Why Quantum Gravity?

Jared Kaplan

Department of Physics and Astronomy, Johns Hopkins University

Abstract

These notes explain why our approach to quantum gravity must be qualitatively different from our treatment of non-gravitational QFTs. We begin by discussing quantum gravity in effective field theory, emphasizing that for all current and planned experiments this is likely a sufficient description of gravity. Then we explain why gravity appears to be very different from the other fundamental forces, requiring a radical new perspective to unite it with quantum mechanics. The vast majority of the material in these notes is not original at all, and was compiled from sources such as Wald's textbook, Ted Jacobson's nice black hole thermodynamics notes, and various old papers; most of it has been known to experts for almost 50 years.

Contents

1 Gravity as an EFT

Historically, some people thought that gravity was different from other QFTs because it's non-renormalizable, and thus requires a UV completion at the Planck scale. This isn't why we are studying it. GR + the standard model are a perfectly good quantum effective field theory, which should be able to make predictions that are accurate enough for all existing and planned experiments. For a detailed exposition, see various notes by John Donoghue on the EFT Description of Gravity (equations in what follows are from his notes). If non-renormalizability were the only problematic feature of quantum gravity, I wouldn't have spent so much time working on it.

As an EFT, we have

$$S = \int d^4x\sqrt{g}\left\{\Lambda + \frac{2}{\kappa^2}R + c_1R^2 + c_2R_{\mu\nu}R^{\mu\nu} + \ldots + \mathcal{L}_{matter}\right\} \tag{1}$$

and all but the CC and E-H term are irrelevant. To see this very explicitly, we can include cR^2 to get a rough EoM

$$\Box h + \kappa^2c^2\Box h = 8\pi GT \tag{2}$$

We can then approximate the resulting short-distance Greens function (ie the potential) via

$$\begin{aligned} G(x) &= \int \frac{d^4q}{(2\pi)^4}\frac{e^{iq\cdot x}}{q^2+\kappa^2cq^4} \\ &= \int \frac{d^4q}{(2\pi)^4}\left[\frac{1}{q^2} - \frac{1}{q^2+1/\kappa^2c}\right]e^{-iq\cdot x} \end{aligned} \tag{3}$$

This then leads to a gravitational potential

$$V(r) = -Gm_1m_2\left[\frac{1}{r} - \frac{e^{-r/\sqrt{\kappa^2c}}}{r}\right] \tag{4}$$

Since $\kappa \sim 10^{-35}$ meters is an incredibly tiny distance, the second term is completely negligable.

Loop effects are equally negligible for foreseeable experiments. Donoghue computes the quantum correction to the potential (just from the E-H term) as

$$V(r) = -\frac{Gm_1m_2}{r}\left[1 - \frac{G\left(m_1+m_2\right)}{rc^2} - \frac{127}{30\pi^2}\frac{G\hbar}{r^2c^3}\right] \tag{5}$$

It's no more difficult in principle to make such predictions in quantum gravity than in eg non-renormalizable theories of pions.

Instead, quantum gravity is interesting because both its gauge redundancy and black hole thermodynamics suggest that at a fundamental level, theories of quantum gravity require a much more radical departure from the familiar world of QFT.

2 Gauss's Law as a First Suggestion of Holography

2.1 Electromagnetism

One of the first things we learn in physics is Gauss's law, which computes the charge enclosed by a surface in classical electromagnetism in terms of the field on that surface:

$$Q = \oint d^2\hat{n}\,(\vec{E}\cdot\hat{n}) \tag{6}$$

This is a consequence of the classical EoM, which say that

$$\nabla^\nu F_{\nu\mu} = J_\mu \quad \Longrightarrow \quad \vec{\nabla}\cdot\vec{E}(x) = \rho(x) \tag{7}$$

Integrating the latter over space, the divergence can be re-written as a surface integral. These beautiful statements have been derived from the classical EoM, so it's not obvious what happens to them in the quantum theory. Let's now argue that they become operator identities!

The charge is defined in terms of the current, which satisfies an EoM

$$J^\mu = \nabla_\nu F^{\nu\mu} \tag{8}$$

This is what gives us Gauss's law. So to elevate it to an operator statement, we just need to recall to what extent the classical EoM hold as exact operator statements. There are many points of view on this question, depending on whether we take canonical quantization or the path integral as our starting point.

In the canonical formalism, the classical equations of motion follow from Poisson brackets with the Hamiltonian. After we quantize, these become the commutation relations that state that H generates time translations. So the EoM are apparently operator statements. In the path integral formalism, we can derive the Schwinger-Dyson equations, which are identical to the classical equations of motion up to contact terms. From the operator perspective, these contact terms correspond to non-trivial commutation relations, and arise directly from the canonical commutation relations for the canonical fields. So the conclusion is that we can view the EoM as operator statements in correlators as long as we account for contact terms in these correlators.

We want to use the EoM to write an operator relation

$$Q = \int d^3x\, J^0(x) = \int d^3x\, \nabla_i F^{0i}(x) = \oint d^2\hat{n}_i\, F^{0i}(x) \tag{9}$$

and its only the use of the EoM in the intermediate step that could be at issue. However, *as long as no operators are inserted on the surface where we are integrating, these equations are identities.* In particular, if we study states created by operators in the past and future, these relations are exact. (If operators are inserted on the surface, then the contact terms account for the extra charge they create). So the quantum charge operator Q can be computed by integrating F^{0i} on a sphere enclosing any given region.

If we care a lot about Q, then this seems to be a very profound statement. In the case of gravity, the analog of Q is H, which is the operator we care about most.

2.2 Linearized Gravity

In General Relativity, the EoM can be written as

$$T_{\mu\nu} = \left(R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} \right) \tag{10}$$

Let's first think about this at the linearized level. In approximately flat spacetime we have

$$g_{\mu\nu} = \eta_{\mu\nu} + \gamma_{\mu\nu} \tag{11}$$

and the linearized Einstein tensor is

$$G_{ab} = \partial^c \partial_{(b} \gamma_{a)c} - \frac{1}{2} \partial^c \partial_c \gamma_{ab} - \frac{1}{2} \partial_a \partial_b \gamma - \frac{1}{2} \eta_{ab} \left(\partial^c \partial^d \gamma_{cd} - \partial^c \partial_c \gamma \right) \tag{12}$$

where $\gamma = \gamma^a_a$. If we use

$$\bar{\gamma}_{ab} = \gamma_{ab} - \frac{1}{2} \eta_{ab} \gamma \tag{13}$$

then the linearized Einstein equations become

$$-\frac{1}{2} \partial^c \partial_c \bar{\gamma}_{ab} + \partial^c \partial_{(b} \bar{\gamma}_{a)c} - \frac{1}{2} \eta_{ab} \partial^c \partial^d \bar{\gamma}_{cd} = 8\pi T_{ab} \tag{14}$$

We can choose an analog of Lorenz gauge

$$\partial^b \bar{\gamma}_{ab} = 0 \tag{15}$$

to make Einstein's equations look a lot like Maxwell's equations:

$$\partial^c \partial_c \bar{\gamma}_{ab} = -16\pi T_{ab} \tag{16}$$

With this simplification, we can now compute the integral of the energy in a region as

$$\begin{aligned} E &= \int d^3x T^{00} = -\frac{1}{16\pi G_N} \int d^3x \partial^c \partial_c \bar{\gamma}^{00} \\ &= -\frac{1}{16\pi G_N} \int d^2\hat{n}^i \partial_i \bar{\gamma}^{00} \\ &= \frac{1}{16\pi G_N} \int d^2\hat{n}^i \partial_i \sum_{j=1}^{3} \bar{\gamma}^{jj} \end{aligned} \tag{17}$$

where we note that $\partial_0 \gamma^{00} = 0$ in our gauge, so we can neglect the time derivative terms.

Thus to lowest order in linearized gravity, the energy in a region is given by a surface integral of the gradient of the metric on the boundary of the region. Getting ahead of ourselves, *this suggests that in the quantum theory the Hamiltonian actually lives at infinity.*

Our derivation has a major limitation – due to our linearized approximation, we have not accounted for the energy of the gravitational field itself. Nevertheless, if our imaginary surface lives very close to infinity, then we expect $\gamma \ll 1$ and the non-linear interactions of the gravitational field will no longer be important. Gravitational binding energies may have a significant effect on the total energy, but they will already be accounted for in the behavior of γ. Thus our final expression should in fact provide a reasonable account of the energy in space, though we are far from justifying it in any sort of rigorous way.

What about Massless Higher Spin Fields?

To preserve gauge invariance and keep higher spin fields massless, we would need the fields to couple to conserved currents of higher spin (if they are to couple at a linearized level), just as A and g couple to J and T. But conserved higher spin currents typically do not exist. This is one way of seeing why we do not experience higher spin forces generalizing gauge and gravitational forces.

3 Black Hole Thermodynamics

This section follows Ted Jacobson's notes very closely.

3.1 Black Hole Basics

Many of the basic features of black holes, such as uniqueness and 'no hair' theorems, the universality of black hole formation, their properties when energy is added or extracted, and (finally) the area theorem seem innocuous at first, but all have striking interpretations via black hole thermodynamics.

