions and bosons, which is necessary if this cancellation is to do any good for quantum gravity. With this tight relation, every correction from the 0-point motion of bosons in the theory tends to be canceled by an opposite correction from the 0-point motion of fermions.

At this point I must digress to discuss the implications supersymmetry has for the cosmological constant Λ. Unbroken supersymmetry is only consistent with Λ being zero or negative, not positive. If it exists in nature then supersymmetry must be badly broken at low energies because unbroken supersymmetry predicts that every known particle has a 'super-partner' (bosonic super-partners for known fermions and fermionic super-partners for known bosons) with the same mass, and not a single one of these super-partners has been observed.

One does not need superstring theory in order to have supersymmetry. In fact, supersymmetric algebras and/or quantum field theories were developed before superstring theory by Golfand and Likhtman [50], by Volkov and Akulov [51], by Volkov and Soroka [52] and by Wess and Zumino [53]. Supersymmetry had also been extended to gravity to produce a class of models known as *supergravity* by Freedman, van Nieuwenhuizen and Ferrara [54] and by Deser and Zumino [55]. (I shall have much more to say about supergravity in the next subsection.) A fact of great importance for this exposition is that the cosmological constant can have any sign in generic theories of gravity but it can only be negative or zero in supergravity and superstring theory. So the only way supergravity or superstring theory can be consistent with the observed acceleration of the current universe [8, 9] is to suppose that we are now in a metastable state which will eventually decay to a universe which is either decelerating or actually contracting.

Recall that the initial string amplitudes had three problems: no fermions, ghosts and a tachyon. We have just seen that superstring theory resolves the first problem and the third one. My University of Florida colleague, Charles Thorn, was among those who proved that the ghosts drop out in certain key dimensions: $D = 26$ dimensions for bosonic string theory and $D = 10$ dimensions for superstrings [56, 57].

The 'no-ghost' theorems actually came in 1972, before the invention of superstring theory, so string theory still had the tachyon problem. It also had massless particles: a spin one particle in the open string amplitudes and a spin two particle in the closed string amplitudes. Although these particles pose no problem for stability, they are no more part of the observed spectrum of strongly interacting particles than tachyons. The funny dimensions dictated by the no-ghost theorems also seemed to preclude using string theory to describe the strong interactions. And recall from section 2.4 that Politzer [14] and Wilzcek and Gross [15] showed in 1973 that the currently accepted theory of the strong interactions gets weaker at high energies, which meant perturbation theory can be used to make predictions. These predictions were confirmed, by which point few people were interested in string theory.

In 1974 Joel Scherk and John Schwarz made a virtue of the massless particles by showing that they could be the photon and the graviton, respectively [58]. A year later Scherk and Schwarz proposed that the six extra dimensions of

(what would become) superstring theory were 'compactified' [59]. A dimension which is compactified does not extend to macroscopic distances the way the three known spatial dimensions do; instead it is rolled up into a circle (or more complicated shape) of radius so small that we cannot observe motion in this direction. With the advent of superstring theory the next year all the ingredients were in place for a potential theory of everything.

At first very few people were interested. The majority of particle physicists were working out the consequences of what would become known as the *standard model*. More and more people also began studying the *grand unified theories* which were suggested by the fact that the three standard model coupling constants become comparable at very high energies. Proper, mainstream particle theorists did *not* worry about quantizing gravity! I recall one of my graduate professors being asked for a reference on supersymmetry and supergravity and replying, with lofty disdain:

> Supersymmetry is one of those subjects which we here at Harvard try to discourage students from studying.

Students who wished to study quantum gravity at Harvard did so through a sort of 'underground railroad' in which Sidney Coleman signed the forms but Stanley Deser was our true adviser.

Supersymmetry was eventually accepted by particle theorists as a way of solving a hierarchy problem somewhat less severe than the one I discussed in section 3.2 with regard to the cosmological constant. Supersymmetry also allowed particle theorists to continue using perturbation theory, which is our chief analytical tool, to very high energies. Quantum gravity got accepted shortly thereafter. Part of the reason for this is that the scale then envisaged for grand unification (10^{16} GeV) is only three orders of magnitude below the Planck scale (10^{19} GeV). Another reason is that the extra components of a higher–dimensional metric (it has $\frac{1}{2}D(D+1)$ components in D spacetime dimensions) could be regarded as the other kinds of bosonic particles, thereby unifying gravity and the other forces. For example, the 15 components of a five-dimensional metric could be regarded as the ten components of our four-dimensional metric, plus a four-component vector potential and a scalar particle.

Superstring theory was rehabilitated because higher-dimensional general relativity is not a very good way of unifying all the forces. For one thing, it was not clear how to make it incorporate fermions correctly. For another, increasing the number of dimensions makes the divergence problem worse. It is easy to understand why from the discussion of section 2.4. Recall that the first order gravitational response to 0-point motion in a static gravitational field of wave number $\vec{p}$ involves a divergent mode sum of the form

$$\int^K \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\mathcal{E}^4}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3} = AK^4 + BK^2p^2 + C\ln(K^2)p^4 + \text{Finite}, \tag{183}$$

where $\mathcal{E}^4$ consists of a variety of terms quartic in $\vec{p}$ and $\vec{k}$, K is the cutoff and A, B and C are numbers of order one. The details of how extra dimensions are compactified do not matter in the regime of large wave numbers which gives rise to the problem. So going to six spacetime dimensions amounts to changing the wave vector in (183) from $\vec{k} = (k_1, k_2, k_3)$ to $\vec{k} = (k_1, k_2, k_3, k_4, k_5)$ to give

$$\int^K \frac{\mathrm{d}^5k}{(2\pi)^5} \frac{\mathcal{E}^4}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3} = AK^6 + BK^4p^2 + CK^2p^4 + D\ln(K^2)p^6 + \text{Finite}. \tag{184}$$

This requires even more unacceptable counterterms than the four-dimensional theory!

Interest in superstring theory surged in 1984 when Green and Schwarz showed that the theory naturally incorporates the right kinds of fermions and is also likely to be finite [60]. The statement about finiteness might seem surprising because the no-ghost theorems require superstring theory to exist in $D = 10$ spacetime dimensions and I have just explained that increasing the dimension makes the divergences of quantum general relativity worse, not better. Of course the explanation is that superstring theory is *not* general relativity; it does not even have the metric as one of its fundamental degrees of freedom. Although superstrings do incorporate gravity, stress energy is not the source of gravity at high wave number, so superstrings violate the 3rd of the four propositions listed at the beginning of section 3.

All of this raises the question of what superstring theory is on the fundamental level. The unsatisfactory answer is that no one knows! Based on the way we have of expressing the perturbative S-matrix one might think the fundamental variable should be a *string field*, that is, a field whose argument is a string's position in spacetime $X^\mu(\sigma)$, just like a normal field could be regarded as depending upon a particle's position x^μ. String field theories can indeed be constructed which reproduce the perturbative string S-matrix [61]. The earliest ones were in a noncovariant formalism due to Kaku and Kikkawa [62], but the 1980s witnessed the development of lovely invariant Lagrangians by Witten [63] and Horowitz *et al* [64]. The trouble with these string field theories is that they must be *nonlocal*, that is, they cannot be expressed in terms of just the dynamical variable and a finite number of its derivatives [65]. This is not some defect of how a string field theory was constructed from the perturbative S-matrix, it *must* be true in order for superstrings to avoid divergences. Instead of the coupling to the energy of a mode, superstring field theories couple to the energy times exponentials of the wave number,

$$\text{General relativity} \Longrightarrow E(\vec{k}) = \hbar c\|\vec{k}\|, \tag{185}$$

$$\text{String field theory} \Longrightarrow \hbar c\|\vec{k}\| \times \mathrm{e}^{-\alpha'\|\vec{k}\|^2}, \tag{186}$$

where α' is the Regge slope parameter. Of course these exponentials make the integrals converge at high $\|\vec{k}\|$, but one has to wonder if they engender new problems.

Anyone who has studied quantum mechanics knows that exponentiating the derivative operator effects a spatial translation,

$$\mathrm{e}^{\Delta x\,\mathrm{d}/\mathrm{d}x} f(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi}\, \mathrm{e}^{\mathrm{i}kx} \times [\mathrm{e}^{\mathrm{i}k\Delta x}\tilde{f}(k)] = f(x+\Delta x). \tag{187}$$

Exponentiating the square of a derivative, as in (186), involves a superposition over translations of all distances out to infinity. And one must of course include *temporal* as well as spatial derivatives or else there would be a massive violation of Lorentz invariance. Recall from section 2.5 that adding even a single higher time derivative results in new degrees of freedom which have the opposite kinetic energy to the lower derivative degrees of freedom. It turns out that the problem grows worse the more derivatives you add: there is an extra degree of freedom for each new time derivative and essentially half of these new degrees of freedom carry negative kinetic energy. That is why we cannot modify the gravitational equations of motion to include the C^2 counterterm or the higher derivative counterterm whose necessity for pure gravity was proven by Goroff and Sagnotti [6, 7]! So is it alright to have exponentials of the derivative operator in string field theory?

Not every nonlocal field theory succumbs to the Ostrogradskian instability [66]. But the nonlocality of string field theory must be restricted to entire functions of the derivative operator or else the conservation of probability would not work out correctly. And the definition of an entire function is that it converges to its Taylor series expansion. This means that string field theory can be viewed as the limit of a sequence of ever higher derivative models, which grow more and more unstable. One sometimes hears the statement that the extra degrees of freedom might decouple in the limit because they get driven to infinite frequency, but recall the fallacy of this argument from section 2.5: it plays on one's intuition, from lower derivative theories, that high frequency modes cannot be excited because there is only a limited amount of free energy. That is only true when all degrees of freedom have positive energy; when there are negative energy degrees of freedom, the high frequency modes can be excited by also exciting modes with the opposite energy. Far from dropping out, high frequency modes of negative energy *dominate* because there are so many more of them[5]!

All of this led David Eliezer and me to conclude in 1989 that string field theory must suffer from the Ostrogradskian instability [68]. It does not show up in perturbation theory simply because the nonlocality is restricted to interactions, but it is present in the full theory. Of course the only reason for studying string field theory is to get a nonperturbative definition of string theory, so our result means either that string theory is wrong or else that some other formalism defines string theory on the nonperturbative level. Most string theorists take the latter view, but there has so far been no other nonperturbative way of defining the formalism in general[6].

It has been notoriously difficult to derive testable predictions from superstring theory. A notable exception to this is the 1988 observation by Antoniadis, Bachas, Lewellen and Tomaras that, if string supersymmetry, which must be broken at the low energy scales we can access in experiments, should happen to be broken *perturbatively*, then there has to be at least one 'large' extra dimension within experimental reach [69]. This was elaborated further [70] and has led to a huge amount of work which I will not attempt to review, but I do wish to relate a funny story about it. The story concerns a 'phenomenologist'—the kind of particle theorist who works closely with experimentalists to explain data—who was frustrated by the sometimes incomprehensible mathematics of string theory. After a seminar by Antoniadis she exclaimed:

> I can't usually understand much of what the string theorists tell me. But this, this looks like a prediction!

That concludes the portion of string history which has the greatest relevance to this paper. Of course there have been many more developments in superstring theory. Some of the principal ones are:

- In 1995 Witten suggested that relations between the various 10-dimensional superstring theories point to an 11-dimensional parent that was dubbed *M-Theory* [71].
- Also in 1995, Polchinski showed that extended objects called *D-branes*, on which open strings can end, are necessary to realize one of the relations between superstrings [72].
- The first microphysical derivation, in 1996, of the Bekenstein–Hawking entropy of a black hole by Strominger and Vafa [73]. This is the most successful application of superstring theory, although one should note that subsequent derivations were made (for example, from loop quantum gravity [74]) and Carlip has suggested that the result follows from conformal symmetry near the horizon rather than from the details of superstring theory [75].
- Maldacena's 1997 conjecture that string theory on one spacetime manifold is equivalent to a certain quantum field theory without gravity on a lower-dimensional manifold [42]. This is a terrifically important insight because it might serve as the long-sought, nonperturbative definition of superstring theory. It has also been applied to nuclear physics and even condensed matter physics with interesting results.
- The 2000 discovery by Busso and Polchinski that vast numbers of discrete choices, called *flux vacua*, can be made in compactifying superstring theory [76]. The number of these choices is estimated to run from between 10^{100} and 10^{500} [77], and no principle is yet known to fix which one is taken.
- The 2003 suggestion by Susskind that the *anthropic principle* might usefully constrain where we are in the vast *string landscape* of flux vacua [78]. The anthropic principle states that physical laws must be such that humans or some other form of intelligent life can exist to observe them.

The last two developments on my list, and the fact that string theory predicts the wrong sign for the cosmological constant, have led to a slackening of interest in superstrings

[5] In view of the recurring efforts to legitimize entire functions of the derivative operator I should observe that, were they allowed, we could solve the divergence problems of quantum gravity without superstrings [67].

[6] Note again that Maldacena's AdS/CFT correspondence [42] is now widely accepted as providing a nonperturbative formulation of string theory (in terms of an ordinary quantum field theory) for the boundary conditions associated with the most symmetric solution of general relativity with a negative cosmological constant.

as a fundamental theory of everything, although there is still much work on applications. (I will discuss one in the next subsection.) In the face of numbers such as 10^{500}, some physicists now speak of a 'theory of anything' rather than a theory of everything. And even veterans of decades of string research find it difficult to accept the anthropic principle. A personal anecdote might best convey the current state of affairs. Early in the spring 2007 semester my University of Florida colleague, Charles Thorn, began a seminar by announcing his belief that:

> String theory is just a technique for summing the leading terms in the $1/N$ expansion of QCD.

After years of hearing more ambitious assessments this was so shocking that I checked to be sure I had understood correctly. Charles confirmed that I had; in his current view, the effort to regard superstrings as a fundamental theory of everything was a blind alley. Later that year I related Charles' pronouncement to string theory colleagues on three continents and solicited their own opinions. About half of them agreed with him, more often the younger people.

String theory has occupied some of the best minds of particle theory for many decades, and it would require more than the space allotted for this paper to do justice to their efforts. Nor do I possess the expertise for it. Let me instead refer the interested reader to some standard books by the leaders of this field. I recommend the texts by Green *et al* [79], by Polchinski [80], by Zwiebach [81], by Kiritsis [82] and by Becker, Becker and Schwarz [83]. In fairness I should also mention popular science books which are critical of string theory by Smolin [84] and by Woit [85].

4.2. On-shell finiteness

Although superstring theory may be short on successful predictions, it does provide a very efficient way of organizing perturbative calculations which involve corrections from many components of tensor fields such as the metric. In fact, the simplest method for computing the lowest order scattering amplitude for 4-gravitons in general relativity is to use superstring theory and then take a certain limit [86]. A few more tricks from superstring theory, and a *lot* of hard work by many physicists over the course of two decades, has led Bern, Dixon and Roiban to conjecture that a version of supergravity might actually be *on-shell finite* to all orders in perturbation theory [87]. (There have also been indications directly from string theory [88].) I shall first explain what this means, what the theory is which may have this property and recent results concerning it. Then I will review the developments that led to this possibility.

Recall that renormalizable quantum field theories have divergences which can be absorbed into redefinitions of free parameters. Once this is done one can compute not only scattering amplitudes but also the expectation values of products of the field operators at different points, which are known as *correlation functions*. In contrast, an on-shell finite theory has scattering amplitudes which are finite, without the need for renormalization, but its correlation functions are not necessarily finite or even renormalizable[7]. On-shell finiteness has particular importance for theories of gravity and supergravity because they cannot be perturbatively renormalizable, but divergences tend to cancel out of their scattering amplitudes. For example, pure gravity, without any matter, is on-shell finite at first order in perturbation theory [2]. One has to go to the second order before divergences occur [6, 7].

