Jerusalem lectures on black holes and quantum information

D. Harlow

Princeton Center for Theoretical Science, Princeton University, Princeton, New Jersey 08540, USA

(published 2 February 2016)

These lectures give an introduction to the quantum physics of black holes, including recent developments based on quantum information theory such as the firewall paradox and its various cousins. An introduction is also given to holography and the anti–de Sitter/conformal field theory (AdS/CFT) correspondence, focusing on those aspects which are relevant for the black hole information problem.

DOI: 10.1103/RevModPhys.88.015002

CONTENTS

I. Introduction 1
A. Conventions 2
II. Classical Black Holes 2
A. The Schwarzschild geometry 2
B. The Kruskal extension 3
C. Penrose diagrams 4
D. Real black holes 6
III. Entanglement in Quantum Field Theory 6
A. Quantum field theory 6
B. Entanglement in the vacuum 8
C. The Rindler decomposition 9
D. Free fields in Rindler space 10
E. Entanglement is important for horizon crossing: An introduction to firewalls 11
IV. Quantum Field Theory in a Black Hole Background 12
A. Two-sided Schwarzschild and the Rindler decomposition 12
B. Schwarzschild modes 13
C. Hawking's calculation of black hole radiation 14
D. Evaporation 15
E. Entropy and thermodynamics 15
F. The information problem 16
G. The brick wall model and the stretched horizon 18
H. The Euclidean black hole 20
V. Unitary Evaporation 21
A. The S matrix 22
B. The Page curve 23
C. Page's theorem 23
D. How hard is it to test unitarity? 25
E. What is a typical microstate? 26
F. Scrambling and recovery of quantum information 26
G. Black hole complementarity 29
VI. Holography and the AdS/CFT Correspondence 30
A. Entropy bounds and the holographic principle 30
B. Statement of the AdS/CFT correspondence 31
1. Anti–de Sitter space 31
2. Conformal field theory 32
3. The dictionary 33
C. Perturbations of the AdS vacuum 34
D. One-sided AdS black holes at fixed energy 36
E. One-sided AdS black holes at fixed temperature and the Hawking-Page transition 37
F. Fields in the AdS-Schwarzschild background 38
G. Collapsing shells and the two-sided AdS wormhole 39
H. The information problem in AdS/CFT 40
I. Unitarity for big AdS black holes 41
J. von Neumann entropy and the Ryu-Takayanagi formula 43
VII. Paradoxes for the Infalling Observer 44
A. The entanglement-monogamy problem 44
B. Firewall typicality 46
C. The creation operator problem 46
D. The Marolf-Wall paradox 47
VIII. Proposals for the Interior 47
A. Complementarity from computational complexity? 47
B. Nonlinearity? 50
C. Postselection? 51
D. Firewalls? 53
Acknowledgments 54
References 55

I. INTRODUCTION

Black holes are fascinating objects in quantum gravity. Starting from fairly mundane initial conditions (such as a collapsing star), nature is able to produce a geometry that amplifies short-distance fluctuations to macroscopic sizes. This "stretching" of spacetime circumvents the Wilsonian decoupling of high-energy physics from low-energy physics, making deep questions of Planck-scale dynamics relevant for low-energy (thought) experiments.[1] Indeed in an extraordinary pair of classic papers, Hawking (1975, 1976) argued first that this stretching of fluctuations causes black holes to evaporate and second that the evaporation process is inconsistent with the quantum mechanical principle that pure states always evolve to other pure states. This conclusion is usually called the *black hole information problem*, and it has instigated a large amount of research in the almost 40 years since Hawking's papers. Is information indeed lost? If not, then what is the nature of the Planckian interference that prevents it? Significant progress has been made on these

[1]This amplification is also present in an expanding universe, and it seems likely that a complete understanding of black hole physics will lead to valuable lessons for quantum cosmology.

 © 2016 American Physical Society

questions, but recent work has emphasized the extent to which we still do not have satisfactory answers to them.

The goal of these lectures is first to give an introduction to as much as is reasonable of the techniques that go into formulating and analyzing these questions, and second to give an overview of the new paradoxes that have led to an explosion of recent work on the subject. I also discuss some ideas that have been proposed to resolve the paradoxes, but I by no means aim at a comprehensive review; I have throughout done my best to prioritize pedagogy over completeness. Of course in a field as chaotic as this one currently is, my views on which material should be included will be somewhat idiosyncratic. As a general rule I attempted to give, or at least sketch, the "real" arguments for things. When the foundations of the subject are under as much doubt as they are here, it is my view that sloppy logic should be avoided as much as possible.

Occasionally some details of the material are new, but I try to not call attention to this since it is awkward and tedious, and in any event my "improvements" are mostly cosmetic.

Not all sections of the notes are equally important in getting to the paradoxes of Sec. VII. Sections II and III are essential, as are Secs. IV.A–IV.F. From there things get more flexible, and Secs. IV.G and IV.H can be skipped on a first reading, as can Secs. V.D–V.F. When the lectures were actually presented I skipped all of Sec. VI, although I would not necessarily recommend that to the reader.

A considerable amount of background material is reviewed in the online Supplemental Material to this article and is referred to throughout [193]. Readers who wish to avoid this cross referencing may instead use the arXiv version of this article, where this material is included as appendixes.

Finally a comment on the target audience for these notes: the lectures were given at the 31st winter school on theoretical physics at Hebrew University in Jerusalem to a diverse audience, including condensed matter, quantum information, and high-energy theorists. The situation might be described by saying that the union of their background knowledge was maximal but the intersection was empty. On behalf of the first two groups I have fairly extensively reviewed rather standard facts about general relativity, black holes, and AdS/CFT. On behalf of the first and third groups I have done the same for some basic results in quantum information theory. Quantum field theory (QFT) is familiar to at least the first and third groups, but even they might not be comfortable with the aspects of it I use here, so on behalf of all three I have reviewed that as well. I hope that this will not cause the resulting size of these notes to deter potential readers from quickly jumping to whatever aspect they find most interesting. My reason for writing these notes is that there did not seem to be a convenient source for many of the things discussed here, some of which are dispersed throughout the literature and some of which are widely known but as far as I know do not appear in print anywhere. The source with the most overlap is probably Susskind and Lindesay (2005), from which I learned many of the things in the earlier sections of these notes. For other shorter reviews, see Preskill (1992), Giddings (1994), and Mathur (2009). I have heard from many people that there is a high barrier of entry to this field, and I hope I have lowered it a bit.

A. Conventions

To simplify equations, starting in Sec. II.B I work in units where the Schwarzschild radius $2GM$ is set to 1. This is a silly convention to use once we begin to consider situations where the black hole mass is time dependent, so it stops in Sec. IV.D. In Sec. V I switch to Planckian units where $8\pi G \equiv \ell_p^2$ is set to 1, in Sec. VI I instead set the anti–de Sitter radius to 1, and from Sec. VII onward the equations are simple enough that I keep all three explicit. Like any civilized person I will of course set $c = \hbar = k_B = 1$ throughout.

I use the symbol Ω in two different contexts; I denote ground state wave functions as $|\Omega\rangle$, and I refer to coordinates on the sphere $\mathbb{S}^d$ as Ω. $d\Omega_d^2$ is the standard "round" metric on $\mathbb{S}^d$, $d\Omega_d$ is the volume element for use in integrals, and

$$\Omega_d \equiv \frac{2\pi[(d+1)/2]}{\Gamma[(d+1)/2]}$$

is the volume.

In multipartite quantum mechanical systems I refer to the subsystems by capital Roman letters such as $A, B, \ldots$, I refer to their associated Hilbert spaces as $\mathcal{H}_A, \mathcal{H}_B, \ldots$, and I call the dimensions of their Hilbert spaces $|A|, |B|, \ldots$.

Except for Sec. VI I work almost exclusively in $3+1$ spacetime dimensions. Results for asymptotically Minkowski black holes in other dimensions can be obtained from the AdS formulas in Sec. VI by taking the limit $r_{\text{ads}} \to \infty$. I will not discuss black holes with charge or angular momentum. These more general black holes provide interesting laboratories for the information problem and reconstructing the interior, but in the end they so far do not seem to add much conceptually.

II. CLASSICAL BLACK HOLES

In this section I review the main properties of classical black holes in general relativity. For readers who are unfamiliar with this theory I give an extremely compact description in Sec. I of the Supplemental Material [193].

A. The Schwarzschild geometry

The Schwarzschild geometry is the unique source-free solution of Einstein's equation with spherical symmetry that approaches ordinary Minkowski space at large distances. By itself the Schwarzschild geometry does not quite describe a black hole in the astrophysical sense, but understanding it is a necessary prerequisite for studying them. Explicitly the spacetime metric for the Schwarzschild geometry is given by

$$ds^2 = -\frac{r-2GM}{r}dt^2 + \frac{r}{r-2GM}dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{1}$$

Here G is Newton's gravitational constant and M is a parameter with units of mass.[2] The quantity in brackets is the unit metric $d\Omega_2^2$ on the two sphere $\mathbb{S}^2$, so the coordinate r parametrizes the proper size of this $\mathbb{S}^2$. $r=0$ and $r=2GM$ are clearly special, and most of the interesting physics of black holes lies in understanding what happens at these two radii.

At $r=0$ the $\mathbb{S}^2$ shrinks to zero size and the metric diverges; this is called the *singularity*. This pathology can be described in a coordinate-invariant way as the divergence of the fully contracted Riemann tensor $R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$. Since the Riemann tensor physically encodes the strength of tidal effects on freely falling objects, this divergence leads to formally infinite tidal forces that would destroy anyone unfortunate enough to find herself in the vicinity of $r=0$. These divergences are probably regulated by Planck-scale physics, but that is little consolation.

The radius $r_s \equiv 2GM$ is called the *Schwarzschild radius*. The metric appears to also be singular here, but we see in the next section that unlike the singularity at $r=0$ this singularity is a spurious artifact of our choice of coordinates. At least classically it does not lead to any locally detectable divergence. Something important globally does happen at $r=r_s$; in Eq. (1) the signs of the coefficients of dr^2 and dt^2 switch. The coordinate r has become timelike, and any test particle which falls into the region with $r<r_s$ will necessarily continue to evolve toward smaller r until it approaches the singularity. Somebody inside of the horizon cannot prevent this for the same reason you cannot prevent yourself moving forward in ordinary time. This is true even for massless particles, so the region behind the horizon is completely invisible to anybody who stays outside at $r>r_s$.[3]

In any spacetime, if the set of points that can ever send a signal to a particular timelike geodesic has a boundary then that boundary is called the *event horizon* of the geodesic. In the Schwarzschild geometry we are especially interested in timelike geodesics that stay outside of the black hole for all times, and the surface $r=r_s$ is the event horizon for any such geodesic. In this case this surface is usually just called the horizon of the black hole.

One very important feature of the Schwarzschild horizon is its *gravitational redshift*. As we approach the horizon, a fixed unit of coordinate time counts for less and less proper time along a curve of fixed r. This means that signals sent with a fixed energy from a point that gets closer and closer to the horizon have lower and lower energy when they reach $r \gg 2GM$. Conversely, a signal propagating away from the horizon that has some fixed energy at $r \gg 2GM$ has higher and higher energy from the point of view of a fixed r observer as we move r closer and closer to $2GM$. This feature seems to allow an observer at $r \gg 2GM$ to be sensitive to very high-energy physics, and it is at the heart of the connection between black hole thought experiments and quantum gravity.

[2]To understand the physical meaning of the parameter M, we can observe that for $r \gg 2GM$ the geodesic equation (6) of the Supplemental Material for a massive nonrelativistic test particle in this geometry reduces to the standard Newtonian equation

$$m\ddot{\vec{x}} = -\frac{GmM}{r^2}\hat{r} \tag{2}$$

for the motion of a particle about a point source of mass M. Many experimental tests of general relativity are based on detecting the $O(r/2GM)$ corrections to this approximation.

[3]These statements can be justified more carefully by studying the geodesic equation in the Schwarzschild metric. One also finds that any massive observer will always reach the singularity in a finite proper time. In fact one survives the longest by not struggling; firing rockets to try to escape just causes you to die faster.

B. The Kruskal extension

The (t, r, Ω) coordinate system we used in the previous section is convenient for thinking about experiments in the $r \gg 1$ region, but it is not well suited for understanding near-horizon physics.[4] A better choice is the Kruskal-Szekeres coordinates. These are motivated by first observing that radial null geodesics in the Schwarzschild geometry can be parametrized as

$$t = \pm r_* + C, \tag{3}$$

where C is some constant of motion and r_* is a new radial coordinate

$$r_* \equiv r + \log(r-1). \tag{4}$$

r_* is called the "tortoise" coordinate, presumably because it fits an infinite coordinate range into a finite geodesic distance. The Kruskal-Szekeres coordinates are then defined as

$$U \equiv -e^{(r_*-t)/2}, \tag{5}$$

$$V \equiv e^{(r_*+t)/2}. \tag{6}$$

By construction they have the property that lines of constant U or V are radial null geodesics. These coordinates have the convenient feature that

$$UV = (1-r)e^r, \tag{7}$$

so the singularity is when $UV=1$ while the horizon is when either U or V is zero. The metric is

$$ds^2 = -\frac{2}{r}e^{-r}(dUdV + dVdU) + r^2 d\Omega_2^2, \tag{8}$$

where r is obtained implicitly from Eq. (7). The off-diagonal nature of the metric may appear unfamiliar, but it can be easily removed by defining yet another set of coordinates

$$\begin{aligned} U &= T - X, \\ V &= T + X, \end{aligned} \tag{9}$$

in terms of which the metric is

[4]From here until Sec. IV.D I work in units where $r_s = 2GM = 1$.

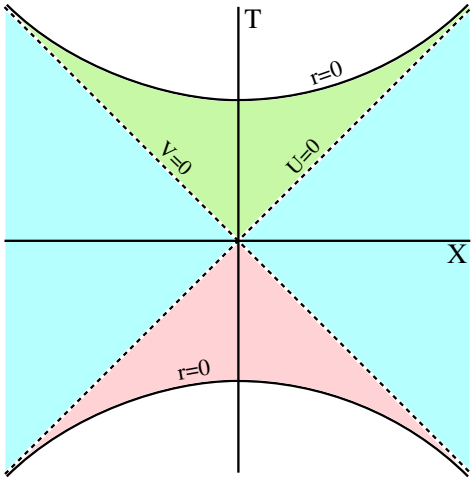


FIG. 1. The XT plane of the Kruskal extension. Lines of constant U and V, or in other words radial null geodesics, are straight lines with slope $\pm\pi/4$. Lines of constant r are hyperboloids centered at the origin, with the blue regions having $r > 1$ and the red/green regions having $r < 1$. The horizons are the dashed lines. The original exterior region is the right light blue wedge, the new exterior is the left blue wedge, the future interior is in green, and the past interior is in red. It is manifest that no radial null geodesic can escape the future interior into one of the blue regions, and it is also clear that no null geodesic connects the right and left blue wedges.

$$ds^2 = \frac{4}{r}e^{-r}(-dT^2 + dX^2) + r^2 d\Omega_2^2. \tag{10}$$

Note that there is now no singularity of any kind at $r = 1$.

Equation (10) defines a geometry over the full XT plane. It is interesting to understand which parts of it correspond to the regions discussed in the previous Sec. II.A. This is illustrated in Fig. 1. The region defined by $r > 1$, $-\infty < t < \infty$ in the old Schwarzschild coordinates is sent to the right blue wedge. In continuing to $r < 1$, however, there is a branch cut in the definition (5) which allows us to either reach the region $T > 0$, $X^2 - T^2 < 0$, shown in green, the region $T < 0$, $X^2 - T^2 < 0$, which is shown in red. The singularity at $r = 0$ is the hyperboloid $X^2 - T^2 = -1$, so it has not one but two connected components, one at the boundary of each of these regions. I refer to these two regions as the future and past interiors, respectively. Finally there is a fourth region, the left blue wedge, which is a second asymptotically Minkowski region in which we can again have $r \gg 1$. Combining all of the regions, we can interpret the full Schwarzschild geometry as a wormhole connecting two asymptotically flat universes, each of them behaving at $r \gg 1$ as if there were a gravitational point source of mass M. The wormhole is nontraversable in the sense that no signal can be sent from one blue region to the other, but observers who jump in from opposite sides are able to meet in the middle and compare notes.

C. Penrose diagrams

In gravitational thought experiments we are often uninterested in the details of the spacetime geometry. We might care only about the causal structure of the spacetime in the sense of which points can receive signals from which other points. If this is the case, then we should be able to throw out some of the irrelevant information in the metric. Indeed the following theorem gives us a useful way to do this:

- *Theorem*: Two spacetimes whose metrics differ only by multiplication by a positive scalar function on the spacetime, that is, which are related as $g'_{\mu\nu}(x) = e^{2\omega(x)}g_{\mu\nu}(x)$ for some smooth real function $\omega(x)$, have the same null geodesics. Timelike and spacelike geodesics in one metric will not necessarily be timelike or spacelike geodesics in the other, but timelike and spacelike curves in one metric will be timelike or spacelike curves in the other.

Two metrics which are related in this way are said to be *conformally equivalent*. The proof of this theorem is not difficult and is left to the interested reader as an exercise.

It was realized long ago by Penrose that this theorem gives an elegant way to represent the asymptotic behavior of spacetimes at large distance (such as the $r \to \infty$ limit in Minkowski space). The idea is to judiciously choose a function $\omega(x)$ that diverges as we approach infinity in just such a way that infinity is brought in to finite proper distance. We may then include infinity as a boundary of the spacetime, turning it into a manifold with boundary. This procedure is called *conformal compactification*.

Conformal compactification is perhaps best understood by studying an example, so let us consider ordinary flat Minkowski space. The usual spacetime metric is

$$ds^2 = -dt^2 + dr^2 + r^2 d\Omega_2^2. \tag{11}$$

There is interesting asymptotic behavior as we take $r \to \infty$ and/or $|t| \to \infty$ that we would like to analyze. The rough idea is to use the function $\arctan(x)$ to "pull in" the boundary to finite distance, but to preserve the simplicity of the causal structure we need to do this in lightlike directions. We thus define

$$\begin{aligned} T + R &\equiv \arctan(t + r), \\ T - R &\equiv \arctan(t - r), \end{aligned} \tag{12}$$

which gives a metric

$$\begin{aligned} ds^2 = &\frac{1}{\cos^2(T+R)\cos^2(T-R)} \\ &\times \left[-dT^2 + dR^2 + \left(\frac{\sin(2R)}{2}\right)^2 d\Omega_2^2\right]. \end{aligned} \tag{13}$$

These new coordinates have ranges $|T \pm R| < \pi/2$, $R \geq 0$, so we can now compactify the spacetime by including the points at the boundary $|T \pm R| = \pi/2$. The prefactor diverges at this boundary, as it must since the boundary is infinitely far away, but we can now use the theorem to define a new spacetime with the same causal structure as Minkowski space by simply removing this prefactor. This construction is illustrated in Fig. 2.

The new boundary is naturally divided into five parts: past and future timelike infinity, marked as $i^{\mp}$ in the figure, past and future null infinity, marked as $J^{\mp}$ in the figure, and spatial infinity, marked as i^0 in the figure. $i^{\mp}$ are where timelike geodesics "come from" and "go to," $J^{\mp}$ are the same for null

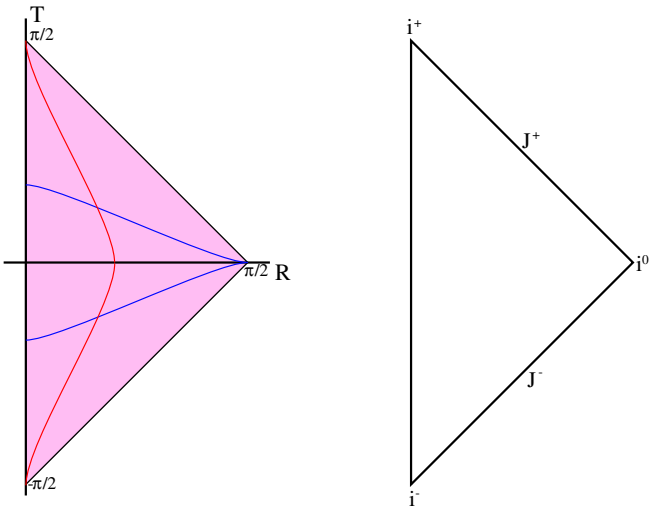


FIG. 2. On the left, the full Minkowski space is the pink wedge in the RT plane. Radial light rays move on lines of slope $\pm\pi/4$. Some slices of constant t are shown in blue and a slice of constant r is also shown in red. On the right we formalize this into a genuine Penrose diagram.

geodesics, and i^0 is where spatial geodesics end. The scattering matrix maps states on $J^-\cup i^-$ to states on $J^+\cup i^+$, with massless particles entering and leaving at $J^\mp$ and massive particles entering and leaving at $i^\mp$. Conserved charges in general relativity such as the total energy or electromagnetic charge are always written as boundary integrals at i^0. The diagram also makes it clear that there are no event horizons in Minkowski space; any timelike geodesic can eventually receive signals from everywhere in the space.

The right-hand diagram in Fig. 2 is our first example of a Penrose diagram. It is an extremely compact way of describing the causal structure of the spacetime without any extraneous details. Just from the diagram we have already seen that this spacetime has no horizons, and that it should have a nice description in terms of an S matrix.