3.1.1 Notion of a BH

In Newtonian physics, we can ask when the escape velocity

$$\frac{1}{2}mv^2 = \frac{GMm}{R} \tag{18}$$

and this occurs when $v = \sqrt{2GM/R}$. Plugging in the speed of light $v = c$ gives R_s, the Schwarzchild radius. Of course in Newtonian physics this wouldn't necessarily mean the BH was inescapable, since you could accelerate with a rocket ship. But it turns out that once you travel to $R < R_s$ in GR, you really can't escape.

It's worth noting that BHs are not elementary particles, because their Compton wavelength $1/M \ll R_s = \sqrt{2GM}$ once $M > M_{pl}$. Only BHs with near Planck scale mass could be elementary particles in this sense. Notice that *reductionism ends at the Planck scale, since you cannot explore distances smaller than $1/M_{pl}$ using high energy collisions.* Higher energy collisions just produce larger and larger BHs.

Relatedly, notice that regions with extremely low density can still form horizons if they are sufficiently large. This is why, very roughly speaking, you cannot have a canonically normalized scalar field interpolate over $\delta\phi \gg M_{pl}$ without forming horizons.

Spherically symmetric asymptotically flat 4d solutions take the Schwarzschild form

$$ds^2 = \left(1 - \frac{r_s}{r}\right) dt^2 - \frac{dr^2}{\left(1 - \frac{r_s}{r}\right)} - r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{19}$$

These coordinates are singular at the horizon so it's better to use (ingoing) Eddington-Finkelstein coordinates

$$ds^2 = \left(1 - \frac{r_s}{r}\right) dv^2 - 2dvdr - r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{20}$$

where $dv = dt + \frac{dr}{1-\frac{r_s}{r}}$. In both cases r tells us about the surface area of spheres. Lines of constant v, θ, ϕ are ingoing radial lightrays, while outgoing lightrays satisfy $\frac{dr}{dv} = \frac{1}{2}\left(1 - \frac{r_s}{r}\right)$. Thus they fail to be outgoing for $r < r_s$. So with positive mass the singularity is causally disconnected from infinity.

Black hole solutions come in very limited families, the most general of which (in 4d) is the Kerr-Newman metric, including both charge and rotation. This is the most general stationary (timelike Killing vector at infinity), asymptotically flat solution.

3.1.2 Singularity Theorems

The singularity at $r = 0$ might have been supposed to be due to spherical symmetry. However, while Newtonian gravity produces $1/r$ potentials, relativistic effects in GR produce a $1/r^3$ potential, which overwhelms any $1/r^2$ angular momentum barrier, providing a physical reason for generic singularity formation.

Penrose proved the existence of singularities using the idea of the trapped surface and the focusing or Raychaudhuri equation:

$$\frac{d}{d\lambda}\rho = \frac{1}{2}\rho^2 + \sigma^2 + R_{ab}k^a k^b \tag{21}$$

Here $\rho \equiv \frac{d}{d\lambda}\log \delta A$, where δA is the change in an infinitesimal cross-sectional area, k^a is a tangent vector to the null geodesic congruence, and σ^2 is the square of the sheer tensor of the congruence. Trapped surfaces are spacelike 2-surface (in 4d) whose ingoing and outgoing null congruences are both converging – ie everything falls in from a trapped surface.

So black hole formation is in many circumstances guaranteed, and is highly universal, independent of the type of matter from which the BH is made.

3.1.3 Energy Extraction

Black holes are fairly unique composite objects (as are states in thermodynamic equilibrium, we observe with hindsight). How do they interact with other systems? Let's study how energy can be given and taken from a BH.

First of all, note that we can extract 100% of the rest mass of a particle by lowering it into a BH on a string. To see this note that $\xi^\mu = \delta^\mu_v$ is a Killing vector for the EF metric, and so $E = m\dot{x}_\mu \xi^\mu = m\dot{x}_v$ is conserved along a geodesic. For a particle at fixed r, Ω we have $\dot{x}_\mu = \xi^\mu/|\xi|$ and so $E = |\xi| m$. Since $|\xi| = \sqrt{1 - r_s/r}$ the energy vanishes at the horizon and is m at infinity. So we can extract all of the rest mass by quasi-statically lowering a particle into a BH.

Extracting energy from a BH is more interesting.

In most physical systems, energy is both conserved and *bounded from below*. The local notion of energy (ignoring gravity, or treating spacetime as a constant background) comes from the time components of the energy momentum tensor $T^{\mu\nu}$. However, the timelike Killing vector ∂_v (from E-F coords) becomes spacelike inside a black hole, meaning that it becomes more like a momentum. And momentum can have either sign (it's not bounded from below). This is related to the possibility of Hawking radiation.

However, this feature also happens outside the horizon for rotating black holes, in what's called the Ergoregion – a place where a timelike Killing vector at infinity becomes spacelike. It's typically a donut-shaped region outside the horizon.

Penrose discovered a process that allows energy extraction from the Ergoregion. We send a particle 0 into the Ergoregion, where it splits into an ingoing particle 2 with negative energy plus an outgoing particle 1 with more energy than the initial particle. Particle 2 consumes some of the angular momentum of the black hole, so it must have opposite angular momentum as that of the BH. For maximum energy extraction, we need to maximize the ratio of energy gained (by the particle) to angular momentum lost (from the BH).

We can understand this using conserved quantities. Let ξ be the time-translation (at infinity) Killing vector field and ψ be the rotation vector field, with corresponding conserved quantities $E = p \cdot \xi$ and $L = -p \cdot \psi$ (negative sign so that L is positive, since ψ is spacelike everywhere). On the horizon both ξ, ψ are spacelike, but since the horizon is null there must exist $\chi = \xi + \Omega\psi$ that is a future-directed null Killing field, which defines Ω as the angular velocity of the horizon.

As the infalling, negative-energy particle 2 crosses the horizon, we must have $p_2 \cdot \chi = E_2 - \Omega L_2 \geq 0$, and we have $L_2 < 0$ so that $E_2/L_2 \leq \Omega$. When we saturate the inequality, particle 2 is null and tangent to the horizon. For this most efficient inequality-saturating process, for the BH itself we have $\delta M = \Omega \delta L$.

It is interesting and important that for maximally efficient energy extraction, *the area of the event horizon does not change*. This follows from the Raychauduri equation noting that $R_{ab} \propto T_{ab}$ via Einstein's equation, and $T_{ab} \propto k_a k_b$ for the infalling particle, where $k_a k_b$ are null vectors tangent to the horizon (the sheer term is higher order because energy is extracted slowly – we only change the horizon infinitesimally).

The same results hold for energy extraction from charge using a charged black hole. Maximally efficiency is $\delta M = V \delta Q$ and this does not change the horizon area.

3.1.4 Area Theorem

In the examples above the most efficient energy extraction occurs when the black hole area is unchanged, and in less efficient processes the area always increases. It was shown by Hawking that in fact the area of an event horizon can never decrease under quite general assumptions. This means all processes are either irreversible or (just barely) reversible and maximally efficient.

Hawking's theorem applies to arbitrary dynamical black holes, for which a general definition of the horizon is needed. The future event horizon of an asymptotically flat black hole spacetime is defined as the boundary of the past of future null infinity, that is, the boundary of the set of points that can communicate with the remote regions of the spacetime to the future. Hawking proved that if $R_{ab} k_a k_b \geq 0$, and if there are no naked singularities (i.e. if "cosmic censorship" holds), the cross sectional area of a future event horizon cannot be decreasing anywhere. The reason is that the focusing equation implies that if the horizon generators are converging somewhere then they will reach a crossing point in a finite affine parameter. But such a point cannot lie on a future event horizon (because the horizon must be locally tangent to the light cones), nor can the generators leave the horizon. The only remaining possibility is that the generators cannot be extended far

enough to reach the crossing point—that is, they must reach a singularity. The singularity may not be naked, i.e. visible from infinity, and we have no good reason to assume clothed (or barely clothed) singularities do not occur.

With a more subtle argument, Hawking showed that convergence of the horizon generators implies the existence of a naked singularity. The basic idea is to deform the horizon cross-section outward a bit from the point where the generators are assumed to be converging, and to consider the boundary of the future of the part of the deformed cross-section that lies outside the horizon. If the deformation is sufficiently small, all of the generators of this boundary are initially converging and therefore reach crossing points and leave the boundary at finite affine parameter. But at least one of these generators must reach infinity while remaining on the boundary, since the deformed cross-section is outside the event horizon. The only way out of the contradiction is if there is a singularity outside the horizon, on the boundary, which is visible from infinity and therefore naked.[1]

We do not have any solid reason to believe that naked singularities do not occur, and yet classical black hole thermodynamics seems to rest on this assumption. Perhaps it is enough for near-equilibrium black hole thermodynamics if naked singularities are not created in quasi-stationary processes.

The area theorem implies that a maximally rotating BH can lose at most $(1 - 1/\sqrt{2})$ of its initial energy, that in a merger of two BHs with equal mass only $(1 - 1/\sqrt{2})$ of the initial energy can be radiated (though if $M_2 \ll M_1$ then almost all of M_2 can be radiated away), and that if two spinning BHs merge almost $1/2$ of the combined initial energy can be radiated away.