Recall that a supersymmetry relates fermions to bosons. This is a very good thing for divergences because bosons contribute a positive 0-point energy whereas fermions contribute a negative 0-point energy, so if one contrives a tight relation between the two kinds of particles then it is conceivable that quantum corrections from fermions will cancel the divergences in quantum corrections from bosons, order-by-order in perturbation theory. Just that seems to be happening in a model called, $N = 4$ *super-Yang–Mills* theory. I should explain that 'Yang–Mills' theories are nonlinear generalizations of electrodynamics which can describe forces such as the weak and strong interactions. They do not have to be supersymmetric, but the term '$N = 4$' means this particular model has the maximum amount of supersymmetry possible for a Yang–Mills theory. Although this supersymmetry almost certainly makes the model finite, it comes at a terrible price: $N = 4$ super-Yang–Mills theory cannot incorporate the known particles of the standard model. Further, the exact cancellation of divergences is only valid as long as the particles are all massless, which is very far from the observed universe in which only the photon, the 'gluon' which carries the strong force, and the graviton appear to be massless.

The gravity model which might be finite is called, '$N = 8$ supergravity.' Just like $N = 4$ super-Yang–Mills theory, it cannot be a realistic model of the universe because its particles are all massless, and they lack other essential properties of the known particles in the standard model. The spectrum of $N = 8$ supergravity consists of: a graviton, 8 spin $\frac{3}{2}$ particles called *gravitinos*, 28 vectors bosons, 56 spin $\frac{1}{2}$ fermions and 70 scalar bosons. The term $N = 8$ means this model has the maximum number of independent supersymmetries for a theory of gravity.

It might seem strange that people are so fascinated with a theory which they know cannot describe the universe. The reason is that supergravity models with less supersymmetry do have a chance to describe physics. Although these models are finite at first order in perturbation theory, and even at second order, it had been believed that they must suffer uncontrollable divergences at third order [89]. Those expectations are based on demonstrating that counterterms exist at higher orders which are not forced to vanish by any known symmetry or dynamical relation. There is a widespread belief in particle physics that any divergence which *can* happen, *must* happen. Weinberg refers to such beliefs as 'folk-theorems', and this one has never been checked because it was just too difficult to make the required computations. The interest in $N = 8$ supergravity derives from the fact that it has recently become

[7] However, one can construct redefinitions of the original fields which give finite correlation functions.

possible to check for divergences in this model *and they are not present* at third order [90].

Of course the third order result prompted supergravity experts to re-examine the old arguments [89] and they soon explained why the divergence fails to occur [91]. They also predicted a new divergence at higher order [91]. There was even a famous bet about this between Zvi Bern and Kelly Stelle. The funny thing is, when the prediction was checked, the theory was again seen to be finite [92]. As a good scientist, Stelle has admitted he lost the bet and is hard at work trying to understand what happened.

The computational technology which made this possible began to be developed in the late 1980s by Zvi Bern and David Kosower. They noticed how much more efficient string theory is than the existing applications of Yang–Mills theory at computing scattering amplitudes. Recall that Yang–Mills theories can describe the strong interactions, which are an important background for very high energy experiments. People needed an efficient way of computing Yang–Mills amplitudes with lots of external particles and Bern and Kosower developed one by taking suitable limits in string theory [93]. By 1990–1991 they had extracted the essential simplifications of string theory and worked out how to do efficient computations directly in Yang–Mills theory [94]. In 1992 they were joined by Lance Dixon and the trio derived a number of useful results about the strong interactions during the mid-1990s [95].

By the turn of the century Bern, Kosower and Dixon had developed a technique they called the 'unitarity method' in which perturbative corrections could be generated from the lowest order results [96]. The idea is a more sophisticated version of the old bootstrap program from the 1960s. If we express the S-matrix as $S = I + iT$, then unitarity implies

$$S^\dagger \times S = I \Longrightarrow -i(T - T^\dagger) = T^\dagger \times T. \tag{188}$$

So if the T matrix consists of a perturbative series expansion, one can determine the Nth order contribution to $(T - T^\dagger)$ from sums of products of lower order terms on the right-hand side of (188). Then the Nth order contribution to $(T + T^\dagger)$ can be inferred by a procedure known as 'bootstrapping', and the stage is set to push one order higher [97–99]. This procedure works best for supersymmetric Yang–Mills theory because then the bootstrap is simplest, and the bootstrap works best of all for $N = 4$ super-Yang–Mills [100].

The reader might wonder why an efficient technique for computing super-Yang–Mills amplitudes has any relevance for supergravity. The connection is provided by the Kawai–Lewellen–Tye (KLT) relations between open and closed string amplitudes [101]. The KLT relations say that a lowest order gravitational scattering amplitude can be factorized into sums of products of Yang–Mills amplitudes. These relations are relatively simple to see from string theory but so difficult to recognize for ordinary field theory that they were only inferred using string theory. Of course superstring theory requires ten dimensions, which we do not want, and it also contains infinite towers of super-massive particles that are present in neither $N = 4$ super-Yang–Mills nor $N = 8$ supergravity. Both problems are avoided by taking the low energy limit [102]. So the key to computing high order perturbative corrections to the scattering amplitudes of $N = 8$ supergravity is first to decompose these amplitudes into products of zeroth order amplitudes using the unitarity method, then evaluate the later in terms of $N = 4$ super-Yang–Mills amplitudes [103].

The most recent results about $N = 8$ supergravity [92] mean something is wrong with our thinking about what sorts of divergences can happen. No one knows how high the cancellations extend in $N = 8$ supergravity or if they might apply to more realistic models. We could be witnessing the start of a revolution. The latest progress on this subject is so new that there are no books or long review papers. I highly recommend the short paper by Bern *et al* [104].

4.3. Asymptotic safety

Recall from section 2.6 that, if one does not consider the R^2 and C^2 counterterms to be part of the gravitational field equations then the divergences of quantum gravity get worse at each order in perturbation theory. The same thing happens if you include the 4th derivative counterterms but regard them as perturbations, rather than as part of the 0th order field equations. In that case each new order in perturbation theory requires the introduction of counterterms which contain two more derivatives of the metric. Accepting this escalating series of counterterms might seem crazy, in view of the fact it could be avoided by simply considering the R^2 and C^2 counterterms to be part of the 0th order equations. However, the procedure has a saving virtue: if we regard the higher derivative counterterms as perturbations then they do not add new degrees of freedom to the theory, and the Ostrogradskian instability of section 2.5 is avoided.

The program of *asymptotic safety* is to accept the escalating series of perturbative counterterms and attempt to show that interesting predictions can be derived from the resulting formalism [105]. To make this approach intelligible I will first explain why regarding the counterterms as perturbative avoids new degrees of freedom. Then I discuss how the counterterms affect the theory's ability to make predictions. Finally, I will review progress on the subject.

The discussion of degrees of freedom is simplest in the context of a one dimensional, point particle whose position as a function of time is $q(t)$. Suppose that the Lagrangian is that of a harmonic oscillator with a higher derivative,

$$L = -\frac{gm}{2\omega^2}\ddot{q}^2 + \frac{m}{2}\dot{q}^2 - \frac{m\omega^2}{2}q^2. \tag{189}$$

Here m is the particle's mass, ω is a frequency and g is a small, positive number I wish to think of as a coupling constant. The Euler–Lagrange equation,

$$\frac{g}{\omega^2}\frac{d^4q}{dt^4} + \ddot{q} + \omega^2 q = 0, \tag{190}$$

has the general initial value solution

$$q(t) = A_+ \cos(k_+ t) + B_+ \sin(k_+ t) + A_- \cos(k_- t) + B_- \sin(k_- t), \tag{191}$$

where the two frequencies and the four combination coefficients are

$$k_\pm = \frac{\omega}{\sqrt{2g}}[1 \mp \sqrt{1-4g}]^{\frac{1}{2}}, \qquad A_\pm = \frac{k_\mp^2 q_0 + \ddot{q}_0}{k_\mp^2 - k_\pm^2},$$
$$B_\pm = \frac{k_\mp^2 \dot{q}_0 + \mathrm{d}^3 q_0/\mathrm{d}t^3}{k_\pm(k_\mp^2 - k_\pm^2)}. \tag{192}$$

The + mode carries positive energy; it is the mode that would appear even for $g = 0$. The $-$mode carries negative energy; it is the new, higher derivative degree of freedom.

Suppose we regard the higher derivative term as a perturbation. That amounts to making the substitution

$$q_{\rm pert} = \sum_{n=0}^{\infty} g^n x_n(t), \tag{193}$$

in the Euler–Lagrange equation (190) and then segregating terms with the same powers of g. The resulting series of equations is

$$\ddot{x}_0 + \omega^2 x_0 = 0, \tag{194}$$

$$\ddot{x}_1 + \omega^2 x_1 = -\frac{1}{\omega^2}\frac{\mathrm{d}^4 x_0}{\mathrm{d}t^4}, \tag{195}$$

$$\ddot{x}_2 + \omega^2 x_2 = -\frac{1}{\omega^2}\frac{\mathrm{d}^4 x_1}{\mathrm{d}t^4}, \tag{196}$$

and so on. Note that the higher derivative terms always enter as sources, evaluated at the lower order solution. They do not appear at 0th order, so one only needs q_0 and $\dot{q}_0$ to get a unique solution $x_0(t)$. The higher derivative term does appear in the 1st order equation, but only in the form of $\mathrm{d}^4 x_0(t)/\mathrm{d}t^4$, which was already fixed at 0th order, so one again needs just q_0 and $\dot{q}_0$ to determine $x_1(t)$. Because the equation is linear the resulting series can be summed up and gives

$$q_{\rm pert} = q_0 \cos(k_+ t) + \frac{\dot{q}_0}{k_+}\sin(k_+ t). \tag{197}$$

Note that the higher derivative term did have an effect, it shifted the frequency of oscillation from the 0th order result of ω to k_+. However, we have avoided the higher derivative degrees of freedom which give rise to the Ostrogradskian instability.

Although explicit results are difficult to obtain when the equations become nonlinear, one can prove that this procedure works generally [68, 106], so it should apply to the higher derivative counterterms of quantum gravity. Of course it cannot serve as a nonperturbative definition of quantum gravity (in this I dispute the official position of asymptotic safety) but I have already explained in section 2.3 that even the asymptotic series solutions one derives from perturbation theory should be wonderfully accurate at low energies. The problem is the escalating series of counterterms, each with its own completely arbitrary, finite part. Those finite parts can only be fixed by requiring some prediction of the theory to agree with measurement. But each time this is done one loses a potential prediction, and it must be done an infinite number of times! In such a case one wonders what the resulting formalism can be used to predict?

One answer to this question was provided by John F Donoghue [107]. He pointed out that all counterterms must take the form of polynomials of the Fourier wave vector $\vec{p}$ such as we found in expression (115). The finite contributions (116) contain some terms of this form—and they will be rendered ambiguous by the arbitrary finite parts of the counterterms. However, there are also terms in (116) of the form $p^4 \ln(L^2 p^2)$ which could never have been part of a counterterm and are *logarithmically enhanced* with respect to the p^4 counterterms in the small p regime which is most accessible to experiment. That is how a unique result was obtained for the first quantum gravitational correction (166) to the Earth's potential [28]. This way of using a nonrenormalizable quantum field theory to make predictions is known as *low energy effective field theory*.

The program of asymptotic safety aims at the even better possibility that we might be able to predict the finite parts of all the counterterms in such a way that quantum general relativity could be used at all energies. Explaining how this works in any detail would go far beyond the scope of this paper but I will try to put the idea across in simple terms. Recall from our discussion of renormalization in section 2.4 that the strengths of interactions change as one varies the energy scale. This *running of coupling constants* has many important consequences for particle physics. For example, forcing the strong interactions into a regime for which perturbation theory is reliable provides our principal check on the theory, and the fact that the various interactions become comparable at a very high energy is the strongest evidence for grand unification. The new counterterms one finds in quantum general relativity vary with the energy scale as well. The hope of asymptotic safety is that this variation might carry each of them (or at least all but a finite number of them) to a unique value called a *nontrivial ultraviolet fixed point*. If this proves to be the case then we would not have to use up an infinite number of predictions in order to determine the finite parts of counterterms. Instead we would just set them to their values at the ultraviolet fixed point and this should be very near to the correct result at sufficiently high energies.

The possibility of asymptotic safety was recognized by Steven Weinberg in 1979 [105]. The equation which governs how coupling constants change is known as the *exact renormalization group equation* and was derived by Christof Wetterich in 1992 [108]. One can never solve this for the infinity of necessary counterterms so what is done instead is to study the way selected combinations of counterterms flow. Exploring these truncations cannot prove asymptotic safety, but it could *disprove* the conjecture, and the results so far are consistent with the existence of an ultraviolet fixed point. It has been computationally impossible to include the first really problematic counterterm of pure gravity—the one whose divergent coefficient was computed by Goroff and Sagnotti [6, 7]. However, a very recent result by Benedetti *et al* [109] shows that there are no problems from the problematic C^2 counterterm which 't Hooft and Veltman found when gravity is minimally coupled to a scalar field [2]. Much recent work on asymptotic safety has been done by Oliver Lauscher and Martin Reuter [110], Roberto Percacci [111] and by Percacci

and Daniele Perini [112]. Interested readers should consult the review by Lauscher and Reuter [113].

4.4. Loop quantum gravity

Recall that relativists suspect the problems of quantum general relativity derive from using perturbation theory. They therefore insist on a very careful formulation which is not based upon perturbing around any background geometry. That has far-reaching consequences. It means that relativists cannot accept, as a complete description of quantum gravity, the asymptotic scattering states employed by particle theorists, nor can they accept the decay rates and cross sections we use as observables nor even our inner product! So the full apparatus of quantum mechanics (states, inner product, observables) must be constructed from the beginning and with an unprecedented level of rigor and generality. It should be noted that we do not possess such a general formulation even for quantum electrodynamics, which is the best understood and most thoroughly tested quantum field theory.

Loop quantum gravity is based upon a Hamiltonian for general relativity which was developed by Abhay Ashtekar in 1986 [114]. Recall that any Hamiltonian formalism has 'coordinates' (the Q's) and conjugate momenta (the P's), and it is usual to consider quantum mechanical states to be functions of the coordinates. The term 'loop' in the name comes from the way Ashtekar's coordinate variables are organized into 'Wilson loops', a sort of line integral (which also involves continuous multiplication of matrices) around closed loops. Under certain assumptions one can show that this is the unique way of organizing the quantum theory [115].

Any Hamiltonian formalism which involves one or more of the fundamental forces will possess *constraint equations*. Solving these at the required level of rigor and generality is a major problem for loop quantum gravity [116]. To understand what a constraint equation is, recall the discussion in section 2.1 concerning the triune nature of fundamental force fields:

- One part of the field can be changed arbitrarily by a symmetry transformation which is associated with the conservation law;
- Another part of the field is completely fixed by its sources and
- The final part consists of independent degrees of freedom.

Solving the constraint equations is what determines the part of the force field which is completely fixed by its sources. The constraint equation of electrodynamics is Gauss's law ($\epsilon_0 \vec{\nabla} \cdot \vec{E} = \rho$), another equation is fixed by current conservation and the remaining two are dynamical. In general relativity four of the ten Einstein equations (9) are constraints, four are fixed by stress–energy conservation and the remaining two are dynamical.