Let us now understand in more detail what happened to the other two dimensions. At each point of the Penrose diagram there is an $\mathbb{S}^2$ which we have suppressed. It is the spherical symmetry of the metric (13) which allows us to do this without losing much information about the spacetime, and indeed we can draw a similar diagram for any spacetime with $\mathbb{S}^2$ symmetry. There are only two ways to have a boundary of a Penrose diagram: one is for the $\mathbb{S}^2$ to have infinite size, as happens at $|T\pm R|=\pi/2$ in our Minkowski diagram. The other way is for the $\mathbb{S}^2$ to collapse to zero size, as happens at $R=0$. In the interior of the diagram, the spacetime is locally just a product manifold where the radius of the $\mathbb{S}^2$ varies as we move around in the diagram; this is often called a *warped product*.[5]

Another useful example of a Penrose diagram is that of de Sitter space, which has metric

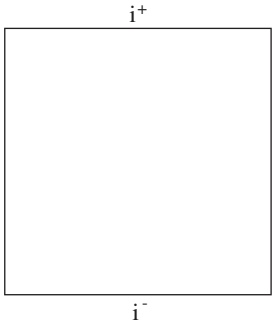


FIG. 3. The Penrose diagram of de Sitter space. The $\mathbb{S}^2$ shrinks to zero size on both the left and right boundaries, while it grows to infinite size at $i^\pm$. Note that here $i^\pm$ are each spacelike surfaces instead of just points; it is this property that leads to the presence of horizons.

$$ds^2=-d\tau^2+\cosh^2\tau d\Omega_3^2. \tag{14}$$

This is a solution of Einstein's equation with positive vacuum energy, and it is a good approximation to the geometry of our Universe today at the largest scales, as well as during cosmological inflation in the past. The spatial geometry is an $\mathbb{S}^3$ which first shrinks exponentially to some minimum size and then expands exponentially. The Penrose diagram of de Sitter space is shown in Fig. 3. From the diagram it is easy to see two important properties of de Sitter space:

- de Sitter space has event horizons. Two observers moving on timelike geodesics, say vertical straight lines in the diagram, will eventually be unable to communicate. Crudely this is because the accelerating expansion of the Universe has caused them to be moving away from each other faster than light.
- de Sitter space has no infinite spatial boundary i^0, nor any separate lightlike infinity $J^\pm$ distinct from $i^\pm$. This is a serious problem for attempts to formulate a quantum theory of de Sitter space, since existing well-defined theories of quantum gravity require at least one of these. In particular there is no straightforward sense in which de Sitter space has an S matrix.

Finally we of course should understand the Penrose diagram of the Schwarzschild geometry. The Kruskal-Szekeres coordinates have already done most of the work for us, since the metric is already in the form (10). From a causal structure point of view the only difference between these (T,X,Ω) coordinates and the (t,r,Ω) coordinates in Minkowski space we just considered are their ranges. In Minkowski space we had $-\infty<t<\infty$ and $r\geq 0$, while for the Kruskal-Szekeres coordinates we have $X^2-T^2>1$. We can use the same compactification transformation as for Minkowski space

$$\begin{aligned} T'+X' &\equiv \arctan(T+X),\\ T'-X' &\equiv \arctan(T-X), \end{aligned} \tag{15}$$

but now instead starting with the diamond $|R\pm T|<\pi/2$ and throwing out the region $R<0$, we instead start with the diamond $|X'\pm T'|<\pi/2$ and throw out the region $|T|>\pi/4$. The resulting Penrose diagram is shown in Fig. 4. This

[5]More generally, in d spacetime dimensions Penrose diagrams are useful anytime we can write the geometry as a warped product of a $d-2$ dimensional space over some subset of the two-dimensional plane.

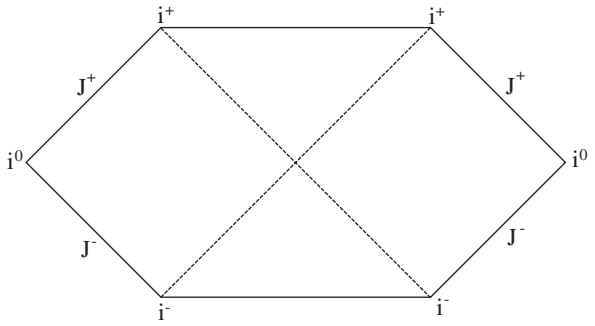


FIG. 4. The Penrose diagram for the Schwarzschild geometry. The $\mathbb{S}^2$ shrinks to zero size only at the singularities at the top and bottom horizontal lines. There are two copies of the asymptotic boundaries of Minkowski space, one on either side. For convenience the horizons are marked with dashed lines.

diagram is rather similar to the Kruskal diagram in Fig. 1, but the spacetime boundaries are now explicitly shown.

D. Real black holes

So far I have not said much about black holes. The reason is that the Schwarzschild geometry, with its two asymptotic infinities and no matter, is not a good description of the black holes that usually form in nature. Real astrophysical black holes result from the gravitational collapse of ordinary matter either at the end of the life of a star or in the center of a galaxy. These processes have some irritating limitations arising from details of particle physics. For example, there is a lower bound on the mass of black holes that can form from stellar collapse: below a mass which is of the order of the solar mass $M_\odot$, gravitational collapse is halted by the formation of a neutron star and no black hole is created.[6] To avoid having to worry about this kind of thing, it is convenient to instead imagine making a black hole out of a spherically symmetric infalling shell of photons.[7] The infalling shell can be very diffuse at the moment when it passes through its own Schwarzschild radius, so there is no obstacle to forming a black hole in this way. In fact we can explicitly construct the geometry in this case by sewing together a piece of Minkowski space and a piece of the Schwarzschild solution. The Penrose diagrams for these two methods of collapse are shown in Fig. 5.

It is interesting to note that in the case of the collapsing photon shell, the horizon extends into the region that is purely

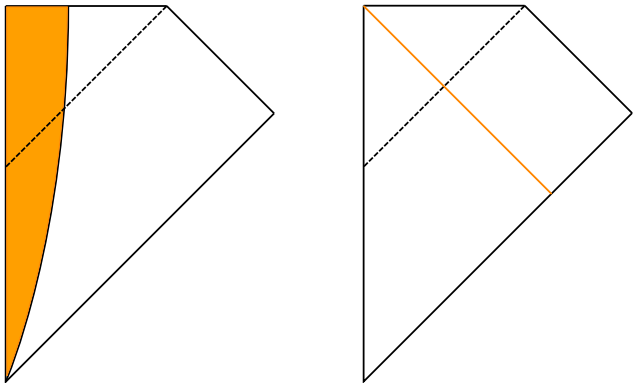


FIG. 5. Classical black hole formation. Left: A black hole forming from the collapse of a cloud of massive particles, shown in orange. Right: A black hole forming from the collapse of a spherical shell of photons. In both cases the top boundary is the singularity, the left boundary is the origin of polar coordinates, and the other boundaries are the usual asymptotic ones for Minkowski space. In the right-hand figure, the geometry above the orange line is exactly a piece from the upper right corner of the Schwarzschild geometry, while below it we have a piece of Minkowski space. As usual the horizon is a dashed line.

Minkowski space. Somebody who was passing through the horizon down at that point would have absolutely no idea that his fate was sealed. In fact we could currently be passing through the horizon of some gigantic yet-to-be-formed black hole, and we would not know. This illustrates the "acausal" nature of horizons; their locations depend on events that have not yet happened.

III. ENTANGLEMENT IN QUANTUM FIELD THEORY

I will now take a break from black holes to recall some basic facts about relativistic quantum field theory.

A. Quantum field theory

A quantum field theory is a particular quantum mechanical system where the Hilbert space can be thought of as an infinite tensor product over all points in space of a finite number of degrees of freedom at each point.[8] The simplest example is a single degree of freedom at each spatial point: a scalar field $\phi(x)$. The Hilbert space is spanned by states $|\phi\rangle$ where the field has a definite value at each point in space. Time evolution is generated by a Hamiltonian H which is usually taken to be a single integral over space of some fairly simple function of the degrees of freedom and their derivatives. For example, a free scalar field of mass m has a Hamiltonian

[6]This bound is fairly easily understood as a consequence of the uncertainty principle. A neutron star is a degenerate Fermi gas of neutrons, so the typical neutron momentum k_n is of the order of the inverse spacing of the neutrons, which is of the order of $N^{1/3}/R$, where $N = M/m_n$ is the total number of neutrons, R is the radius of the star, M is its mass, and m_n is the mass of a neutron. We can relate M and R by equating the gravitational energy GM^2/R and the kinetic energy Nk_n^2/m_n. Demanding that the neutrons stay nonrelativistic, i.e., that $k_n \lesssim m_n$, gives $M \lesssim m_p(m_n/m_p)^2 \approx M_\odot$ for the existence of a neutron star, where m_p is the Planck mass. Determining the $O(1)$ coefficient requires more work; see, for example, Bombaci (1996), where a bound between $1.5M_\odot$ and $3M_\odot$ is quoted.

[7]If you do not want to assume the existence of photons then you can use gravitons instead.

[8]The infinite number of points on a spatial slice is the source of the well-known UV (short-distance) and IR (long-distance) divergences of quantum field theory. IR divergences can be regulated by working in finite volume, while UV divergences can be controlled by instead considering a theory with degrees of freedom only on some fine spatial lattice of points. It is rather awkward to carry this around explicitly, so for the most part I will not attempt to. I will bring back the UV cutoff occasionally when it is needed to regulate a divergence.

$$H = \frac{1}{2}\int d^3x[\pi(x)^2 + \vec{\nabla}\phi(x)\cdot\vec{\nabla}\phi(x) + m^2\phi(x)^2]. \quad (16)$$

Here $\pi(x)$ is the canonical momentum $-i\delta/\delta\phi(x)$ conjugate to ϕ; together they obey

$$\begin{aligned}[\phi(x),\pi(y)] &= i\delta^3(x-y),\\ [\phi(x),\phi(y)] &= 0,\\ [\pi(x),\pi(y)] &= 0.\end{aligned} \quad (17)$$

Note that these commutation relations are consistent with the statement that fields at different points act on different tensor factors of the Hilbert space. This Hamiltonian follows from the Lorentz-invariant action

$$S = -\frac{1}{2}\int d^4x(\partial_\mu\phi\partial^\mu\phi + m^2\phi^2). \quad (18)$$

In conventional quantum mechanics we are often interested in explicitly describing the ground state wave function $|\Omega\rangle$.[9] This is possible for the free massive scalar field; indeed one can show (Weinberg, 1995) that

$$\langle\phi|\Omega\rangle \propto e^{-(1/2)\int d^3xd^3y\phi(x)\phi(y)K(x,y)}, \quad (19)$$

where

$$\begin{aligned}K(x,y) &= \int\frac{d^3k}{(2\pi)^3}e^{i\vec{k}\cdot(\vec{x}-\vec{y})}\sqrt{\vec{k}^2+m^2}\\ &= \frac{m}{2\pi^2 r}\frac{d}{dr}\left(\frac{1}{r}K_{-1}(mr)\right).\end{aligned} \quad (20)$$

Here $r \equiv |x-y|$ and K_{-1} is a modified Bessel function (Abramowitz and Stegun, 19645). This expression is not particularly useful, however, and it is rather inconvenient to generalize to theories with interactions.

Rather than trying to write down the vacuum wave functional explicitly it is usually more fruitful to instead study the vacuum expectation values of products of Heisenberg picture fields $\phi(t,x) \equiv e^{iHt}\phi(x)e^{-iHt}$. For the free massive scalar field we can write a simple expression for the solution of this operator equation of motion:

$$\phi(t,x) = \int\frac{d^3k}{(2\pi)^3}\frac{1}{\sqrt{2\omega_k}}[e^{i(\vec{k}\cdot\vec{x}-\omega_k t)}a_{\vec{k}} + e^{-i(\vec{k}\cdot\vec{x}-\omega_k t)}a^\dagger_{\vec{k}}], \quad (21)$$

where $\omega_k \equiv \sqrt{\vec{k}^2+m^2}$. The creation and annihilation operators $a_{\vec{k}}$, $a^\dagger_{\vec{k}}$ obey

$$\begin{aligned}[a_{\vec{k}}, a^\dagger_{\vec{k}'}] &= (2\pi)^3\delta^3(\vec{k}-\vec{k}'),\\ [a_{\vec{k}}, a_{\vec{k}'}] &= 0,\\ [a^\dagger_{\vec{k}}, a^\dagger_{\vec{k}'}] &= 0,\\ [H, a_{\vec{k}}] &= -\omega_k a_{\vec{k}}.\end{aligned}$$

The vacuum state $|\Omega\rangle$ is annihilated by all of the $a_{\vec{k}}$, and low-lying excitations are created by acting on the vacuum with $a^\dagger_{\vec{k}}$'s.

More abstractly what this expression says is that we look for a complete basis of positive-frequency solutions[10] $f_n(x)$ to the wave equation

$$(\partial_\mu\partial^\mu - m^2)f(t,x) = 0. \quad (22)$$

We do not want solutions that grow at infinity, so we want them to in some sense be normalizable. The best choice is to have the solutions f_n be orthonormal in the Klein-Gordon (KG) norm

$$(f_1,f_2)_{\mathrm{KG}} \equiv i\int d^3x(f_1^*\dot{f}_2 - \dot{f}_1^* f_2). \quad (23)$$

To each f_n we then associate an annihilation operator a_n, and we express the field as

$$\phi = \sum_n (f_n a_n + f_n^* a_n^\dagger). \quad (24)$$

The choice of normalization ensures that a_n and $a_n^\dagger$ have the standard algebra. The solutions f_n are typically referred to as "modes." In Eq. (21) we chose a plane-wave set of modes, which are delta-function normalized, but we could also have chosen some other set. This more abstract formalism will be very useful in thinking about black holes.

Finally we look at some vacuum expectation values in the free massive theory. These are typically called correlation functions. The one-point functions vanish trivially:

$$\langle\Omega|\phi(t,x)|\Omega\rangle = 0. \quad (25)$$

This follows from the action of the creation and annihilation operators on the vacuum, but it is also a consequence of the translation invariance of the vacuum. The two-point function is more interesting, when the two points are at equal times it is given by

$$\langle\Omega|\phi(0,x)\phi(0,y)|\Omega\rangle = \frac{1}{4\pi^2}\frac{m}{|x-y|}K_1(m|x-y|). \quad (26)$$

This correlation function scales as $1/|x-y|^2$ for $|x-y| \ll m^{-1}$, while for $|x-y| \gg m^{-1}$ it goes as $e^{-m|x-y|}$. When a correlation function falls exponentially with separation, the decay constant, here m^{-1}, is often called the correlation length. Note that if $m = 0$, the correlation length

[9] In relativistic quantum field theory one often refers to the ground state as the vacuum. I use the two terms interchangeably.

[10] Somewhat confusingly, positive frequency means time dependence of the form $e^{-i\omega t}$ with $\omega > 0$.

is infinite and we have power-law behavior all the way out. In this case the theory is sometimes said to be "gapless," since there are excited states with energies arbitrarily close to the ground state energy.

In fact, a massless scalar field enjoys a symmetry larger than just the relativistic Poincaré group; it is invariant under the conformal group. Among other things this bigger symmetry group includes rescalings of spacetime $x'^{\mu} = \lambda x^{\mu}$, so the theory is scale invariant. A quantum field theory with this larger symmetry group is called a conformal field theory or CFT.

Although the correlation function (26) is valid only in free scalar field theory, its basic structure of short-distance power-law divergence and long-distance decay is expected to be true in any relativistic QFT.

We also eventually consider the two-point function when the two fields have different times, but since fields at timelike separation do not necessarily commute we need to be more careful about their ordering. A particularly nice object is the "time-ordered" two-point function $\langle\Omega|T\phi(t,x)\phi(t',y)|\Omega\rangle$, which is defined by ordering the fields in such a way that their time increases as we go to the left. This can be interpreted as a "transition amplitude," where we act on the vacuum with a field ϕ at some time, evolve forward for a while, and then compute the projection onto the state we would have gotten by acting on the vacuum at the later time. For our free massive scalar theory the time-ordered two-point function is

$$\langle\Omega|T\phi(t,x)\phi(t',y)|\Omega\rangle = \frac{1}{4\pi^2}\frac{m}{\sqrt{|x-y|^2-(t-t')^2+i\epsilon}} \times K_1\left(m\sqrt{|x-y|^2-(t-t')^2+i\epsilon}\right), \tag{27}$$

where ϵ is a positive infinitesimal quantity which is there to remind us which branch of the square root we should take when the points are timelike separated. We should always take $\epsilon \to 0$ in any observable.

B. Entanglement in the vacuum

We saw in Sec. III.A that the ground state of a relativistic QFT has nonzero correlation between field operators at spatially separated points.[11] One convenient way to interpret this correlation is as an illustration of the *entanglement* of different regions of spacetime in the vacuum of a relativistic QFT.

For comparison recall the two-qubit state[12]

$$|\Psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle). \tag{28}$$

This state is entangled in the sense that the full state is pure while the reduced state on either qubit is mixed, but we can also illustrate this using correlation functions. Consider the Pauli operators X_i, Y_i, and Z_i acting on the spins. Their one-point functions in the state $|\Psi\rangle$ are all zero, but they have two-point functions

$$\langle\Psi|X_1X_2|\Psi\rangle = 1,$$
$$\langle\Psi|Y_1Y_2|\Psi\rangle = -1,$$
$$\langle\Psi|Z_1Z_2|\Psi\rangle = 1.$$

This is the same qualitative behavior for one- and two-point functions that we saw in QFT, provided that we make the connection that the different qubits correspond to different regions in the QFT and the Pauli operators correspond to fields.

More explicitly we imagine decomposing the Hilbert space of the quantum field theory into a tensor product of the local field degrees of freedom in a region A and its complement B. By studying a two-point function with one field in A and the other in B, we can learn about the entanglement between A and B. In fact we saw that at short distances the correlation functions are divergent, so if we allow our two points to approach each other at the interface between region A and region B, as illustrated in Fig. 6, the correlation becomes infinite. This must mean that in some formal sense there is an infinite amount of entanglement between neighboring regions in the ground state of a relativistic quantum field theory.

Another fairly rigorous illustration of the entangled nature of the vacuum state in relativistic QFT is the Reeh-Schlieder theorem (Streater and Wightman, 2000), which says that for any region A, by acting on the vacuum $|\Omega\rangle$ with operators located in that region, we can produce a set of states which is dense in the full Hilbert space of the QFT. In other words, by acting on the vacuum with some operators localized in this classroom we can create the moon, or the Andromeda galaxy. This is possible because of the highly entangled nature of the vacuum. Were the field theory degrees of freedom in the vicinity of the moon in a product state with the field degrees of freedom here, then no operators we act with here could do anything there.[13]

[11]The presence of interesting correlation in the ground state is a special property of quantum field theories. The ground state of a noninteracting nonrelativistic gas of massive particles just has all the particles sitting on the floor.

[12]Readers unfamiliar with qubit notation should see Sec. II of the Supplemental Material [193]. Also for a more detailed discussion of the definition and meaning of entanglement see Sec. III.

[13]To see the connection with entanglement more clearly, consider the state $|\Psi\rangle = \sum_{ab} C_{ab}|a\rangle|b\rangle$ in a tensor product Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$. For simplicity say that the dimensionalities of $\mathcal{H}_A$ and $\mathcal{H}_B$ are equal. Then if the matrix C_{ab} is invertible, it is easy to see that we can produce any state in the Hilbert space by acting on $|\Psi\rangle$ with an operator which acts nontrivially only on $\mathcal{H}_A$. That C_{ab} is invertible is a statement about entanglement. In quantum field theory both dimensionalities are infinite so one has to work harder to prove the theorem, but this example captures the basic point. Note that a common confusion here is that if we were restricted to only acting with unitary operators on $\mathcal{H}_A$ the theorem would be false, since they would not be able to change the amount of entanglement. It is essential that we are allowed to use nonunitary operators such as projections.

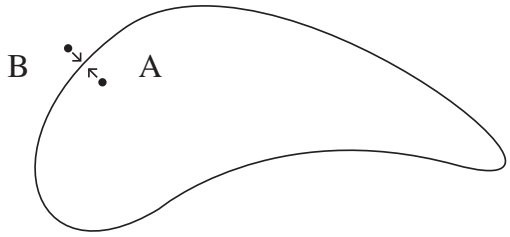


FIG. 6. Decomposing a QFT into a tensor product of the fields in region A and the fields in region B. The two-point function of fields at the indicated points diverges as they approach each other, indicating UV-divergent entanglement between the two factors. The boundary between the two regions is often called the entangling surface.

C. The Rindler decomposition

In Sec. III.B we demonstrated vacuum entanglement in QFT in a rather general sense, but we want to have a more quantitative understanding of who exactly is entangled with whom and by how much. To this end, it is extremely useful to introduce the Rindler decomposition of Minkowski space.

The basic idea is to pick one of the three spatial coordinates, say x, and then divide the Hilbert space of the field theory into a factor $\mathcal{H}_R$ acted on by fields with $x > 0$ and a factor $\mathcal{H}_L$ acted on by fields with $x < 0$. We then want to find a nice basis of states for each factor in which we can decompose the vacuum. For this purpose it is convenient to introduce the Lorentz boost operator K_x that mixes x and t, while acting trivially on y and z. This operator will exist in any relativistic QFT, for example, in our free massive theory it is given by

$$K_x = \frac{1}{2}\int d^3x[x(\dot{\phi}^2 + \vec{\nabla}\phi\cdot\vec{\nabla}\phi + m^2\phi^2) + t\dot{\phi}\partial_x\phi]. \quad (29)$$

This operator appears time dependent, but in the Heisenberg picture the time dependence of the fields cancels the explicit time dependence. The action of the boost operator in the xt plane is shown in Fig. 7. As the figure makes clear, there are four different regions on which the boost operator has a well-defined action. On the right Rindler wedge shown in blue, it evolves forward in time from one black line to another. On the left Rindler wedge also shown in blue, it evolves backward in time along the same lines. In the future and past wedges, shown in green and red, respectively, its action is spacelike. The resemblance of this figure to Fig. 1 is not a coincidence. Indeed a large percentage of the confusions one encounters about black holes is eventually resolved by understanding the Rindler decomposition better.