3.2 Classical BH Thermodynamics

Previously we saw that BHs have properties that seem analogous to thermodynamics if we equate the event horizon area with an entropy. On dimensional grounds, this requires $\hbar$ (and G_N and c) in order to relate $S \propto A$, and of course it will also turn out to require $\hbar$ to obtain a non-vanishing temperature.

3.2.1 Four Laws of BH Mechanics

We already saw that when $dA = 0$ so that the area doesn't change, the mass of a BH obeys

$$dM = \Omega dJ + \Phi dQ \tag{22}$$

where we change angular momenta and charge and Ω and Φ are angular velocity and electric potential at the horizon. This looks a lot like the first law of thermodynamics with $dQ = TdS$ missing.

[1] Essentially the same argument as the one just given also establishes that an outer trapped surface must not be visible from infinity, i.e. must lie inside an event horizon. This fact is used sometimes as an indirect way to probe numerical solutions of the Einstein equation for the presence of an event horizon. Whereas the event horizon is a nonlocal construction in time, and so can not be directly identified given only a finite time interval, a trapped surface is defined locally and may be unambiguously identified at a single time. Assuming cosmic censorship, the presence of a trapped surface implies the existence of a horizon.

That missing term is

$$\frac{\kappa}{8\pi G}dA \tag{23}$$

where κ is the surface gravity of the horizon. For a stationary BH, if we assume that the event horizon is a Killing horizon, so that the null horizon generators are symmetries, then κ is the magnitude of the gradient of the norm of the horizon generating Killing field at the horizon

$$\kappa^2 \equiv \nabla^a|\chi|\nabla_a|\chi| \tag{24}$$

where χ^a itself is the Killing vector field. Equivalently, κ is the magnitude of the acceleration wrt Killing time of a stationary zero-angular momentum particle just outside the horizon. This is the force per unit mass that must be applied at infinity in order to hold the particle on its path (but not the tension in a string attached to the particle near the particle, which diverges at the horizon).

Amusingly, in the absence of angular momentum the surface gravity is $1/(4M)$, which is the same as the Netwonian surface gravity at the Schwarzchild radius.

The surface gravity is always constant over the horizon of a stationary black hole. This is the **zeroth law**, as it dictates that in equilibrium the quantity analogous to the temperature is uniform. The constancy of κ can be proved without any field equations if the horizon is a Killing horizon and the BH is static or axisymmetric and $t-\phi$ reflection symmetric. Alternatively, it can be proved using stationarity, the Einstein equations, and the dominant energy condition (a very strong assumption). It's interesting to consider the rate of approach to equilibrium as well, as this is analogous to thermalization.

The **first law** states

$$dM = \frac{\kappa}{8\pi G}dA + \Omega dJ + \Phi dQ \tag{25}$$

for infinitesimal quasi-static changes, so that the BH in question remains stationary (in equilibrium). This equation acquires additional terms if stationary matter other than electromagnetic fields are present. Note that κ, Ω, Φ must all be constant on the horizon of a stationary black hole (ie we're in equilibrium).

We can understand the first law via heat flow. Imagine dropping some mass into the BH using the flux of energy $T_{ab}\xi^a$. Then via Einstein's equations we have

$$\begin{aligned}
\Delta M &= (\kappa/8\pi G)\int R_{ab}k^a k^b \lambda d\lambda dA \\
&= (\kappa/8\pi G)\int \frac{d\rho}{d\lambda}\lambda d\lambda dA \\
&= (\kappa/8\pi G)\int (-\rho) d\lambda dA \\
&= (\kappa/8\pi G)\Delta A
\end{aligned} \tag{26}$$

where we have used the infinitesimal focusing equation (21) and an integration by parts (boundary terms vanish by stationarity).

Of course the **Second Law** is Hawking's area theorem, stating that in fact $\Delta A \geq 0$ (assuming Cosmic Censorship and an unproven energy condition). We will revisit it in a moment to include matter entropy.

There is also a **Third Law** stating that the surface gravity cannot be reduced to zero in a finite number of steps; this has been precisely formulated and proven by Israel. Extremal black holes have zero temperature and surface gravity (but finite entropy), so the third law says that it's very hard to make an exactly extremal BH. An interesting example is for a spinning BH – if you try to drop a spin on the axis of a spinning, near-extremal BH, then you face a repulsive gravitational spin-spin force. (Apparently no one has investigated adding a charge to a spinning BH; this might be fun.)

3.2.2 Generalized 2nd Law

Bekenstein proposed a generalized 2nd law of the form

$$\delta\left(S_{outside} + \frac{\eta A}{\hbar G}\right) \geq 0 \tag{27}$$

where η is some constant. It turns out $\eta = 1/4$. This really equates area with entropy. Note that the BH entropy is infinite when $\hbar$ or $G \to 0$.

At the classical level it seems we can add entropy to the BH without increasing its area, by eg lowering a box slowly in. But it's unclear if a classical analysis is sufficient, since the BH entropy diverges in the classical limit.

When we include EFT corrections to General Relativity, the laws of BH Thermodynamics can change. There is a modified proposal for the entropy in this situation, and in special cases one can prove that it's non-decreasing, but it's unclear when and why. BH Thermodynamics beyond Einstein gravity has been a fruitful area for research.

3.3 A Summary of Quantum BH Thermodynamics

Incorporating QM into BH Thermodynamics completes the story, as we will see.

The historical route to Hawking's discovery is worth mentioning. After the Penrose process was invented, it was only a short step to consider a similar process using waves rather than particles [Zel'dovich, Misner], a phenomenon dubbed "super-radiance". Quantum mechanically, supperradiance corresponds to stimulated emission, so it was then natural to ask whether a rotating black hole would spontaneously radiate [Zel'dovich, Starobinsky, Unruh]. In trying to improve on the calculations in favor of spontaneous emission, Hawking stumbled onto the fact that even a non-rotating black hole would emit particles, and it would do so with a thermal spectrum at a temperature

$$T_H = \frac{\hbar\kappa}{2\pi} \tag{28}$$

This history also explains why Hawking radiation is often viewed as pair creation – for rotating black holes, this is a valid perspective. We can make a pair conserving Killing energy and angular

momentum, as in the ergoregion there are negative energy states for real particles. Then the negative energy particle can later fall in to the BH. In the non-rotating case the ergoregion only exists beyond the horizon, so pair creation must exactly straddle the horizon.

Hawking radiation can be derived in a variety of ways. The simplest is to use the KMS condition (ie thermal states are periodic in Euclidean time). Hawking's original paper, which is quite read-able, derives the effect from a much more complicated abstract scattering experiment. There are also other derivations that attempt to make the calculation look more like pair creation. I won't go into any of the derivations here, because I want to focus on BH Thermodynamics itself, but we may discuss them later.

3.3.1 Revisiting the 2nd Law

The generalized 2nd law might fail if A does not increase enough to compensate for entropy dropped into a BH, and perhaps it could fail due to Hawking radiation. Does it?

Massless radiation has an energy density $\frac{1}{4}T^4$ and entropy $\frac{1}{3}T^3$, so that $dS = \frac{4}{3}\frac{dE}{T}$. But since $dM = -dE$, we see that with $dS_{BH} = -dE/T$, the total entropy increases when BHs emit Hawking radiation. This is an over-simplification however. It has apparently been checked in many cases that instead of 4/3 one can obtain different factors, always > 1 (due to grey-body factors), but apparently there is not a completely general argument of this form.

It's worth realizing that the argument above implies that the entropy of the final state radiation is of order the entropy of the BH. During evaporation BHs emit $\sim S$ quanta of radiation with energy $1/R_s$ each, over a time period of order SR_s. This way of stating these quantities is valid in any number of spacetime dimensions.

Box-lowering led Bekenstein the propose the Bekenstein bound on entropy, $S \leq 2\pi ER$. This is a very interesting inequality since it doesn't involve G_N! A version of it was proven by Casini and Huerta. But the Bekenstein bound is not needed to avoid violations of the 2nd law when lowering boxes into BHs.

Unruh and Wald argued that the Hawking or Unruh radiation near the horizon creates a buoyant force on the box (since it is lowered in, it is accelerating), because the box sees a larger temperature on its lower side. Note that the entropy S_{box} must be less than the entropy of thermal radiation with the same volume and entropy, since a thermal state maximizes entropy. This means that the entropy of the box is less than or equal to the entropy of Unruh radiation that it displaces as it's lowered into the BH.

One can also attempt to mine energy from BHs, though there are surprising limits to this process due to the strength of materials. The most efficient mining procedure is to thread the BH with strings.

3.4 From BH Thermodynamics to Holography

Taking BH Thermodynamics very seriously, we're led to expect that the maximum number of states in a region bounded by an area A should be less than $A/4$ in Planck units. If the region is large, this suggests that the states of the universe are really holographic, and the fundamental degrees of freedom really live on areas rather than inside volumes.

This idea is quite radical, and naively seems like it cannot be reconciled with the apparently locality of physics. For if the fundamental DoF live on areas, then we cannot think of quantum gravity as an 'Aether Theory', even though QFT very much seems to be a theory of the Aether. That is, in non-gravitational QFT we can think of spacetime as though it's filled with little bits of quantum field that live at every point in space, and evolve with time (just in a Lorentz invariant way, ie respecting special relativity). This isn't very different from eg the little bits of information that are stored by magnets in a hard drive. And a crucial feature of this (correct) picture of QFT is that *interactions are local in space*, and can just be visualized as nearest-neighbor interactions among a finite number of DoF localized at each point in space.