In the Hamiltonian formalism of loop quantum gravity one does not think of the constraint equations as determining field operators, the way I did in section 2, but rather as restricting how state wave functions can depend upon the coordinates and also what operators can be observed. It is simple enough to work this out perturbatively to a given order, but that sort of perturbative solution is exactly what relativists wish to avoid! This makes things tougher but perhaps not impossible. The constraint equations of electrodynamics and the other fundamental forces have been solved nonperturbatively, as have all but one of the quantum gravitational constraints. Unfortunately, there is not yet a general solution to the final constraint.

The absence of explicit states or observables makes it difficult to compute much in full loop quantum gravity. In a number of symmetry reduced models (see, e.g. [117]) the program has been completed and the resulting quantum theory sheds light on important conceptual issues such as the meaning of time and dynamics in background independent physics and the likely fate of the most interesting classical singularities in quantum gravity. Work has also been done on developing approximation procedures to construct explicit observables [118].

A measure of the difficulty involved is the effort that went into deriving the free graviton propagator [119]. This is a completely trivial exercise for particle theorists but it required several years of hard work using loop quantum gravity precisely because the concepts involved in its construction are highly nontrivial in a background independent approach. At the current stage, it also seems to involve a number of choices which indicates that the formalism is not yet complete.

The field of loop quantum gravity has grown to include work on quantum cosmology, black hole entropy, the issue of information loss, coupling to matter, path integral quantization and the breaking or deformation of Poincaré invariance. Because this approach is far removed from my own area of expertise it is best that interested readers consult recent reviews by the leaders of the field. I recommend the papers by Ashtekar and Lewandowski [120], by Smolin [121] and the books by Rovelli [122] and Thiemann [123].

4.5. Causal dynamical triangulations

I have admitted my bias as a particle theorist who loves being able to get approximate results using perturbation theory, and I have tried to be honest about the possibility that these results might be erroneous, as so many of my relativist friends believe. However, exact calculations are unlikely to be attainable for quantum gravity, so the most fruitful way of questioning perturbation theory is to develop better approximation techniques. An example of work along these lines is the program of *causal dynamical triangulations* by Jan Ambjorn, Jerzy Jurkiewicz and Renate Loll.

The idea behind causal dynamical triangulations is to numerically simulate a formal expression of quantum general relativity in terms of an infinite number of integrations, one for each of the ten components of the metric at each point in spacetime. Most everyone is familiar with the use of computers to numerically evaluate a finite number of integrals. Of course there are an infinite number of points in continuum spacetime, and no computer can simulate more than a finite number of integrations. So one first discretizes the formal representation using a technique devised in 1961 by Tullio Regge for approximating classical general relativity without the use of coordinates [124]. Then the issue becomes how

to take the continuum limit so that one plausibly recovers the correct theory.

Recovering the correct continuum theory is highly nontrivial! In much simpler models for which the answer was known, physicists discovered that one typically has to make parameters change with the level of discretization in sensitive ways. And sometimes, one must introduce new parameters which do not seem to be present in the continuum theory. Of course we do not know exactly what the result should be for quantum gravity but we do know it should have four spacetime dimensions on macroscopic scales. With the most straightforward quantum and computer adaptations of Regge's approach [125, 126] there does not seem to be any way of taking a continuum limit for pure gravity which will have this property [127, 128], although the issue is still open when matter is added [129].

To surmount the problem Ambjorn and Loll developed a two-dimensional variation of Regge's discretization in which time is given a preferred status [130]. Subsequent work with Jurkiewicz produced a four dimensional model [131] which seems to remain four-dimensional as the number of elements in the triangulation increases [132]. On large scales the resulting spacetime looks like the most symmetric solution of the classical Einstein equations with a positive cosmological constant [132]. It is still not clear how to take the continuum limit or if that limit even exists. And it is of course unknown if the continuum limit will reproduce Newtonian gravitation at large distances or contain gravitational radiation. One should not expect progress too soon; it required decades of labor to attain better than 10% accuracy with numerical simulations of strong interactions in the low energy regime for which perturbation theory fails [133]. A very recent review of causal dynamical triangulations is [134].

5. Cosmology

For years quantum gravity was a realm of speculation in which it was even respectable to deny the need for a quantum theory of gravity. That has changed recently with the advent of the first recognizable quantum gravitational data and it has revolutionized the field. These data come from cosmology, and many more are likely on the way, so any discussion of quantum gravity must include cosmology. I shall begin by introducing the classical metric which describes cosmology, and explaining how the Einstein equations constrain this metric. Then I discuss why primordial inflation is necessary and the simplest class of models which provide the stress energy to support it. The key point of this section is why quantum gravitational effects are enhanced during inflation and why the results on cosmological perturbations [22, 23] represent the first ever recognized quantum gravitational data.

5.1. FRW geometry

On the largest scales the universe seems to have no special origin or special directions [135, 136]. These properties are known are *homogeneity* and *isotropy*, respectively. The universe also seems to be devoid of spatial curvature [22]. The spacetime geometry consistent with these three features is characterized by the following invariant element,

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a^2(t)\,\mathrm{d}\vec{x}\cdot\mathrm{d}\vec{x}. \tag{198}$$

(The acronym 'FRW' is formed from the names of three cosmologists: Alexander Friedman, Howard Percy Robertson and Arthur Geoffrey Walker.) The coordinate t represents physical time, the same as it does in flat space. However, the physical distance between $\vec{x}$ and $\vec{y}$ is not given by their Euclidean norm, $\|\vec{x}-\vec{y}\|$, but rather by $a(t)\|\vec{x}-\vec{y}\|$. Because it converts coordinate distance into physical distance $a(t)$ is known as the *scale factor*.

Although the scale factor is not directly measurable, three simple observable quantities can be constructed from it,

$$z \equiv \frac{a_0}{a(t)} - 1, \qquad H(t) \equiv \frac{\dot{a}}{a},$$
$$q(t) \equiv -\frac{a\ddot{a}}{\dot{a}^2} = -1 - \frac{\dot{H}}{H^2}. \tag{199}$$

The *redshift* z gives the proportional increase in the wavelength of light emitted at time t and received at the current time, t_0. (I am ignoring the special relativistic Doppler shift.) Redshift is often used to measure cosmological time, even for epochs from which we detect no radiation. The *Hubble parameter* $H(t)$ gives the rate at which the universe is expanding. Its current value is $H_0 = (70.5 \pm 1.3)\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \simeq 2.3 \times 10^{-18}\,\mathrm{Hz}$ [22]. The *deceleration parameter* $q(t)$ is less well measured. Observations of type Ia supernovae are consistent with a current value of $q_0 \simeq -0.6$ [8, 9].

Astronomers infer H_0 and q_0 by constructing *Hubble plots*. Suppose the light from a distant star contains a distinctive absorption line measured at the wavelength λ. If the same line occurs at wavelength $\lambda_{\rm E}$ on Earth, we say the star's redshift is $z = \lambda/\lambda_{\rm E} - 1$. One can also measure the flux of energy $\mathcal{F}$ from the star. If we understand the star well enough to know it should emit radiation at luminosity $\mathcal{L}$ (that is why Type Ia supernovae are important) we can infer its luminosity distance $d_{\rm L}$, which is the distance the star would be at if the scale factor was one,

$$\mathcal{F} = \frac{\mathcal{L}}{4\pi d_{\rm L}^2} \Longrightarrow d_{\rm L} = \sqrt{\frac{\mathcal{L}}{4\pi\mathcal{F}}}. \tag{200}$$

A Hubble plot is a graph of z versus $d_{\rm L}$ for many distant stars.

Stars throughout the universe move with respect to their local environments at typical velocities of about 10^{-3} the speed of light c. This motion gives rise to a special relativistic Doppler shift of $\Delta z \sim \pm 10^{-3}$. If spacetime was not expanding, this shift would be the only source of nonzero z, and averaging over many stars at the same luminosity distance would give zero redshift. That is just what happens for stars within our galaxy. However, the luminosity distances of stars in distant galaxies are observed to grow approximately linearly with their redshifts,

$$c^{-1}H_0 d_{\rm L} = z + \tfrac{1}{2}(1-q_0)z^2 + O(z^3). \tag{201}$$

One really does get H_0 from the slope, although inferring q_0 requires extending the plot to $z \sim 1$, at which point

the expansion breaks down and one must use the Einstein equations.

I cannot forbear to comment on the inconvenient minus sign in the definition (199) of $q(t)$. It was placed there because almost all theoretical physicists were certain the current universe must be decelerating before the measurement was done in 1998. (Checking this belief was not a priority; the head of one of the two teams who measured q_0 told me he was not even funded at the time!) Physicists like to define the parameters of equations to be positive so that one can infer general trends at a glance. Generations of physicists have cursed Benjamin Franklin for proposing the arbitrary sign convention that lead to electrons—which are the principal charge carrier of our electrical industry—having negative charge! But the universe played a trick on us and it is actually accelerating, rather than decelerating, so future generations will curse the minus sign we clever theorists inserted in the definition (199) of $q(t)$. Aside from general amusement, this tale should serve to caution readers about placing too much confidence (which means, any confidence at all) in the pronouncements of the scientific establishment on issues that have not yet been subjected to experimental and observational scrutiny.

5.2. Einstein's equations for FRW

Homogeneity and isotropy restrict the stress–energy tensor to only an energy density $\rho(t)$ and a pressure $p(t)$,

$$T_{00} = \rho(t), \qquad T_{0i} = 0, \qquad T_{ij} = p(t) g_{ij}, \tag{202}$$

where i and j are spatial indices. In this geometry Einstein's equations take the form

$$3H^2 - c^2\Lambda = \frac{8\pi G}{c^2}\rho, \tag{203}$$

$$-2\dot{H} - 3H^2 + c^2\Lambda = \frac{8\pi G}{c^2}p. \tag{204}$$

It is usual to redefine the energy density and pressure so as to absorb the cosmological constant,

$$\rho \longrightarrow \rho + \frac{c^4\Lambda}{8\pi G} \text{ and } p \longrightarrow p - \frac{c^4\Lambda}{8\pi G}. \tag{205}$$

When this is done, the current energy density is

$$\rho_0 = \frac{3c^2H_0^2}{8\pi G} \simeq 8.5 \times 10^{-10}\,\mathrm{J\,m^{-3}}. \tag{206}$$

This is the rest mass energy of about 5.7 Hydrogen atoms per cubic meter.

By differentiating (203) and then adding $3H$ times (203) plus (204), we derive a relation between the energy density and pressure known as stress–energy conservation,

$$\dot{\rho} = -3H(\rho + p). \tag{207}$$

If we also assume a constant equation of state, $w \equiv p(t)/\rho(t)$, then relation (207) can be used to express the energy density in terms of the scale factor,

$$\rho(t) = \rho_1 \left(\frac{a(t)}{a_1}\right)^{-3(1+w)}. \tag{208}$$

The substitution of (208) in (203) gives

$$a(t) = a_1[1 + \tfrac{3}{2}(1+w)H_1(t-t_1)]^{\frac{2}{3(1+w)}}. \tag{209}$$

The cases of $w = +\frac{1}{3}$, 0, $-\frac{1}{3}$ and -1 correspond to radiation, nonrelativistic matter, spatial curvature and vacuum energy (which includes the cosmological constant), respectively,

$$\text{Radiation} \Longrightarrow \rho \propto a^{-4}, \qquad a(t) \propto (H_1 t)^{\frac{1}{2}}, \tag{210}$$

$$\text{Non-relativistic matter} \Longrightarrow \rho \propto a^{-3}, \quad a(t) \propto (H_1 t)^{\frac{2}{3}}, \tag{211}$$

$$\text{Curvature} \Longrightarrow \rho \propto a^{-2}, \qquad a(t) \propto H_1 t, \tag{212}$$

$$\text{Vacuum energy} \Longrightarrow \rho \propto 1, \qquad a(t) \propto e^{H_1 t}. \tag{213}$$

The actual universe seems to be composed of at least three of the pure types, so the scale factor does not have a simple time dependence. However, as long as each type is separately conserved, we can use (208) to conclude that

$$\rho(t) = \frac{\rho_{\rm rad}}{a^4(t)} + \frac{\rho_{\rm mat}}{a^3(t)} + \frac{\rho_{\rm cur}}{a^2(t)} + \rho_{\rm vac}. \tag{214}$$

As the universe expands, the relative importance of the four types changes. Whenever a single type predominates, we can infer $a(t)$ from (209). This different dependence is one reason it makes sense to think of an early universe dominated by radiation (210), evolving to a universe dominated by nonrelativistic matter (211). It is also how one can understand that the current universe seems to be making the transition to domination by vacuum energy (213).

Under certain conditions there can be significant energy flows between three of the pure types of stress energy. For example, as the early universe cooled, massive particles changed from behaving like radiation to behaving like nonrelativistic matter. This change would increase $\rho_{\rm mat}$ and decrease $\rho_{\rm rad}$ in equation (214). The parameter that cannot change is that of the spatial curvature, $\rho_{\rm cur}$. I should not actually have regarded spatial curvature as a type of stress energy, but rather as an additional parameter in the homogeneous and isotropic metric (198). I avoided this complication because the extra terms it gives in the Einstein equations (203) and (204) can be subsumed into the energy density and pressure, and because the measured value of $\rho_{\rm cur}/a_0^2$ is consistent with zero [22].

5.3. Primordial inflation

The cosmology in which a radiation dominated universe evolves to matter domination is a feature of what is known as the Big Bang scenario. Although strongly supported by observation [135, 136], the composition of ρ at the start of radiation domination (at which the scale factor is $a_{\rm rad}$) does not seem natural,

$$\frac{\rho_{\rm rad}}{a_{\rm rad}^4} \gg \rho_{\rm vac}, \qquad \frac{\rho_{\rm rad}}{a_{\rm rad}^4} \gg \frac{\rho_{\rm cur}}{a_{\rm rad}^2}. \tag{215}$$

It might be expected instead that each of the three terms was comparable, in which case the universe would quickly

become dominated by vacuum energy. There is no accepted explanation for the first inequality of (215) or for the seeming coincidence that $\rho_{\rm mat}/a_0^3 \sim \rho_{\rm vac}$. However, the second inequality of (215) finds a natural explanation in the context of *primordial inflation*.

Inflation is defined as a phase of accelerated expansion, that is, $q(t) < 0$ with $H(t) > 0$. From the current values of the cosmological parameters one can see that the universe is in such a phase now. Recall from section 3.2 that explaining why this is happening is one thing a theory of quantum gravity might do. However, for now I wish to discuss primordial inflation, which is conjectured to have occurred at something like 10^{-37} s after the beginning of the universe with a Hubble parameter 55 orders of magnitude larger than it is today. I will be more specific later about what might cause inflation but, for now, let us assume it is driven by a vacuum energy $\rho'_{\rm vac} \gg \rho_{\rm vac}$ (remember it is only $\rho_{\rm cur}$ which cannot change) and that it begins at scale factor $a_{\rm inf}$. If the energy densities of curvature and vacuum energy are comparable at the beginning of inflation then their ratio by the onset of radiation domination is

$$\frac{\rho_{\rm cur}}{a_{\rm inf}^2 \rho'_{\rm vac}} \sim 1 \Longrightarrow \frac{\rho_{\rm cur}}{a_{\rm rad}^2 \rho'_{\rm vac}} \sim \left(\frac{a_{\rm inf}}{a_{\rm rad}}\right)^2 . \tag{216}$$

In the next subsection I will explain why this number is smaller than about 10^{-51}. Inflation makes the other types of stress energy even smaller, but there are mechanisms through which the primordial vacuum energy $\rho'_{\rm vac}$ can be converted into matter and radiation. This process, which I will not discuss, is known as *reheating*.