In order to understand how to express the vacuum state in the basis of boost eigenstates in the left and right Rindler wedges, it is extremely useful to introduce the Euclidean path integral. Recall that, in any quantum system, one way to find the ground state is simply to act on any generic state $|\chi\rangle$ with e^{-HT}, where T is some long time. More precisely, we have

$$\langle\phi|\Omega\rangle = \frac{1}{\langle\Omega|\chi\rangle}\lim_{T\to\infty}\langle\phi|e^{-HT}|\chi\rangle. \quad (30)$$

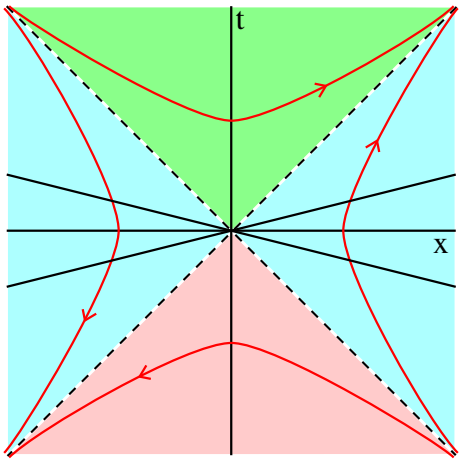


FIG. 7. The Rindler decomposition of Minkowski space. The orbits of the boost operator are shown in red, and slices of Rindler time are shown as black straight lines. The right and left Rindler wedges are shown in blue, the future wedge is shown in green, and the past wedge is shown in red.

In the Euclidean path integral formalism, this means that we can compute this wave functional as[14]

$$\langle\phi|\Omega\rangle \propto \int_{\hat{\phi}(t_E=-\infty)=0}^{\hat{\phi}(t_E=0)=\phi}\mathcal{D}\hat{\phi}e^{-I_E}, \quad (31)$$

where I_E is the Euclidean action, obtained from the usual one by analytic continuation $t \to -it_E$. For the free massive scalar field it is

$$I_E[\hat{\phi}] = \frac{1}{2}\int d^3x dt_E[(\partial_{t_E}\hat{\phi})^2 + (\vec{\nabla}\hat{\phi})^2 + m^2\hat{\phi}^2]. \quad (32)$$

For simplicity I choose the early time boundary conditions to be $\phi = 0$. In the free massive case it is possible to evaluate this path integral explicitly to obtain Eq. (19). In understanding the entanglement of the vacuum, however, it is more convenient to evaluate it in a way that continues to work when interactions are present. The method is shown in Fig. 8. It is based on the observation that the boost operator K_x in the Euclidean plane generates rotations in the xt_E plane, as can be seen from analytically continuing its action on t and x. So instead of evaluating the path integral from $t_E = -\infty$ to 0, we evaluate it along the angular direction over an angle π. From the point of view of the path integral this is just a trivial change of variables, but it makes clear that there is an alternative Hilbert space interpretation of the same path integral

$$\langle\phi_L\phi_R|\Omega\rangle \propto \langle\phi_R|e^{-\pi K_R}\Theta|\phi_L\rangle_L, \quad (33)$$

where K_R is the restriction of K_x to the right Rindler wedge and Θ is an antiunitary operator that exists in all quantum field

[14]In this expression $\hat{\phi}$ is a field history depending on both position and time, while ϕ is the field configuration at $t = 0$ and is thus time independent.

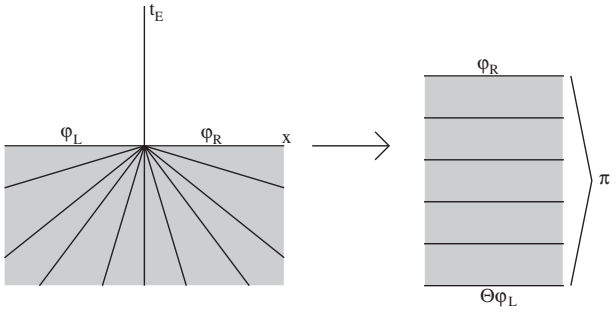


FIG. 8. Evaluating the Euclidean path integral representation of the vacuum wave functional.

theories and is usually called CPT. Its action on a Heisenberg picture scalar field is[15]

$$\Theta^\dagger \Phi(t,x,\vec{y})\Theta = \Phi^\dagger(-t,-x,\vec{y}), \tag{34}$$

although for a real scalar we can ignore the dagger since the field is Hermitian; it gives a natural map between $\mathcal{H}_L$ and $\mathcal{H}_R$. It is needed in Eq. (33) for two reasons: first to allow us to reinterpret the left-hand side of Eq. (33) as a matrix element just in $\mathcal{H}_R$ and second because in the left-hand diagram of Fig. 8 ϕ_L is playing the role of a final state while in the right-hand diagram we need an initial state. We can then evaluate Eq. (33) by inserting a complete set of K_R eigenstates:

$$\begin{aligned}\langle \phi_L \phi_R|\Omega\rangle &\propto \sum_i e^{-\pi\omega_i}\langle i|\Theta|\phi_L\rangle\langle\phi_R|i\rangle_R \\ &\propto \sum_i e^{-\pi\omega_i}\langle\phi_L|i^*\rangle_L\langle\phi_R|i\rangle_R,\end{aligned} \tag{35}$$

where the antiunitarity[16] of Θ is used and

$$|i^*\rangle_L = \Theta^\dagger|i\rangle_R. \tag{36}$$

Thus we arrive at the following simple expression for the vacuum state:

$$|\Omega\rangle = \frac{1}{\sqrt{Z}}\sum_i e^{-\pi\omega_i}|i^*\rangle_L|i\rangle_R. \tag{37}$$

Note that the entanglement between the left and right wedges is now completely manifest. We can also compute the reduced density matrix for the right wedge:

$$\rho_R = \frac{1}{Z}\sum_i e^{-2\pi\omega_i}|i\rangle_R\langle i|. \tag{38}$$

This is nothing other than the thermal density matrix with temperature

$$T = \frac{1}{2\pi}. \tag{39}$$

It may appear mysterious that the temperature (39) is dimensionless; later I will comment on the physical meaning of this.

D. Free fields in Rindler space

To get some more intuition for the Rindler decomposition, we can study it in the free massive scalar theory. The basic idea is to find a set of modes f_n which are better suited to the Rindler decomposition than the usual plane waves, and then expand the field in creation and annihilation operators for them. To accomplish this, it is convenient to introduce new coordinates for the left and right Rindler wedges[17]:

$$\begin{aligned}x &= e^{\xi_R}\cosh\tau_R = -e^{-\xi_L}\cosh\tau_L, \\ t &= e^{\xi_R}\sinh\tau_R = e^{-\xi_L}\sinh\tau_L.\end{aligned} \tag{40}$$

These coordinates have the property that evolving with the boost operator K_x is just a translation of τ_R forward in time and a translation of τ_L backward in time. The $\xi_{L,R}$ coordinates label hyperboloids that are orbits of K_x; they are also trajectories of constant proper acceleration. The coordinate ranges are $-\infty < \xi_{L,R} < \infty$, $-\infty < \tau_{L,R} < \infty$, and they cover the left and right Rindler wedges, respectively, but not the future or past Rindler wedges. The surfaces $\xi_R = -\infty$, $\xi_L = \infty$ are usually called the Rindler horizons, although they are not actually event horizons according to the definition given in Sec. II.A. The metric in these coordinates is

$$\begin{aligned}ds^2 &= e^{2\xi_R}(-d\tau_R^2 + d\xi_R^2) + d\vec{y}^2 \\ &= e^{-2\xi_L}(-d\tau_L^2 + d\xi_L^2) + d\vec{y}^2,\end{aligned} \tag{41}$$

where for convenience y and z are combined into a vector $\vec{y}$. The idea is then to look for solutions of the massive wave equation of the form

$$f_{R/L\omega k} = e^{-i\omega\tau_{R/L}}e^{i\vec{k}\cdot\vec{y}}\psi_{R/Lk\omega}(\xi_{R/L}), \tag{42}$$

where $\omega > 0$ and the wave equation implies that $\psi_{Rk\omega}$ and $\psi_{Lk\omega}$ obey

$$\begin{aligned}[-\partial_{\xi_R}^2 + (m^2+\vec{k}^2)e^{2\xi_R} - \omega^2]\psi_{Rk\omega} = 0, \\ [-\partial_{\xi_L}^2 + (m^2+\vec{k}^2)e^{-2\xi_L} - \omega^2]\psi_{Lk\omega} = 0.\end{aligned} \tag{43}$$

Formally these equations are nothing but the Schrödinger equations for a nonrelativistic particle in an exponential potential. It can be solved explicitly in terms of Bessel functions, but for our purposes it is sufficient simply to observe that the normalizable solutions [in the KG norm (23)]

[15] In even numbers of spacetime dimensions one usually combines this with a rotation of the $\vec{y}$ directions to invert them as well, but this is the definition that works in any dimension.

[16] Remember that for an antilinear operator A the adjoint is defined as $\langle x|A^\dagger y\rangle = \langle y|Ax\rangle$. If A is antiunitary then $\langle Ax|Ay\rangle = \langle y|x\rangle$.

[17] In this transformation I suppressed an arbitrary choice of length scale, which I call ℓ, which is needed to make the units work out. I restore it at the end of this section.

will oscillate at sufficiently negative ξ_R (or sufficiently positive ξ_L), and will decay exponentially at sufficiently positive ξ_R (or sufficiently negative ξ_L). We can thus think of the modes as being "confined" to be near the horizon, with lower energy and/or higher tranverse momentum modes being confined more strongly.

Of course the point of introducing these modes is that they have definite boost energy: ω in the right wedge and $-\omega$ in the left wedge. As such, if we expand the field in terms of them as

$$\phi = \sum_{k,\omega} (f_{R\omega k} a_{R\omega k} + f_{L\omega k} a_{L\omega k} + f^*_{R\omega k} a^\dagger_{R\omega k} + f^*_{L\omega k} a^\dagger_{L\omega k}), \tag{44}$$

then the creation operators $a^\dagger_{L,R\omega k}$ create states of definite boost energy on the Rindler vacuum $|0\rangle$, which is defined as being annihilated by the annihilation operators $a_{L,R,\omega k}$. We can thus rewrite Eq. (37) as a product state over all modes

$$|\Omega\rangle = \bigotimes_{\omega,k} \left[\sqrt{1 - e^{-2\pi\omega}} \sum_n e^{-\pi\omega n} |n\rangle_{L\omega(-k)} |n\rangle_{R\omega k} \right], \tag{45}$$

where n labels the number of particles on top of the Rindler vacuum in each mode. The sign flip for k comes from the CPT conjugation in Eq. (37).[18] As expected, in the reduced density matrix on either side each mode is thermally occupied with temperature (39).

For future reference note that the state (45) is annihilated by the operators

$$\begin{aligned} c_{1\omega k} &\equiv \frac{1}{\sqrt{1 - e^{-2\pi\omega}}} (a_{R\omega k} - e^{-\pi\omega} a^\dagger_{L\omega(-k)}), \\ c_{2\omega k} &\equiv \frac{1}{\sqrt{1 - e^{-2\pi\omega}}} (a_{L\omega k} - e^{-\pi\omega} a^\dagger_{R\omega(-k)}), \end{aligned} \tag{46}$$

so we can think of the vacuum state as the joint zero eigenspace of the "number" operators $c^\dagger_{1\omega k} c_{1\omega k}$ and $c^\dagger_{2\omega k} c_{2\omega k}$. States where these oscillators are excited are excited states with respect to the Minkowski Hamiltonian H.[19]

We now come back to the meaning of the temperature (39). In defining the dimensionless Rindler coordinates $\xi_{R,L}$ and $\tau_{R,L}$, an arbitrary choice of length scale ℓ is suppressed. Had they been defined to have units of length, the temperature would have been $1/2\pi\ell$. But what is the meaning of ℓ? Indeed it is straightforward to see that it is the inverse proper acceleration of an observer at $\xi_R = 0$, who also happens to have τ_R as her local proper time. Energy defined with respect to τ_R is thus the energy that such an observer would define in her locally inertial frame. This strongly suggests that any observer with acceleration $a = 1/\ell$ should actually perceive a temperature

$$T_{\text{Unruh}} = \frac{\hbar a}{2\pi k_B c}, \tag{47}$$

where the dimensionful constants have been restored. This is called the Unruh effect and is one of the more striking results of relativistic QFT. Note that both relativity and quantum mechanics are important; either $c \to \infty$ or $\hbar \to 0$ would send T_{Unruh} to zero. You may wonder if this temperature is actually real; in fact one can build a model of a detector that couples to the scalar field and set it on an accelerating trajectory, and indeed one finds that it clicks just as if it were in the presence of thermal fluctuations at temperature T_{Unruh}. For more details about this, as well as other interesting observations about the physics of the Rindler decomposition, see Unruh and Wald (1984).

E. Entanglement is important for horizon crossing: An introduction to firewalls

Before finally getting back to black holes, I need to make one final point about the Rindler decomposition. So far I have mostly focused on the left and right Rindler wedges, but of course the future and past wedges are also interesting. In particular, the entanglement of the Minkowski vacuum $|\Omega\rangle$ across the entangling surface $x = 0$ is essential for having a smooth transition from either the left or the right Rindler wedges to the future interior. To see this, imagine that instead of the ground state $|\Omega\rangle$, we put the system into the mixed state

$$\rho = \rho_L \otimes \rho_R, \tag{48}$$

where $\rho_{L,R}$ are the thermal density matrices obtained on either side by tracing out the other in the vacuum $|\Omega\rangle$. For any observer who stays in the left or right Rindler wedges, this state is indistinguishable from the vacuum. But in fact it has infinite energy: the Hamiltonian includes a gradient term that is divergent if the field is discontinuous at $x = 0$. More precisely if the left and right wedges are completely uncorrelated, as in the state (48), then the typical difference between neighboring fields on either side is of the order of the typical field fluctuation, which is given by $1/\epsilon$, where ϵ is a UV length cutoff, so we have

$$\partial_x \phi|_{x=0} \propto \frac{1}{\epsilon^2}. \tag{49}$$

The gradient term in the Hamiltonian then produces a contribution

$$dx \int d^2y (\partial_x \phi)^2 \propto \epsilon \int d^2y \frac{1}{\epsilon^4} = \frac{A}{\epsilon^3}, \tag{50}$$

where I have replaced $dx \to \epsilon$. There is thus a large concentration of energy sitting at $x = 0$, waiting to annihilate anybody who tries to jump through the Rindler horizon

[18] More explicitly, Θ sends $\tau_R \to -\tau_L$, $\xi_R \to -\xi_L$, and $\vec{y} \to \vec{y}$. Applying this to the right-hand modes Eq. (42) sends positive-frequency modes to negative-frequency modes, so to get the coefficient of an annihilation operator we need to take the complex conjugate. This flips the sign of $\vec{k}$.

[19] A more technical way of seeing this is to observe that the modes annihilated by the c's have only positive frequency with respect to the Minkowski time t, so they must agree with the usual plane-wave modes about the definition of the ground state.

into the future wedge. This type of thing has recently been given a new name: a *firewall*.[20]

One can study this more concretely using a model detector. I will not describe this in detail here, but one can see without too much difficulty that even fairly mild decorrelation already leads to $O(1)$ probability that the detector clicks as it crosses the horizon. For a related calculation see Sec. III.D of Giddings (2006).

An important point here is that although entanglement is necessary for a smooth infalling experience through the Rindler horizon, it is not sufficient. Heuristically, imagine that the states $(1/\sqrt{2})(|00\rangle \pm |11\rangle)$ both had smooth horizons. By the linearity of quantum mechanics this would mean that the product states $|00\rangle$ and $|11\rangle$ must as well, but we just saw that no product state possibly can have a smooth horizon in QFT. Not only do we need entanglement to get a smooth horizon, it needs to be the right entanglement.[21]

IV. QUANTUM FIELD THEORY IN A BLACK HOLE BACKGROUND

We have now discussed classical black holes and quantum field theory; it is time to combine them. To really analyze the problem correctly, we need to treat both the metric and the matter fields quantum mechanically in a full theory of quantum gravity. I will discuss how this might be done in Sec. VI, but first discuss the much simpler problem of quantum field theory in a fixed black hole background. Physically this is the limit where we send the Newton constant G to zero and the black hole mass M to infinity in such a way that the Schwarzschild radius $r_s = 2GM$ is fixed. Stated in terms of dimensionless quantities, we send $M/m_p \to \infty$, where $m_p \equiv 1/\sqrt{8\pi G}$ is the Planck mass, and then study observables whose length scale is of the order of r_s in this ratio. In this limit there are no gravitational interactions, and the metric can be viewed as a fixed external field.[22]

This limit is quite reasonable from the point of view of astrophysical black holes: for a solar mass black hole we are working to leading order in

$$\frac{m_p}{M} \approx 10^{-38}, \tag{51}$$

which is not a bad thing to base a perturbation theory on. The Schwarzschild radius is of the order of kilometers, which is indeed the type of scale at which we could imagine doing experiments. Of course detecting individual photons whose wavelength is a kilometer is no picnic, but nobody credible ever said life would be easy.

A. Two-sided Schwarzschild and the Rindler decomposition

Before proceeding further, we need to decide which of the two geometries discussed in Sec. II.D to study: the full two-sided Schwarzschild geometry or the one-sided collapse geometry that is only Schwarzschild after the infalling matter has gone in? Both are interesting, but it is easier to begin with the two-sided case to avoid the complications of the infalling shell.

From the Kruskal and Szekeres expression for the metric (10) it is clear that, for $r \approx 1$ and sufficiently small angular displacements, the Schwarzschild geometry resembles the region of Minkowski space that is near the Rindler horizon in the Rindler decomposition. More explicitly, in the right exterior ($r > 1$) and using the tortoise coordinate

$$r_* = r + \log(r-1), \tag{52}$$

we have

$$ds^2 = \frac{r-1}{r}(-dt^2 + dr_*^2) + r^2 d\Omega_2^2. \tag{53}$$

If we define $y_1 \equiv \theta_1$, $y_2 \equiv \theta_2$, where θ_1 and θ_2 are two orthogonal angular coordinates on the sphere, then for $r \approx 1$ we have

$$ds^2 \approx e^{r_*-1}(-dt^2 + dr_*^2) + d\vec{y}^2. \tag{54}$$

This is very reminiscent of the right Rindler wedge metric (41), and indeed if we define

$$\begin{aligned} r_* &= 2\xi_R + 1 - \log 4, \\ t &= 2\tau_R, \end{aligned} \tag{55}$$

they become equivalent. We can extend this argument to the other three parts of the geometry; the upshot is illustrated in Penrose diagrams in Fig. 9.

This observation makes it plausible that, whatever initial state we pick for the quantum field in the Schwarzschild geometry, if it is to locally look like the Minkowski vacuum near the interface of the left and right exteriors, they need to be thermally entangled just as in the Rindler decomposition. We can determine the temperature with respect to Schwarzschild time, the natural time for obervations made at $r \gg 1$, from the relationship (55) between Schwarzschild and Rindler time; apparently

$$T_{\text{Hawking}} = \frac{T_{\text{Unruh}}}{2} = \frac{1}{4\pi r_s}. \tag{56}$$

I temporarily restored the Schwarzschild radius $r_s = 2GM$. In fact for pedagogical purposes we can restore all of the dimensionful constants:

[20]We can also study this firewall by computing the expectation value of the number operator $c^\dagger_{1\omega k}c_{1\omega k}$ for the modes defined by Eq. (46). In the state (48) a short calculation shows that $\langle c^\dagger_{1\omega k}c_{1\omega k}\rangle = 2/(e^{\pi\omega} - e^{-\pi\omega})^2$, so for $\omega \lesssim 1$ the mode will have an $O(1)$ excitation. Since there are many such modes, the state will be quite singular.

[21]One way to think about this is that the divergence of the two-point function of ϕ as we bring the points together is precisely the divergent correlation between the left and right which is necessary to avoid a large expectation value for the gradient. If we correlate the fields in the wrong manner, for example, if we anticorrelate them, then the gradient will still be large.

[22]This is not quite true; free gravitons are still present in this limit, but we can just treat them as another matter field.

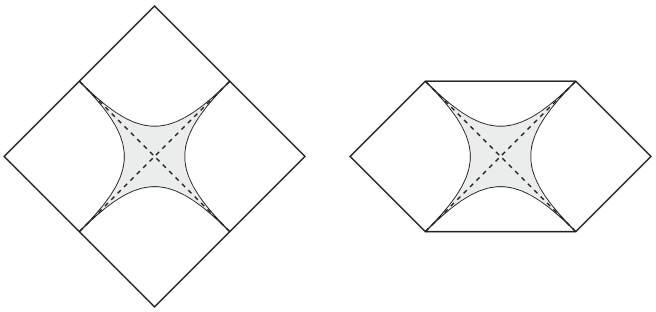


FIG. 9. The shaded region of the Rindler Penrose diagram on the left well approximates the shaded region of the Schwarzschild Penrose diagram on the right. The Rindler Penrose diagram is different from the Minkowski Penrose diagram in Fig. 2, even though they represent the same spacetime. The reason is that at each point I have suppressed an $\mathbb{R}^2$ instead of an $\mathbb{S}^2$.

$$T_{\text{Hawking}} = \frac{\hbar c^3}{8\pi k_B GM}. \tag{57}$$

Note that the temperature decreases as the black hole increases in size; for example, for a solar mass black hole this temperature is of the order of 10^{-7} K, which is fairly cold.

This global pure state for the Schwarzschild geometry, where the two exteriors are thermally entangled as in Rindler space, is called the Hartle-Hawking (HH) state (or the Hartle-Hawking-Israel state) (Hartle and Hawking, 1976; Israel, 1976). We describe it more explicitly in Sec. IV.H.

B. Schwarzschild modes

Now we study free fields in the Schwarzschild geometry. To begin we need to find a simple set of modes f that solve the free scalar equation of motion

$$\frac{1}{\sqrt{-g}}\partial_\mu(\sqrt{-g}g^{\mu\nu}\partial_\nu\phi) = m^2\phi, \tag{58}$$

where $g_{\mu\nu}$ is the Schwarzschild metric and g is its determinant. We can then express the field in terms of these modes as in Eq. (24) and study its properties in an appropriate quantum state such as the Hartle-Hawking state.