But if the fundamental DoF live on areas (perhaps at infinity), and not within the volume of space, then nearest-neighbor interactions don't make sense anymore. So why are the laws of physics local at all? Why do forces get weaker when objects are far apart? Why does the notion of distance in space even make sense?

It's an interesting historical note that post-AdS/CFT, both BH Thermodynamics, the fact that 'Energy lives at infinity in GR', and the gauge redundancies of GR all suggest a holographic viewpoint. Nowadays these ideas seem like fantastic motivations for holography. Yet for a long while they were not at all convincing, perhaps because the conclusion seems so radical.

4 Gauge Redundancy

The diffeomorphisms / coordinate transformation gauge redundancy can be confusing – Einstein was confused[2] about it for 3 years!

From a field theory viewpoint, GR has a gauge redundancy because the graviton is massless. From a geometric viewpoint, it's due to the fact that manifolds and metrics related by a diffeomorphisms are geometrically identical. Let's make these ideas explicit, since they are such an important feature of GR.

4.1 Global Symmetry and Gauge Redundancy for a $U(1)$

As a warm-up, consider the situation where we have a $U(1)$ global symmetry. We assume that we have a *matter action* $S_M[A_\mu;\phi]$, where A_μ is a *background* gauge field and ϕ are the matter fields charged under the $U(1)$. Because we have introduced A_μ, the matter action will actually be gauge invariant when we simultaneously gauge-transform ϕ and also transform the background A_μ. The matter action excludes terms that depend only on A_μ, such as the gauge kinetic term, as A is just a background field. Now we can define

$$J_\mu = \frac{\delta S_M}{\delta A_\mu} \tag{29}$$

[2]Google Einstein's Hole Argument, or look at pages 47-50 of Rovelli's book: http://www.cpt.univ-mrs.fr/ rovelli/book.pdf

and prove that it is a conserved current. Let us consider the variation of the S_M under gauge transformations. By gauge invariance (and without using the EoM) this is

$$0 = \int \frac{\delta S_M}{\delta A_\mu}\delta A_\mu + \int \frac{\delta S_M}{\delta \phi}\delta\phi \tag{30}$$

where δA_μ and $\delta\phi$ are their variations under a gauge transformation (for example $\phi \to e^{i\alpha(x)}\phi$, so $\delta\phi = i\alpha(x)\phi(x)$). Now if we impose the EoM for the matter fields ϕ, the second term must vanish, since the equations of motion set to zero all variations with respect to ϕ.

But we know that under a gauge transformation, $A_\mu \to A_\mu + \nabla_\mu\alpha$, so this means that

$$\begin{aligned} 0 &= \int \frac{\delta S_M}{\delta A_\mu}\delta A_\mu \\ &= \int d^d x J_\mu(x)\nabla^\mu\alpha(x) \end{aligned} \tag{31}$$

for any gauge transformation $\alpha(x)$, which means that $\nabla^\mu J_\mu = 0$ identically. Note that J_μ had to come from the matter action alone, because if we had included the gauge field action as well, J_μ would have vanished identically on the gauge field EoM.

So far we have treated A_μ as a background field. In that case, different configurations of the dynamical fields are all inequivalent – there is no gauge redundancy. More explicitly: gauge transformations relate a unique configuration of the dynamical fields plus a specific background field choice to a different configuration of both the dynamical fields and the background field. Since gauge transformations change the background field, they are not a redundancy of the physical states.

If instead we make A_μ a dynamical gauge field (eg by path-integrating over A_μ), then the gauge transformation becomes a redundancy of the description, since it relates many distinct configurations of dynamical fields (now that A_μ is dynamical!). Additionally, the redundancy can be understood from the particle physics perspective as a way of eliminating the longitudinal mode of the photon.

4.2 Gauge Redundancy for GR

4.2.1 Diffeomorphisms and the Conservation of $T_{\mu\nu}$ without Dynamical Gravity

An important and elementary result states that if a non-gravitational field theory is diffeomorphism invariant, then the energy momentum tensor will be covariantly conserved. Let's review the derivation, and then we'll try to discuss what it means. The argument mirrors the $U(1)$ case above.

Consider any matter action S_M in a general *fixed* spacetime background g_{ab} (S_M does not include any terms that depend only on the metric). We define the energy momentum tensor as[3]

$$T_{ab} \equiv -\frac{1}{8\pi}\frac{1}{\sqrt{-g}}\frac{\delta S_M}{\delta g^{ab}} \tag{32}$$

This differs from the definition most often given in flat space QFT, where instead T_{ab} gathers together the 4 conserved currents associated with spacetime translations (which are a symmetry in

[3]Conventions for numerical coefficients were lifted from Wald; I'm not being careful with them.

Poincaré invariant theories, but not in general spacetimes). Our definition guarantees that T_{ab} will be symmetric and conserved, whereas only the latter property is obvious when defining T_{ab} using translations in flat spacetime. Our proof below that T_{ab} is conserved should make it clear how T_{ab} relates to translations in the flat space limit.

Now we will assume that the matter action $S_M[g_{ab};\phi]$ is diffeomorphism invariant when we transform both the fields and the non-dynamical background metric g_{ab}. For infinitessimal diffeomorphisms, we can write the formal relation

$$0 = \int \frac{\delta S_M}{\delta g^{ab}} \delta g^{ab} + \int \frac{\delta S_M}{\delta \phi} \delta\phi \tag{33}$$

where δg^{ab} and $\delta\phi$ are the variations of the metric and the matter fields 'ϕ' under the diffeomorphism.

When the matter fields satisfy their equations of motion, the second term above vanishes identically. The variation of the metric under a diffeomorphism is

$$\delta g^{ab} = \nabla^{(a} v^{b)} \tag{34}$$

where v^a is a vector field, and the parentheses denote symmetrization. We derived this above when we discussed Lie derivatives. This means that

$$\begin{aligned} 0 &= \int d^d x \sqrt{-g} T_{ab} \nabla^{(a} v^{b)} \\ &= \int d^d x \sqrt{-g} \nabla^a T_{ab} v^b \end{aligned} \tag{35}$$

which implies that

$$\nabla^a T_{ab} = 0 \tag{36}$$

so the stress-energy tensor is covariantly conserved as a consequence of diffeomorphism invariance. In flat space, infinitessimal diffeomorphisms include infinitessimal translations, so our derivation connects with translation invariance.

4.2.2 Diffeomorphism Non-Invariant Examples

We tend to live in a cloistered world where we rarely encounter theories that violate diffeomorphism invariance, so let's write some down to make sure we understand what's going on. The deSitter metric is

$$ds^2 = dt^2 - e^{2t} d\vec{x}^2 \tag{37}$$

If we write a free field theory in this spacetime and *make the coordinate dependence explicit, rather than expressing it in terms of* g_{ab}, we have

$$S = \int d^3x dt \, \frac{e^{3t}}{2} \left((\partial_t \phi)^2 - e^{-2t} (\partial_i \phi)^2 \right) \tag{38}$$

and it is obvious that the action depends explicitly on time. This explicit time dependence implies that the (naive) energy will not be conserved. In fact we can re-write the deSitter metric as

$$ds^2 = \frac{1}{\tau^2}(d\tau^2 - d\vec{x}^2) \tag{39}$$

in which case the action would take the form

$$S = \int d^3x d\tau \frac{1}{\tau^2} \left(\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi\right) \tag{40}$$

where $\eta^{\mu\nu}$ is the flat space metric. So we again have explicit time dependence (destroying energy conservation), even though we wrote the action as a function of a flat spacetime metric $\eta^{\mu\nu}$. This version of the theory will not be invariant if we attempt to perform a diffeomorphism on $\eta_{\mu\nu}$.

However, we can of course re-write the theory using the full spacetime metric as

$$S = \frac{1}{2}\int d^3x dt\sqrt{-g}g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi \tag{41}$$

Now the theory is clearly diffeomorphism invariant. The point is that in a diff invariant theory all dependence on the coordinates must be through the covariant spacetime fields (which are tensors), the spacetime metric (with invariant contractions of indices), and the volume form $d^dx\sqrt{g}$. It is the presence of 'bare' coordinates the ruins diff invariance.

Diffeomorphism invariance now guarantees that the stress tensor

$$T_{\mu\nu} = \nabla_\mu\phi\nabla_\nu\phi - \frac{1}{2}g_{\mu\nu}g^{ab}\nabla_a\phi\nabla_b\phi \tag{42}$$

must be (covariantly!) conserved and in particular, it implies that there will be a conserved energy given by

$$H = \int d^3x\sqrt{-g}\,T^{00} \tag{43}$$

The key point here is the word *covariantly*. We have that $\nabla^\mu T_{\mu\nu} = 0$ due to the ϕ equations of motion, but this only holds for the covariant derivative associated to the full spacetime metric.