Inflation also explains how the large scale universe became so nearly homogeneous and isotropic. This explanation is crucial because gravity makes even tiny inhomogeneities grow, and the process has had 13.7 billion years to operate. It is believed that the galaxies of today's universe had their origins in quantum fluctuations of magnitude $\Delta\rho/\rho \simeq 10^{-5}$, which occurred during the last 60 e-foldings of inflation. The imprint of these fluctuations in the cosmic microwave background has recently been imaged with unprecedented accuracy by the WMAP satellite [22]. They also show up in large scale structure surveys [23]. If this view is correct, these observations represent the first quantum gravitational data.

5.4. The smoothness problem

There are many reasons for believing that the very early universe underwent a phase of primordial inflation [137]. I will confine myself to reviewing how inflation resolves the *smoothness problem*. This can be summed up in the question, why does the large scale universe possess such a simple geometry (198)?

To understand the problem we need to compare the distance light can travel from the beginning of the universe to the time of some observable event, with the distance it can travel from then to the present. Recall from special relativity that light rays travel along paths of zero invariant length, $ds^2 = 0$. From the invariant element (198) we see that the radial position of a light ray obeys, $dr = \pm c\, dt/a(t)$. The minus sign gives the *past light-cone* of the point $x^\mu = (ct_0, \vec{0})$, whereas the plus sign gives the *future light-cone* of a point $x^\mu = (ct_{\rm beg}, \vec{0})$ at the beginning of the universe,

$$R_{\rm past} = \int_{t_{\rm obs}}^{t_0} dt\, \frac{c}{a(t)}, \qquad R_{\rm future} = \int_{t_{\rm beg}}^{t_{\rm obs}} dt\, \frac{c}{a(t)}. \tag{217}$$

We can observe thermal radiation from the time of decoupling ($z_{\rm dec} \simeq 1089$) whose temperature is isotropic to one part in 10^5. This is a much higher degree of thermal equilibrium than exists in the air in any office! Unless the universe simply began this way—which seems unlikely—this equilibrium must have been established by processes acting at or below the speed of light. In other words, we must have $R_{\rm future} > R_{\rm past}$.

Suppose that, during the period $t_1 \leqslant t \leqslant t_2$, the deceleration parameter is constant $q(t) = q_1$. In that case we can obtain explicit expressions for the Hubble parameter and the scale factor in terms of their values at $t = t_1$,

$$H(t) = \frac{H_1}{1 + (1+q_1) H_1 (t - t_1)} \qquad \text{and}$$

$$a(t) = a_1 [1 + (1+q_1) H_1 (t - t_1)]^{\frac{1}{1+q_1}} . \tag{218}$$

These expressions permit us to evaluate the fundamental integral involved in the past and future light cones (217),

$$\int_{t_1}^{t_2} dt\, \frac{c}{a(t)} = \frac{c}{a_1 H_1 q_1} [1 + (1+q_1) H_1 (t - t_1)]^{\frac{q_1}{1+q_1}} \Big|_{t_1}^{t_2}$$

$$= \frac{c}{q_1} \left\{ \frac{1}{a_2 H_2} - \frac{1}{a_1 H_1} \right\} . \tag{219}$$

Although q_0 is negative, this is a recent event ($z \simeq 1$) which followed a long period of nearly perfect matter domination with $q = +\frac{1}{2}$. Much before the time of matter–radiation equality ($z_{\rm eq} \simeq 3300$) the universe was almost perfectly radiation dominated, which corresponds to $q = +1$. To simplify the computation we will ignore the recent phase of acceleration and also the transition periods,

$$a(t) H(t) = a_0 H_0 \begin{cases} \sqrt{1+z} & \forall\, z \leqslant z_{\rm eq} \\ \dfrac{1+z}{\sqrt{1+z_{\rm eq}}} & \forall\, z \geqslant z_{\rm eq} . \end{cases} \tag{220}$$

The cosmic microwave radiation was emitted within about a hundred redshifts of $z_{\rm dec} < z_{\rm eq}$, so the past light cone is

$$R_{\rm past} = \frac{2c}{a_0 H_0} \left\{ \frac{1}{\sqrt{1+0}} - \frac{1}{\sqrt{1+z_{\rm dec}}} \right\} \simeq \frac{2c}{a_0 H_0} . \tag{221}$$

The future light-cone derives from both epochs and it depends slightly upon the beginning redshift, $z_{\rm beg}$

$$R_{\rm future} = \frac{2c}{a_0 H_0} \left\{ \frac{1}{\sqrt{1+z_{\rm dec}}} - \frac{1}{\sqrt{1+z_{\rm eq}}} \right\}$$

$$+ \frac{c}{a_0 H_0} \left\{ \frac{\sqrt{1+z_{\rm eq}}}{1+z_{\rm eq}} - \frac{\sqrt{1+z_{\rm eq}}}{1+z_{\rm beg}} \right\} . \tag{222}$$

One maximizes $R_{\rm future}$ by taking $z_{\rm beg} \to \infty$, but it is not enough,

$$\lim_{z_{\rm beg} \to \infty} R_{\rm future} \simeq \frac{c}{a_0 H_0} \left\{ \frac{2}{\sqrt{z_{\rm dec}}} - \frac{1}{\sqrt{z_{\rm eq}}} \right\} . \tag{223}$$

Under the assumption of $q = +1$ before $z_{\rm eq}$ we are forced to conclude that the two-dimensional surface we can see from the time of decoupling consists of about

$$\left(\frac{R_{\rm past}}{R_{\rm future}}\right)^2 \simeq \frac{z_{\rm dec}}{\left[1 - \frac{1}{2}\left(\frac{z_{\rm dec}}{z_{\rm eq}}\right)^{\frac{1}{2}}\right]^2} \simeq 2200, \tag{224}$$

regions which cannot have exchanged even a photon since the beginning of time! So how did these 2200 different regions reach equilibrium to one part in 10^5?

The problem grows worse the earlier one believes the universe was homogeneous. For example, the seven lightest nuclear species were almost all produced during Nucleosynthesis at about $z_{\rm nuc} \simeq 10^9$ and their isotopic abundances seem to be uniform over the observed universe. For Nucleosynthesis the radii of the past and future light-cones are

$$R_{\rm past} = \frac{2c}{a_0 H_0}\left\{\frac{1}{\sqrt{1+0}} - \frac{1}{\sqrt{1+z_{\rm dec}}}\right\} + \frac{c}{a_0 H_0}\left\{\frac{1}{\sqrt{1+z_{\rm eq}}} - \frac{\sqrt{1+z_{\rm eq}}}{1+z_{\rm nuc}}\right\} \simeq \frac{2c}{a_0 H_0}, \tag{225}$$

$$R_{\rm future} = \frac{c}{a_0 H_0}\left\{\frac{\sqrt{1+z_{\rm eq}}}{1+z_{\rm nuc}} - \frac{\sqrt{1+z_{\rm eq}}}{1+z_{\rm beg}}\right\} \simeq \frac{c}{a_0 H_0}\frac{\sqrt{z_{\rm eq}}}{z_{\rm nuc}}. \tag{226}$$

So the assumption of $q = +1$ for all time before $t_{\rm eq}$ implies that the number of causally disconnected regions at the time of Nucleosynthesis which comprise the current universe is about

$$\left(\frac{R_{\rm past}}{R_{\rm future}}\right)^2 \simeq \frac{4 z_{\rm nuc}^2}{z_{\rm eq}} \simeq 10^{15}. \tag{227}$$

The corresponding numbers for the phase of quark-gluon plasma ($z \sim 10^{12}$) and the electroweak phase transition ($z \sim 10^{15}$) are 10^{21} and 10^{27}, respectively. One could quibble about the extent to which we know the universe was homogeneous at these times but it seems obvious something is very wrong with the assumption of positive deceleration throughout cosmic history.

This embarrassment resulted from the fact that the upper limit of integration dominates $R_{\rm future}$ for positive deceleration. Inflation solves the problem by positing a very early epoch of negative deceleration. Let us suppose the deceleration is $q = -1$ for $z > z_{\rm rad}$. In that case our simplified model of cosmic history (220) generalizes to

$$a(t)H(t) \simeq a_0 H_0 \begin{cases} \sqrt{1+z} & 0 < z < z_{\rm eq}, \\ \dfrac{z}{\sqrt{z_{\rm eq}}} & z_{\rm eq} < z < z_{\rm rad}, \\ \dfrac{z_{\rm rad}^2}{\sqrt{z_{\rm eq}}\, z} & z_{\rm rad} < z < z_{\rm inf}. \end{cases} \tag{228}$$

Now let us compute the past and future light-cones of some event at redshift z in the radiation dominated period. Of course our previous result of $R_{\rm past} \simeq 2c/a_0 H_0$ is still valid. However, the initial phase of acceleration makes a profound change in the future light-cone. During acceleration it is the lower limit of integration which dominates the future light-cone, and the result can be made as large as desired simply by increasing the redshift $z_{\rm inf}$ at the beginning of inflation. Under the assumption that $z_{\rm inf} \gg z_{\rm rad} \gg z$ we get

$$R_{\rm future} \simeq \frac{c}{a_0 H_0}\left\{\frac{\sqrt{z_{\rm eq}}}{z} - \frac{\sqrt{z_{\rm eq}}}{z_{\rm rad}}\right\} - \frac{c}{a_0 H_0}\left\{\frac{\sqrt{z_{\rm eq}}}{z_{\rm rad}} - \frac{\sqrt{z_{\rm eq}} z_{\rm inf}}{z_{\rm rad}^2}\right\}, \tag{229}$$

$$\simeq \frac{c}{a_0 H_0}\frac{\sqrt{z_{\rm eq}} z_{\rm inf}}{z_{\rm rad}^2}. \tag{230}$$

Inflation explains the smoothness of the large scale universe by supposing that everything we now see derived from a region which was small enough for causal processes to make it homogeneous and isotropic. Then inflation stretched it out and the various pieces slowly came back into contact, after inflation, still looking very much alike. In fact, the inhomogeneities we now see on less than cosmic scales derived from almost 14 billion years of gravitational collapse operating on a universe that was smooth to one part in 10^5 just after inflation.

The usual assumption that $z_{\rm rad} \simeq 10^{26}$ derives from supposing that radiation domination commences at a scale of about 10^{13} GeV. If we require $R_{\rm future} \gtrsim R_{\rm past}$ then $z_{\rm inf} \gtrsim 10^{51}$. (Many models of inflation vastly exceed this minimum.) Note that primordial inflation not only solves the smoothness problem but it also explains why the spatial curvature (216) is small, which is an observed fact [22].

5.5. *Slow roll scalar-driven inflation*

The case for an early period of accelerated expansion is very strong, and the idea was suggested even before the advent of inflation [138]. However, a completely satisfactory mechanism for causing accelerated expansion has yet to be identified, either for primordial inflation or for the current phase. Guth's proposal [21] failed to have a satisfactory ending but this was quickly corrected by the *slow roll scalar* models proposed by Linde [139] and by Albrecht and Steinhardt [140]. Although many other classes of models have since then been devised, and none are without problems, these are the simplest and I will describe them.

Slow roll scalar models are based upon a hypothetical spin zero (scalar) field called *the inflaton* $\varphi(t, \vec{x})$. (Not having a good candidate for this field from fundamental theory is one problem with these models.) The Lagrangian for $\varphi(t, \vec{x})$ also involves the metric $g_{\mu\nu}(t, \vec{x})$,

$$L = \int {\rm d}^3x \sqrt{-\det(g_{\alpha\beta})} \times \left\{-\frac{1}{2}\sum_{\mu=0}^{3}\sum_{\nu=0}^{3}\frac{\partial\varphi}{\partial x^\mu}\frac{\partial\varphi}{\partial x^\nu}\left(g^{-1}\right)^{\mu\nu} - V(\varphi)\right\}, \tag{231}$$

where $\det(g_{\alpha\beta})$ is the determinant of the metric and g^{-1} is its matrix inverse. Note how the metric, which is the true measure of lengths and times, modifies the infinitesimal coordinate

volume d^3x and the derivatives. This is typical of the way it couples to matter in general relativity.

The stress–energy tensor of the inflaton is

$$T_{\mu\nu} = \frac{\partial\varphi}{\partial x^\mu}\frac{\partial\varphi}{\partial x^\nu} - \frac{1}{2}g_{\mu\nu}\sum_{\rho=0}^{3}\sum_{\sigma=0}^{3}\frac{\partial\varphi}{\partial x^\rho}\frac{\partial\varphi}{\partial x^\sigma}(g^{-1})^{\rho\sigma} - g_{\mu\nu}V(\varphi). \quad (232)$$

One can see that the term involving the scalar potential has the same form as the cosmological constant Λ in the Einstein equations (9). By itself, this contribution would tend to make the universe accelerate, but one must also reckon with the kinetic terms which involve derivatives. In order for inflation to start, the scalar potential energy must dominate over its kinetic energy throughout a region somewhat larger than light can cross from the beginning of the universe. This is not anywhere near as bad as the terrific mismatches we found in the previous subsection but it does concern inflationary cosmologists because such an initial condition can only have been an accident. Estimating how unlikely this accident is depends upon what one believes about how the universe began, and also involves tricky questions about how to compute probabilities. However, all estimates give very small numbers (I have heard 10^{-120}) for the chances of it happening.

Rather than start from an inhomogeneous configuration I will simply assume the initial condition was homogeneous and isotropic. I will also assume spatial flatness, which would in any case result, approximately, from a long phase of inflation. This means the metric takes the FRW form (198) and that the scalar depends only upon time, $\varphi(t,\vec{x}) \rightarrow \varphi_0(t)$. The nontrivial Einstein equations are

$$3H^2 - c^2\Lambda = \frac{8\pi G}{c^2}\left[\frac{\dot{\varphi}_0^2}{2c^2} + V(\varphi_0)\right], \quad (233)$$

$$-2\dot{H} - 3H^2 + c^2\Lambda = \frac{8\pi G}{c^2}\left[\frac{\dot{\varphi}_0^2}{2c^2} - V(\varphi_0)\right]. \quad (234)$$

The deceleration parameter (times H^2) can be computed from a linear combination of these equations,

$$qH^2 = -\dot{H} - H^2 = \frac{8\pi G}{3c^2}\left[\frac{\dot{\varphi}_0^2}{c^2} - V(\varphi_0) - \frac{c^4\Lambda}{8\pi G}\right]. \quad (235)$$

So the condition for accelerated expansion is

$$V(\varphi_0) + \frac{c^4\Lambda}{8\pi G} > \frac{\dot{\varphi}_0^2}{c^2}. \quad (236)$$

The equation for the homogeneous scalar $\varphi_0(t)$ is

$$\ddot{\varphi}_0 + 3H\dot{\varphi}_0 + c^2V'(\varphi_0) = 0, \quad (237)$$

where $V'(\varphi) \equiv \partial V/\partial\varphi$. The simplest model of scalar-driven inflation, and one which is still consistent with all data [22], consists of a constant plus a quadratic term

$$V(\varphi) = V_0 + \frac{1}{2}\left(\frac{mc}{\hbar}\right)^2\varphi^2 \Longrightarrow V'(\varphi) = \left(\frac{mc}{\hbar}\right)^2\varphi. \quad (238)$$

Substituting (238) into (237) gives the equation for a damped harmonic oscillator,

$$\ddot{\varphi}_0 + 3H\dot{\varphi}_0 + \left(\frac{mc^2}{\hbar}\right)^2\varphi_0 = 0. \quad (239)$$

The term $3H\dot{\varphi}_0$ in equations (237) and (239) is known as *Hubble friction*. For sufficiently large $H(t)$ one can see that it makes the scalar over-damped so that it slowly rolls down the quadratic potential (238). That makes the scalar kinetic energy small, which will enforce condition (236) for accelerated expansion provided two more conditions hold.