We focus on the right exterior of the Schwarzschild geometry, covered by the coordinates (t, r, Ω). We look for solutions of the form

$$f_{\omega\ell m} = \frac{1}{r}Y_{\ell m}(\Omega)e^{-i\omega t}\psi_{\omega\ell}(r). \tag{59}$$

As in the Rindler wedge, we can write an effective Schrodinger equation for $\psi_{\omega\ell}$. It is again convenient to work in the tortoise coordinate, in terms of which we have

$$-\frac{d^2}{dr_*^2}\Psi_{\omega\ell} + V(r)\Psi_{\omega\ell} = \omega^2\Psi_{\omega\ell}, \tag{60}$$

with

$$V(r) = \frac{r-1}{r^3}\left(m^2r^2 + \ell(\ell+1) + \frac{1}{r}\right). \tag{61}$$

In solving this equation we express r implicitly in terms of r_*.

A considerable amount of the physics lies in Eq. (61) for the effective potential. We first consider the mass m; it is quite reasonable to assume for simplicity that the Compton wavelength $1/m$ is either much larger or much smaller than the Schwarzschild radius. We are mostly interested in the case where it is much larger, in which case we can set the mass to zero, but I will briefly comment on the massive case. For $r \gg 1$ the potential goes to a constant m^2, so massive modes will propagate near infinity only if $\omega \geq m$. Since we are taking $m \gg 1$, this means that any modes whose energy ω is of the order of the Schwarzschild radius will be confined very near the horizon, having only exponentially small tails at infinity. Since we already concluded that the temperature of the black hole is of the order of the inverse Schwarzschild radius, we will indeed mostly be interested in modes with $\omega \approx 1$. For this reason massive particles are usually not of much interest in black hole physics; from now on we restrict to the case of $m^2 = 0$.

In the massless case, the asymptotic behavior of the potential is

$$V \approx \begin{cases} \frac{\ell(\ell+1)}{r_*^2} & r_* \to \infty, \\ (\ell^2+\ell+1)e^{r_*-1} & r_* \to -\infty, \end{cases} \tag{62}$$

so it vanishes polynomially in r_* at spatial infinity and exponentially in r_* near the horizon. In between the two regions there is a barrier. For $\ell \gg 1$ the peak of the barrier is at $r = 3/2$ and the height is of the order of ℓ^2. This potential is plotted for the first few ℓ in Fig. 10.

For modes whose energy is less than the height of the barrier, in particular, those with energy of the order of the temperature, we can think of the barrier as dividing the black hole exterior into two regions. For $r \gg 3/2$ the geometry is weakly curved, and the propagating modes are the usual ones of Minkowski space. For $1 < r < 3/2$, there are also propagating modes but they are mostly confined to be near the horizon. This region is sometimes called the "thermal atmosphere," since these modes will typically be occupied by a Boltzmann distribution with temperature $T_{\text{Hawking}} \approx 1$.

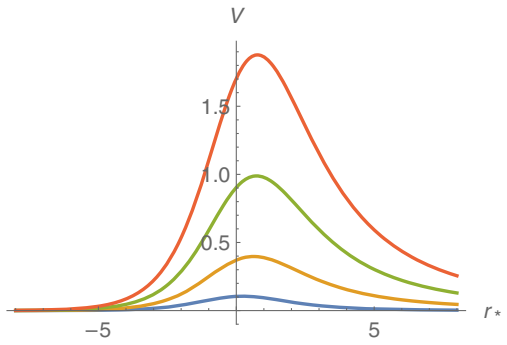


FIG. 10. Plots of V as a function of r_* for $\ell = \{0, 1, 2, 3\}$.

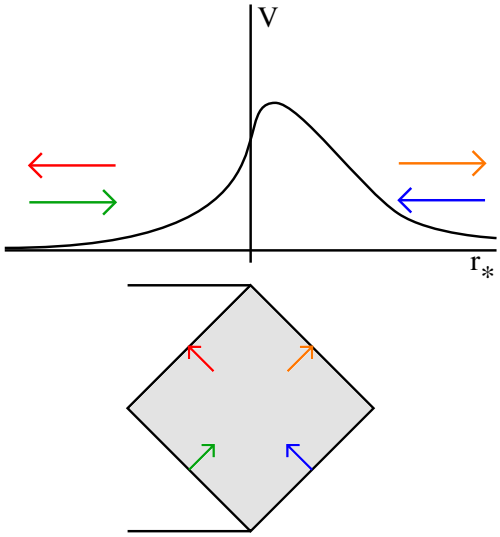


FIG. 11. The spacetime interpretation of scattering in the effective Schrodinger potential. Incoming and outgoing modes are shown above on the potential and below on a section of the Schwarzschild Penrose diagram, with the region $-\infty < r_* < r_*$ shaded in gray.

Recently people have also started to call the near-horizon region "the zone."

From the effective Schrodinger equation point of view, it is clear that the relation between these two regions is a scattering problem. This is illustrated in Fig. 11. Modes can come in either from the white hole horizon at $r_* \to -\infty$ or from J_- at $r_* \to \infty$, and they can go out either through the black hole horizon at $r_* \to -\infty$ or through J_+ at $r_* \to \infty$. To get a unique mode, we need to pick boundary conditions of some sort. One obvious choice is send in particles from the right and then see if they are absorbed by the black hole. The absorption probability is just the transmission coefficient of this Schrodinger problem. This type of mode has most of its support out in the asymptotic $r \gg 1$ region. Another choice is to say that there is no flux coming in from the right, but allow some to come in from the left. This is a flux coming in out of the white hole and mostly being reflected off of the barrier back into the black hole, with a small amount tunneling through the barrier and transmitting out to infinity. Modes of the latter type have most of their support in the near-horizon region and are often called "zone modes" or "modes in the zone."

To understand which modes are important, we need some sort of prescription for the initial quantum state of the field we build out of creation and annihilation operators for them. In the Hartle-Hawking state both are excited with a thermal distribution at temperature T_{Hawking}. There is a steady flux of thermal particles going in and coming out at each end of the wormhole. The one-sided black hole made from collapse is more subtle, and I turn to it now.

C. Hawking's calculation of black hole radiation

We now focus on a black hole created from the collapse of a matter. The Penrose diagram is shown in Fig. 12.

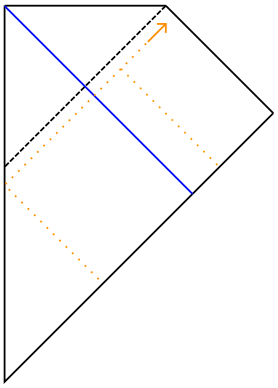


FIG. 12. The geometry for a black hole made from collapse. The collapsing shell is now shown in blue, and the backward evolution of an outgoing wave packet at late times is shown in orange. Part of it is scattered off of the potential barrier, and part of it goes back through the photon shell.

The mode solutions discussed in the previous section now apply only above the shell. Below the shell they must be matched onto solutions of the ordinary Minkowski wave equation such as plane waves. This matching will be nontrivial; since there is no global time-translation symmetry, positive-frequency modes with respect to Schwarzschild time above the shell will evolve back to mixtures of positive and negative frequency modes with respect to the Minkowski time below the shell. In fact the negative frequency part comes entirely from the part of the solution that propagates back through the photon shell. The part that reflects off of the barrier while still in the Schwarzschild region will conserve energy and stay positive frequency.

We then define the quantum state in such a way that other than the shell which created the black hole, at early times the Minkowski modes below the shell are not excited. In other words their annihilation operators will annihilate the state. Because of the mixing between positive and negative frequency just described, the modes which are positive frequency in the Schwarzschild region will be excited in this state. To figure this out one needs to know how to relate the creation and annihilation operators for the Minkowski modes below the shell to the creation and annihilation operators for the Schwarzshild modes above the shell. The details are not hard but are slightly technical; they are explained in Hawking's original paper (Hawking, 1975) [and nicely reviewed in Wald (2010)]. The result is that in this choice of initial state the energy flux in a band of late-time outgoing modes $f_{\omega \ell m}$ with width $d\omega$ is

$$\frac{dE}{dt} = \frac{\omega d\omega}{2\pi} \frac{P_{abs}(\omega, \ell)}{e^{\beta\omega} - 1}. \tag{63}$$

Here $\beta = 1/T_{\text{Hawking}}$, and $P_{abs}(\omega, \ell)$ is the absorption probability for a "blue" mode of this frequency and angular momentum to be transmitted through the barrier from the right in Fig. 11. It is often called a gray-body factor, since other than this factor Eq. (63) is just the standard formula for radiation from a blackbody at temperature $T = \beta^{-1}$ into the

vacuum.[23] We saw earlier that the height of the potential barrier grows like ℓ^2, so the emission is dominated by the modes with the very lowest angular momentum.

We are thus led to the following picture of the quantum state of a field near a black hole formed from collapse: after an initial ringdown period, the modes in the atmosphere are in a quasistationary state where they are excited thermally, with the low-ℓ modes gradually carrying energy out to infinity by tunneling through the barrier.

Any massless field that happens to be around will carry away energy in this manner, but those with higher spin will carry less (Page, 1976). A simple way to understand this is that particles with spin can only radiate into modes with $\ell > 0$ (dipole radiation for photons, quadrupole for gravitons, etc.), so they encounter higher potential barriers. The largest fraction of the energy is thus carried away by whatever the lowest spin massless particles are; in our Universe it would be photons. A smaller, but still $O(1)$ fraction would be radiated into gravitons.

Although I will not use it in what follows, there is a heuristic explanation of Hawking radiation that is occasionally brought up. The idea is that entangled pairs of particles are constantly jumping into existence near the horizon via vacuum fluctuations, and sometimes one of them falls in and one of them gets out. This cartoon has several problems if it is taken literally, among them that the "particles" involved have wavelengths comparable to the size of the black hole and that the Hawking process is not really stochastic, and in my view it tends to create more confusion than it resolves.

D. Evaporation

In the QFT-in-curved-spacetime limit considered so far, the mass of the black hole, and therefore its energy, is infinite. The black hole can radiate away a constant energy flux forever without decreasing in size. This is clearly unphysical, but to fix it we need to restore dynamical gravity. Since the black hole mass will now be time dependent, it is absurd to continue setting the Schwarszchild radius $r_s = 2GM$ to 1.

Before discussing the decrease of the mass due to evaporation, I briefly comment on the fact that we should think of the mass M of a black hole as its energy. So far we somewhat cavalierly defined the energy as the generator of t translations in the Schwarzschild geometry, but in general relativity such a coordinate-dependent definition cannot really be correct. There is a long and interesting story about this which I will not get into, but the upshot is that in a proper Hamiltonian formulation of general relativity in asymptotically flat space the energy is defined as a certain boundary integral on a two-sphere at $r \to \infty$, or more rigorously as a boundary integral on i^0 in the Penrose diagram (Regge and Teitelboim, 1974). It is usually called the ADM energy, in honor of Arnowitt, Deser, and Misner who were the first to write it down and realize its relevance to a Hamiltonian formulation (Arnowitt, Deser, and Misner, 2008). For our purposes though, it is enough just to know that the ADM energy is conserved and that a black hole of mass M formed from the collapse of a shell does have ADM energy M.

We are now in a position to estimate the lifetime of a black hole of mass M. The total energy flux leaving the black hole is

$$\frac{dE}{dt} = \sum_{\ell,m} \int_0^\infty \frac{d\omega}{2\pi} \frac{\omega P_{abs}(\omega,\ell)}{e^{\beta\omega} - 1}. \tag{64}$$

Unfortunately computing P_{abs} involves solving the differential equation (60), which cannot be done analytically. Page (1976) solved it numerically to compute exact lifetimes, but for our purposes a simple estimate is enough. For $\omega \lesssim 1/r_s$ we can approximately solve the equation to find $P_{abs}(\omega, \ell = 0) \sim (\omega r_s)^2$. Higher ℓ modes have P_{abs} proportional to some higher power of (ωr_S) in the same limit, so roughly we can neglect the $\ell > 0$ terms in the sum and use this approximation for $\ell = 0$ to find

$$\frac{dE}{dt} \approx \frac{C}{r_s^2}, \tag{65}$$

where C is some constant that this crude method is unable to compute. Amusingly this is consistent with what a naive application of the Stefan-Boltzmann law $dE/dAdt = \sigma T^4$ would predict, although the constant factor is different. The mass of the black hole as a function of time thus obeys the differential equation

$$\frac{dM}{dt} = -\frac{C}{(GM)^2}, \tag{66}$$

so a black hole of initial mass M will evaporate in time

$$t_{\text{evap}} \sim G^2 M^3. \tag{67}$$

For a solar mass black hole this time is of the order of 10^{66} yr. This can be compared, for example, to the age of the Universe, which is about 13.8×10^9 yr.[24] It seems that if we are ever to do experiments to test this prediction, we had better find a way to make black holes which are smaller than those produced astrophysically.

Hawking evaporation means that our old Penrose diagram of Fig. 5 needs to be revisited; I show the improved diagram in Fig. 13.

E. Entropy and thermodynamics

If a black hole has a temperature and an energy, it must also have an entropy. Recall that the standard definition of temperature in statistical mechanics is

[23] Including this factor it is the standard formula for radiation from an imperfect blackbody that has some probability to absorb incoming modes. It is derived by demanding that the absorption and emission probabilities are related in such a way that the body can be in thermal equilibrium with an external radiation field.

[24] As a side comment, this should more accurately be called the time since the beginning of nucleosynthesis. The actual age of the Universe is unknown and could quite possibly be infinite.

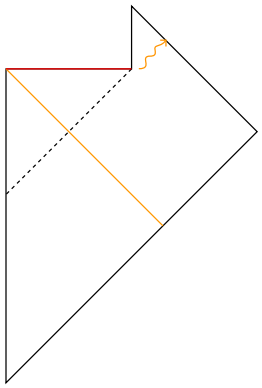


FIG. 13. The Penrose diagram for an evaporating black hole formed by a collapse of a shell of photons. The Hawking radiation seems to all come out at once, but this is only an illusion arising from the conformal transformation. The singularity is shown in red, and the two vertical black lines both represent a nonsingular origin of polar coordinates. The dashed line is the event horizon, which exists even though the black hole evaporates since there are still points that cannot send messages to i^+.

$$\frac{dS}{dE} = \frac{1}{T}. \tag{68}$$

For the black hole we have $T = 1/8\pi GM$, so identifying $M = E$ and assuming that $S(E = 0) = 0$ we find

$$S = \frac{A}{4G} = 2\pi \frac{A}{\ell_p^2}, \tag{69}$$

where $A = 4\pi r_s^2$ is the area of the horizon and $\ell_p = \sqrt{8\pi G}$ is the Planck length. This entropy is enormous, of the order of 10^{78} for a solar mass black hole, which is much larger than the entropy of the Sun, which is of the order of 10^{60}. In fact the entropy of the entire observable Universe excluding black holes, which is dominated by the cosmic microwave background photons, is of the order of $(10^{10}\ \text{pc}/1\ \text{mm})^3 \approx 10^{87}$, while the entropy of a single 10^6 solar mass black hole like that in the center of the Milky Way is of the order of 10^{88}. The largest supermassive black holes have masses of the order of 10^{10} solar masses, and thus entropies of the order of 10^{96}.

Historically it was actually the black hole entropy which was discovered before the temperature. In classical general relativity, one can prove under rather general assumptions that the area of an event horizon can never decrease in time (Hawking, 1971). This property is reminiscent of the second law of thermodynamics, and if we formally define an entropy proportional to the horizon area and a temperature of the order of $1/r_S$, then the first law of thermodynamics $dM = TdS$ is also satisfied (Bardeen, Carter, and Hawking, 1973; Bekenstein, 1973). Bardeen, Carter, and Hawking (1973) viewed this as only a mathematical analogy, but it was the point of view of Hebrew University's own Jacob Bekenstein that this entropy should actually represent the statistical entropy of the black hole in the sense of counting the number of ways the black hole could have formed (Bekenstein, 1973, 1974). Bekenstein argued the entropy should be given by some constant multiple of the horizon area in Planck units and provided evidence for this by considering various thought experiments where an entropic system is thrown into a black hole. He saw that in each case the black hole entropy defined in this way always increased more than the exterior entropy decreased from losing the system. Bekenstein called this observation the *generalized second law* and conjectured that it was true in general.[25] It was at this point that Hawking's paper on the temperature appeared, closing the circle. For these reasons Eq. (69) is usually called the Bekenstein-Hawking entropy.

The idea that the Bekenstein-Hawking entropy counts microstates has found quite strong support in string theory, a set of ideas that has produced many insights into quantum gravity in the last few decades (Polchinski, 1998a, 1998b). General arguments based on counting the states of a long vibrating string are able to produce the area scaling of Eq. (69) in a wide variety of situations (Susskind, 1993; Horowitz and Polchinski, 1997), and in certain special supersymmetric cases (Strominger and Vafa, 1996) one is actually able to sharpen these arguments enough to compute the numerical coefficient analogous to the 1/4 in Eq. (69).

F. The information problem

We have now seen that black holes behave like thermodynamic objects in many respects, and it seems almost irresistible, following Bekenstein, to take the point of view that the Bekenstein-Hawking formula indeed counts the logarithm of the number of microstates of a black hole of a given size. For a moment say that we nonetheless insist that in fact black holes are really only distinguished by their mass (and charge and angular momentum if we had included these), as suggested by general relativity. We then find ourselves in tension with a basic principle of physics: if we know the state of a system at some time, we should be able to infer its initial state by running the dynamics backward. This proposal thus is really saying that by creating a black hole, we destroy most of the information about how we made it. Since we would like to preserve this principle, we are naturally led to assume that indeed black holes have microstates. In one of the most remarkable papers in the last 50 yr however, Hawking brilliantly argued that, once we allow the black hole to evaporate, this assumption is not sufficient to avoid the destruction of information.

Before explaining Hawking's argument, it is worth saying a bit more about the principle of information conservation. In classical mechanics time evolution is generated by Hamiltonian flow on phase space, which can always be reversed by changing the sign of the Hamiltonian. More prosaically we can just solve the equations of motion backward. Similarly in quantum mechanics, time evolution is described as unitary evolution in Hilbert space, which can again be reversed by switching the sign of the Hamiltonian. In both cases the principle applies only if the system is isolated;

[25]The name is a bit misleading. If we take Bekenstein's suggestion seriously that black hole entropy is real, then his "generalized" second law is really just the ordinary second law.

otherwise, information can leak out. The quantum case is slightly counterintuitive since the measurement process is nondeterministic, but measurement always involves coupling the system being measured to some external apparatus. The evolution of the joint system remains unitary and deterministic.

Hawking's point then was that his calculation of black hole radiation described in Sec. IV.C has the property that the outgoing radiation is completely independent of the details of the initial state of photons. This was made more explicit by Wald (1975), where it was emphasized that the emission rate (63) arises from a diagonal density matrix

$$\rho \propto \bigotimes_{\omega,\ell,m} \left(\sum_n |n\rangle\langle n|_{\omega,\ell,m} P_{abs}(\omega,\ell) e^{-\beta\omega n} \right) \tag{70}$$

for (late-time wave packets of) the Schwarzschild modes $f_{\omega,\ell,m}$. This should be reminiscent of our Rindler result (45) that the reduced density matrix for the right (or left) Rindler wedge is just a thermal density matrix, but here it leads to catastrophic consequences once we turn gravity back on.

Consider a black hole that was formed by a shell of matter in some pure quantum state $|\Psi\rangle$. As time goes on, the quantum state of the radiation field outside becomes more and more mixed, which we can quantify by saying that its entanglement entropy is increasing.[26] This may not seem so bad at first, since after all in looking at the late radiation we are looking just at the part of the state which is outside of the black hole horizon. As the black hole evaporates, however, it decreases in size until at some point it becomes Planckian. Equation (70) is supposed to be accurate up to corrections of the order of m_p/M, so until this point the entanglement entropy of the radiation field outside continues to increase. At this point one of two things must happen:

(1) The evaporation stops, and the Planck-sized object just sits around. This possibility is called a *remnant*. For the total state to remain pure, as required if the evolution of the system is to be unitary, the remnant must have an extraordinary amount of entanglement entropy. Even before it becomes Planck sized its entanglement entropy would need to exceed the Bekenstein-Hawking value, which would violate the state-counting interpretation of Eq. (69).

(2) The black hole finishes evaporating into ordinary quanta such as photons and gravitons. Energy conservation prevents the final burst of quanta from containing nearly enough entanglement entropy to purify the earlier radiation, so the end result of the evaporation process is a mixed state of the radiation field whose entropy is of the order of the initial black hole horizon area in Planck units.

Hawking argued for option (2), claiming that the process of black hole formation and evaporation cannot be described by a unitary map from the ingoing shell to the outgoing radiation, since this would have resulted in a pure quantum state of the radiation. Moreover, since different initial states result in the same final state, he claimed that black holes violate the principle of information conservation. Option (2) is thus usually referred to as *information loss*.

Basic physical principles are not often discarded, and this should only be done once it is clear that we have no other choice. How might information loss be avoided? Option (1) is in principle possible, but is rather difficult since it requires objects with finite energy but an infinite number of states and has rarely been taken seriously (Preskill, 1992; Giddings, 1995; Susskind, 1995b). Most people who are unwilling to accept information loss have instead gone for a third option:

(3) Equation (70) is correct only in a coarse-grained sense; the Hawking radiation does not actually come out in a mixed state. The information is carried out in subtle correlations between the Hawking photons, and the final state of the evaporation is a pure state of the radiation field. Because it is a complicated state any small subsystem looks thermal, justifying the approximate validity of Eq. (70) if we do not look at too many photons at once. There is a complete basis of such pure states whose dimensionality is the exponential of the Bekenstein-Hawking formula, which thus can indeed be interpreted as counting microstates.