So as final examples note that these actions

$$\begin{aligned}
&\int d^3x dt\sqrt{-g}f(t,x)g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi, \\
&\int d^3x dt\sqrt{-g}g^{\mu\nu}\phi\partial_\mu\partial_\nu\phi \\
&\int d^3x dt\sqrt{-g}\left(g_{02}\right)^3 g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi
\end{aligned} \tag{44}$$

are not diff invariant, although the first two can be minorly modified to restore diff invariance. If we include non-covariant derivatives, explicit coordinate dependence, or uncontracted indices in the action, then we will not have a conserved energy-momentum tensor, as you can easily check. The point of diffeomorphism invariance is that the action must be geometric, in the sense that all coordinate dependence occurs through the volume element, metric, and covariant derivatives.

4.2.3 Diffeomorphisms as a Redundancy in Gravity

When we studied matter in a fixed metric, the laws of physics were invariant to diffeomorphisms. But diffs were not a redundancy of the description, because they changed both the dynamical matter fields and the non-dynamical background metric. In other words, diffeomorphisms equate different dynamical field configurations that live in different environments, rather than different field configurations in the same environment.

Now we will allow the gravitational field g_{ab} to fluctuate. Once we make g_{ab} dynamical, diffeomorphisms become a redundancy of description, because they equate distinct configurations of the dynamical fields. This also means that diffeomorphisms should be viewed as a redundancy for g_{ab}, since without the dynamical metric, the redundancy disappears.

To dispel confusion, let's give an extremely explicit example. Let's say we have $g_{ab}(x)$ and a matter field $\phi(x)$ in 1+1 dimensions with coords (t, x). We can define $x = y^3$ as a change of coordinates, and identify

$$\phi^{new}(t, y) \equiv \phi(t, y^3) \tag{45}$$

and for the metric

$$\begin{aligned} ds^2 = g_{ab}dx^a dx^b &= g_{tt}(t,x)dt^2 + g_{tx}(t,x)dtdx + g_{xx}(t,x)dx^2 \\ &= g_{tt}(t,y^3)dt^2 + 3y^2 g_{tx}(t,y^3)dtdy + 9y^4 g_{xx}(t,y^3)dy^2 \end{aligned} \tag{46}$$

so that

$$g^{new}_{ty}(t,y) \equiv 3y^2 g_{tx}(t,y^3), \quad g^{new}_{yy}(t,y) \equiv 9y^4 g_{xx}(t,y^3) \tag{47}$$

These new fields ϕ^{new}, g^{new}_{ab} are totally different functions as compared to ϕ, g_{ab} but they have the same physical content. For a general $x(y)$, if we have a diff $y(x)$, then we take

$$\begin{aligned} \phi^{new}(y) &\equiv \phi(x(y)) \\ g^{new}_{ab}(y) &\equiv g_{cd}(x(y))\frac{dx^c}{dy^a}\frac{dx^d}{dy^b} \end{aligned} \tag{48}$$

These new fields $g^{new}_{ab}(y), \phi^{new}(y)$ are physically equivalent to the old fields, even though they are numerically different. Even more generally, we have the tensor transformation rule

$$T'^{\mu'_1\cdots\mu'_k}{}_{\cdots\nu'_l}(x') = \sum_{\mu_1,\ldots,\nu_l=1}^{n} T^{\mu_1\cdots\mu_k}{}_{\cdots\nu_l}(x)\frac{\partial x'^{\mu'_i}}{\partial x^{\mu_1}}\cdots\frac{\partial x^{\nu_l}}{\partial x'^{\nu'_l}} \tag{49}$$

when we change from the coordinate system x to x'.

In gauge theories of spin 1 particles, A_μ isn't observable, since it's not gauge invariant. Similarly, diff gauge redundancies mean that we can't measure g_{ab}, because its value at a point in spacetime changes under the gauge redundancy. But g_{ab} is not special in this regard, as all of the tensor fields in our theory will change under diffeomorphisms. This means that they are also not physical/observable,

since for example ϕ^{new} differs functionally from the original ϕ. So all of the objects we usually imagine measuring, including all of the local quantum fields, are not actually observable in quantum gravity.

Instead, physical observables must be defined with reference to the state of the system. For example, at the classical level we can define the curvature (but not the metric) at a point where two particles intersect, since this will be invariant to the diffeomorphism gauge redundancy.

4.2.4 As a Property of Massless Spin 2 Particles

A traceless symmetric tensor in 4d has $4 \times 3/2 - 1 = 5$ degrees of freedom. This is the number of degrees of freedom for a massive spin 2 particle. At the most direct (undergraduate) level, this is the fact that we have spin z components $2, 1, 0, -1, -2$. But massless spin ℓ particles only have 2 degrees of freedom. Thus the gauge redundancy serves the purpose of eliminating the extra 3 DoF. This is the (opposite of) the Higgs mechanism, where massless particles need to 'eat' DoF to become massive. It's most directly seen by studying the little group.

Why 4 gauge redundancies if we only need to kill 3 DoF? Roughly speaking this is because we need to eliminate all the DoF in a massive vector field, which has 4 components, one of which is non-dynamical.

4.3 What's Redundant? Classical vs Quantum?

At the classical level, and in perturbation theory around a classical solution, the gauge redundancy of GR may not be a major conceptual problem. But it does seem to present a major challenge at the non-perturbative quantum level. Let's unpack this a bit.

If we can do semiclassical perturbation theory with $G_N \to 0$, where some semiclassical energies $E \to \infty$ so that $G_N E$ is held fixed, and quantum effects are expanded as fluctuations around the semiclassical background, then we're in the realm[4] of EFT. Semiclassical perturbation theory simplifies our task because we can use the heavy background as a reference frame, and fix the gauge with respect to it. Nevertheless, it can still be confusing to define physical measurements for local observers once we include quantum fluctuations.

And things are much more confusing in the full quantum theory, where we allow large fluctuations in the metric.

I should emphasize that by definition, gauge transformations must vanish at infinity, so there is no redundancy (no gauge symmetry) in the holographic description of quantum gravity. That is, there is no redundancy in the CFT of AdS/CFT, nor is there any in the S-Matrix, which is the holographic dual of quantum gravity in flat spacetime.

This means that we can anchor ourselves to infinity, and study physics within a causal diamond. But without perturbation theory, we have no a priori way to distinguish different spacelike surfaces (Cauchy surfaces) within a causal diamond.

[4]It's important to realize that EFT is really a long-distance expansion, not a low-energy expansion. EFTs can correctly describe very large energies as long as the energy is spread among many DoFs.

5 Canonical Gravity

5.1 ADM Variables and Their Geometry

To provide a Hamiltonian formulation of GR we need (1) a notion of constant time hypersurfaces foliating spacetime and (2) a set of canonical coordinates and momenta, which will presumably involve components of the metric and its time derivatives. Less formally, we would like to cleanly separate 'time' from 'space' in our dynamical spacetime. This is nicely explained in Wald, from which much of the following is drawn.

Let us consider a general spacetime metric, and foliate it with Cauchy surfaces Σ_t parameterized by a global time function t, which sets a time for every point in spacetime. The time function determines a time-like vector field t^a satisfying

$$t^a \nabla_a t = 1 \tag{50}$$

Note that t^a will not in general be orthogonal to the Σ_t. So let n^a be a unit vector field orthogonal to Σ_t. We can write a spatial $d-1$-dimensional metric on Σ_t as

$$h_{ab} = g_{ab} + n_a n_b \tag{51}$$

This metric has rank $d-1$ because n_a is a null vector of this metric (in the linear algebra sense); it is a metric on the Σ_t because vectors in Σ_t have the same length according to h_{ab} as they do according to g_{ab}.

Now we can decompose t^a into parts normal and parallel to Σ_t; the decomposition defines the lapse and shift functions

$$\begin{aligned} N &= -t^a n_a = (n^a \nabla_a t)^{-1} \\ N^a &= h^a_b t^b = t^a + n^a (n_b t^b) \end{aligned} \tag{52}$$

The lapse function measures the rate of flow of proper time with respect to coordinate time t as one moves normally to Σ_t, whereas the shift determines how we 'shift' along Σ_t as we move forward in time. It may be helpful to note that if $t^a = n^a$ then $N = 1$ and $N^a = 0$.

The full spacetime metric can be written in terms of the lapse and shift functions N and N_i and the spatial metric h_{ab}. In this well-known ADM form the metric is

$$ds^2 = -N^2 dt^2 + h_{ij}(dx^i + N^i dt)(dx^j + N^j dt) \tag{53}$$

We can also think of this as the metric h_{ij} within Σ_t, plus time-space and time-time components

$$ds^2 = h_{ij} dx^i dx^j + 2N^i dx_i dt + \left(N^i N^j h_{ij} - N^2\right) dt^2 \tag{54}$$

The first two terms have an obvious interpretation. For the last, note that

$$h_{ij} N^a N^b = h_{ij} t^i t^j \tag{55}$$

so the last term can be viewed as the norm of t^i written in terms of N^i and N. Another way to derive this form of the metric is to note that

$$n^a = \frac{t^a - N^a}{N} \tag{56}$$

and so we can write the inverse metric as

$$g^{ab} = h^{ab} - \frac{(t^a - N^a)}{N}\frac{(t^b - N^b)}{N} \tag{57}$$

Inverting the inverse metric gives us back the original metric in the ADM form, where $t^a dt_a = 1$ determines what we mean by 'dt'.