The first of these extra conditions is that the constant V_0 must be chosen to almost cancel the cosmological constant,

$$V_0 = -\frac{c^4\Lambda}{8\pi G} + \rho_{\rm vac}, \quad (240)$$

where $\rho_{\rm vac} \sim 6\times10^{-10}\,{\rm J\,m^{-3}}$ is the currently observed value of the vacuum energy [8, 9]. Of course the absence of any explanation for this choice is the problem of the cosmological constant that I discussed in section 3.2. The second extra condition is that the scalar must start with a large enough initial value, and a small enough initial time derivative, so that (236) holds initially. The absence of any explanation for this is part of what is known as the *initial condition problem.*

If one assumes these two conditions then Hubble friction causes the scalar to slowly roll down its potential. As it rolls, the Hubble parameter grows smaller, which makes Hubble friction less effective. Eventually the system becomes under-damped and begins oscillating. If the scalar is coupled to other fields (the quantum corrections from which must be prevented from distorting its potential too much—another problem with this class of models!) then this phase of oscillations can result in a hot, radiation dominated universe [141]. By choosing the initial value of the scalar sufficiently large one can make the phase of inflation last arbitrarily long, although making it too large can force the system to a regime known as *eternal inflation* in which quantum fluctuations actually push the scalar up its potential.

I have described scalar slow roll models to show that primordial inflation can be supported in a relatively simple way, even if this requires a number of arbitrary assumptions. There are other models, none of which is without problems. Although I very much doubt any of these models is correct, I consider the case for an early phase of accelerated expansion to be to be overwhelming. Finding a realistic model which causes this phase is one of the things I hope quantum gravity can do.

5.6. The strength of quantum effects during inflation

Leonard Parker was the first to give a quantitative assessment of how spacetime expansion affects quantum processes [142] but one can understand some of the things he found in a qualitative way. In particular, I will try to explain three crucial facts:

- The expansion of spacetime strengthens quantum effects;
- This strengthening is greatest during accelerated expansion;

- Just as in flat space, the largest quantum effects derive from the lightest particles *provided* they can avoid being driven to zero amplitude by the expansion of the universe.

Recall from section 2 that quantum fields obey exactly same equations as their classical counterparts, so one can understand quantum effects as the classical response to the 0-point motion which is required by the uncertainty principle. For example, we saw that vacuum polarization works the same way as classical polarization in a medium if one simply accepts that each mode of the electron field has 0-point motion. The amount of 0-point motion a field experiences is controlled by its free field mode functions, and whatever increases the amount of 0-point motion will strengthen quantum effects. For example, we observed from the oscillatory factors of $e^{iEt/\hbar}$ in the electron mode functions (41) that the combination of an electron of wave vector $\vec{k}$ and a positron of wave vector $\vec{p}-\vec{k}$ can only remain coherent for a time Δt of about

$$\Delta t \sim \frac{\hbar}{E(\vec{k}) + E(\vec{p}-\vec{k})} \qquad \text{where}$$

$$E(\vec{k}) \equiv \sqrt{m^2c^4 + \hbar^2 c^2 \|\vec{k}\|^2}. \tag{241}$$

That is the only quantum mechanics one needs. The remainder of the computation consists of using the Lorentz force law to find the classical polarization induced by such a pair being acted upon for time Δt by a static electric field $\tilde{\vec{E}}(\vec{p})$,

$$e\Delta\vec{x} \sim \frac{e^2c^2\Delta t^2 \tilde{\vec{E}}(\vec{p})}{E(\vec{k}) + E(\vec{p}-\vec{k})} \sim \frac{e^2c^2\hbar^2 \tilde{\vec{E}}(\vec{p})}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3}. \tag{242}$$

Summing over all modes $\vec{k}$ gives (up to a factor of $\frac{8}{3}$) the actual first order result (99) and (100) for the vacuum polarization due to a static field,

$$\tilde{\vec{P}}(\vec{p}) \sim \int \frac{d^3k}{(2\pi)^3} \frac{e^2\hbar^2c^2 \tilde{\vec{E}}(\vec{p})}{[E(\vec{k}) + E(\vec{p}-\vec{k})]^3}. \tag{243}$$

The reason spacetime expansion strengthens quantum effects is that physical wave numbers redshift. Because the FRW geometry is invariant under spatial translations, particles are still labeled by constant wave vectors $\vec{k}$. However, because $\|\vec{k}\| = 2\pi/\lambda$ is the inverse of a coordinate wavelength, not the physical wavelength $a(t)\lambda$, it is really the combination $\vec{k}/a(t)$ which tends to enter physical expressions. The actual mode functions of a massive particle are complicated but it is not a bad approximation to think of the term $E(\vec{k})\Delta t/\hbar$ in the phase of a mode function generalizing to

$$\frac{E(\vec{k})\Delta t}{\hbar} \Longrightarrow \int_t^{t+\Delta t} dt' \frac{E(t',\vec{k})}{\hbar} \qquad \text{where}$$

$$E(t,\vec{k}) = \sqrt{m^2c^4 + \hbar^2c^2\|\vec{k}\|^2/a^2(t)}. \tag{244}$$

The expansion of $a(t')$ always makes the accumulated phase smaller than it would have been for constant scale factor, so the mode can persist longer and we see why the expansion of spacetime strengthens quantum effects.

Just as in flat space (that is, $a(t) = 1$) particles with the smallest masses remain coherent longest, which is why almost all vacuum polarization comes from electrons and positrons, even though there are many other charged particle fields. Setting $m = 0$ in expression (244) gives the same integral (217) we saw in section 5.4,

$$\lim_{m\to 0} \int_t^{t+\Delta t} dt' \frac{E(t',\vec{k})}{\hbar} = c\|\vec{k}\| \times \int_t^{t+\Delta t} dt' \frac{1}{a(t')}. \tag{245}$$

Now recall the key distinction between accelerated expansion and deceleration:

- For deceleration ($q(t) > 0$) the integral (245) grows without bound as Δt increases; but
- For acceleration ($q(t) < 0$) the integral (245) approaches a constant as Δt goes to infinity.

The first fact means that modes in a decelerating universe must eventually become decoherent, even though they persist longer than for a static universe. The second fact means that the 0-point motion of a sufficiently long wavelength mode can persist forever during inflation.

This is not quite the end of the story because almost all particles develop a symmetry known as *conformal invariance* when they become massless. This symmetry causes their mode functions to fall off like powers of the scale factor $a(t)$. For example, if we set the electron mass to zero in (41) and then account for the FRW geometry, the electron mode function becomes

$$\varepsilon_i(t,\vec{x};\vec{k},s) = \sqrt{\frac{\hbar}{2k}}\, e^{-ikct + i\vec{k}\cdot\vec{x}}\, u_i(\vec{k},s), \tag{246}$$

$$\longrightarrow [a(t)]^{-\frac{3}{2}} \sqrt{\frac{\hbar}{2k}} e^{-ikc\int_0^t dt'/a(t') + i\vec{k}\cdot\vec{x}} u_i(\vec{k},s), \tag{247}$$

where the spinor wave function $u_i(\vec{k},s)$ is unchanged from its flat space value (37) with $m = 0$. Of course this decreasing amplitude tends to suppress 0-point motion, which weakens quantum effects.

To complete the third point on my list I need to show that there are massless particles whose mode functions avoid being driven to zero. It turns out there are two such particles:

- Scalars like the inflaton (231) but with zero potential $V(\varphi)$ and
- Gravitons.

It also turns out that the mode functions of gravitons obey the same equations as those of the massless inflaton [143], so I will specialize to the latter.

We can understand the physics of massless inflatons by specializing the Lagrangian (231) to the FRW geometry (198) but still considering $\varphi(t,\vec{x})$ to be an arbitrary function of space and time, and then using Parseval's theorem,

$$L = \int d^3x a^3(t) \left\{ \frac{1}{2}\left(\frac{\dot{\varphi}(t,\vec{x})}{c}\right)^2 - \frac{1}{2}\left\| \frac{\vec{\nabla}\varphi(t,\vec{x})}{a(t)} \right\|^2 \right\}, \tag{248}$$

$$= \int \frac{d^3k}{(2\pi)^3} a^3(t) \left\{ \frac{1}{2c^2} |\dot{\tilde{\varphi}}(t,\vec{k})|^2 - \frac{\|\vec{k}\|^2}{2a^2(t)} |\tilde{\varphi}(t,\vec{k})|^2 \right\}. \tag{249}$$

It follows that the field at each Fourier wave vector $\vec{k}$ behaves as an independent harmonic oscillator with time-dependent mass and frequency,

$$m(t) = m_0 a^3(t) \qquad \text{and} \qquad \omega(t) = \frac{kc}{a(t)}. \tag{250}$$

If we use $q(t)$ to represent the position of such an oscillator its Lagrangian and energy are

$$L(t) = \frac{m(t)}{2}[\dot{q}^2(t) - \omega^2(t)q^2(t)]$$

$$\Longrightarrow E(t) = \frac{m(t)}{2}[\dot{q}^2(t) + \omega^2(t)q^2(t)]. \tag{251}$$

All the usual theorems of quantum mechanics apply to this system. In particular, the state with minimum energy at any fixed time t must be

$$E_{\min}(t) = \frac{\hbar kc}{2a(t)}. \tag{252}$$

However, the time dependence of the scale factor means that the minimum energy state at one time is not the same at other times! The usual choice for the ground state under these circumstances is called *Bunch–Davies vacuum* [144] and it corresponds to the state which was minimum energy in the distant past.

Let us find an expression for $q(t)$, analogous to (25), in terms of the raising and lowering operators $\alpha^\dagger$ and α for Bunch–Davies vacuum $|\Omega\rangle$. It takes the form

$$q(t) = \alpha\varepsilon(t) + \alpha^\dagger\varepsilon^*(t) \qquad \text{where}$$

$$[\alpha, \alpha^\dagger] = 1 \qquad \text{and} \qquad \alpha|\Omega\rangle = 0. \tag{253}$$

The mode functions obey,

$$\ddot{\varepsilon}(t) + 3H(t)\dot{\varepsilon}(t) + \left(\frac{kc}{a(t)}\right)^2 \varepsilon(t) = 0 \qquad \text{and}$$

$$\varepsilon(t)\dot{\varepsilon}^*(t) - \dot{\varepsilon}(t)\varepsilon^*(t) = \frac{\mathrm{i}\hbar}{m_0 a^3(t)}. \tag{254}$$

These equations are too difficult to solve for general scale factor $a(t)$ but the solution for $H(t) = H_I$ (which corresponds to the $q(t) = -1$ paradigm we employed for primordial inflation in section 5.4) is

$$a(t) \propto \mathrm{e}^{H_I t} \Longrightarrow \varepsilon(t)$$

$$= \sqrt{\frac{\hbar H_I^2}{2m_0 c^3 k^3}}\left[1 - \frac{\mathrm{i}ck}{H_I a(t)}\right]\exp\left[\frac{\mathrm{i}ck}{H_I a(t)}\right]. \tag{255}$$

Now evaluate the expectation value of the energy operator in Bunch–Davies vacuum, first for arbitrary $a(t)$ and then for (255),

$$\langle\Omega|E(t)|\Omega\rangle = \frac{1}{2}m_0 a^3(t)\left[\dot{\varepsilon}(t)\dot{\varepsilon}^*(t) + \left(\frac{kc}{a(t)}\right)^2 \varepsilon(t)\varepsilon^*(t)\right], \tag{256}$$

$$\longrightarrow \frac{\hbar kc}{2a(t)} + \frac{\hbar H_I^2 a(t)}{4ck} = \frac{\hbar kc}{a(t)}\left[\frac{1}{2} + \left(\frac{H_I a(t)}{2kc}\right)^2\right]. \tag{257}$$

Recall that this system has the energy eigenstates of a harmonic oscillator with energies $(N + \frac{1}{2})\hbar\omega(t)$ at any fixed instant in time. So it is completely valid to regard the final expression in (257) as giving the occupation number,

$$N(t) = \left[\frac{H_I a(t)}{2kc}\right]^2. \tag{258}$$

Hence we see that Bunch–Davies vacuum starts out empty. Expression (255) reveals that for $kc \gg H_I a(t)$ the mode function oscillates and falls off like $1/a(t)$, which is how a massless, conformally invariant scalar behaves. As the scale factor grows, the occupation number (258) increases, and both the oscillations in the mode function and the decrease in its amplitude slow down. A key event is first horizon crossing when $kc = H(t)a(t)$, at which point the occupation number becomes order one and the mode function begins approaching a constant. After that the occupation number becomes exponentially large, as does the amount of energy in this single mode.

One might worry that the behavior I have just sketched is very special to the case of $q(t) = -1$ but that is not true. As long as the universe is accelerating the product $H(t)a(t)$ grows, and one can see from equation (254) that modes which start with $kc \gg H(t)a(t)$ oscillate and fall off,

$$kc \gg H(t)a(t)$$

$$\Longrightarrow \varepsilon(t) \sim \sqrt{\frac{\hbar}{2m_0 kc}}\,\frac{1}{a(t)}\exp\left[-\mathrm{i}kc\int_0^t \mathrm{d}t'\,\frac{1}{a(t')}\right]. \tag{259}$$

They are also drawn towards first horizon crossing. If inflation persists long enough for them to reach it, the $(kc/a)^2$ term of (254) drops out and one can see that the mode function approaches a constant. So the behavior I found for (255) is actually generic to any inflating geometry.

Of course this is why massless inflatons and gravitons show enhanced quantum effects during inflation. Before closing I should comment on the sheer wonder of what we found in expression (257). This represents the 0-point energy of one single mode! Of course one must divide it over the vast 3-volume of the inflating universe, but summing over the modes which have experienced first horizon crossing gives a small, macroscopic effect. Another important comment is that this result derives from the long wavelength regime in which perturbative quantum general relativity should be valid, even if quantum gravity is described, on the fundamental level, by some other theory, or if perturbative methods do not give the correct asymptotic series. My final comment is that there is no natural mechanism for keeping the inflaton massless, and if it develops a large mass that will suppress quantum effects during inflation the same way it would in flat space. By contrast, nothing needs to be done to keep the graviton massless, so this long wavelength regime during inflation is a natural place to look for quantum gravitational effects.

5.7. Quantum gravitational data

I have already discussed scalar-driven inflation and why one expects enhanced quantum effects during this epoch.

In this subsection I will sketch the theory behind the first quantum gravitational observables ever measured, which are cosmological perturbations. The original work on tensor perturbations was done by Starobinsky [145]; Mukhanov and Chibisov did the first calculation of scalar perturbations [146]. For more details I recommend the excellent recent text by Mukhanov [141].