Option (3) may obviously seem like the best of the three, but it is more radical than it appears at first. Equation (70) seems to follow from very widely held assumptions about the validity of quantum field theory on scales that are large compared to the Planck scale. If it is wrong, then should not this violation of quantum field theory be detectable in other ways? Since all three options thus have unappealing features, this state of affairs is referred to as the black hole information problem.

There is however at least one reason why one might expect substantial corrections to Eq. (70) ('t Hooft, 1985; Unruh, 1994; Corley and Jacobson, 1996). Returning to Fig. 12, consider the limit where we have the outgoing mode coming out later and later. If it comes out more than a time[27]

$$t_{\text{scr}} \equiv r_s \log \frac{r_s}{\ell_p} \tag{71}$$

after the initial shell falls in, then as we evolve it back in time its collision with any of the photons that make up the initial shell happens at a center of mass energy that is greater than the Planck scale. Intuitively this is because near the horizon in tortoise coordinates the proper distance to the horizon behaves like e^{r_*}, so from Eq. (3) we see that wave packets which come out of the potential barrier at times that are later than Eq. (71) after the formation of the black hole will have come out from within a Planckian distance of the horizon. In Hawking's calculation these modes are being "pulled out of the vacuum" near the horizon from an imaginary reservoir of trans-Planckian degrees of freedom; this is called the trans-Planckian problem.

[26] For more details on the basic properties of pure and mixed states, as well as entanglement entropy, see Sec. III of the Supplemental Material [193].

[27] In this name "scr" stands for "scrambling." The reason is that this is the time scale it takes for perturbations of the black hole to die down to Planckian size. It is the time it takes them to be "scrambled" by the black hole horizon. I discuss this further in Sec. V.F.

To see the trans-Planckian collision with the initial shell more explicitly, we can note from Eq. (3) that an ingoing null geodesic that crosses the potential barrier at time $t = 0$ intersects an outgoing geodesic that crosses the potential barrier at time $t_{\text{out}} \gg r_s$ at

$$r_{\text{collision}} \approx 1 + e^{-t_{\text{out}}/2r_s}. \tag{72}$$

To compute the energy of the collision we need the four-momenta of the two photons in Schwarzschild coordinates, which are given by[28]

$$p_{\pm}^{\mu} = E_{\pm}\left(\frac{r}{r - r_s}, \pm 1, 0, 0\right), \tag{73}$$

where $E_{\pm}$ are the energies of the massless particles. The center of mass energy of the collision is then

$$E_{\text{c.m.}} = \sqrt{-\tfrac{1}{2}p_{+}^{\mu}p_{-}^{\nu}g_{\mu\nu}} \approx \sqrt{E_{+}E_{-}}e^{t_{\text{out}}/4r_s}. \tag{74}$$

The outgoing photon will always have $E_{+} \sim 1/r_s$, which is also a lower bound for E_{-} since otherwise the infalling photon would not have fit into the black hole in the first place, so we see that indeed after a time (71) the center of mass collision energy is Planckian.

There has been much debate over whether or not the trans-Planckian problem is a serious criticism of Hawking's argument for information loss. There is a well-known candidate rebuttal called the "nice-slice" argument (Polchinski, 1995). Basically one argues that we can perform Hawking's calculation in Kruskal coordinates where the trans-Planckian origin of the Schwarzschild modes is absorbed into the standard renormalization of quantum field theory in curved spacetime. Although this prescription gives a well-defined procedure that reproduces Hawking's calculation for the late-time state of the radiation, in my view it is not completely satisfactory since it does not get rid of the fact that projecting onto possible final states of the late-time Hawking radiation produces states with a genuine high-energy collision in the past. The renormalization procedure used in the nice-slice argument does involve making an arbitrary choice about physics at high-energy scales, and, unlike in most situations where quantum field theory is used, the redshifting of the black hole geometry allows low-energy properties of the state at later times to depend on this choice.[29] I believe that the trans-Planckian problem gives a plausible excuse for why Hawking's calculation might not be completely correct for times longer than t_{scr}.

You might ask why we do not conclude from this that Eq. (70) is completely wrong for $t > t_{\text{scr}}$, for example, the black hole could explode at $t = t_{\text{scr}}$? This is actually ruled out experimentally; the black hole in the center of our galaxy would have already exploded, but we could imagine something more mild.[30] Anything along these lines besides option (3) however would mean that we should not really think of the black hole as a complex thermal system with entropy S_{BH}. The apparent successes of black hole thermodynamics would be a mirage. Aesthetically this is rather unappealing, since so far all cases of systems that behave thermodynamically ultimately have their explanations in statistical mechanics. This would also be in tension with the string theory microstate counting arguments mentioned in Sec. IV.F, and we will see powerful evidence in favor of unitary evaporation with approximate thermality from the AdS/CFT correspondence in Sec. VI. For now I will thus provisionally adopt the point of view that option (3) is correct.

In the final two parts of this section I discuss two important aspects of QFT in black hole backgrounds which do not quite fit into the main flow but are in my opinion too important to leave out. Casual or first-time readers may wish to skip these final two sections and proceed to Sec. V.

G. The brick wall model and the stretched horizon

To get some intuition for black hole thermodynamics, it is interesting to study the thermodynamics of a scalar quantum field in a black hole background.[31] We have already seen that the region outside the horizon can be understood in terms of the modes $f_{\omega\ell m}$, and Hawking's analysis says that all of these modes are excited thermally. We then roughly have the total energy

$$E = \sum_{\omega\ell m} \frac{\omega}{e^{\beta\omega} - 1} \tag{75}$$

and entropy

$$S = \sum_{\omega\ell m}\left[\frac{\beta\omega}{e^{\beta\omega} - 1} - \log\left(1 - e^{-\beta\omega}\right)\right], \tag{76}$$

where β is the inverse temperature. These expressions are not really well defined; it is not clear what is meant by $\sum_{\omega}$. It cannot just be $\int d\omega$, since this would not have the right units. This ambiguity reflects something physical; these quantities are both UV and IR divergent. The IR divergence arises from the infinite volume of flat space in the region where $r \to \infty$, and the UV divergence arises from the near-horizon region. The former can be regulated by putting the black hole in a large box, while the latter are presumably regulated by some sort of Planckian physics close to the horizon.

[28]This is determined by finding an appropriate affine parametrization of the null geodesics in Eq. (3), as discussed in Sec. I of the Supplemental Material [193].

[29]It is sometimes argued that the adiabatic theorem of quantum mechanics prevents us from making other choices that would lead to different results for the Hawking radiation (Polchinski, 1995), but the adiabatic theorem applies only to the global conserved energy, not to the center of mass energy of localized excitations. Projections on the late Hawking radiation will not appreciably change the energy, since any localized excitations that are created this way are redshifted by the horizon.

[30]For another observational argument, the trans-Planckian problem also exists in the early univere during inflation, but nothing too terrible seems to have come out of it.

[31]In this section I again set $r_s = 1$ to simplify formulas.

As a simple model, 't Hooft suggested implementing the near-horizon cutoff by simply putting Dirichlet boundary conditions $\phi = 0$ on a surface a Planckian distance from the horizon ('t Hooft, 1985). More explicitly one demands the field vanishes at $r = r_{\min}$, which we can express in terms of the proper distance ϵ from the horizon as

$$r_{\min} \equiv 1 + \frac{\epsilon^2}{4}. \tag{77}$$

This is called the "brick wall" model. It is not a good model from the point of view of the full black hole geometry; we do not think the infalling observer will actually encounter a brick wall. The actual cutoff imposed by quantum gravity is undoubtably more subtle. Nonetheless the brick wall does give a physically motivated way to discretize the sum over ω, allowing estimates for Eqs. (75) and (76).

The IR box is a little awkward to deal with explicitly, but it can be dispensed by instead including only the modes which have no incoming piece outside of the barrier; these are the zone modes of Sec. IV.B. By doing this we are throwing out the contributions to the thermal energy and entropy of the radiation field in the region $r \gg 1$, but these should not be considered as part of the black hole anyway. The quantization of ω is also messy to derive in detail, but there is a simple way to estimate it. Returning for a moment to tortoise coordinates, the brick wall is at

$$r_{*\min} \approx 2\log\frac{\epsilon}{2}. \tag{78}$$

I focus on modes with $\ell \gg 1$ and $\omega \lesssim 1$, since these will dominate the thermodynamic ensemble.[32] The turning point where the potential barrier becomes important and the mode begins to decay exponentially is at

$$r_{*\text{turn}} \approx 2\log\frac{\omega}{\ell}. \tag{79}$$

We can thus approximate this mode problem as the Schrodinger problem of a particle in a box of size

$$\Delta r_* \approx 2\log\frac{2\omega}{\ell\epsilon}, \tag{80}$$

where I neglected various order 1 factors. We then have a quantization condition

$$\omega_n \approx \frac{\pi n}{2\log(2\omega_n/\ell\epsilon)}, \tag{81}$$

which allows the replacement

$$\sum_{\omega\ell m} f(\omega) \approx 4\int_0^\infty \frac{d\omega}{2\pi} f(\omega) \int_0^{\frac{2\omega}{\epsilon}} d\ell(2\ell+1)\log\frac{2\omega}{\ell\epsilon}$$
$$\approx \frac{8}{\epsilon^2}\int_0^\infty \frac{\omega^2 d\omega}{2\pi} f(\omega). \tag{82}$$

We can finally then compute the energy and entropy

$$E \approx \frac{r_s}{960\pi\epsilon^2},$$
$$S \approx \frac{r_s^2}{180\epsilon^2}, \tag{83}$$

where I have set $\beta = 1/T_{\text{Hawking}} = 4\pi r_s$ and restored r_s. These results are manifestly divergent as we remove the short-distance cutoff ϵ, but on physical grounds we should probably not take ϵ to be any smaller than the Planck length $\ell_p = \sqrt{8\pi G}$. Indeed if we choose $180\pi\epsilon^2 = G \sim \ell_p^2$, we then have

$$E = \frac{3}{8}M,$$
$$S = \frac{A}{4G}. \tag{84}$$

With this choice the entropy of the field is the full Bekenstein-Hawking entropy S_{BH} of the black hole, and its energy is comparable to the full black hole energy M. Taking $\epsilon \gtrsim \ell_p$ is thus essential in ensuring that the fields do not carry more energy and entropy than the black hole itself.

In fact in this model we essentially replaced the black hole with the field theory degrees of freedom near the horizon, and there is nothing left for the black hole to do. Of course this choice of ϵ was rather arbitrary, and by making it larger we can imagine a separation of degrees of freedom into QFT degrees of freedom at distances greater than ϵ from the horizon and "quantum gravity" degrees of freedom closer in. In this more general model, one imagines a dynamical membrane a Planck distance away from the horizon, which is usually called the *stretched horizon*. The stretched horizon then carries most of the black hole entropy and energy, and in a unitary theory it absorbs infalling matter and reemits it later in scrambled form. In this picture the stretched horizon is in thermal equilibrium with the QFT modes in the atmosphere, and evaporation happens because these modes occasionally tunnel out to infinity.

Clearly any such model cannot be taken too seriously; a real brick wall or stretched horizon would be detectable by an infalling observer who crosses the horizon. This may seem like a trivial comment, but as seen in Sec. VII, harmonizing a unitary description of black hole evaporation with a smooth experience for an infalling observer is proving more difficult than many people expected. In any event these models make it clear that there is a somewhat arbitrary distinction between which degrees of freedom are counted as being "part of the black hole" and which are "part of the atmosphere."

[32] Modes with larger ω are Boltzmann suppressed since we will take $\beta = 1/T_{\text{Hawking}} \sim 1$, and modes with low ℓ are entropically suppressed.

H. The Euclidean black hole

I now return to the full two-sided Schwarzschild geometry of Fig. 1.[33] In the Rindler decomposition of Minkowski space, we found the ground state by evaluating the Euclidean path integral (31). We want to do the same for quantum fields in the Schwarzschild geometry, but to do this we need to identify an appropriate Euclidean version of the geometry to evaluate the path integral on. This geometry should be a spherically symmetric solution of the Euclidean Einstein equation $R_{\mu\nu} = 0$ with asympotically flat boundary conditions. Since we already know that the Schwarzschild geometry is the unique such geometry in Lorentzian signature, it is natural to find its Euclidean version by simple analytic continuation $t \to t_E$:

$$ds^2 = \frac{r-1}{r} dt_E^2 + \frac{r}{r-1} dr^2 + r^2 d\Omega_2^2. \tag{85}$$

This geometry still appears to be singular at $r = 0$ and $r = 1$, but here a surprise is in order. If we define a new coordinate

$$d\rho = \sqrt{\frac{r}{r-1}} dr, \tag{86}$$

then near the horizon ($\rho \to 0$) we have

$$\begin{aligned} r &\approx 1 + \frac{\rho^2}{4}, \\ ds^2 &\approx d\rho^2 + \frac{1}{4}\rho^2 dt_E^2 + d\Omega_2^2. \end{aligned} \tag{87}$$

The first two terms look very much like the origin of polar coordinates in $\mathbb{R}^2$, and in fact if we were to decide to have t_E be an angular variable with period 4π, then the apparent singularity at $\rho = 0$ would be resolved by the geometry capping off smoothly. Remarkably, this means that the curvature singularity at $r = 0$ has been completely excised, and we are left with an entirely nonsingular geometry. This is illustrated in the upper part of Fig. 14.

By cutting this Euclidean geometry in half and evaluating the Euclidean path integral over the lower half, we can define the Hartle-Hawking wave functional for the Lorentzian Schwarzschild geometry (Hartle and Hawking, 1976; Israel, 1976). By the same argument we used for the Rindler decomposition, this allows a simple explicit expression for the quantum state of the fields:

$$|\Psi\rangle_{\mathrm{HH}} \propto \sum_i e^{-\beta E_i/2} |i^*\rangle_L |i\rangle_R. \tag{88}$$

Here i labels eigenstates of the Schwarzschild Hamiltonian in the left and right exteriors and $*$ indicates CPT conjugation.[34] As in the Rindler case, this derivation does not assume that the quantum fields are free.

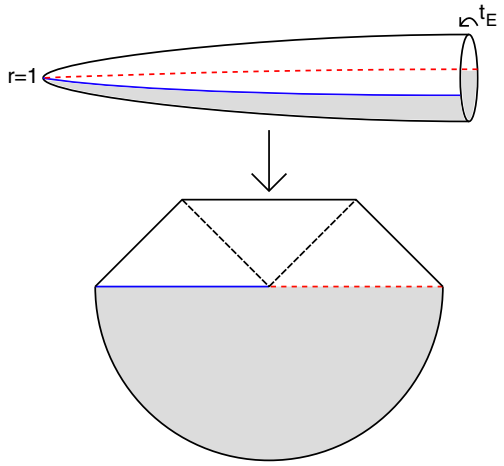


FIG. 14. The Hartle-Hawking construction. The upper diagram is the Euclidean Schwarzschild "cigar," with the two-sphere suppressed. The wave functional of the quantum fields in the Lorentzian Schwarzschild geometry is constructed by evaluating the Euclidean path integral on the lower half of the cigar and then using this as initial data, as shown in the lower diagram.

Tracing out one of the two sides we immediately see that in the Hartle-Hawking state the reduced density matrix on either side is the thermal density matrix, with temperature

$$T_{\mathrm{Hawking}} = \frac{1}{4\pi} \tag{89}$$

as expected. Indeed this type of state was suggested outside of the context of black holes as a method for doing thermal field theory calculations (Takahashi and Umezawa, 1996), and in this broader context it is usually called the *thermofield double state*.

This derivation of the black hole temperature is much more compact than that presented in Sec. IV.C, and one might expect that a similar derivation is possible for the entropy. This is indeed the case (Gibbons and Hawking, 1977), but the analysis is more subtle and involves some assumptions. It also involves more advanced geometry. The basic idea is to interpret the path integral over the full Euclidean Schwarzschild geometry as a partition function. In ordinary field theory this is the standard observation that

$$Z(\beta) \equiv \mathrm{Tr} e^{-\beta H} = \int \mathcal{D}\phi \int_{\hat{\phi}(t_E=0)=\phi}^{\hat{\phi}(t_E=\beta)=\phi} \mathcal{D}\hat{\phi} e^{-I_E}, \tag{90}$$

or in other words that the partition function can be computed from a Euclidean path integral that is periodic in Euclidean time. Once we have computed the partition function it is straightforward to compute the energy and entropy via the standard formulas

$$\begin{aligned} E &= -\frac{Z'}{Z}, \\ S &= \beta E + \log Z. \end{aligned} \tag{91}$$

We want to compute the black hole entropy along similar lines, but it is clear that in order to get the Bekenstein-Hawking entropy $S = A/4G$ it is insufficient to integrate

[33] In this section evaporation is unimportant, so I continue to use units where $r_s = 1$.

[34] CPT here acts as $\Theta^\dagger \Phi_L(t, r, \Omega)\Theta = \Phi_R^\dagger(-t, r, \Omega)$. In the free case the modes which are entangled are related by $m \to -m$.

only over a single scalar field. There is nowhere for the G to come from. This is not a surprise; in computing the partition function we must take into account all dynamical fields, including the metric. Once we start integrating over metrics, it is not immediately clear what family of geometries to integrate over. Roughly we want them to be asymptotically flat with periodic Euclidean time. To make this condition precise it is convenient to introduce an explicit boundary in the spacetime by cutting off the geometry at some finite two-sphere radius r_c. We then integrate over compact geometries whose only boundary has topology $\mathbb{S}^1 \times \mathbb{S}^2$ and induced metric

$$ds^2 = dt_E^2 + r_c^2 d\Omega_2^2. \tag{92}$$

Here the periodicity of t_E is taken to be β. Two such geometries are the piece of the Euclidean Schwarzschild geometry with $r \leq r_c$, which I call g_{sch}, and ordinary Euclidean $\mathbb{S}^1 \times \mathbb{R}^3$ in cylindrical coordinates

$$ds^2 = dt_E^2 + dr^2 + r^2 d\Omega_2^2, \tag{93}$$

again with $r \leq r_c$ and t_E periodicity β, which I call g_{flat}.

It so happens that the leading order behavior in M/m_p of the partition function is not affected by matter fields, so we want to compute the path integral

$$Z(\beta) = \int \mathcal{D}g e^{-I_E} \tag{94}$$

over the set of Euclidean compact geometries with boundary metric (92). The Euclidean action is

$$I_E = -\frac{1}{16\pi G}\int_{\mathcal{M}} d^4x\sqrt{g}R - \frac{1}{8\pi G}\int_{\partial\mathcal{M}} d^3x\sqrt{\gamma}K, \tag{95}$$

where $\mathcal{M}$ denotes some compact manifold with a boundary $\partial\mathcal{M}$. The topology of this manifold is not fixed; we should sum over different topologies. In particular, g_{sch} has the topology of a disk times a two-sphere while g_{flat} has the topology of a circle times a three-dimensional ball. $\gamma_{\mu\nu}$ is the induced metric on the boundary, and K is the trace $\gamma^{\mu\nu}K_{\mu\nu}$ of the extrinsic curvature tensor $K_{\mu\nu} = \nabla_\mu n_\nu \equiv \partial_\mu n_\nu - \Gamma^\alpha{}_{\nu\mu} n_\alpha$, where n_ν is the outward-pointing normal vector at the boundary. This extra term is necessary to include in the gravitational action when a boundary is present if we want to fix the induced geometry on the boundary (Gibbons and Hawking, 1977). To leading order in the m_p/M expansion we can compute this path integral by saddle-point techniques:

$$Z[\beta] \approx \sum_{g_{\text{cl}}} e^{-I_E[g_{\text{cl}}]}, \tag{96}$$

where g_{cl} correspond to geometries which solve the classical equations of motion. Here there are two contributions: g_{sch} and g_{flat}, and their Euclidean actions are

$$\begin{aligned} I_E[g_{\text{flat}}] &= -\frac{\beta r_c}{G} \\ I_E[g_{\text{sch}}] &= I_E[g_{\text{flat}}] + \frac{\beta^2}{16\pi G}. \end{aligned} \tag{97}$$

The dominant contribution to Eq. (96) thus comes from g_{flat}. This is not surprising. We already know that black holes in asymptotically flat space evaporate, so it must be that a gas of radiation in flat space dominates the thermal ensemble. If we are nonetheless interested in understanding the subleading contribution of the black hole to the ensemble, it is natural to look at the contribution to the partition function from g_{sch}. Even in the black hole geometry however there is a large contribution from thermal excitations of gravitons far away from the black hole, which we do not want to include as part of the black hole. We can remove both of these effects by including only g_{sch} in the sum over solutions and subtracting from its action the action of g_{flat}, so we find that the partition function of the black hole only is (Gibbons and Hawking, 1977)[35]

$$Z_{\text{BH}}(\beta) \approx e^{-\beta^2/16\pi G}. \tag{98}$$

From Eq. (91) we then have

$$\begin{aligned} E &= M, \\ S &= \frac{A}{4G}, \end{aligned} \tag{99}$$

as expected.

It is instructive to compare what happened here to the brick wall model of the previous section. We could have also included a scalar field here, whose saddle point would have been $\phi = 0$, and its one-loop determinant would have produced a UV-divergent contribution to the partition function that would match the brick wall model result. In the Euclidean formalism however we interpret this contribution (and a similar one from gravitons) as a renormalization of G, which combines with the "bare" contribution from the gravitational action in such a way that the entropy becomes $A/4G$ with the renormalized G (Susskind and Uglum, 1994; Demers, Lafrance, and Myers, 1995). This is a nontrivial statement since the renormalization of G in a given cutoff scheme can also be computed by using Feynman diagrams for gravitational scattering (Demers, Lafrance, and Myers, 1995). It guarantees that the Bekenstein-Hawking formula for the entropy is independent of the arbitrariness of how we divide up the black hole and the atmosphere. Although it is somewhat mysterious, the Euclidean gravity path integral is apparently a quite powerful method for extracting the thermodynamic properties of gravitational systems.