Interpreting t as the global time coordinate, we can use h_{ij} as a canonical 'Q' coordinate in the Lagrangian or Hamiltonian formalism. Note that the lapse and shift are not dynamical, because they only define what it means to go 'forward in time'; no time derivatives of N or N^i will appear in the Einstein-Hilbert action. The extrinsic curvature of the constant time hypersurfaces is

$$K_{ab} = h_a^c \nabla_c n_b = \frac{1}{2}\mathcal{L}_n h_{ab} \tag{58}$$

and so the extrinsic curvature is, roughly speaking, a kind of 'time derivative' of the metric. We can view $(\Sigma_t, h_{ab}, K_{ab})$ as classical initial (though not necessarily canonical) data for GR. By this I mean that specifying them at one time should allow us to use the EoM (Einstein's equations) to determine them in the future.

5.2 Hamiltonian Formulation

The ADM variables are useful for obtaining an Hamiltonian formulation of GR, since to define a Hamiltonian we need to fix a notion of time.

We can then choose the spatial metric h_{ab}, the lapse N, and the shift vector $N_a = h_{ab}N^b$ as our field variables, rather than using the full inverse spacetime metric g^{ab} directly. Note that we obtain the inverse h^{ab} on the Cauchy surfaces Σ_t using the fact that $h^{ab}\nabla_b t = 0$. Note that this means that

$$\sqrt{-g} = N\sqrt{h} \tag{59}$$

and this appears as the volume element for spacetime integration.

To begin with we will ignore boundary terms. The scalar curvature in the E-H action can be written as

$$R = 2\left(G_{ab}n^a n^b - R_{ab}n^a n^b\right) \tag{60}$$

and we can write

$$G_{ab}n^a n^b = \frac{1}{2}\left[{}^{(3)}R - K_{ab}K^{ab} + K^2\right] \tag{61}$$

where K_{ab} is the extrinsic curvature of a Cauchy surface Σ_t, with $K = K_a^a$.

To simplify the other term we use

$$\begin{aligned} R_{ab}n^a n^b &= R^c_{acb}n^a n^b \\ &= -n^a\left(\nabla_a\nabla_c - \nabla_c\nabla_a\right)n^c \\ &= \left(\nabla_a n^a\right)\left(\nabla_c n^c\right) - \left(\nabla_c n^a\right)\left(\nabla_a n^c\right) \\ &\quad - \nabla_a\left(n^a\nabla_c n^c\right) + \nabla_c\left(n^a\nabla_c n^c\right) \\ &= K^2 - K_{ac}K^{ac} - \nabla_a\left(n^a\nabla_c n^c\right) + \nabla_c\left(n^a\nabla_a n^c\right) \end{aligned} \tag{62}$$

and the last two terms are total derivatives. So we find that the Einstein-Hilbert action is

$$\mathcal{L}_G = \sqrt{h}N\left[{}^{(3)}R + K_{ab}K^{ab} - K^2\right] \tag{63}$$

The extrinsic curvature is defined by

$$K_{ab} = h^c_a\nabla_c n_b = \frac{1}{2}\mathcal{L}_n h_{ab} \tag{64}$$

so it is the Lie derivative of the spatial metric of Σ_t in the direction n^a orthogonal to the surface. In ADM variables we can write it as

$$\begin{aligned} K_{ab} &= \frac{1}{2}\xi_n h_{ab} = \frac{1}{2}\left[n^c\nabla_c h_{ab} + h_{ac}\nabla_b n^c + h_{cb}\nabla_a n^c\right] \\ &= \frac{1}{2}N^{-1}\left[Nn^c\nabla_c h_{ab} + h_{ac}\nabla_b\left(Nn^c\right) + h_{cb}\nabla_a\left(Nn^c\right)\right] \\ &= \frac{1}{2}N^{-1}h^c_a h^d_b\left[\mathcal{L}_t h_{cd} - \mathcal{L}_N h_{ca}\right] \\ &= \frac{1}{2}N^{-1}\left[\dot{h}_{ab} - D_a N_b - D_b N_a\right] \end{aligned} \tag{65}$$

where D_a is the derivative operator on the Cauchy surface compatible with h_{ab}. Thus we can write the E-H action in terms of h_{ab}, N, N^a, as was first done by ADM 58 years ago. Notice that time derivatives of N, N^a do not occur, as D_a is a purely spatial derivative.

The canonical momentum in the ADM formalism is

$$\pi^{ab} = \frac{\partial\mathcal{L}_G}{\partial\dot{h}_{ab}} = \sqrt{h}\left(K^{ab} - Kh^{ab}\right) \tag{66}$$

There are no momenta for N, N^a as they are not dynamical variables, but only give constraints. Dropping a boundary term, the Hamiltonian density is then

$$\begin{aligned} \mathcal{H}_G &= \pi^{ab}\dot{h}_{ab} - \mathcal{L}_G \\ &= -h^{1/2}N^{(3)}R + Nh^{-1/2}\left[\pi^{ab}\pi_{ab} - \frac{1}{2}\pi^2\right] + 2\pi^{ab}D_a N_b \\ &= h^{1/2}\left\{N\left[-{}^{(3)}R + h^{-1}\pi^{ab}\pi_{ab} - \frac{1}{2}h^{-1}\pi^2\right] - 2N_b\left[D_a\left(h^{-1/2}\pi^{ab}\right)\right]\right\} \end{aligned} \tag{67}$$

and Hamilton's equations are equivalent to Einstein's equations in the vacuum.

However, we also have constraints from the variation of N, N^a which happen to take the form

$$\begin{aligned} h^{-1}\pi^{ab}\pi_{ab} - {}^{(3)}R - \frac{1}{2}h^{-1}\pi^2 &= 0 \\ D_a\left(h^{-1/2}\pi^{ab}\right) &= 0 \end{aligned} \tag{68}$$

The second constraint is equivalent to the statement that spatial diffeomorphisms – ie diffeomorphisms of the Cauchy surface Σ_t itself – are a gauge redundancy. Very specifically

$$h_{ab} \to h_{ab} + D_{(a}v_{b)} \tag{69}$$

is the redundancy, and so the Poisson bracket of π with the gauge term should vanish. After an integration by parts this is exactly the second constraint above.

However, the first constraint remains, as a result of the diffeomorphism gauge redundancy in the time direction. We will discuss it further below. First let's discuss boundary terms (see Regge and Teitelboim's 1974 paper for many details).

The bulk Hamiltonian is linear in N, N^a and so on the EoM, including the constraints, it seems that $H = 0$! But one might wonder if this strange feature is resolved by boundary terms. In order to obtain a well-defined variational principle, we do need to add a boundary term to the E-H action

$$S_{\text{full}} = S_{\text{E}-\text{H}} + 2\int_{\partial M} K \tag{70}$$

involving the extrinsic curvature of the boundary. However, if we carefully include all of the boundary terms from the derivation above along with the extrinsic curvature term, then we find that the total energy of a closed universe (no spatial boundary, eg the universe is sphere) is still zero. In the QM theory, this would mean that the operator $H = 0$ and the Schrodinger equation is completely vacuous!

In the case of a universe with a boundary, after carefully incorporating boundary terms we obtain a total Hamiltonian of the form

$$H_{\text{full}} = H_{\text{E}-\text{H}} + \int_{\partial\Sigma_t}\left(\frac{\partial h_{\mu\nu}}{\partial x^\nu} - \frac{\partial h_{\nu\nu}}{\partial x^\mu}\right)r^\mu \tag{71}$$

where the second term is the only non-vanishing term when we evaluate on the EoM. This is the more precise version of our Gauss's Law (17) analysis from the beginning of these notes, though the final expression is identical. It's also possible to compute a total momentum and even an angular momentum of the entire universe using similar methods (see Regge and Teitelboim for details).

Thus these (boundary) terms define the total energy in the universe! Our discussion may appear classical, but in a QM treatment we would simply elevate Poisson brackets to commutators and derive equivalent operator equations. So at roughly the same time that Bekenstein, Hawking, and others were discovering BH thermodynamics, physicists studying semiclassical GR were learning that the Hamiltonian of the universe – the operator that generates time translations – only depends on the behavior of the gravitational field at infinity.

5.3 Diffeomorphisms in Time and the Wheeler-DeWitt Equation

In a closed universe, we have found that the full Hamiltonian is zero on the EoM. This is a consequence of the constraints. However, one of the constraints is explicitly associated with time diffeomorphisms, and takes the form

$$\frac{1}{h}\pi^{ab}\pi_{ab} - {}^{(3)}R - \frac{1}{2h}\pi^2 \;\; = \;\; 0 \tag{72}$$

It's the result of varying wrt the lapse N, where we recall that π_{ab} are the canonical momenta dual to h_{ab}. This constraint is the classical version of the so-called Wheeler-DeWitt equation

$$H\Psi = 0 \tag{73}$$

in a closed universe (ie ignoring boundary terms).

What does this equation mean? How can we get from it to our usual expectations for semiclassical and perturbative physics? Let's explore these questions using a few representative toy models.