Cosmological perturbations are spacetime dependent fluctuations of the full scalar and metric fields around the spatially homogeneous background values described in section 5.5,

$$\varphi(t,\vec{x}) = \varphi_0(t) + \delta\varphi(t,\vec{x}), \tag{260}$$

$$g_{\mu\nu}(t,\vec{x}) = \bar{g}_{\mu\nu}(t) + h_{\mu\nu}(t,\vec{x}). \tag{261}$$

Here $\bar{g}_{\mu\nu}(t)$ stands for the FRW metric (198),

$$\bar{g}_{00}(t) = -1, \qquad \bar{g}_{0i}(t) = 0, \qquad \bar{g}_{ij}(t) = a^2(t)\delta_{ij}. \tag{262}$$

It is typical to express the graviton using just scalar and tensor fields because the linearized Einstein equations cause the vector fields to vanish

$$h_{00}(t,\vec{x}) = -2\phi(t,\vec{x}), \tag{263}$$

$$h_{0i}(t,\vec{x}) = -c\frac{\partial}{\partial x^i}B(t,\vec{x}), \tag{264}$$

$$h_{ij}(t,\vec{x}) = -2a^2(t)\psi(t,\vec{x})\delta_{ij} - 2c^2\frac{\partial}{\partial x^i}\frac{\partial}{\partial x^j}E(t,\vec{x}) + a^2(t)h^{\rm TT}_{ij}(t,\vec{x}). \tag{265}$$

Here the dynamical graviton field $h^{\rm TT}_{ij}$ is both transverse and traceless,

$$\sum_{i=1}^{3}\frac{\partial}{\partial x^i}h^{TT}_{ij}(t,\vec{x}) = 0 = \sum_{i=1}^{3}h^{\rm TT}_{ii}(t,\vec{x}). \tag{266}$$

Dynamical gravitons are invariant under linearized coordinate transformations and their spatial Fourier transforms obey the same equation (254) as a massless inflaton,

$$\left[\left(\frac{\partial}{\partial t}\right)^2 + 3H(t)\frac{\partial}{\partial t} + \frac{k^2c^2}{a^2(t)}\right]\tilde{h}^{\rm TT}_{ij}(t,\vec{k}) = 0. \tag{267}$$

We have already seen that for $kc \gg H(t)a(t)$ the solutions are oscillatory with amplitudes that fall off like $1/a(t)$. Long after first horizon crossing, in the regime for which $kc \ll H(t)a(t)$ the term proportional to $(kc/a)^2$ drops out and the solution goes to a linear combination of a constant and the falling indefinite integral

$$\int^t {\rm d}t'\frac{1}{a^3(t')}. \tag{268}$$

The following combinations of the scalar fields are invariant under linearized coordinate transformations:

$$\Phi(t,\vec{x}) \equiv \phi(t,\vec{x}) - \frac{\partial}{\partial t}[B(t,\vec{x}) - \dot{E}(t,\vec{x}) + 2H(t)E(t,\vec{x})], \tag{269}$$

$$\Psi(t,\vec{x}) \equiv \psi(t,\vec{x}) + H(t)[B(t,\vec{x}) - \dot{E}(t,\vec{x}) + 2H(t)E(t,\vec{x})], \tag{270}$$

$$\Xi(t,\vec{x}) \equiv \delta\varphi(t,\vec{x}) - \dot{\varphi}_0(t)\Big[B(t,\vec{x}) - \dot{E}(t,\vec{x}) + 2H(t)E(t,\vec{x})\Big]. \tag{271}$$

The linearized Einstein equations for indices $\mu = i$ and $\nu = j$ imply $\Psi = \Phi$, whereupon the linearized $\mu = 0$, $\nu = 0$ and $\mu = 0$, $\nu = i$ equations become

$$\left[6H\frac{\partial}{\partial t} + 6H^2 - \frac{2c^2\nabla^2}{a^2}\right]\Phi = \frac{8\pi G}{c^4}\left[\dot{\varphi}_0^2\Phi - \dot{\varphi}_0\dot{\Xi} - c^2V'(\varphi_0)\Xi\right], \tag{272}$$

$$-\frac{2}{a}\frac{\partial}{\partial x^i}\Big[\dot{\Phi} + H\Phi\Big] = -\frac{1}{a}\frac{\partial}{\partial x^i}\left[\frac{8\pi G}{c^4}\,\dot{\varphi}_0\Xi\right]. \tag{273}$$

Of course equation (273) relates the scalar perturbation to the scalar part of the metric perturbation,

$$\Xi(t,\vec{x}) = \frac{c^4}{4\pi G}\left[\frac{\dot{\Phi}(t,\vec{x}) + H(t)\Phi(t,\vec{x})}{\dot{\varphi}_0(t)}\right]. \tag{274}$$

Substituting this in (272), taking the spatial Fourier transform, and making some simplifications eventually results in what has been termed the *Mukhanov equation* for $\tilde{\Phi}(t,\vec{k})/\dot{\varphi}_0(t)$,

$$\left[\left(\frac{\partial}{\partial t}\right)^2 + H\frac{\partial}{\partial t} + \frac{k^2c^2}{a^2(t)} - \left(\frac{\ddot{\theta}(t) + H(t)\dot{\theta}(t)}{\theta(t)}\right)\right] \times\left(\frac{\tilde{\Phi}(t,\vec{k})}{\dot{\varphi}_0(t)}\right) = 0, \tag{275}$$

where the function

$$\theta(t) \equiv \frac{H(t)}{a(t)\dot{\varphi}_0(t)}. \tag{276}$$

The solution of (275) is qualitatively similar to that of the massless inflaton. In the regime $kc \gg H(t)a(t)$ the field oscillates and its amplitude falls off. Long after first horizon crossing we can again drop the term proportional to $(kc/a)^2$ and the solution becomes a linear combination of $\theta(t)$ and the indefinite integral

$$\theta(t)\int^t {\rm d}t'\frac{1}{a(t')\theta^2(t')}. \tag{277}$$

It is this second solution which dominates long after first horizon crossing.

For both the scalar perturbations of $\tilde{\Phi}(t,\vec{k})$ and the tensor perturbations of $\tilde{h}^{\rm TT}_{ij}(t,\vec{k})$ we can get good approximate solutions for the mode functions in the regime $kc \gg H(t)a(t)$. Canonical quantization fixes their normalization. Long after first horizon crossing we can again get good solutions, one of which approaches a constant with the other falling. The strength of the each perturbation is quantified by its *power spectrum*. If the spatial Fourier transform of some quantum

field $F(t,\vec{x})$ approaches a constant we define its power spectrum $\mathcal{P}_F(k)$ as

$$\langle\Omega]|\tilde{F}(t,\vec{k})\tilde{F}^{\dagger}(t,\vec{k}')|\Omega\rangle \equiv \frac{\mathcal{P}_F(k)}{4\pi k^3} \times (2\pi)^3\delta^3(\vec{k}-\vec{k}'). \tag{278}$$

It is quite a challenge to connect the early, oscillatory regime—for which the normalization is known—to the late time regime, long after first horizon crossing [147]. Approximate results can be obtained by crudely matching the early and late time solutions at the time t_k for which wave number k experiences first horizon crossing,

$$kc \equiv H(t_k)a(t_k). \tag{279}$$

These mode-matching results are

$$\mathcal{P}_\Phi(k) \sim \frac{GH^2(t_k)}{1+q(t_k)}, \tag{280}$$

$$\mathcal{P}_h(k) \sim GH^2(t_k), \tag{281}$$

where we recall from expression (199) that $q(t_k)$ is the deceleration parameter at first horizon crossing.

There is also a conceptual challenge to understand the manner in which the quantum fluctuations of primordial inflation become the approximately classical perturbations we observe at much later stages in cosmic history. This is the subject of some controversy, and viewpoints differ depending upon how one interprets quantum mechanics. It is too late in an already long paper to inflict a lengthy discussion of these issues so let me instead point to the excellent papers by Polarski, Starobinsky and their collaborators [148].

It should be noted that (280) and (281) are the *primordial* power spectra. What is actually measured is the result of complicated but quite well understood physics that occurs after inflation, during the subsequent epochs of radiation domination and matter domination. This is included through a known *transfer function*, so the challenge for fundamental theory is to compute the primordial power spectra. It should also be noted that the primordial power spectra are almost independent of k because $H(t)$ is nearly constant during inflation. The data are therefore typically organized into scalar and tensor *spectral indices*,

$$n_s(k) \equiv 1 + \frac{\partial \ln[\mathcal{P}_\Phi(k)]}{\partial \ln(k)}, \tag{282}$$

$$n_t(k) \equiv \frac{\partial \ln[\mathcal{P}_h(k)]}{\partial \ln(k)}, \tag{283}$$

reported at a fiducial wave number. Lately there has been an effort to report the first derivative of the scalar spectral index as well [22].

One can see from expressions (280) and (281) that the scalar power spectrum is enhanced by a factor of $1/(1+q)$, relative to the tensor one. Because q is very near -1 during inflation this is a large enhancement, and that is why the tensor signal has so far not been detected. Getting a nonzero measurement for $\mathcal{P}_h(k)$ is very important because it would tell us the scale of primordial inflation. It would also be the first proof that gravitons exist and are quantized.

I want to emphasize that primordial perturbations are a quantum gravitational effect. Many quantum gravity experts dismiss them because we have so far not detected the tensor signal from gravitons, but this is wrong-headed and equivalent to ignoring the solar system tests of classical general relativity because they concern that part of the metric field which is fixed by matter. The way quantum matter gravitates is as much a quantum gravitational effect as the way classical matter gravitates is a classical gravitational effect. Further, it tells us precisely what we most wish to know. There is no problem with pure gravitons at the lowest order in perturbation theory; the problem at lowest order comes from matter, and the scalar perturbation signal probes precisely this effect at an energy scale potentially as high as 10^{13} GeV. If I had been told only one of the two perturbation spectra could be measured and asked to choose which one, I would have picked the scalar. This is priceless information, although it would be better if we has a unique theory of inflation which made specific predictions. As it is, one can accommodate almost any data by changing the model of inflation, and none of the models proposed so far is very compelling.

I should like to close with two more comments, one about experiment and the other about theory. First, there are a lot more data on the way. The European Space Agency has launched the Planck satellite which will try to resolve the primordial tensor spectrum. Planck may be able to measure the tensor-to-scalar ratio r to an accuracy of 0.05, compared with the present bound of $r < 0.22$ [22]. Current ground-based experiments are aiming for a sensitivity of 0.01, and future satellite missions such as BPoL and CMBPoL are hoping for an accuracy of better than one part in a thousand. The most exciting idea to me personally is the proposal to use the 21 cm line out to huge redshifts such as $z \sim 50$ [149]. Foregrounds will be an enormous problem, and it will require decades of labor, but there is potentially enough data present to determine r to one part in 10^8!

My comment about theory is that we are just beginning to understand how to compute quantum corrections during inflation. Nothing like the asymptotic scattering theory of flat space quantum field theory yet exists for this environment. Results such as (280) and (281) receive quantum corrections which Weinberg has studied [150]. He found a peculiar thing: although these corrections are suppressed by a very small factor of $GH^2 \lesssim 10^{-12}$, in addition to the factor of GH^2 already present in the lowest order result (280), they contain time dependent enhancement factors that grow like $\ln[a(t)/a(t_k)]$, where t_k is the time at which the mode experienced first horizon crossing. Nick Tsamis and I have found similar factors in quantum gravitational corrections to the metric and to the graviton self-energy [151], and Shun-Pei Miao and I have found them in quantum gravitational corrections to fermion mode functions [152]. For the modes whose spatial variation we can resolve today there have been at most about 120 e-foldings, so these *infrared logarithms* represent a terrific enhancement, but not enough to make corrections observable. (But they *might* become observable if the 21 cm observations fulfill their promise!) However, this is just because one insists on being able to resolve the spatial

variation. If one studies something which is spatially constant, like quantum corrections to the vacuum energy, then there can be infrared logarithms from modes which experienced first horizon crossing early during a very long period of inflation, and there seems no limit on the size of the effect they might give. Which is interesting, because it just so happens that there is this little problem understanding the vacuum energy.

6. Conclusions

This paper began with a list of seven questions. I will devote a paragraph to each question and the answers that have been developed. Then I will make a few additional points and comment on the future.

What is the distinction between classical physics and quantum physics that makes general relativity give such a wonderful classical theory of gravity and such a problematic quantum one? There is no difference between the equations of motion for Heisenberg field operators and those of the corresponding classical theory. Nor is there any distinction in what it means to solve those equations; in both cases the general solution consists of expressing the dynamical variable at any time in terms of its initial values. These initial values are the fundamental degrees of freedom of physics, and it is usual to label them by their Fourier wave number $\vec{k}$. The key distinction between classical and quantum is what the initial values are: in classical physics they are numbers and each of them can be set to any value; in quantum physics each mode $\vec{k}$ contains one or more pairs of noncommuting operators which obey the uncertainty principle. One consequence of this is that no mode can have less than a minimum 0-point energy. When we say that general relativity gives a superb classical theory of gravity, we mean that its results are indistinguishable from nature when we set all the large $\|\vec{k}\|$ modes to zero. There is no problem doing that classically, but it is not permitted in quantum mechanics. It is the influence of these high $\|\vec{k}\|$ modes which makes general relativity so problematic as a quantum theory.

Why do we have to quantize gravity? Because part of any force field is entirely determined by its sources, and the matter fields which source gravity are indisputably quantum mechanical, whether or not gravitons are quantized or even exist. It used to be argued that we could take the metric field to be sourced by the expectation value of the matter stress–energy tensor but that view is not tenable under the simplest interpretation of the observed anisotropies in the cosmic ray microwave background. If the predicted tensor component of these anisotropies can be imaged there will be direct evidence for the existence and quantum nature of gravitons as well.

Why do quantum field theories have divergences? Because continuum field theories have an infinite number of modes, and the 0-point motion of each one of them contributes a little to typical quantum effects. Spacetime may well be discrete on some level but the expansion of the universe by the staggering factor of $e^{120} \sim 10^{52}$ (if we accept the reality of primordial inflation) means that this discreteness cannot be responsible for keeping quantum gravitational effects so small.

Why are the divergences of quantum general relativity worse than those of the other forces? Because the other forces couple to charges which are the same for all modes, whereas gravity couples to stress energy, which grows with the wave number. This means that first order quantum corrections to the gravitational field equations produce divergences not just on terms which have two derivatives of the metric or zero derivatives of it, but also on terms which contain four derivatives of the metric. Two terms of this type are possible, the 'R^2 counterterm' and the 'C^2 counterterm.'

How bad is the problem? Adding the R^2 and C^2 counterterms to the gravitational field equations would allow the divergences of quantum gravity to be renormalized to all orders, the same way as with other forces [5]. However, increasing the number of derivatives in a field equation introduces new degrees of freedom. The new degree of freedom associated with the R^2 counterterm is a positive energy particle with spin zero, which poses no essential problem. Unfortunately, the new degrees of freedom associated with the C^2 counterterm comprise a negative energy particle with spin two, which would make the universe blow up instantly. The C^2 counterterm is necessary at first order in gravity plus scalar particles such as the Higgs [2]. It is also needed for gravity plus electromagnetism [3] and for gravity plus the weak or strong interactions [4]. It is not necessary at first order for pure gravity in four spacetime dimensions [2], but this theory requires an equally unacceptable counterterm at second order [6, 7]. Hence we must either add unacceptable counterterms, which gives a finite theory that is virulently unstable, or else low order perturbation theory makes divergent predictions.

What are the main schools of thought about quantizing gravity and why do they disagree? The problem with perturbative quantum general relativity arises from a conflict between four things: continuum field theory, quantum mechanics, general relativity and perturbation theory. Because the first two items on this list are not likely to be at fault the schools of thought on quantum gravity differ on which of the last two they suspect. Particle theorists come from a long tradition of quickly exploiting perturbation theory to get results and then rejecting models which fail to measure up. This brought great success with the standard model, so it seems reasonable to particle theorists that they should trust perturbation theory and reject general relativity. All particle theory fixes involve adding new degrees of freedom. Relativists come from a long tradition in which general relativity was many times alleged to have problems that always disappeared when a sufficiently careful analysis was made. So it makes sense to relativists that they should reject perturbation theory and instead focus on a painstakingly rigorous formulation of quantum general relativity. All relativist fixes involve the correct asymptotic expansion including terms which are not analytic in Newton's constant G.