V. UNITARY EVAPORATION

In this section I explore some consequences of the assumption that black hole evaporation is unitary. In this

[35]This argument is admittedly rather vague. We see in Sec VI that for black holes in anti–de Sitter space an analogous argument can be justified more rigorously.

section I work in Planckian units, where $8\pi G = 1$. Up to the order of 1 constants, this means that we can collect the results of the previous section as

$$T \sim \frac{1}{M}, \tag{100}$$

$$S \sim M^2, \tag{101}$$

$$t_{\text{evap}} \sim M^3. \tag{102}$$

Starting in this section I begin to use more techniques from quantum information theory, so one unfamiliar with these techniques should consult the Supplemental Material [193] as needed.

A. The S matrix

So far we have mostly been working in the limit of quantum field theory in curved spacetime, with the gravitational coupling G taken to zero compared to any other scale in the problem. In this limit we are able to make precise sense out of local ideas such as a field operator at a particular point in the spacetime. Operationally this is possible because in this limit one can imagine arbitrarily precise rods and clocks, which can be used to determine the background as accurately as one would like. Once we allow nonzero G, however, we immediately run into the issue that any apparatus we might like to use will necessarily backreact on the geometry. This is a consequence of the "universal" nature of gravitational interactions. It does not seem to be possible to invent matter that does not couple to gravity.[36] It thus seems likely that in an exact theory of quantum gravity we are going to need some formulation in which local field operators are emergent and approximate notions.[37]

In asymptotically flat spacetimes, such as those relevant for the formation and evaporation of black holes we have considered so far, there is a very natural candidate for an exact quantity to study in place of the correlation functions of local operators that one naturally studies in quantum field theory. This quantity is the S matrix. It is defined as a linear map from initial states on $i_- \cup J_-$ to final states on $i_+ \cup J_+$ with the property that

$$P(\chi|\psi) = |\langle\chi|S|\psi\rangle|^2. \tag{103}$$

In other words the probability of finding an "out" state $|\chi\rangle$ given an "in" state $|\psi\rangle$ is given by the absolute value squared of the matrix elements of the S matrix. Formally we can think of the S matrix as

$$S = e^{-i\infty H}, \tag{104}$$

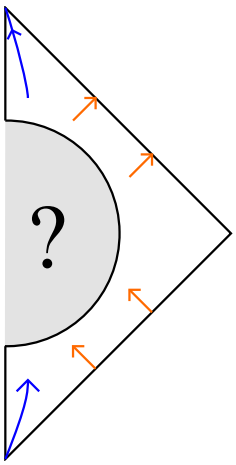


FIG. 15. The S matrix. We send in massless (orange) particles and massive particles (blue) from past infinity; they interact in some complicated way, and then more particles come back out at future infinity. The S matrix encodes the transition probabilities of this process via Eq. (103).

although this expression needs to be taken with a grain of salt. It is not hard to see that the S matrix must be a unitary map for Eq. (103) to make sense, which is consistent with Eq. (104).[38] I illustrate the basic idea of the S matrix in Fig. 15.

Before the S matrix can really have any nontrivial content, we of course need to specify the meanings of its labels. In other words, are there natural bases for the spaces of in and out states that have simple physical interpretations? Here we are in luck; as we approach infinity any incoming or outgoing particles which are around become more and more widely separated, so gravitational (and any other) interactions between them become irrelevant and we can really make sense out of them as exact quantum states. So in fact we expect the Hilbert space of states at past or future infinity to simply have the structure of a free quantum field theory. States are labeled by how many particles there are of such-and-such type and such-and-such momentum or spin, making sure to include the appropriate boson or fermion statistics.[39]

That the unitary S matrix should be an exact observable of asymptotically flat quantum gravity has been confirmed in the one real example we know so far of such a theory (in a number of dimensions large enough to have black holes): the Banks-Fischler-Shenker-Susskind (BFSS) matrix model (Banks *et al.*, 1997). Among other things this model provided the first example of a unitary theory which is expected to have black holes. I will not discuss it in any detail because AdS/CFT gives a broader and easier set of examples.

[36]This is actually not quite true; there exist things called topological field theories where one indeed has fields that do not couple to the metric. The simplest example is the Chern-Simons theory in $2+1$ dimensions. This only seems to be possible however if none of the fields which are present interact with the metric. One cannot have fields which do not interact with the metric interacting with fields that do. The reason is that any such interaction would induce interaction with the metric even if it was not there before.

[37]I discuss this more in Sec. VI.A.

[38]In Hawking's attempts to study information loss he introduced the idea of a "$" matrix, which is a linear map from density operators to density operators. Today this kind of thing is called a superoperator or a quantum channel, and it often appears in discussions of noise and communications.

[39]It is clear that this is a subtle claim to make precise. Even in quantum field theory one has to make sure that one picks the right set of asymptotic particles, and the choice is not always obvious.

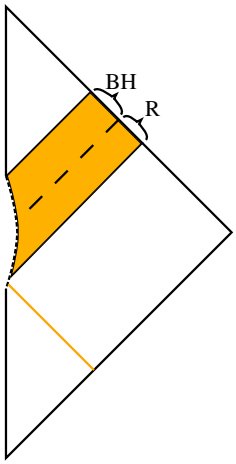


FIG. 16. The S matrix for a black hole (BH) that is made from an initial orange shell of photons and then decays into a Hawking cloud. The horizon is shown as the short-dashed line; it forms and then dissipates. It is often convenient to split the Hawking radiation into an early part R and a late part BH, represented by the long-dashed line. The names are meant to suggest that at some time the photons in R were already out in the radiation while the photons in BH had yet to be emitted from the black hole.

B. The Page curve

For our purposes we are interested in a particular type of scattering process, one where we create a black hole from infalling matter and then watch it evaporate. This is illustrated in Fig. 16.

In this setting we want to get a sense of how the quantum state of the black hole and its radiation evolves with time. One fairly rigorous way to do this is by making a split of the Hawking radiation into "late" and "early," as shown in Fig. 16. This gives a tensor product decomposition of the out state of the Hawking radiation into a bipartite system

$$\mathcal{H}_{\text{out}} = \mathcal{H}_R \otimes \mathcal{H}_{BH}, \tag{105}$$

and we can study how the reduced density matrix on R or on BH depends on when we do the decomposition.

One thing that is particularly interesting to compute is the entanglement entropy S_R as a function of time. The plot of this function is called the "Page curve," in honor of its inventor[40] (Page, 1993b, 2013). Now we think about what to expect. At the beginning of the experiment the black hole is in a pure state, so the radiation field is trivial and has $S_R = 0$. As the black hole begins to radiate S_R will start increasing. At some point however it must turn over and come back to zero, since once all the radiation is out it must again be in a pure state. I show a plot of one possibility in Fig. 17.

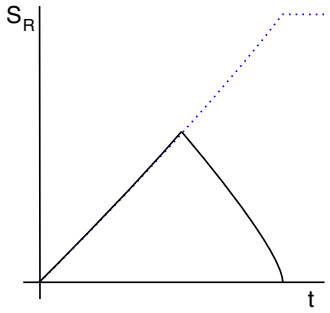


FIG. 17. Page's suggestion for a Page curve, in black. S_R increases as the black hole evaporates until a time of the order of $t_{\text{evap}}/2$, at which point we have $S_R \approx S_0/2$, where S_0 is the initial coarse-grained entropy of the black hole. It then decreases back to zero as the black hole evaporates. A more detailed analysis (Page, 2013) based on the arguments in Sec. V.C puts the turnover at $t = 0.54 t_{\text{evap}}$ and at entropy $S_R = 0.6 S_0$. For comparison I show Hawking's proposal as the dotted blue line, where S_R continues to increase until at the end of the evaporation it reaches the full initial coarse-grained entropy, violating unitarity.

In fact, Andy Strominger has argued that being able to compute the Page curve in some particular theory is what it means to have solved the black hole information problem. Even in AdS/CFT or the BFSS model we are far from being able to really do this. Nonetheless there are some fairly compelling arguments, perhaps not surprisingly due to Don Page, about what its basic form should be. I will now explain these arguments, but first we need a bit of technology from quantum information theory.

C. Page's theorem

Consider a bipartite system

$$\mathcal{H}_{\mathcal{AB}} = \mathcal{H}_A \otimes \mathcal{H}_B. \tag{106}$$

Without loss of generality I will take $|A| \leq |B|$, whereby $|\cdot|$ I mean the dimensionality of the indicated system. We say that the system is maximally entangled if the state ρ_{AB} is pure but the state ρ_A obtained from it by partial trace is proportional to the identity operator on $\mathcal{H}_A$.

Page's theorem (Page, 1993a) then says that a randomly chosen pure state in $\mathcal{H}_{AB}$ is likely to be very close to maximally entangled as long as $|A|/|B| \ll 1$. In order to state the theorem precisely, I need to describe first how to choose a random pure state and second how to quantify what "close" means.

Choosing a random pure state is accomplished by acting on any particular state $|\psi_0\rangle$ with a random unitary matrix:

$$|\psi(U)\rangle \equiv U|\psi_0\rangle. \tag{107}$$

U is chosen from the group-invariant Haar measure. We can compute the reduced density matrix $\rho_A(U)$ by tracing out B.

[40] In fact, because of the UV divergences of quantum field theory, what one really plots is the renormalized entanglement entropy $S_R^{\{\text{ren}\}} \equiv S_R - S_R^{\{\text{vac}\}}$, where $S_R^{\{\text{vac}\}}$ is the entanglement entropy of the vacuum. For notational simplicity I will ignore this subtlety in what follows, but if you are worried about it an excellent laboratory for convincing yourself it is ok is the "moving mirror" model of black hole evaporation (Holzhey, Larsen, and Wilczek, 1994). In this model the Page curve is computable analytically, and one can see that it has the basic features suggested here.

Closeness of states is defined using the operator trace norm, also called the L_1 norm, which is defined for any operator M as

$$\|M\|_1 \equiv \mathrm{tr}\sqrt{M^\dagger M}. \tag{108}$$

The trace norm distance between two density matrices ρ and σ is then defined as $\|\rho-\sigma\|_1$.[41] We are also interested in the L_2 norm, defined as

$$\|M\|_2 \equiv \sqrt{\mathrm{tr}M^\dagger M}. \tag{109}$$

It is not too hard to show that these obey

$$\|M\|_2 \le \|M\|_1 \le \sqrt{N}\|M\|_2, \tag{110}$$

for any operator M, where N is the dimensionality of the Hilbert space.

We are now in a position to state (a version of) Page's theorem.

Theorem: For any bipartite Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$, we have

$$\int dU \|\rho_A(U) - \frac{I_A}{|A|}\|_1 \le \sqrt{\frac{|A|^2-1}{|A||B|+1}}. \tag{111}$$

For intuition we can slightly weaken the bound by writing $\sqrt{|A|/|B|}$ on the right-hand side. Page's theorem then says that once $|B|$ is significantly larger than $|A|$, the typical deviation of ρ_A from the maximally mixed state is extremely small. For example, say that the two systems are sets of qubits; if B has 10 more qubits than A, then the typical deviation from maximal entanglement is bounded by 2^{-5}.

The proof goes as follows:

$$\left(\int dU \|\rho_A(U) - \frac{I_A}{|A|}\|_1\right)^2 \le \int dU \left(\|\rho_A(U) - \frac{I_A}{|A|}\|_1\right)^2 \le |A| \int dU \left(\|\rho_A(U) - \frac{I_A}{|A|}\|_2\right)^2, \tag{112}$$

with the first inequality following from Jensen's inequality and the second following from Eq. (110). We can then evaluate the integral over U exactly, using unitary matrix technology which is developed in Sec. IV of the Supplemental Material [193], and take the square root of both sides to find the right-hand side of Eq. (111). Note that this version of the theorem does not have any assumptions about either $|A|$ or $|B|$ being large, although clearly to get $|A|/|B| \ll 1$ we want $|B| \gg 1$.

Page actually stated his theorem in terms of the entanglement entropy S_A instead of the trace norm. This version is derived by a similar argument: First defining

$$\Delta\rho_A \equiv \rho_A - \frac{I_A}{|A|}, \tag{113}$$

we have

$$\begin{aligned}\int dU S_A &= -\int dU \mathrm{Tr}\rho_A \log\rho_A \\ &= \mathrm{Tr}\left[\left(\frac{I_A}{|A|}+\Delta\rho_A\right)\right. \\ &\quad \left.\times\left(\log|A| - |A|\Delta\rho_A + \frac{1}{2}|A|^2\Delta\rho_A^2 + \cdots\right)\right] \\ &= \log|A| - \frac{|A|}{2}\int dU \mathrm{Tr}\Delta\rho_A^2 + \cdots \\ &= \log|A| - \frac{1}{2}\frac{|A|}{|B|} + \cdots,\end{aligned} \tag{114}$$

where $\cdots$ indicate terms that are smaller in the limit $|A|$, $|B| \gg 1$. Abstractly I prefer the trace norm version, both because the trace norm is a better measure of the distance between states and because no limit is necessary, but the entropy version is also useful.

Indeed we can now use Page's theorem to justify the proposed form of the Page curve shown in Fig. 17. The idea is that black hole evaporation is such a complex process that it is plausible to assume that the pure state of R and BH together is random, up to the basic constraints imposed by energy conservation and causality. At early times the coarse-grained entropy $S_R^{\{\mathrm{coarse}\}} = \log|R|$ of the Hawking radiation is small compared to the coarse-grained entropy $S_{\mathrm{BH}}^{\{\mathrm{coarse}\}} = \log|BH|$ of the black hole. More explicitly, most of the Hawking radiation is emitted into photons in the lowest ℓ modes, which we can approximately think of as a $(1+1)$-dimensional photon gas. Its coarse-grained entropy at early times is then

$$S_R^{\{\mathrm{coarse}\}} \propto tT, \tag{115}$$

where t is the time since the black hole began evaporating and $T \sim 1/M$ is the temperature of the black hole. The coarse-grained entropy of the black hole is just the Bekenstein-Hawking entropy, which is of the order of M^2, so we see that for $t \ll M^3$ we have $S_R^{\{\mathrm{coarse}\}} \ll S_{BH}^{\{\mathrm{coarse}\}}$. By Page's theorem we then have[42]

[41] The motivation for this definition is as follows: say that $\|\rho-\sigma\|_1 < \epsilon$. Then for any projection operator Π we also have $\mathrm{tr}[\Pi(\rho-\sigma)] < \epsilon$, so the probability for any experimental outcome differs between ρ and σ by at most ϵ. Also beware that it is somewhat common to define a trace distance $D(\rho,\sigma) \equiv (1/2)\|\rho-\sigma\|_1$, but this factor of 2 would make our lives more difficult.

[42] A subtlety in applying Page's theorem here is that the subspace of states of fixed energy in $\mathcal{H}_{BH} \otimes \mathcal{H}_R$ does not have a tensor product form, so to respect energy conservation we should not really average over all states in the product Hilbert space. In the limit of weak interactions between A and B one can modify the proof of Page's theorem to deal with this. The basic steps are the same but the notation is a bit heavier since we need to ensure that the random unitary acts only within the subspace of fixed total energy. As you might expect, the result is that the reduced density matrix for the smaller system A is very close to the *thermal* density matrix $(1/Z_A)e^{-\beta H_A}$, with the inverse temperature β chosen so that A has the right expectation value for its remaining energy. The question of how weak the interactions need to be is a subtle one, but a good rule of thumb is that the unperturbed thermal expectation value of the perturbation to the Hamiltonian should be small compared to the unperturbed expectation value of the unperturbed Hamiltonian.

$$S_R \approx S_R^{\{\text{coarse}\}}, \qquad t \ll M^3, \tag{116}$$

so for a while the Page curve grows linearly in t. Eventually we have $S_R^{\{\text{coarse}\}} \approx S_{BH}^{\{\text{coarse}\}}$, which is defined to happen at the "Page time" t_{Page}. At this point S_R is some order one fraction of the original coarse-grained entropy S_0 of the black hole. After the Page time we can apply Page's theorem in the other direction, so we expect to have $S_{BH} \approx S_{BH}^{\{\text{coarse}\}}$. Since the total state is pure we must have $S_{BH} = S_R$ at all times (this follows from the Schmidt decomposition described in Sec. III of the Supplemental Material [193]), so we now have

$$S_R \approx S_{BH}^{\{\text{coarse}\}} \propto S_0\left(1 - \frac{t}{t_{\text{evap}}}\right)^{2/3}, \qquad t_{\text{page}} < t < t_{\text{evap}}. \tag{117}$$

Thus by using Page's theorem we reproduced the qualitative features of Fig. 17. By being more careful about details of the evaporation, such as gray-body factors and the number and helicities of the available massless particles, one can work out more of the quantitative details about exactly when the curve turns over and at what value of the entropy (Page, 2013). The results for a four-dimensional Schwarzschild black hole radiating into photons and gravitons are quoted in the caption of Fig. 17.

Intuitively we can think of what is going on as follows: at the beginning of the evaporation process the radiation that comes out is entangled with the remaining black hole. But eventually it must start coming out entangled with the earlier radiation, since eventually the final state of the radiation must be pure. It is only once we are past the Page time that we can think of the quantum information about the initial state as having started to come out.

D. How hard is it to test unitarity?

I now consider a different question; say that we are convinced theoretically that black hole evaporation is unitary. Is there a good way to test this experimentally? There are a host of practical difficulties that would need to be dealt with. First we would need to figure out how to reliably make black holes in a laboratory setting. They will need to be considerably smaller than the black holes that are produced astrophysically, for example, a black hole that evaporates in a year has a mass of the order of 10^9 kg and a radius of the order of 10^{-18} m. Second we would need to be able to directly manipulate individual photons (and possibly gravitons) in the Hawking radiation. The energies of these photons and gravitons for this black hole would be of the order of 100 GeV. Most worrying, the entropy is of the order of 10^{34}, so we have 10^{34} particles to deal with. We can improve this by making the black hole even smaller, but that comes with its own technical challenges. These obstacles are worrisome to say the least, but even if we were able to surmount them there is still a purely quantum mechanics question of how we should go about verifying the unitarity of black hole evaporation, say for a black hole whose entropy is only of the order of 10^{10}.

We thus imagine that we are successfully able to create a black hole and capture the full quantum state of the Hawking radiation, which we can then transfer to the memory of a quantum computer.[43] We could of course determine the quantum state explicitly by sampling from an ensemble of exponentially many (in the entropy) identically prepared black holes, but this is rather impractical.[44] If we grant ourselves the ability to simulate the black hole formation and evaporation process on a quantum computer in polynomial time,[45] then a simpler option is to transfer the final state of the Hawking radiation to a quantum computer and then apply the time reverse of this simulation. What emerges should then be the initial state of the black hole. We could easily check this, for example, by entangling the first few qubits in the collapse with some reference system and then checking if the output of the computation still possesses that entanglement. A downside of this method is that it requires us to know the S matrix, but an upside is that we can therefore check if we got it right.

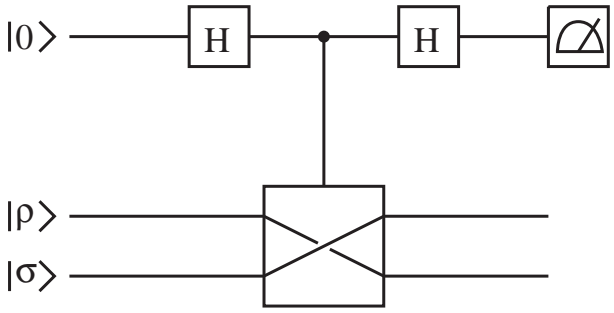


FIG. 18. The swap test. This circuit measures the expectation value of the swap operator in the state $\rho \otimes \sigma$. The circuit works by starting with an extra qubit in the state $|0\rangle$, applying the Hadamard transformation that exchanges the Z and X eigenbases, applying a controlled swap which exchanges the two lower states only if the upper qubit is $|1\rangle$, reapplying the Hadamard, and then measuring the Z operator for extra qubit. One can convince themself (homework) that the expectation value for this last measurement is $\text{Tr}\rho\sigma$.

Alternatively Hayden and Preskill (2007) suggested an experiment that does not require a time-reversed simulation. The idea is to prepare two identical black holes and then capture their radiation in the memory of a quantum computer. We then pass these two states through the quantum algorithm shown in Fig. 18. This algorithm measures the unitary "swap" operator that exchanges the two systems, so it returns ± 1 with expectation value $\text{Tr}\rho\sigma$, where ρ and σ are the (possibly mixed) quantum states of the two Hawking clouds. This expectation value is close to zero unless the two states are both

[43] Those unfamiliar with quantum computation and quantum circuits should see Sec. V of the Supplemental Material for a brief introduction [193].

[44] People with a particle-physics background sometimes talk about "measuring high-point correlation functions" to test unitarity; that is another version of this brute-force algorithm.

[45] This is probably possible for the quantum gravity theories we know about such as the BFSS matrix model or AdS/CFT (Feynman, 1982; Lloyd *et al.*, 1996; Jordan, Lee, and Preskill, 2011), and if you believe in what Scott Aaronson and Stephen Jordan called in their lectures the "extended Church-Turing-Deutsch hypothesis" then it must be possible.

equal and pure.[46] This is thus like trying to distinguish a fair coin from a weighted one that almost gives heads. An $O(1)$ number of trials is sufficient to determine if the states are pure and equal to high probability, which would confirm the unitarity of the evaporation. Somewhat miraculously this is possible without knowing either the final state of the Hawking radiation or the dynamics of the black hole. This protocol is called the *swap test* (Buhrman *et al.*, 2001).