5.3.1 A Toy Model 'Parameterizing' Time

Let's first consider the simplest toy context where $H\Psi = 0$, and where it's easy to interpret the result and its relationship with constraints. In this model time is 'parameterized'. Consider a QM model with action

$$S = \int dt\dot{T}(t)\left(\frac{\dot{q}^2}{2\dot{T}^2} - V(q)\right) \tag{74}$$

where we treat $T(t)$ as a field to be path-integrated over, and t is a mere parameter-time. Note that $\frac{\dot{q}}{\dot{T}} = \partial_T q$ and $dt\dot{T} = dT$, and so we could eliminate the parameter time t entirely in favor of T as an integration variable. If we did this, then we would simply have a conventional action for one DoF.

Now let's canonically quantize this theory, keeping $T(t)$ and pretending that it's a canonical variable. We find that

$$\begin{aligned} P_T &\equiv \frac{\partial L}{\partial \dot{T}} = -\frac{\dot{q}^2}{2\dot{T}^2} - V(q) \\ P_q &\equiv \frac{\partial L}{\partial \dot{q}} = \frac{\dot{q}}{\dot{T}} \end{aligned} \tag{75}$$

This means that the Hamiltonian vanishes when evaluated on the EoM, as

$$H = P_T\dot{T} + P_q\dot{q} - L = 0 \tag{76}$$

Fortunately, not all is lost, as the 'constraint' given by the definition of P_T can actually be interpreted as the usual Schrodinger equation, since

$$\begin{aligned} & P_T \equiv i\frac{\partial}{\partial T} = -P_q^2 - V(q) \\ \Longrightarrow \quad & -i\frac{\partial}{\partial T}\Psi = \left[P_q^2 + V(q)\right]\Psi \end{aligned} \tag{77}$$

just says that T evolution is generated by what we would have usually called the Hamiltonian of the theory without T. So we have obtained the usual Schrodinger equation. Of course this was inevitable since our theory was equivalent to a standard theory with T playing the role of time.

The Hamiltonian version of these statements views $P_T, T; P_q, q$ as the dynamical variables (ie the relations between momenta and coordinates follow from the Hamiltonian formalism, so we do not impose them from the begining). This means that we just have

$$H = \dot{T}\left(P_T + \tilde{H}(q, P_q)\right) \tag{78}$$

where $\tilde{H}$ is the usual Hamiltonian $\frac{1}{2}P_q^2 + V(q)$. In the usual ADM formalism we would call $\dot{T} = N$, the lapse. Hamilton's equations are

$$\begin{aligned} \dot{T} &\equiv \frac{\partial H}{\partial P_T} = N \\ \dot{P}_T &\equiv -\frac{\partial H}{\partial T} = 0 \\ \dot{q} &\equiv \frac{\partial H}{\partial P_q} = N\frac{\partial \tilde{H}}{\partial P_q} \\ \dot{P}_q &\equiv -\frac{\partial H}{\partial q} = -N\frac{\partial \tilde{H}}{\partial q} \end{aligned} \tag{79}$$

These need to be supplemented by the constraint $H = 0$ (which comes from varying H with respect to N), after which they reproduce the usual EoM. The constraint is needed because we cannot eliminate $\dot{T}$ using P_T, ie we cannot solve for both $\dot{q}$ and $\dot{T}$ in terms of P_T, P_q. This is because T isn't an independent degree of freedom.

5.3.2 A Better Toy Model

The reason we were able to obtain a nice Schrodinger equation in the Hamiltonian in the toy model above is that the constraint (from varying wrt $\dot{T}$) was linear in the momentum P_T, and so P_T could act as $-i\partial_T$ in the Schrodinger equation.

Unfortunately, this situation does not arise in GR. As we see in equation (72), the Wheeler-DeWitt equation is quadratic in the canonical momenta of GR, so we cannot understand it as a re-writing of a conventional Schrodinger equation. Time diffeomorphisms are more non-trivial.

However, we can mimic perturbative GR quite well with a slightly more complicated toy model, where we view its mass parameter $M \gg 1$ as the analog of M_{pl}. Consider the action

$$S = \int dt N \left(\frac{\dot{q}^2}{2N^2} - V(q, a) + M\frac{\dot{a}^2}{2N^2} - MU(a)\right) \tag{80}$$

with N playing the role of the lapse. The canonical momenta for q, a are

$$P_q = \frac{\dot{q}}{N}, \quad P_a = M\frac{\dot{a}}{N} \tag{81}$$

and the Hamiltonian is

$$H = N\left[\frac{1}{2}P_q^2 + \frac{1}{2M}P_a^2 + V(q,a) + MU(a)\right] \tag{82}$$

So the EoM for the Lagrange multiplier N sets

$$H\psi = 0 \tag{83}$$

as in the WdW equation. Explicitly the constraint dictates that

$$\left[-\frac{1}{2}\partial_q^2 - \frac{1}{2M}\partial_a^2 + V(q,a) + MU(a)\right]\Psi = 0 \tag{84}$$

A challenge is that this equation is quadratic in q and a derivatives, and so neither variable seems to be a natural time coordinate. We cannot immediately interpret this as a Schrodinger equation.

But when $M \gg 1$, the a DoF behaves nearly classically, and so can serve as a clock. The key idea is that in the semiclassical approximation, the wavefunction partially factors in a way that lets us use the value of a as a temporal reference point.

We can solve for the large M behavior of the wavefunction as

$$\psi(a) \approx U^{-1/4} e^{\pm iM\int^a \sqrt{U(x)}dx} \tag{85}$$

via the WKB approximation. Then we can write

$$\psi(q,a) = \psi_{WKB}(a)\chi(q,t(a)) \tag{86}$$

where we define

$$\pm\sqrt{U(a)}\frac{dt}{da} = -1 \tag{87}$$

and have the Schrodinger equation

$$i\partial_t\chi(q,t) = \left[-\frac{1}{2}\partial_q^2 + V(q,a(t))\right]\chi(q,t) \tag{88}$$

where the time t has been defined in terms of the heavy 'clock' DoF.

To further illustrate this, let's consider the case where U is a constant as an explicit example, and go through the math. We have

$$\psi_{WKB}(a) = U^{-1/4} e^{\pm iM\sqrt{U}a} \tag{89}$$

There is no time dependence here, rather this is just a wavefunction for a particle with a constant energy. Plugging the combined wavefunction into the WdW equation and dividing by the WKB wavefunction gives

$$\pm i\sqrt{U}\partial_a\chi(q,t(a)) - \frac{1}{2}\partial_q^2\chi(q,t(a)) - \frac{1}{2M}\partial_a^2\chi(q,t(a)) + V(q,a)\chi(q,t(a)) = 0 \tag{90}$$

As promised, the first term, where the ∂_a^2 derivatives act on both the ψ_{WKB} and χ wavefunctions, turn the WdW equation into a Schrodinger equation for χ. Re-writing, we have

$$\mp i\partial_t\chi(q,t) = -\frac{1}{2}\partial_q^2\chi(q,t(a)) + V(q,a)\chi(q,t(a)) - \frac{1}{2M}\left(\frac{1}{U}\partial_t^2\chi(q,t(a)) \pm \frac{U'(t)}{2U^2}\partial_t\chi(q,t(a))\right)$$

In the case where U is constant the final term vanishes, but in general it will be part of the correction to the WKB approximation. The first term in parentheses is universal, and represents a $1/M$ correction to our Schrodinger equation.

5.3.3 Wheeler-DeWitt

I don't know how to define time or local observables in quantum gravity in the absence of a semiclassical perturbative expansion, or some sort of background of objects to use as reference points. In AdS/CFT one can attempt to make reference to the boundary, though this seems very unnatural for bulk observers.

The perturbative interpretation we have provided, 'the WKB Interpretation', is somewhat controversial, but it would appear to be sufficient for perturbative EFT around a semiclassical gravitational background. In that case, by taking the large M_{pl} limit we reduce canonical gravity with the WdW equation to the usual picture of quantum fluctuations around a classical solution, where the physics of the semiclassical solution (and some gauge fixing) can provide us with a notion of time. As far as I know no one has explicitly spelled out how this works in detail, connecting WdW with a saddle point expansion of a path integral, but it seems that this should be possible, if notationally inconvenient. Perhaps the nearest attempt is in an old paper by Banks, Fischler, and Susskind from 1985.

Typical discussions of WdW most often focus on 'mini-superspace', where you study quantum cosmology and reduce to just two DoF, much like the QM model from the last section. Interpretations also tend to be similar, but they often don't reference perturbation theory as we have done.

6 Symmetries in General Relativity

When discussing symmetries, it is important to differentiate between the symmetries of the theory and the symmetries of some particular state.

In QFTs it is easy to mistake these two concepts because the fields fill spacetime, forming a backdrop that we are liable to take for granted. In General Relativity the spacetime manifold itself becomes a dynamical variable, so that the entire background geometry[5] is state-dependent. So we

[5]As an example of the kind of confusion I have in mind: you might have been tempted to assume that gravity in Minkowski space has Poincaré symmetry, because the metric

$$ds^2 = -dt^2 + dx_i^2 \tag{91}$$

is translation and Lorentz invariant. But the words 'gravity in Minkowski space' do not actually make sense – we can talk about very small gravitational perturbations about Minkowski space, but this is just an expansion about one

face a problem: how can we understand the symmetries of our theory of spacetime without referring to any particular spacetime?