What would we do with the theory of quantum gravity if we had it? The strength of quantum gravity corrections from a mode of energy E seems to be roughly $GE^2/\hbar c^5 \sim (E/10^{19}\,\mathrm{GeV})^2$. Although we cannot access interesting

energies in the laboratory, nature reaches these scales in four cases: the initial singularity, the final stages of black hole collapse, the final stages of black hole evaporation, and during primordial inflation. Ideas about the first three are still speculative but the simplest interpretation of current data is that the primordial perturbations in the gravitational potential of the Universe derive from quantum matter fluctuations during primordial inflation.

Establishing phenomenological contact with a weak interaction typically involves exploiting its special properties. The unique properties of gravity are:

1. One of the gravitational parameters is the cosmological constant Λ;
2. It determines the maximum speed at which signals can propagate;
3. Gravitons have zero mass without being driven to zero amplitude by the expansion of the universe and
4. The gravitational interaction energy is negative.

Physicists suspect that a successful theory of quantum gravity will explain the enormous disparity between the measured value of the cosmological constant (168) and the natural scales of fundamental theory. The use of first order perturbation theory in the context of point 2 leads to finite predictions for the blurring of images, fluctuations in luminosity and a broadening of spectral lines. Some of these effects may be observable in the not-too-distant future. Point 3 suggests that inflationary cosmology is a natural venue to search for quantum gravitational effects. And point 4 is the basis for the old dream of being able to compute the masses of fundamental particle from the interactions of their own force fields.

An interesting and possibly significant fact strikes one about all four of the current approaches to quantizing gravity reviewed in section 4. Each of them involves negative energy in some form:

- The particle theorists' dreams concerning superstrings and on-shell finiteness exploit the negative 0-point energy of fermionic superpartners;
- The higher counterterms allowed in asymptotic safety would induce negative energy degrees of freedom if they were treated nonperturbatively; and
- The relativists' dream that quantum general relativity will regulate its own ultraviolet divergences relies upon negative gravitational interaction energies.

Perhaps this is more than a curiosity.

My answer to the question posed by the title of this paper is that we are very far from having a complete quantum theory of gravity. A measure of the reliability of current thought (including my own) is this quotation from a renowned string theorist in reaction to the initial observations which indicated the universe is accelerating [8]:

> I'm sure the data is wrong because string theory predicts a negative cosmological constant.

I recount these words not because string theorists are bad physicists but rather because they are among the best our species has ever produced. That even they failed points up the folly of trying to guess natural law with nothing more to go on than mathematics and aesthetics. If that was not obvious two and a half decades ago it is surely beyond dispute now.

Neither the nature of the problem nor the prescription for its inevitable solution are unique in man's long struggle to understand the universe. The historian Colin McEvedy had this to say about the intellectual stagnation of late Roman civilization [153]:

> Speculation ran way beyond the testable and dwindled into metaphysics; technology remained tradition-bound and sluggish. Only the evolution of a scientific stance—one foot inside the boundary of the known, the other just outside—could have guaranteed the superiority, and consequently the integrity, of Mediterranean society, and the world was still too young for that.

The good news for quantum gravity is that we shall not have to endure centuries of darkness until a more powerful mode of thought emerges from the ashes of our failures. The process of achieving a scientific stance is underway; the first data have been taken, and many more are coming. Understanding what they have to teach us will likely be a long and painful process, and I expect that most of what we currently believe will need to be abandoned. But the outcome cannot be in doubt and those who finally win it for us will write in the book of human history.

Acknowledgments

Quantum gravity is a strange land whose features have only been glimpsed by the hardiest of theoretical physicists. One does not attempt the journey without guides, and I was fortunate to have the best: Stanley Deser and Bryce DeWitt. One also is not advised to go alone, and I have enjoyed more than three decades of friendship and collaboration with Nikolaos Tsamis. Among many other things, I am grateful to him for hosting me at the lovely University of Crete during the composition of this paper. And I must thank the following colleagues for advice on the manuscript: Jan Ambjorn, Abhay Ashtekar, Dimitri Bourilkov, Cecile DeWitt, Pedro Ferreira, Gary Horowitz, Elias Kiritsis, Costas Kounnas, Renate Loll, Shun-Pei Miao, Sohyun Park, Tomislav Prokopec, Martin Reuter, Frank Sauressig, Jan Smit, Lee Smolin, Theodore Tomaras, Pierre Vanhove and Zvi Bern. This work was partially supported by NSF grants PHY-0653085 and PHY-0855021, by FQXi grant RFP2-08-31 and by the Institute for Fundamental Theory at the University of Florida.

References

[1] DeWitt B S 1967 *Phys. Rev.* **160** 1113–48
DeWitt B S 1967 *Phys. Rev.* **160** 1195–239
DeWitt B S 1967 *Phys. Rev.* **160** 1239–56
[2] 't Hooft G and Veltman M 1974 *Ann. Inst. Henri Poincaré* **XX** 69–94
[3] Deser S and van Nieuwenhuizen P 1974 *Phys. Rev. Lett.* **32** 245–7
Deser S and van Nieuwenhuizen P 1974 *Phys. Rev.* D **10** 401–10

Deser S and van Nieuwenhuizen P 1974 *Phys. Rev.* D **10** 411–20
[4] Deser S, Tsao H-S and van Nieuwenhuizen P 1974 *Phys. Lett.* B **50** 491–3
Deser S, Tsao H-S and van Nieuwenhuizen P 1974 *Phys. Rev.* D **10** 3337–42
[5] Stelle K S 1977 *Phys. Rev.* D **16** 953–69
[6] Goroff M H and Sagnotti A 1985 *Phys. Lett.* B **160** 81–6
Goroff M H and Sagnotti A 1986 *Nucl. Phys.* B **266** 709–36
[7] van de Ven A E M 1992 *Nucl. Phys.* B **378** 309–66
[8] Riess A G *et al* 1998 *Astron. J.* **116** 1009–38 (arXiv:astro-ph/9805201)
Perlmutter S *et al* 1999 *Astrophys. J.* **517** 565–86 (arXiv:astro-ph/9812133)
[9] Wang Y and Mukherjee P 2006 *Astrophys. J.* **650** 1–6 (arXiv:astro-ph/0604051)
Alam U, Sahni V and Starobinsky A A 2007 *J. Cosmol. Astropart. Phys.* JCAP02(2007)011 (arXiv:astro-ph/0612381)
[10] Lamb W E and Retherford R C 1947 *Phys. Rev.* **72** 241–3
[11] Will C M 1981 *Theory and Experiment in Gravitational Physics* (Cambridge: Cambridge University Press)
[12] Abramowitz M and Stegun I 1964 *Handbook of Mathematical Functions* (New York: Dover)
[13] Arnowitt R, Deser S and Misner C W 1960 *Phys. Rev. Lett.* **4** 375–7
Arnowitt R, Deser S and Misner C W 1960 *Phys. Rev.* **120** 313–20
Arnowitt R, Deser S and Misner C W 1960 *Phys. Rev.* **120** 321–4
Arnowitt R, Deser S and Misner C W 1965 *Ann. Phys.* **33** 88–107
[14] Politzer D H 1973 *Phys. Rev. Lett.* **30** 1346–9
[15] Wilczek F and Gross D H 1973 *Phys. Rev. Lett.* **30** 1343–6
[16] Schwinger J 1961 *J. Math. Phys.* **2** 407–32
[17] Mahanthappa K T 1962 *Phys. Rev.* **126** 329–40
Bakshi P M and Mahanthappa K T 1963 *J. Math. Phys.* **4** 1–11
Bakshi P M and Mahanthappa K T 1963 *J. Math. Phys.* **4** 12–6
[18] Keldysh L V 1964 *Zh. Eksp. Teor. Fiz.* **47** 1515–27
[19] Chou K C, Su Z B, Hao B L and Yu L 1986 *Phys. Rep.* **118** 1–131
Jordan R D 1986 *Phys. Rev.* D **33** 444–54
Calzetta E and Hu B L 1987 *Phys. Rev.* D **35** 495–509
[20] Ostrogradsky M 1850 *Mem. Ac. St. Petersbourg* VI **4** 385
[21] Guth A H 1981 *Phys. Rev.* D **23** 347–56
[22] WMAP Collaboration (Komatsu E *et al*) 2009 *Astrophys. J. Suppl.* **180** 330–76
[23] Percival W J *et al* 2007 *Mon. Not. R. Astron. Soc.* **381** 1053–66 (arXiv:0705.3323)
[24] Bourilkov D 2000 *Phys. Rev.* D **62** 076005 (arXiv:hep-ph/0002172)
Bourilkov D 2001 *Phys. Rev.* D **64** 071701 (arXiv:hep-ph/0104165)
[25] Hubble E 1929 *Proc. Natl Acad. Sci. USA* **15** 168–73
[26] Vainshtein A I 1972 *Phys. Lett.* B **39** 393–4
Deffayet C, Dvali G R, Gabadadze G and Vainshtein A I 2002 *Phys. Rev.* D **65** 044026 (arXiv:hep-th/0106001)
Babichev E, Deffayet C and Ziour Z 2009 *Recovering General Relativity From Massive Gravity* (arXiv:0907.4103)
[27] Coleman S R 1979 *Subnucl. Ser.* **15** 805
[28] Bjerrum-Bohr N E J, Donoghue J F and Holstein B R 2003 *Phys. Rev.* D **67** 084033
Bjerrum-Bohr N E J, Donoghue J F and Holstein B R 2005 *Phys. Rev.* D **71** 069903 (arXiv:hep-th/0211072) (erratum)
[29] Hawking S W and Ellis G F R 1973 *The Large Scale Structure of Space-Time* (Cambridge: Cambridge University Press)
[30] Borde A, Guth A H and Vilenkin A 2003 *Phys. Rev. Lett.* **90** 151301 (arXiv:gr-qc/0110012)
[31] Weinberg S 1995 *Quantum Theory of Fields Vol 1: Foundations* (Cambridge: Cambridge University Press)
[32] Carroll S M 2001 *Living Rev. Rel.* **4** 1 (arXiv:astro-ph/0004075)
[33] Albrecht A J *et al* 2006 *Report of the Dark Energy Task Force* (arXiv:astro-ph/0609591)
[34] Jackson J D 1999 *Classical Electrodynamics* 3rd edn (New York: Wiley)
[35] Thompson R T and Ford L H 2006 *Phys. Rev.* D **74** 024012 (arXiv:gr-qc/0601137)
Borgman J and Ford L H 2004 *Phys. Rev.* D **70** 064032 (arXiv:gr-qc/0307043)
[36] Pauli W 1956 *Helv. Phys. Acta Suppl.* **4** 69
Deser S 1957 *Rev. Mod. Phys.* **29** 417–23
[37] Ford L H and Woodard R P 2005 *Class. Quantum Grav.* **22** 1637–47 (arXiv:gr-qc/0411003)
[38] Wu C H and Ford L H 2001 *Phys. Rev.* D **64** 045010 (arXiv:quant-ph/0012144)
[39] Schwinger J 1948 *Phys. Rev.* **73** 416
[40] Weisskopf V S 1939 *Phys. Rev.* **56** 72–85
[41] Carlip S 2001 *Rep. Prog. Phys.* **64** 885–942 (arXiv:gr-qc/0108040)
[42] Maldacena J M 1998 *Adv. Theor. Math. Phys.* **2** 231–52 (arXiv:hep-th/9711200)
[43] Veneziano G 1968 *Nuovo Cimento* A **57** 190–7
[44] Virasoro M A 1969 *Phys. Rev.* **177** 2309–11
[45] Goebel C J and Sakita B 1969 *Phys. Rev. Lett.* **22** 257–60
Chan H M 1969 *Phys. Lett.* B **28** 425–8
[46] Shapiro J A 1970 *Phys. Lett.* B **33** 361–2
[47] Ramond P 1971 *Phys. Rev.* D **3** 2415–8
[48] Neveu A and Schwarz J H 1971 *Nucl. Phys.* B **31** 86–112
[49] Gliozzi F, Scherk J and Olive D 1977 *Phys. Lett.* B **65** 282–6
Gliozzi F, Scherk J and Olive D 1977 *Nucl. Phys.* B **122** 253–90
[50] Golfand Yu A and Likhtman E P 1971 *JETP Lett.* **13** 323–6
[51] Volkov D V and Akulov V P 1972 *JETP Lett.* **16** 438–40
Volkov D V and Akulov V P 1973 *Phys. Lett.* B **46** 109–10
[52] Volkov D V and Soroka V A 1973 *JETP Lett.* **18** 312–4
Volkov D V and Soroka V A 1974 *Theor. Math. Phys.* **20** 829–34
[53] Wess J and Zumino B 1974 *Nucl. Phys.* B **70** 39–50
Wess J and Zumino B 1974 *Phys. Lett.* B **49** 52–4
[54] Freedman D Z, van Nieuwenhuizen P and Ferrara S 1976 *Phys. Rev.* D **13** 3214–8
[55] Deser S and Zumino B 1976 *Phys. Lett.* B **62** 335–7
[56] Brower R C 1972 *Phys. Rev.* D **6** 1655–62
[57] Goddard P and Thorn C B 1972 *Phys. Lett.* B **40** 235–8
[58] Scherk J and Schwarz J H 1974 *Nucl. Phys.* B **81** 118–44
[59] Scherk J and Schwarz J H 1975 *Phys. Lett.* B **57** 463–6
[60] Green M B and Schwarz J H 1984 *Phys. Lett.* B **149** 117–22
[61] Giddings S B and Wolpert S A 1987 *Commun. Math. Phys.* **109** 177–90
[62] Kaku M and Kikkawa K 1974 *Phys. Rev.* D **10** 1110–33
Kaku M and Kikkawa K 1974 *Phys. Rev.* D **10** 1823–43
[63] Witten E 1986 *Nucl. Phys.* B **268** 253–94
[64] Horowitz G T, Lykken J, Rohm R and Strominger A 1986 *Phys. Rev. Lett.* **57** 283–6
[65] Gross D J and Jevicki A 1987 *Nucl. Phys.* B **283** 1–49
Gross D J and Jevicki A 1987 *Nucl. Phys.* B **287** 225–50
[66] Bennett D L, Nielsen H B and Woodard R P 1998 *Phys. Rev.* D **57** 1167–70 (arXiv:hep-th/9707088)
[67] Evens D, Moffat J W, Kleppe G and Woodard R P 1991 *Phys. Rev.* D **43** 499–519
Kleppe G and Woodard R P 1992 *Nucl. Phys.* B **388** 81–112 (arXiv:hep-th/9203016)

[68] Eliezer D A and Woodard R P 1989 *Nucl. Phys.* B **325** 389–469

[69] Antoniadis I, Bachas C, Lewellen D C and Tomaras T N 1988 *Phys. Lett.* B **207** 441–6

[70] Antoniadis I 1990 *Phys. Lett.* B **246** 377–84

[71] Witten E 1996 Some comments on string dynamics *STRINGS 95: Future Perspective in String Theory* ed I Bars *et al* (Ridge Edge, NJ: World Scientific) pp 501–23 (arXiv:hep-th/9507121)

[72] Polchinski J 1995 *Phys. Rev. Lett.* **75** 4724–7 (arXiv:hep-th/9510017)

[73] Strominger A and Vafa C 1996 *Phys. Lett.* B **379** 99–104 (arXiv:hep-th/9601029)

[74] Ashtekar A, Baez J, Corichi A and Krasnov K 1998 *Phys. Rev. Lett.* **80** 904–7 (arXiv:gr-qc/9710007)

[75] Carlip S 2002 *Phys. Rev. Lett.* **88** 241301 (arXiv:gr-qc/0203001)

[76] Bousso R and Polchinski J 2000 *J. High Energy Phys.* JHEP06(2000)006 (arXiv:hep-th/0004134)