These algorithms inconveniently require us to really capture the full state of the evaporation. As soon as we start making mistakes, for example, because it is rather difficult to detect gravitons, they do not work. In particular, if we try to use the swap test to compare subsystems of the two states then the states would be mixed and as just argued the test would fail. To deal with missing (or corrupted) radiation we need to do some kind of quantum error correction. A brief introduction to quantum error correction and its computational complexity is given in Sec. IV of Harlow and Hayden (2013). The upshot is that quantum error correction is possible provided that we do not lose too much of the radiation. We can lose or corrupt up to almost half if we know which part is affected, while if we do not know we can lose or corrupt up to a quarter. Recalling our observation in Sec. IV.C that black holes radiate less into higher-spin particles, losing gravitons is thus not really a principled obstruction: even if we do not measure any of them we will be able to correct for the loss. Unfortunately it seems likely that this error correction procedure will take exponential time in the number of affected qubits, so if an $O(1)$ fraction of the radiation is lost then the correction procedure will probably take a time which is exponential in the black hole entropy. To deal with this exponential we thus need to take the entropy to be at most $O(10-20)$, which corresponds to a black hole whose radius is only slightly larger than the Planck scale.

At present it thus seems that testing the unitarity of Hawking radiation experimentally will be outside of our technological capability for the foreseeable future, even though there seems to be no in principle obstruction preventing it. Does this mean we should not worry about it? Not necessarily; remember that thought experiments about trains moving close to the speed of light were very helpful to Einstein in understanding relativity, even though we still do not have trains that move at almost the speed of light. The black hole information problem is a litmus test for theories of quantum gravity. Any consistent theory of quantum gravity must either give us an answer or explain why the question does not make sense, and once we have found such a theory it may be experimentally testable in other ways which today are unimaginable.

[46]One way to see this is to observe that $\mathrm{Tr}\rho\sigma$ defines an inner product on the space of Hermitian operators. We can then apply the Cauchy-Schwarz inequality to see that

$$(\mathrm{Tr}\rho\sigma)^2 \leq \mathrm{Tr}(\rho^2)\mathrm{Tr}(\sigma^2). \tag{118}$$

The right-hand side will be small unless both ρ and σ are close to being projection operators or in other words are close to being pure. Moreover the inequality is saturated only if they are equal.

E. What is a typical microstate?

So far I have been discussing black holes made out of a fairly rapid collapse of a matter or photon shell. But what fraction of the total number of microstates can we make this way? We can estimate this as follows: a $(3+1)$-dimensional photon gas at temperature T and in volume V has energy and entropy

$$E \sim VT^4, \qquad S \sim VT^3, \tag{119}$$

so if we imagine forming the black hole by compressing a gas of energy $E = M$ into a volume $V \approx r_s^3 \sim M^3$, then the entropy is

$$S \sim M^{3/2} \sim A^{3/4}, \tag{120}$$

where A is the area of the horizon. Comparing this to the full black hole entropy $S \sim A$, we see that we can make only a small fraction of the microstates this way. How do we make the rest? In statistical mechanics there is a standard way of answering this question. All known laws of physics are CPT invariant, so a typical microstate must have coarse-grained behavior which is time symmetric. In other words, to make a black hole in a typical member of its microcanonical ensemble of dimensionality $e^{2\pi A}$, we must slowly build it up over a time of the order of M^3 by sending in low-ℓ photons at the Hawking temperature in such a way that no radiation comes out until we are done. This is an entropy-decreasing process; it looks like the time reverse of the usual Hawking evaporation. Once we finish building the typical black hole, it will evaporate in the usual manner and the whole process will look time symmetric.

The geometry of the interior in a typical microstate is somewhat mysterious. The Penrose diagram of Fig. 13 is not time symmetric, and if we try to make it so then we invariably end up drawing a past singularity as well as a future singularity. Do these two singularities meet in the middle? Is there a piece of smooth geometry between them? Or could it be the case that there is no global geometry describing the interior and exterior of a typical state? We will see in Sec. VII that there are indeed some reasons to suspect that there may not be a smooth interior for typical states, although it is too early to reach a definite conclusion.

F. Scrambling and recovery of quantum information

Say that we throw a quantum diary into a black hole. How long do we have to wait before its contents comes out in the Hawking radiation? This question was studied by Hayden and Preskill (2007); I here give a sketch of their arguments. The problem can be broken into two parts:

(1) How long does it take the black hole to absorb the information?
(2) Once the information has been absorbed, how much radiation needs to come out before we can recover it?

Before answering these questions, I first say a little about what it means to recover quantum information.

Say we have a quantum system in a state

$$|\psi\rangle = \sum_i C_i |i\rangle, \tag{121}$$

where $|i\rangle$ is some complete basis for a Hilbert space of dimensionality 2^n. For example, it could be a a state of a system of n qubits. Fully describing this quantum state requires specifying 2^{n-1} complex numbers (the C_i's modulo normalization), but this is not what is usually meant by the "quantum information" contained in the state. Classically an n-bit string is specified by only n bits of information, and this should be true for quantum information as well. The exponential comes from trying to classically describe n qubits. If I want to give a quantum state to you, it would be extremely inefficient for me to write down all of the C_i's, send them to you in the mail, and then have you prepare a system in the state $|\psi\rangle$. I should really just send you the quantum state itself. In other words quantum information is carried by qubits, not bits. When we talk about sending or recovering quantum information, this is what we always mean: we have some set of physical operations which at the end of the day enable the transportation of an arbitrary quantum state $|\psi\rangle$ of some number of qubits from one place to another, without anybody having to measure it.[47]

The ability to send and receive quantum information is significantly hampered by the "no-cloning" theorem of quantum mechanics (Dieks, 1982; Wootters and Zurek, 1982). This theorem says that it is impossible to find a system C such that, after adjoining it to an arbitrary quantum state $|\psi\rangle$ times an empty register $|0\rangle$ of the same dimensionality, we have time evolution

$$|\psi\rangle|0\rangle|\phi\rangle_C \to |\psi\rangle|\psi\rangle|\phi'\rangle_C. \tag{122}$$

The proof follows immediately by contradiction when we try to use this evolution and the linearity of quantum mechanics to clone the state $(1/\sqrt{2})(|\psi\rangle + |\chi\rangle)$ for some $|\chi\rangle$ orthogonal to $|\psi\rangle$. This means that in some sense we can think of quantum information as being conserved: I can send a quantum state to you only if I lose it myself.[48]

There is a convenient method for determining whether or not a particular protocol successfully transfers quantum information from one place to another. Say that we have a procedure which for any $|\psi\rangle$ implements

$$|\psi\rangle_A|0\rangle_B \to |0\rangle_A|\psi\rangle_B. \tag{123}$$

This protocol transfers quantum information from system A to system B. Now introduce an additional auxiliary system C, of the same dimensionality as A and B, and maximally entangle it with A. By linearity we then have the evolution

$$\frac{1}{\sqrt{|A|}}\sum_i |i\rangle_A|0\rangle_B|i\rangle_C \to |0\rangle_A \frac{1}{\sqrt{|A|}}\sum_i |i\rangle_B|i\rangle_C. \tag{124}$$

The evolution thus transfers the *purification* of C from A to B.[49] More generally we might imagine an evolution

$$|\psi\rangle_A|0\rangle_B \to |0\rangle_A U_B|\psi\rangle_B, \tag{125}$$

where U_B is some unitary transformation on B. This is typically still counted as successfully transferring the quantum information, since after all the receiver can get back to the previous evolution by acting with $U_B^\dagger$. The evolution once we maximally entangle A with C is then

$$\frac{1}{\sqrt{|A|}}\sum_i |i\rangle_A|0\rangle_B|i\rangle_C \to |0\rangle_A \frac{1}{\sqrt{|A|}}\sum_i U_B|i\rangle_B|i\rangle_C. \tag{126}$$

In either case we can theoretically test if the transfer was successful by comparing the final states ρ_{AC} and $\rho_A \otimes \rho_C$ in trace norm. They will be close to equal if and only if the purification has been successfully switched.

We now apply this discussion to the quantum diary falling into a black hole.[50] Following Hayden and Preskill we first assume that the black hole absorbs the diary instantly, neglecting question (1) above. We can model this as follows. We begin with a diary D maximally entangled with a reference system S. We then throw the diary into a black hole B, which we model by acting on the joint BD system with a random unitary U. We can also include the radiation process as part of this unitary, so we reinterpret the BD system after U has acted as a tensor product of some Hawking radiation R and a remaining black hole B'. The question is then how large R has to be before we can recover the quantum diary, or in other words how long we have to wait before

$$\|\rho_{SB'} - \rho_S \otimes \rho_{B'}\|_1 \ll 1. \tag{127}$$

If black hole B started in a pure state, then we can think of the diary as being an extra piece of the infalling matter that created

[47] An important point here is that the state does not have to be carried by the same physical qubits at the end as it was in the beginning. For example, quantum teleportation (Bennett *et al.*, 1993) is a famous protocol by which we can send an arbitrary state of n qubits from one place to another by exchanging only n classical bits of information.

[48] You might object that I can easily prepare multiple copies of a quantum state for which I know the C_i's. This is true, but missing the point. In sending classical information it is not necessary to know the message to send it; for example, it could be in a sealed envelope or be encrypted. What we want is a single procedure that works for a single copy of any state of the qubits. Measuring the state to determine the C_i's and then sending them does not count, since it requires many initial copies of the same state.

[49] See Sec. III of the Supplemental Material [193] for more discussion on purification.

[50] We can phrase the following discussion more rigorously in terms of decomposing the states at past and future null infinity, as done in Sec. V.B, but I will not attempt it. The new subtlety is that, in arguing that the matrix U that appears below is unitary and does not act on E or S, we need to use the *cluster decomposition property* of the S matrix. This property is a crude form of locality which says that the S matrix approximately factorizes for widely separated systems. It is definitely true in quantum field theory and is usually expected to be true even in quantum gravity; see, for example, Fitzpatrick, Kaplan, and Walters (2014) for a recent discussion in AdS/CFT. To justify this discussion, we need to study this more quantitatively to make sure it works.

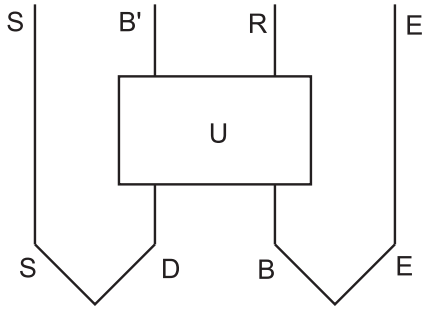


FIG. 19. The Hayden-Preskill experiment. We throw a diary D entangled with a reference system S into a black hole B that is already maximally entangled with its early Hawking radiation E. The diary is absorbed by the black hole, which then partially evaporates into a remaining black hole B' and some more radiation R. The question is to find when the purification of S is transferred to ER.

the black hole. From Page's theorem we then already know the answer for how long we have to wait until information gets out. Until the Page time the Hawking radiation is maximally mixed and carries no quantum information.

What Hayden and Preskill realized however is that the situation is more interesting if we wait until after the Page time to throw in the diary. In this situation the black hole is already maximally entangled with its early radiation E; see Fig. 19. Using the same random unitary technology used above in proving Page's theorem, one can then show that (Hayden and Preskill, 2007)

$$\int dU \|\rho_{SB'} - \rho_S \otimes \rho_{B'}\|_1 \leq \sqrt{\frac{(|D|^2-1)(|B'|^2-1)}{|D|^2|E|^2-1}} \approx \frac{|D|}{|R|}, \tag{128}$$

where in the approximation I have used that $|B'||R| = |E||D|$ and assumed all systems are large enough we can ignore the 1's. Intuitively this result says the following: say that the number of bits radiated after throwing in the diary is c bits more than were contained in the diary itself. Then the right-hand side is 2^{-c}, which rapidly becomes extremely small. So the information essentially comes out as fast as it possibly could. This result led Hayden and Preskill to describe "old" black holes as "information mirrors." Beware however that, as in Eq. (125), a unitary transformation on ER will in general be necessary to put the quantum state of the diary back into usable form. This unitary transformation might be quite difficult to perform, and indeed as we discuss later this is probably the case (Harlow and Hayden, 2013).

Finally we return to the question of how long it takes the black hole to absorb the diary and reemit it, or in other words how long it takes U to act. A first guess is to treat the diary as a massless particle sent inward from some fixed radius r_0, and then ask how much Schwarzschild time goes by before it reaches the stretched horizon at $r - r_s \approx \ell_p^2/r_s$ (recall that the stretched horizon is the surface a Planckian proper distance outside the actual horizon). From Eq. (3) for the trajectory of a radial null geodesic, we see that the time is

$$\Delta t \approx t_{\text{scr}} \equiv r_s \log \frac{r_s}{\ell_p}, \tag{129}$$

which is the same time scale encountered previously around Eq. (71) in our discussion of the trans-Planckian problem. Here we are interpreting it as the time for the diary to appear to have been completely thermalized from the point of view of an outside observer. Using the same calculation as done below Eq. (71) any signal sent from the diary at that point would need to have super-Planckian energy to be distinguishable from the thermal atmosphere. The time scale (129) is also the time it takes for geometric perturbations of the black hole to ring down to Planckian amplitude (Price and Thorne, 1986).[51] For these reasons the time scale t_{scr} is usually called the *scrambling time*.

In fact, scrambling is a technical notion in quantum information theory. We say that a piece of quantum information is scrambled into a system if it cannot be recovered from any subfactor of the system that is smaller than some order one fraction of the whole. Hayden and Preskill's result (128) shows that a random unitary accomplishes this, and indeed were the transformation U not to have this property it would not necessarily be the case that we could recover the diary from RE since there might be some information left in B. It is not immediately clear that the evolution of a black hole for a time of the order of t_{scr} is really "sufficiently random" to scramble the system in this technical sense, but Hayden and Preskill provided plausible evidence for this from the theory of quantum circuits.[52] What they pointed out is that a level of scrambling which is sufficient for Eq. (128) to hold can be produced by much smaller quantum circuits than the exponential-sized ones which would be needed to produce Haar-typical U's. For a system of n qubits there exist families of quantum circuits called ϵ-approximate unitary two-designs, with the property that Eq. (128) holds to within accuracy ϵ if we replace the average over all unitaries by an average just over one of these families (Dankert *et al.*, 2009). Moreover these circuits have a depth which scales like $O(\log n \log 1/\epsilon)$. Hayden and Preskill then modeled the black hole horizon as a set of S qubits positioned at the stretched horizon. They further imagined that any two of the qubits can interact pairwise via two-qubit gates, and that each layer of the circuit requires one Planck unit of proper time to execute. A time step of the order of ℓ_p near the horizon is redshifted to a time step of the order of r_s in Schwarzschild time, so the total execution time for an ϵ-approximate unitary two-design on these qubits will be of the order of $r_s \log S$, which is consistent with Eq. (129).[53]

The scrambling time (129) is quite short compared to the evaporation time. Note that the Planck scale appears only inside of the logarithm, so even for a solar mass black hole the

[51]The reason for this is straightforward; the scrambling time is the time it takes to function $r_s e^{-t/r_s}$ to become of the order of ℓ_p.

[52]See Sec. V of the Supplemental Material [193] for an introduction to quantum circuits.

[53]Actually Hayden and Preskill did not quite get this right, since they imposed locality on the qubit interactions and then had to cancel it by having the circuit run faster. This was corrected by Sekino and Susskind (2008).

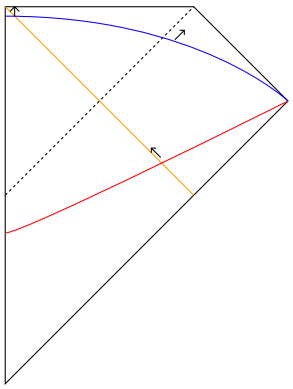


FIG. 20. Black hole cloning. If we believe that Hawking evaporation is unitary, then the quantum information on the red slice about how the black hole was made is apparently present at two places on the blue slice.

scrambling time is only about 10^{-4} s. It is also quite fast compared to what is expected for more conventional systems with comparable entropy, and in the aftermath of the Hayden and Preskill paper there was quite a bit of activity aimed at understanding what sorts of physical systems might realize this "fast scrambling" (Barbon and Magan, 2011, 2012; Lashkari *et al.*, 2013; Sekino and Susskind, 2008). More recently Shenker and Stanford were able to independently derive the scrambling time (129) for certain "large" black holes by using the AdS/CFT correspondence (Shenker and Stanford, 2013, 2014). I describe this in more detail in Sec. VI.J.

G. Black hole complementarity

So far in this section we have only studied the physics of the S matrix. The interior region of the black hole has disappeared from view. This was not an accident; essentially everything we really know about the quantum mechanics of black holes rests on the precise formalism of the S matrix (or its AdS equivalent that I introduce in the next section). Unfortunately there does not seem to be any simple way to relate the experience of an infalling observer who crosses the horizon to the properties of the S matrix. What then are we to do? It seems clear to me at least that we cannot really claim to understand black holes until we understand the experience of the infalling observer.

The black hole interior has been an extremely hot subject in the last year or two, which I will return to in the final two sections of this article, but as a teaser I will now discuss what up until recently was the best-known attempt to grapple with the implications of unitary evaporation for the experience of an infalling observer. This is an analysis by Susskind and Thorlacius (1994) of the possibility of seeing quantum cloning by jumping into a black hole.

The setup is shown in Fig. 20. In order to describe what is seen by the infalling observer in a quantum mechanical manner, we apparently need to introduce some type of degrees of freedom that live behind the horizon. How these degrees of freedom relate to those that describe the in and out states at past and future infinity is of course an important question that any theory of the interior needs to address, but for now I will try not to do this. The strongest assumption we could make is that there is a valid low-energy field theory description of the quantum state throughout the red and then the blue spatial slices in the figure. These types of slices are sometimes called nice slices, since they seem to be the kind of thing you would want to draw to get a complete Hamiltonian description of the spacetime including the interior.

What Susskind and Thorlacius emphasized, following a suggestion of Preskill, is that if we believe the evaporation to be unitary, this description of the slices cannot be consistent with quantum mechanics. The reason is that the quantum information on the red slice that is in the infalling shell apparently is cloned on the blue slice. We could throw in arbitrary quantum states and they would both stay inside in the shell and reappear in the Hawking radiation. This type of evolution violates the quantum no-cloning theorem described in the previous section. It is inconsistent with the linearity of quantum mechanics. Both pieces of quantum information seem to be real. Somebody could jump in riding the shell and find the state up in the left corner or somebody could stay outside and see the information come out in the Hawking radiation. What Susskind and Thorlacius pointed out [and which was later refined by Hayden and Preskill (2007)] is that no *single* observer can see both copies. In the diagram this is essentially obvious by causality, since anybody who waits around long enough (i.e., after a Page time) to get out the quantum information will then jump in far too late to be able to see the infalling shell on this slice. We could try to avoid this by bending the blue slice back down toward the horizon, but this causes an enormous redshift of the infalling shell from the point of view of the person jumping in late, which prevents her from seeing the second copy. This argument was significantly improved by Hayden and Preskill, who pointed out that a much more stringent test arises if the infalling observer waits until a Page time has passed and then throws in a diary as discussed previously. We saw that in this case the information now comes out after a time which is only of the order of the scrambling time $M \log M$, which is much shorter. Hayden and Preskill nonetheless showed that the observer is still just barely not able to see any quantum cloning.

You might be tempted to say that a contradiction is a contradiction, regardless of whether somebody sees it or not. And indeed this contradiction certainly does show that imagining that there is unitary quantum mechanical evolution on these nice slices is inconsistent with having the evaporation process be unitary. But Susskind and Thorlacius instead interpreted the inability to see the potential cloning as an argument against using these slices in the first place. If nobody can see the quantum state on the whole slice, why should it exist? They argued that quantum mechanics only needs to describe the experiences of individual observers, who are appropriately restricted by causality in what they can do. They called this idea *black hole complementarity* (Susskind, Thorlacius, and Uglum, 1993; Susskind and Thorlacius, 1994; Kiem, Verlinde, and Verlinde, 1995; Lowe *et al.*, 1995).

If this seems like it is getting too philosophical to you, I am sympathetic. After all nobody ever managed to really implement black hole complementarity into an actual theory, so its consistency (or indeed its precise definition) was never clear. But it is worth saying that this type of argument does have at

least one very successful historical analog: the Heisenberg uncertainty principle. Heisenberg's realization that it is operationally impossible to measure both the position and momentum of a particle could have been dismissed by a hide-bound "classical physicist" as follows: "Of course the particle has both a position and a momentum, that is part of what makes it a particle. What do I care if you can't figure out how to measure it"! This would have been a profoundly wrong retort. The correct interpretation is that Heisenberg's operational limitation is an essential part of the consistency of a new type of theory where "particle" no longer means what it used to. For black hole complementarity we have yet to find the new theory, and indeed recent work that I discuss in Secs. VII and VIII suggests that more new ideas are needed before such a theory can be found, but black hole complementarity may yet play an important role in its eventual formulation.

VI. HOLOGRAPHY AND THE AdS/CFT CORRESPONDENCE

The debate over whether or not black hole evaporation is unitary persisted for a while, but major progress in string theory in the 1990s managed to eventually push most people in the field into the prounitarity camp [including Hawking himself (Hawking, 2005)]. The first important part of this was the microscopic calculations of black hole entropy in string theory already mentioned in Sec. IV.E, but the main source of the shift was the explicit realization in AdS/CFT (Gubser, Klebanov, and Polyakov, 1998; Witten, 1998a; Maldacena, 1999) of the holographic principle ('t Hooft, 1993; Susskind, 1995a). In this section I give a brief overview of these ideas, focusing mostly on material relevant for understanding the black hole information problem. It will by no means be complete; AdS/CFT is an extensive subject. Only two years after its discovery the standard introduction (Aharony *et al.*, 2000) already weighed in at 261 pages. Interested readers are encouraged to turn there for more details on AdS/CFT, although many things described in this section were not known at the time of that review. In this section I will restore the Planck scale and instead mostly work in units where the AdS radius r_{ads} is set to 1. Given the wide variety of interesting AdS/CFT examples in various spacetime dimensions, in this section I will mostly work in $d+1$ spatial dimensions instead of just $3+1$.

The precision of the AdS/CFT correspondence means that this section will necessarily have a higher density of technical results than the previous ones. The upshot is that the correspondence provides compelling evidence for the unitarity of black hole evaporation.