To truly understand symmetries in GR, we will need to take a few detours. In the process we'll gain a better understanding of how gravitational theories can be given holographic descriptions.

6.1 Penrose Diagrams

6.2 Asymptotic Symmetries

6.3 AdS_3 and Virasoro

7 The Temperature of a Horizon

7.1 KMS Condition and Geometry

7.2 Rindler Space and Unruh Radiation

7.3 Black Hole Temperature

7.4 Analysis of a Detector

7.5 Other Derivations of Hawking Radiation

7.6 DeSitter Horizons

A Vector Fields, Diffeomorphisms, and Isometries

A.1 Vectors and Diffeomorphisms

It will be useful to take advantage of the fact that there is a one-to-one correspondence between directional derivatives and vectors. Given a vector v^μ (in a coordinate basis) in the tangent space to a manifold M at some point p, we can define the directional derivative

$$v^\mu \partial_\mu \tag{92}$$

This takes functions $f : M \to R$ to functions. Thus we can (re-)define tangent vectors at $p \in M$ as the space of linear maps that obey the Leibnitz rule, so

$$v(af + bg) = av(f) + bv(g) \tag{93}$$

and

$$v(fg) = v(f)g + fv(g) \tag{94}$$

particular classical solution to the EoM. Since the metric $g_{\mu\nu}$ is a dynamical variable, it is state dependent. We can only conclude that Poincaré symmetry is the symmetry of one particular state where $\langle g_{\mu\nu} \rangle = \eta_{\mu\nu}$; the existence of this state tells us nothing about the symmetries of the *theory* itself.

One can easily prove that v has the usual properties of vectors from this definition.

On a manifold M, a *vector field* is a linear map taking functions on M to functions on M. Intuitively, the vector field is just a smooth specification of vectors at each point in M.

A diffeomorphism is a smooth map of manifolds $\phi : M \to N$ that is bijective and that has a smooth inverse. It 'pulls back' a function $f : N \to R$ to a function $f \cdot \phi : M \to R$. Similarly, it 'carries along' tangent vectors at $p \in M$ to tangent vectors at $\phi(p) \in N$, because if $g : N \to R$, we can compute $v(g \cdot \phi)$. This defines a map $\phi^* : V_p \to V_{\phi(p)}$ via

$$(\phi^* v)(g) = v(f \cdot \phi) \tag{95}$$

The map ϕ^* is linear and can be viewed as the derivative of ϕ, since it takes any vector in the tangent space of M to a vector in the tangent space of N.

We can also 'pull back' dual vectors (formally defined as linear maps from vectors to real numbers) at $\phi(p) \in N$ to $p \in M$ via

$$(\phi_* \mu)_a v^a = \mu_a (\phi^* v)^a \tag{96}$$

The 'carry along' operation we defined above on vectors has been used to define the pull back of dual vectors. We can extend this to general tensors (defined as multilinear maps, of course).

If ϕ is a diffeomorphism then we can use ϕ^{-1} to define the pull back and carry forward operations. One can show that $\phi_* = (\phi^{-1})^*$, so the pull back and carry forward are equivalent. Note that this isn't yet enough by itself to relate vectors and dual vectors; we need a metric for that.

Now if $\phi : M \to M$ and T is a tensor field on M, then we can compare T with $\phi^* T$. In particular we can ask if $T = \phi^* T$, or in other words, we can ask if ϕ is a symmetry transformation for the tensor field T. A symmetry transformation that leaves the metric $g_{\mu\nu}$ invariant is called an *isometry*. You might naively think that isometries are symmetries in GR. We will eventually see that isometries have something to do with symmetries in GR, but we're no where near the end of the story. At a conceptual level, since isometries just leave $g_{\mu\nu}$ invariant, they can only be symmetries associated with a particular state in GR, but not a true symmetry of the full gravitational theory. That said, isometries do define the symmetries associated with physics in a fixed background metric, as we will see very soon.

A.2 Infinitesimal Diffeomorphisms and Lie Derivatives

Now let's consider a 1-parameter family of diffeomorphisms $\phi_t : M \to M$, where ϕ_0 is the identity. We can relate ϕ_t to a vector field by looking at $\phi_t(p)$ as a curve in M and finding the tangent vector to that curve. If we do this for all p, we get a vector field v.

Conversely, given a vector field v on M, we can look for the integral curves of v, or the family of curves with tangent $v(p)$ at each $p \in M$. It's easy to see that we can do that if we pick a coordinate system; then we just get an ordinary differential equation for the curves.

Now we can combine the idea of a 1-parameter family of diffeomorphisms with that of the pullback in order to define *Lie Derivatives*. If we use some ϕ_t^* to carry along a tensor field $T^{a_1 \cdots a_k}_{b_1 \cdots b_l}$, then we can make the comparison

$$\mathcal{L}_v T^{a_1 \cdots a_k}_{b_1 \cdots b_l} = \lim_{t \to 0} \frac{\phi^*_{-t} T^{a_1 \cdots a_k}_{b_1 \cdots b_l} - T^{a_1 \cdots a_k}_{b_1 \cdots b_l}}{t} \tag{97}$$

which defines the Lie derivative of T with respect to the vector field v. It tells us how a tensor field changes as we move along the flow of a vector field. Note that for functions

$$\mathcal{L}_v f = v(f) \tag{98}$$

where by the latter we are interpreting the vector field as a directional derivative. To analyze the action in general it's helpful to introduce a coordinate system where t-translations just correspond to shifting the first coordinate x^1; in that case the Lie derivative simply becomes ∂_{x^1}.

Note that this means that when we act with $\mathcal{L}_v$ on another vector field w, we get

$$\mathcal{L}_v w = [v, w] \tag{99}$$

in some particular coordinate system, such as that adapted to v. But since both the commutator of vector fields and the Lie derivative can be defined in a coordinate independent way, these quantities must be identical in general. This means that as expected, $\mathcal{L}_v v = 0$.

To compute the explicit action of $\mathcal{L}_v T^{a_1 \cdots a_k}_{b_1 \cdots b_l}$ we can work backwards from the information we already have, noting that T can be dotted into a combination of vectors and dual vectors to leave us with a scalar function on M. For example we know that when applied to a function

$$\begin{aligned} \mathcal{L}_v(\mu_a w^a) &= v^b \nabla_b (\mu_a w^a) \\ &= v^b w^a \nabla_b \mu_a + v^b \mu_a \nabla_b w^a \end{aligned} \tag{100}$$

and when applied to a vector field w we get the commutator

$$\mathcal{L}_v w = v^b \nabla_b w^a - w^b \nabla_b v^a \tag{101}$$

so that from the Liebnitz rule

$$\mathcal{L}_v(\mu_a w^a) = w^a \mathcal{L}_v \mu_a + \mu_a \mathcal{L}_v w^a \tag{102}$$

we deduce that for a dual vector

$$\mathcal{L}_v \mu = v^b \nabla_b \mu_a + \mu_b \nabla_a v^b \tag{103}$$

is the Lie derivative. We can easily go on to compute the action of the Lie derivative on a general tensor; the result is

$$\mathcal{L}_v T^{a_i}_{b_j} = v^c \nabla_c T^{a_i}_{b_j} + T^{a_i}_{b_k, c} \nabla_{b_j} v^c - T^{a_k, c}_{b_j} \nabla_c v^{a_i} \tag{104}$$

where there's a sum on each upper and lower index in the second and third terms.

If $\mathcal{L}_v T^{a_1 \cdots a_k}_{b_1 \cdots b_l} = 0$ then we say that the tensor field T doesn't change when it's transported along v. Thus we can view v and its associated diffeomorphism as a symmetry of the tensor field T on M.

A.3 Algebras of Vector Fields

What if vector fields v and w are both symmetries of some tensor field $T^{a_1 \cdots a_k}_{b_1 \cdots b_l}$? Then it's also true that their commutator vector field $[v, w]$ will be a symmetry, since it can only translate T by a combination of v and w. Thus for a given tensor field T we can talk about the *Lie algebra of vector fields that leave it invariant*.

Vector fields that leave the metric tensor g_{ab} invariant are called *Killing vector fields*. These generate one-parameter families of isometries. A vector field is a Killing field iff

$$\begin{aligned} \mathcal{L}_v g_{ab} &= v_c \nabla^c g_{ab} + g_{cb} \nabla_a v^c + g_{ac} \nabla_b v^c \\ &= \nabla_a v_b + \nabla_b v_a = 0 \end{aligned} \tag{105}$$

for the covariant derivative ∇_a compatible with the metric g_{ab} (so the covariant derivative of the metric vanishes by definition). Sometimes we talk about conformal killing vector fields, which do not leave the metric invariant, but act on it to give a tensor proportional to g_{ab}. In equations, $\nabla_a v_b + \nabla_b v_a \propto g_{ab}$. These are one way of defining the conformal group in flat spacetime.

A useful property of Killing vector fields is that if we take any geodesic in the manifold, and u^a *is a vector tangent to the geodesic, then* $v_a u^a$ *is constant along the geodesic*. When we study non-gravitational physics in a fixed spacetime geometry, we use this property to define the spacetime symmetries of the theory.