[77] Ashok S and Douglas M R 2004 *J. High Energy Phys.* JHEP01(2004)060 (arXiv:hep-th/0307049)

[78] Susskind L 2007 The anthropic landscape of string theory *Universe or Multiverse?* ed B Carr (Cambridge: Cambridge University Press) pp 247–66 (arXiv:hep-th/0302219)

[79] Green M, Schwarz J H and Witten E 1987 *Superstring Theory, Vol 1: Introduction* (Cambridge: Cambridge University Press)
Green M, Schwarz J H and Witten E 1987 *Superstring Theory, Vol 2: Loop Amplitudes, Anomalies and Phenomenology* (Cambridge: Cambridge University Press)

[80] Polchinski J 1998 *String Theory, Vol 1: An introduction to the Bosonic String* (Cambridge: Cambridge University Press)
Polchinski J 1998 *String Theory, Vol 2: Superstring Theory and Beyond* (Cambridge: Cambridge University Press)

[81] Zwiebach B 2004 *A First Course in String Theory* (Cambridge: Cambridge University Press)

[82] Kiritsis E 2007 *String Theory in a Nutshell* (Princeton, NJ: Princeton University Press)

[83] Becker K, Becker M and Schwarz J H 2007 *String Theory and M-Theory: A Modern Introduction* (Cambridge: Cambridge University Press)

[84] Smolin L 2006 *The Trouble with Physics: The Rise of String Theory, the Fall of Science, and What Comes Next* (New York: Houghton Mifflin)

[85] Woit P 2006 *Not Even Wrong—The Failure of String Theory and the Search for Unity in Physical Law* (New York: Basic Books)

[86] Sannan S 1986 *Phys. Rev.* D **34** 1749–58

[87] Bern Z, Dixon L J and Roiban R 2007 *Phys. Lett.* B **644** 265–71 (arXiv:hep-th/0611086)

[88] Chalmers G 2007 On the finiteness of $N = 8$ quantum supergravity (arXiv:hep-th/0008162)
Berkovits N 2007 *Phys. Rev. Lett.* **98** 211601 (arXiv:hep-th/0609006)
Green M B, Russo J G and Vanhove P 2007 *J. High Energy Phys.* JHEP0702(2007)099 (arXiv:hep-th/0610299)
Green M B, Russo J G and Vanhove P 2007 *Phys. Rev. Lett.* **98** 131602 (arXiv:hep-th/0611273)

[89] Deser S, Kay J H and Stelle K S 1977 *Phys. Rev. Lett.* **38** 527–30
Ferrara S and Zumino B 1978 *Nucl. Phys.* B **134** 301–26
Marcus N and Sagnotti A 1985 *Nucl. Phys.* B **256** 77–108
Howe P S and Stelle K S 1989 *Int. J. Mod. Phys.* A **4** 1871–912

[90] Bern Z, Carrasco J J, Dixon L J, Joansson H, Kosower D A and Roiban R 2008 *Phys. Rev.* D **7** 105019 (arXiv:0808.4112)
Bern Z, Carrasco J J, Dixon L J, Joansson H, Kosower D A and Roiban R 2007 *Phys. Rev. Lett.* **98** 161303 (arXiv:hep-th/0702112)

[91] Bossard G, Howe P S and Stelle K S 2009 *Gen. Rel. Grav.* **41** (arXiv:0901.4661)

[92] Bern Z, Carrasco J J, Dixon L J, Johansson H and Roiban R 2009 *Phys. Rev. Lett.* **103** 081301 (arXiv:0905.2326)

[93] Bern Z and Kosower D A 1988 *Phys. Rev.* D **38** 1888–92

[94] Bern Z and Kosower D A 1992 *Nucl. Phys.* B **379** 451–561
Bern Z and Kosower D A 1991 *Nucl. Phys.* B **362** 389–448
Bern Z and Kosower D A 1991 *Phys. Rev. Lett.* **66** 1669–72

[95] Bern Z, Dixon L J and Kosower D A 1996 *Annu. Rev. Nucl. Part. Sci.* **46** 109–48 (arXiv:hep-ph/9602280)
Bern Z, Dixon L J and Kosower D A 1995 *Nucl. Phys.* B **437** 259–304 (arXiv:hep-ph/9409393)
Bern Z, Dixon L J and Kosower D A 1994 *Nucl. Phys.* B **412** 751–816 (arXiv:hep-ph/9306240)
Bern Z, Dixon L J and Kosower D A 1993 *Phys. Rev. Lett.* **70** 2677–80 (arXiv:hep-ph/9302280)
Bern Z, Dixon L J and Kosower D A 1993 *Phys. Lett.* B **302** 299–308
Bern Z, Dixon L J and Kosower D A 1993 *Phys. Lett.* B **318** 648 (arXiv:hep-ph/9212308) (erratum)

[96] Bern Z, Dixon L J and Kosower D A 2000 *J. High Energy Phys.* JHEP01(2000)027 (arXiv:hep-ph/0001001)
Bern Z, Dixon L J and Kosower D A 1998 *Nucl. Phys.* B **513** 3–86 (arXiv:hep-ph/9708239)

[97] Bern Z, Dixon L J and Kosower D A 2006 *Phys. Rev.* D **73** 065013 (arXiv:hep-ph/0507005)

[98] Bern Z, Bjerrum-Bohr N E J, Dunbar D C and Ita H 2005 *J. High Energy Phys.* JHEP11(2005)027 (arXiv:hep-ph/0507019)

[99] Berger C F, Bern Z, Dixon L J, Forde D and Kosower D A 2007 *Phys. Rev.* D **75** 016006 (arXiv:hep-ph/0607014)
Berger C F, Bern Z, Dixon L J, Forde D and Kosower D A 2006 *Phys. Rev.* D **74** 036009 (arXiv:hep-ph/0604195)

[100] Bern Z, Dixon L J, Dunbar D C and Kosower D A 1997 *Phys. Lett.* B **394** 107–15 (arXiv:hep-ph/9611127)
Bern Z, Dixon L J, Dunbar D C and Kosower D A 1995 *Nucl. Phys.* B **435** 59–101 (arXiv:hep-ph/9409265)
Bern Z, Dixon L J, Dunbar D C and Kosower D A 1994 *Nucl. Phys.* B **425** 217–60 (arXiv:hep-ph/9403226)

[101] Kawai H, Lewellen D C and Tye S H H 1986 *Nucl. Phys.* B **269** 1–23

[102] Bern Z, Dixon L J, Dunbar D C, Perelstein M and Rozowsky J S 1998 *Nucl. Phys.* B **530** 401–56 (arXiv:hep-th/9802162)

[103] Bern Z, Bjerrum-Bohr N E J and Dunbar D C 2005 *J. High Energy Phys.* JEHP05(2005)056 (arXiv:hep-th/0501137)

[104] Bern Z, Carrasco J J M and Johansson H 2009 Progress on ultraviolet finiteness of supergravity (arXiv:0902.3765)

[105] Weinberg S 1979 Ultraviolet divergences in quantum theories of gravitation *General Relativity: An Einstein Centenary Survey* ed S W Hawking and W Israel (Cambridge: Cambridge University Press) pp 790–831

[106] Jaén X, Llosa J and Molina A 1986 *Phys. Rev.* D **34** 2302–11

[107] Donoghue J F 1994 *Phys. Rev. Lett.* **72** 2996–9 (arXiv:gr-qc/9310024)
Donoghue J F 1994 *Phys. Rev.* D **50** 3874–88 (arXiv:gr-qc/9405057)

[108] Wetterich C 1993 *Phys. Lett.* B **301** 90

[109] Benedetti D, Machado P F and Saueressig F 2009 Taming perturbative divergences in asymptotically safe gravity (arXiv:0902.4630)

[110] Lauscher O and Reuter M 2005 *J. High Energy Phys.* JHEP10(2005)050 (arXiv:hep-th/0508202)
Lauscher O and Reuter M 2002 *Phys. Rev.* D **66** 025026 (arXiv:hep-th/0205062)

Lauscher O and Reuter M 2002 *Int. J. Mod. Phys.* A **17** 993–1002 (arXiv:hep-th/0112089)
Lauscher O and Reuter M 2002 *Class. Quantum Grav.* **19** 483–92 (arXiv:hep-th/0110021)
Lauscher O and Reuter M 2002 *Phys. Rev.* D **65** 25013 (arXiv:hep-th/0108040)
[111] Percacci R 2007 Asymptotic safety (arXiv:0709.3851)
Percacci R 2006 *Phys. Rev.* D **73** 041501 (arXiv:hep-th/0511177)
[112] Percacci R and Perini D 2003 *Phys. Rev.* D **68** 044018 (arXiv:hep-th/0304222)
Percacci R and Perini D 2003 *Phys. Rev.* D **67** 081503 (arXiv:hep-th/0207033)
[113] Lauscher O and Reuter M 2007 *Lect. Notes Phys.* **721** 265–85
[114] Ashtekar A 1986 *Phys. Rev. Lett.* **57** 2244–7
Ashtekar A 1987 *Phys. Rev.* D **36** 1587–602
[115] Lewandowski J, Okolow A, Sahlmann H and Thiemann T 2006 *Commun. Math. Phys.* **267** 703–33 (arXiv:gr-qc/0504147)
[116] Nicolai H, Peeters K and Zamaklar M 2005 *Class. Quantum Grav.* **22** R193–R247 (arXiv:hep-th/0501114)
[117] Ashtekar A 2009 *Gen. Rel. Grav.* **41** 707–41 (arXiv:0812.0177)
[118] Dittrich B 2006 *Class. Quantum Grav.* **23** 6156–84 (arXiv:gr-qc/0507106)
Dittrich B 2007 *Gen. Rel. Grav.* **39** 1891–927 (arXiv:gr-qc/0411013)
[119] Rovelli C 2006 *Phys. Rev. Lett.* **97** 151301 (arXiv:gr-qc/0508124)
Bianchi E, Modesto L, Rovelli C and Speziale S 2006 *Class. Quantum Grav.* **23** 6989–7028 (arXiv:gr-qc/0604044)
[120] Ashtekar A and Lewandowski J 2004 *Class. Quantum Grav.* **21** R53–R152 (arXiv:gr-qc/0404018)
[121] Smolin L 2006 The main postulates and results of loop quantum gravity *Deserfest: A Celebration of the Life and Works of Stanley Deser* ed J T Liu *et al* (Hackensack, NJ: World Scientific) pp 266–302
[122] Rovelli C 2004 *Quantum Gravity* (Cambridge: University Press)
[123] Thiemann T 2007 *Introduction to Modern Canonical Quantum General Relativity* (Cambrdige: Cambrdige University Press)
[124] Regge T 1961 *Nuovo Cimento.* **19** 558–71
[125] Agishtein M E and Migdal A A 1992 *Nucl. Phys.* B **385** 395-412 (arXiv:hep-lat/9204004)
[126] Hamber H W 1993 *Nucl. Phys.* B **400** 347–89
[127] Bialas P, Burda Z, Krzywicki A and Petersson B 1996 *Nucl. Phys.* B **472** 293–308 (arXiv:hep-lat/9601024)
[128] de Baker B V 1996 *Phys. Lett.* B **389** 238–42 (arXiv:hep-lat/9603024)
[129] Egawa H S, Horata S and Yukawa T 2002 *Nucl. Phys. Proc. Suppl.* **106** 971–3 (arXiv:hep-lat/0110042)
[130] Ambjorn J and Loll R 1998 *Nucl. Phys.* B **536** 407–34 (arXiv:hep-th/9805108)
[131] Ambjorn J, Jurkiewicz J and Loll R 2000 *Phys. Rev. Lett.* **85** 924–7 (arXiv:hep-th/0002050)
[132] Ambjorn J, Jurkiewicz J and Loll R 2004 *Phys. Rev. Lett.* **93** 131301 (arXiv:hep-th/0404156)
[133] Durr S *et al* 2008 *Science* **322** 1224–7 (arXiv:0906.3599)
[134] Ambjorn J, Jurkiewicz J and Loll R 2009 Quantum gravity as sum over spacetimes (arXiv:0906.3947)
[135] Kolb E W and Turner M S 1990 *The Early Universe* (Redwood City, CA: Addison-Wesley)
[136] Weinberg S 2008 *Cosmology* (Oxford: Oxford University Press)
[137] Linde A D 1990 *Particle Physics and Inflationary Cosmology* (Chur, Switzerland: Harwood)
[138] Brout R, Englert F and Gunzig E 1978 *Ann. Phys.* **115** 78–106
Brout R, Englert F and Spindel P 1979 *Phys. Rev. Lett.* **43** 417–20
Starobinsky A A 1980 *Phys. Lett.* B **91** 99–102
Kazanas D 1980 *Astrophys. J.* **241** L59–L63
Stao K 1981 *Phys. Lett.* B **99** 66–70
[139] Linde A 1982 *Phys. Lett.* B **108** 389–93
[140] Albrecht A and Steinhardt P J 1982 *Phys. Rev. Lett.* **48** 1220–3
[141] Mukhanov V 2005 *Physical Foundations of Cosmology* (Cambridge: Cambridge University Press)
[142] Parker L 1969 *Phys. Rev.* **183** 1057–68
[143] Grishchuk L P 1975 *Sov. Phys.—JETP* **40** 409–15
[144] Birrell N D and Davies P C W 1982 *Quantum Fields in Curved Space* (Cambridge: Cambridge University Press)
[145] Starobinsky A A 1979 *JETP Lett.* **30** 682–5
[146] Mukhanov V F and Chibisov G V 1981 *JETP Lett.* **33** 532–5
[147] Wang L M, Mukhanov V F and Steinhardt P J 1997 *Phys. Lett.* B **414** 18–27 (arXiv:astro-ph/9709032)
[148] Polarski D and Starobinsky A A 1996 *Class. Quantum Grav.* **13** 377–92 (arXiv:gr-qc/9504030)
Lesgourgues J, Polarski D and Starobinsky A A 1997 *Nucl. Phys.* B **497** 479–510 (arXiv:gr-qc/9611019)
Kiefer C, Polarski D and Starobinsky A A 1998 *Int. J. Mod. Phys.* D **7** 455–62 (arXiv:gr-qc/9802003)
Kiefer C, Lesgourgues J, Polarski D and Starobinsky A A 1998 *Class. Quantum Grav.* **15** L67–L72 (arXiv:gr-qc/9806066)
Kiefer C, Lohmar I, Polarski D and Starobinsky A A 2007 *Class. Quantum Grav.* **24** 1699–718 (arXiv:astro-ph/0610700)
[149] Furlanetto S R, Oh S P and Briggs F H 2006 *Phys. Rep.* **433** 181–301 (arXiv:astro-ph/0608032)
[150] Weinberg S 2006 *Phys. Rev.* D **74** 023508 (arXiv:hep-th/0605244)
Weinberg S 2005 *Phys. Rev.* D **72** 043514 (arXiv:hep-th/0506236)
[151] Tsamis N C and Woodard R P 1996 *Nucl. Phys.* B **474** 235–48 (arXiv:hep-ph/9602315)
Tsamis N C and Woodard R P 1997 *Ann. Phys.* **253** 1–54
Tsamis N C and Woodard R P 1996 *Phys. Rev.* D **54** 2621–39
[152] Miao S P and Woodard R P 2006 *Phys. Rev.* D **74** 024021 (arXiv:gr-qc/0603135)
Miao S P and Woodard R P 2006 *Class. Quantum Grav.* **23** 1721–62 (arXiv:gr-qc/0511140)
[153] McEvedy C 1967 *The Penguin Atlas of Ancient History* (Harmondsworth: Penguin Books)