A. Entropy bounds and the holographic principle

Perhaps surprisingly, the idea that black holes have microstates and are described by unitary dynamics has interesting implications for the statistical mechanics of other systems.[54]

For the first example, say that we have an object of linear size L and energy E. By lowering it to within a proper distance of order L of the horizon of a black hole whose Schwarzschild radius r_s is much greater than L, we can use the gravitational redshift to decrease the energy of the object as seen from infinity by a factor of the order of L/r_s. If we then drop the object into the black hole, the black hole mass changes by $\Delta M = LE/r_s$. The change in the Bekenstein-Hawking entropy of the black hole is

$$\Delta S_{\text{bh}} \propto LE. \tag{130}$$

This experiment potentially is a challenge to the second law of thermodynamics. If the increase in the black hole entropy is less than the entropy of the system we threw in then the total entropy would decrease. This led Bekenstein to conjecture that for any system we must have

$$S < CLE, \tag{131}$$

where C is some $O(1)$ coefficient not determined by this argument (Bekenstein, 1981). This conjecture is called the *Bekenstein bound*, and it has the rather surprising feature that, although it was motivated from an argument about black holes, the Planck scale does not appear on either side of the inequality. *A priori* there does not seem to be an independent reason for it to be true, so it at first appears to be a nontrivial constraint on the type of nongravitational systems that can be consistently coupled to gravity. Indeed there has been quite a lot of controversy in the literature, both about whether or not the bound is true and whether or not it needs to be true to preserve the second law (Unruh and Wald, 1982; Marolf, Minic, and Ross, 2004; Marolf and Sorkin, 2004). Most of the controversy stems from the issue that the precise definitions of the quantities S, L, and E appearing in the bound are not clear. Recently however Casini (2008) gave a simple and elegant proof that a precise version of the bound holds in any relativistic quantum field theory.[55] This makes it clear that the Bekenstein-Casini bound is not really a constraint on matter theories, but it is nonetheless a deep and surprising property they possess which does not seem particularly natural unless we think about black holes.

To get a bound that involves the Planck scale explicitly, Susskind suggested a different thought experiment (Susskind, 1995a). Consider a stationary object of entropy S and energy E, which is contained in a sphere of area A. If we assume that it is not a black hole, then E must be less than the mass M of a black hole with horizon area A. Now consider the process where we collapse a spherical shell of matter onto this object

[54]For a broad review of the ideas in this section, see Bousso (2002).

[55]The proof of the Bekenstein-Casini bound is based on the positivity of *relative entropy*, which for any two density matrices is defined as $S(\rho|\sigma) \equiv \text{tr}\rho\log\rho - \text{tr}\rho\log\sigma$. Casini interprets the left-hand side of the bound as the renormalized von Neumann entropy $-\text{tr}\rho_V\log\rho_V + \text{tr}\rho_V^0\log\rho_V^0$ of some region V in an excited state ρ_V, where ρ_V^0 is the reduced density matrix of the ground state in the same region. The right-hand side of the bound is interpreted as the renormalized expectation value $\text{tr}\rho_V K - \text{tr}\rho_V^0 K$ of the modular Hamiltonian $K \equiv -\log\rho_V^0$ in the excited state ρ. The bound then follows immediately from the positivity of $S(\rho_V|\rho_V^0)$. I encourage the reader to read his paper for the details; the motivation and discussion are transparent, as is the explanation of why the theorem avoids various potential objections such as increasing the number of species.

whose energy is $M - E$; this results in the formation of a black hole of mass M and area A. In order for this process to not violate the second law we apparently need

$$S \leq \frac{A}{4G}, \tag{132}$$

which is called the *holographic entropy bound*. Roughly it says that the maximal amount of entropy in a spacetime region scales with the area of the boundary of the region. From a quantum field theory point of view this is surprising. Typically we are used to entropies scaling extensively with the volume of a spacetime region. For example, if we imagine a lattice of Planckian spacing with some finite number of degrees of freedom at each point on the lattice, then the logarithm of the Hilbert space dimension would scale with the volume of the lattice in Planck units. So the holographic entropy bound is saying that the number of degrees of freedom in spacetime is much less than we might have naively thought; if you try to excite more you make a black hole. The bound (132) as stated has several problems (Bousso, 2002), but a more general covariant version of it (Bousso, 1999a, 1999b) has survived many quantitative tests and been proven to hold in a wide variety of classical and semiclassical situations (Flanagan, Marolf, and Wald, 2000; Wall, 2010, 2012a; Bousso *et al.*, 2014).

Clearly if the holographic entropy bound is correct then there is a large amount of nonlocality in whatever the correct theory of quantum gravity is. Indeed the area scaling of the entropy led 't Hooft and Susskind ('t Hooft, 1993; Susskind, 1995a) to conjecture that a true theory of quantum gravity must in some sense live in one fewer dimensions than naively expected; Susskind called this idea the *holographic principle*.

B. Statement of the AdS/CFT correspondence

The holographic principle has so far had its most precise realization in the widely celebrated anti–de Sitter/conformal field theory correspondence. AdS/CFT was originally discovered by studying the low-energy limit of brane systems in string theory (Maldacena, 1999). In the most well-known example, one looks at a stack of N $D3$ branes in type IIB string theory in ten dimensions.[56] At large N the branes backreact and produce a nontrivial geometry which approaches five-dimensional anti–de Sitter space times an $\mathbb{S}^5$ in the vicinity of the branes. The AdS radius in Planck units is of the order of $N^{1/4}$. At any N, however, the region near the branes has a low-energy description given by a particular $(3+1)$-dimensional conformally invariant quantum field theory called maximally supersymmetric SU(N) Yang-Mills theory, with a gauge coupling constant given by $g_{\mathrm{YM}}^2 = 4\pi g$, where g is the string coupling constant. In the region of overlapping validity these two theories must be the same, and since the latter one makes sense at any N (and g) it is natural to conjecture that the equivalence holds at finite N (and g). For reasons we will soon see, the AdS description is often called the "bulk" theory while the CFT is called the "boundary" theory.[57] Rather than unpack the details of this argument, however, with the benefit of hindsight I will instead present AdS/CFT as a self-consistent framework. I will state the correspondence precisely below, but I will first briefly introduce the ingredients.

1. Anti–de Sitter space

So far in this article we considered geometries which asymptote to ordinary flat Minkowski space. In the presence of a nonzero vacuum energy, however, Minkowski space is not a solution of Einstein's equations. If the vacuum energy is negative, the simplest solution is anti–de Sitter space, which in $d+1$ spacetime dimensions has metric

$$ds^2 = -\left[1 + \left(\frac{r}{r_{\mathrm{AdS}}}\right)^2\right]dt^2 + \frac{dr^2}{1 + (r/r_{\mathrm{AdS}})^2} + r^2 d\Omega_{d-1}^2. \tag{133}$$

Here we have $t \in (-\infty, \infty)$, $r \in [0, \infty)$. The length r_{AdS} is related to the vacuum energy density ρ_0 as

$$\frac{1}{r_{\mathrm{AdS}}^2} = -\frac{16\pi G \rho_0}{d(d-1)}. \tag{134}$$

For the remainder of this section I will work in units where $r_{\mathrm{AdS}} = 1$. This geometry manifestly has the property that for $r \ll 1$ it resembles Minkowski space in spherical coordinates. It is not obvious in this presentation, but it has an isometry group SO$(d,2)$ which is large enough to send any point in the spacetime to any other point; the geometry is homogeneous. As $r \to \infty$ it does not approach Minkowski space, so it has its own interesting boundary structure. As usual we can describe this more intuitively with a Penrose diagram, which can be derived by introducing a new coordinate $r = \tan\rho$:

$$ds^2 = \frac{1}{\cos^2\rho}[-dt^2 + d\rho^2 + \sin^2\rho d\Omega_{d-1}^2]. \tag{135}$$

We have $\rho \in [0, \pi/2)$, so we can conformally compactify by discarding the diverging prefactor and including the boundary at $\rho = \pi/2$. The diagram is shown in Fig. 21.

The main lesson of the AdS Penrose diagram is that we should think of AdS as a box. Massless particles sent out from the center get all the way out to the boundary and back in a finite proper time π, as seen by somebody at the center.[58] This is also true for massive particles; say you are floating in the center of AdS and you throw a ball away from you. It will go out some finite

[56] This argument may be difficult for someone unfamiliar with string theory to follow, but there are only so many things I can review. These objects are defined, for example, by Polchinski (1998b), but do not worry, they will not be on the final.

[57] This terminology is convenient but it can be misleading. The situation here is different from an often-encountered one in condensed matter physics, where there are "edge modes" at the boundary of some system that also has quasiparticle excitations in its interior. In that type of system, the two types of excitations exist in the same theory. In AdS/CFT, the two theories are equivalent, and we should use either one description or the other but not both.

[58] Provided that we choose reflecting boundary conditions at $r = \infty$, as we probably should if we want to view AdS as a closed system.

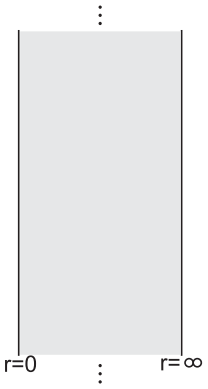


FIG. 21. The Penrose diagram for anti–de Sitter space. The left side is the origin of polar coordinates at $r = 0$ and the right side is the timelike spatial boundary of AdS at $r = \infty$. The time coordinate t is the proper time at $r = 0$, so we see immediately that signals can be sent out to the boundary and return in a finite amount of proper time at the center. The diagram continues infinitely far to the past and future so it is not really compact. This can be fixed by an additional coordinate transformation, but the current form of the diagram is actually more useful so people usually do not do this.

distance (unlike a massless particle it will not make it all the way to the boundary), but eventually it will turn around and return to you after a time of the order of 1 in AdS units. These observations are formalized in the statement that the boundary is *timelike*. It has the topology of $\mathbb{R} \times \mathbb{S}^{d-1}$ with the $\mathbb{R}$ being temporal.

As in flat space there is a natural notion of asymptotically AdS spacetime (Henneaux and Teitelboim, 1985). I will not define this rigorously, but roughly a spacetime is asymptotically AdS if its only boundary is timelike and in the vicinity of that boundary the geometry approaches that of AdS near $r \to \infty$. In a theory of quantum gravity with nontrivial states there will be backreaction, so this type of geometry will need to be included to get anything interesting. These states will lie in representations of the AdS symmetry $\mathrm{SO}(d, 2)$ in the same way that in Minkowski space excitations can be characterized by their Lorentz transformation properties. One of the generators of this symmetry is the AdS version of the ADM Hamiltonian H, which generates translations of the boundary coordinate t, so quantum gravity in asymptotically AdS space is a strong candidate for a closed Hamiltonian system. Moreover, since excitations reach the boundary and return in finite time, we can think of it as "gravity in a box"; in particular, we should expect that the spectrum of H is discrete.

2. Conformal field theory

A conformal field theory is a relativistic quantum field theory which is also invariant under a larger set of spacetime transformations, the conformal group, which is generated by the usual Poincaré transformations, rescalings of the coordinates $x'^\mu = \lambda x^\mu$, and special coordinate transformations

$$x'^\mu = \frac{x^\mu + a^\mu x^2}{1 + 2x_\nu a^\nu + a^2 x^2}. \tag{136}$$

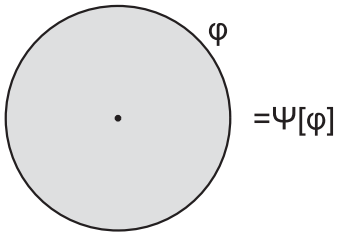


FIG. 22. The state-operator correspondence. The path integral over the ball $\rho < 1$ with boundary condition ϕ and an operator $\mathcal{O}$ at $\rho = 0$ computes a wave functional $\Psi[\phi]$ which has energy Δ if $\mathcal{O}$ has dimension Δ. Moreover given such a state, we can construct the operator by evolving the state radially inward assuming no operators are present until we are left with something at the center that must be local.

More abstractly the conformal group is defined as the set of transformations of Minkowski $\mathbb{R}^d$ which preserve angles but not necessarily lengths. It is isomorphic to $\mathrm{SO}(d, 2)$, which already suggests a connection to AdS_{d+1}.

The simplest example of a CFT is a free massless scalar field, which in $3 + 1$ Minkowski space is invariant under the dilatation transformation

$$\begin{aligned} x'^\mu &= \lambda x^\mu, \\ \phi'(x') &= \lambda^{-1}\phi(x). \end{aligned} \tag{137}$$

CFTs have many interesting properties which I do not have time to go into, but two are crucial for us. First in any CFT we can always find a special set of local operators, called primary operators, which transform simply under conformal transformations.[59] In particular under dilatations they transform as

$$\mathcal{O}'(x') = \lambda^{-\Delta}\mathcal{O}(x), \tag{138}$$

where Δ is the conformal dimension of the primary operator $\mathcal{O}$. In a unitary conformal field theory Δ is real and positive, and if $\mathcal{O}$ is a scalar operator it obeys $\Delta \geq (d-2)/2$. Derivatives $\partial^n\mathcal{O}$ of a primary operator are called descendants; they have conformal dimension $\Delta + n$, meaning that they rescale as $\lambda^{-\Delta-n}$ under dilatations, but they are no longer primary. Primary operators have simple correlation functions, for example, in any CFT a scalar primary $\mathcal{O}$ of dimension Δ has a time-ordered two-point function[60]

$$\langle\Omega|T\mathcal{O}(x,t)\mathcal{O}(0,0)|\Omega\rangle = \frac{1}{(|x|^2 - t^2 + i\epsilon)^\Delta}. \tag{139}$$

Second it is often interesting to study a CFT on the "cylinder" $\mathbb{R} \times \mathbb{S}^{d-1}$, with metric

[59]More precisely a local operator is primary if when it is located at $x = 0$ its commutators with the special conformal generators are zero.

[60]Remember that time ordering means that operators at earlier times appear to the right of operators at later times. ϵ is a positive infinitesimal quantity which fixes the interpretation of the branch cut. For the special case of a massless free field in $3 + 1$ dimensions this formula follows from the $m \to 0$ limit of Eq. (27).

$$ds^2 = -dt^2 + d\Omega_{d-1}^2. \tag{140}$$

As in our previous discussion of the Hilbert space of quantum field theory in Sec. III.A, a natural basis of states for this system is field configurations on $\mathbb{S}^{d-1}$. Another natural basis is the set of eigenstates of the Hamiltonian H that generates t translation. In fact in CFTs there is a natural bijection between local operators of dimension Δ and these energy eigenstates. The bijection is based on the observation that Euclidean $\mathbb{R}^d$

$$ds^2 = d\rho^2 + \rho^2 d\Omega_{d-1}^2 \tag{141}$$

is conformally equivalent to $\mathbb{R} \times \mathbb{S}^{d-1}$ in the sense of Sec. II.C. Indeed the coordinate transformation $\rho = e^\tau$ puts the metric (141) into the form

$$ds^2 = e^{2\tau}(d\tau^2 + d\Omega_{d-1}). \tag{142}$$

Note that dilatations of ρ become time translations of τ, which after analytic continuation becomes t. In a conformal field theory we can essentially ignore conformal equivalences in the metric. They do not affect angles and they should basically leave the theory invariant.[61] The bijection then works by using the Euclidean path integral on the ball $\rho < 1$ in $\mathbb{R}^d$ with a local operator of dimension Δ at $\rho = 0$ to construct a quantum state of energy Δ at $t = 0$ on $\mathbb{R} \times \mathbb{S}^{d-1}$, as shown in Fig. 22. This is called the state-operator correspondence.

3. The dictionary

We now state the AdS/CFT correspondence. The modern version is as follows (Heemskerk *et al.*, 2009):

- Any relativistic conformal field theory on $\mathbb{R} \times \mathbb{S}^{d-1}$ with metric (140) can be interpreted as a theory of quantum gravity in an asymptotically $\mathrm{AdS}_{d+1} \times M$ spacetime. Here M is some compact manifold that may or may not be trivial.

This statement is really something of a definition; at the moment we do not know an alternative precise theory of quantum gravity in asymptotically AdS space, so we are instead using something well defined (the CFT) to say what it is we mean by quantum gravity in AdS space. In order for this definition to be good, we need to address several issues:

(1) What is the detailed map between the theories? Given an AdS quantity we are interested in, how do we compute it in the CFT?
(2) In what cases does the CFT lead to a gravity theory with a good semiclassical description? In other words, under what circumstances is the AdS radius large in Planck units?

Question (1) is answered by the "dictionary" of AdS/CFT: the list of CFT expressions for interesting bulk quantities. I describe here some of the most important entries in the dictionary. First of all the Hilbert space of physical states of the bulk is by definition identical to the CFT Hilbert space. Moreover symmetry generators of $\mathrm{SO}(d,2)$ in the CFT are identified with the corresponding bulk symmetry generators of asymptotically AdS space. In particular the Hamiltonian is the same on both sides. Quantities which depend only on the space of states and the Hamiltonian, for example, the thermal partition function or the free energy at finite temperature, are thus computed by their CFT expressions by definition. We also want to have something like a CFT expression for a local bulk field, but as already argued in Sec. VI.A local bulk fields should not exist in a true theory of quantum gravity such as AdS/CFT is claiming to provide. In Sec. V.A, however, we argued that states at past and future infinity in Minkowski space could be described in terms of free fields. Along similar lines we might guess that in AdS we should be able to make precise sense of the boundary limits of local bulk fields. Indeed the AdS/CFT dictionary postulates, for example, that if $\mathcal{O}$ is a scalar primary CFT operator then there is a bulk scalar field ϕ such that

$$\lim_{r\to\infty} r^\Delta \phi(t, r, \Omega) \equiv \mathcal{O}(t, \Omega). \tag{144}$$

In other words if we extrapolate a bulk field to the boundary, stripping off a normalization factor, then we get a quantity which is exactly described in the CFT as a primary operator with dimension Δ.[62] In an expectation value of products of bulk operators, the limit in Eq. (144) should be taken simultaneously for all operators together.[63]

As an example of the extrapolate dictionary, consider a free massive scalar field in AdS. Its time-ordered two-point function is[64]

[61] More carefully one can show that for correlation functions of scalar primary operators we have

$$\langle \mathcal{O}_1(x_1)\cdots\mathcal{O}_n(x_n)\rangle_{e^{-2\omega}g} = e^{\Delta_1\omega(x_1)+\cdots+\Delta_n\omega(x_n)} e^{A[g,\omega]} \langle \mathcal{O}_1(x_1)\cdots\mathcal{O}_n(x_n)\rangle_g, \tag{143}$$

where $A[g,\omega]$ is some known local functional of g and ω that weakly depends on the theory under consideration but is independent of which operators appear in the correlation function. It is called the conformal anomaly functional and is nonzero only when d is even. A similar equivalence holds for wave functionals computed from path integrals on conformally equivalent geometries.

[62] This version of the dictionary, sometimes called the "extrapolate" dictionary, was first proposed by Banks *et al.* (1998). Its equivalence to another version (Gubser, Klebanov, and Polyakov, 1998; Witten, 1998a), sometimes called the "differentiate" dictionary, was shown to all orders in perturbation theory in Harlow and Stanford (2011).

[63] Other choices of how we scale the operators will lead to CFT correlation functions computed on spaces other than Eq. (140). There are several interesting options, for example, we can instead get CFT correlation functions on $\mathbb{R}^d$ ("AdS-Poincaré slicing") or on $\mathbb{R} \times \mathbb{H}^{d-1}$ ("AdS-Rindler slicing"), but I will not describe these in detail here.

[64] See, for example, Burgess and Lutken (1985). The most elegant way to derive this involves starting with the Euclidean Green's function on the hyperbolic disk, which depends only on the geodesic distance and obeys the wave equation, and analytically continuing. For an example see an analogous calculation for dS space in Appendix B of Harlow and Stanford (2011).

$$\langle\Omega|T\phi(x)\phi(x')|\Omega\rangle = (2\Delta - d)^{-1}2^{-2\Delta}\pi^{-d/2}\frac{\Gamma(\Delta)}{\Gamma(\Delta - d/2)} \times u^{\Delta}F\left(\Delta, \Delta + \frac{1-d}{2}, 2\Delta - d + 1, u\right), \tag{145}$$

where

$$u = \frac{2}{1 + \cosh \ell}, \tag{146}$$

with ℓ the geodesic distance between the two points and F a hypergeometric function. Δ is related to the mass m as

$$\Delta = \frac{d}{2} + \frac{1}{2}\sqrt{d^2 + 4m^2}. \tag{147}$$

In terms of the coordinates in Eq. (133), we have

$$u = \frac{1}{1 - rr'\cos\alpha + \sqrt{(1+r^2)(1+r'^2)}\cos\left[(t-t')(1-i\epsilon)\right]}. \tag{148}$$

Here α is the angle between the two points on $\mathbb{S}^{d-1}$ and ϵ is an infinitesimal positive quantity that picks out the right branch of u^{Δ}. Applying the dictionary (144) we then have

$$\langle\Omega|T\mathcal{O}(t,\alpha)\mathcal{O}(t',0)|\Omega\rangle \propto \left(\frac{1}{\cos\left[(t-t')(1-i\epsilon)\right] - \cos\alpha}\right)^{\Delta}, \tag{149}$$

which is the correct two-point function in conformal field theory on $\mathbb{R}\times\mathbb{S}^{d-1}$ for a scalar operator of dimension Δ.[65]

The extrapolate dictionary works not only for scalars, but indeed any CFT has a unique energy momentum tensor $T_{\mu\nu}$ which is a spin two primary operator of dimension d. Its bulk dual is the metric tensor, which of course should exist in any theory of gravity. Moreover if the CFT has global symmetries then by Noether's theorem it must have conserved currents of dimension $d-1$, and these are dual to gauge fields in the bulk. There are also other interesting items in the dictionary, among them Wilson lines (Maldacena, 1998) and von Neumann entropy (Hubeny, Rangamani, and Takayanagi, 2007; Lewkowycz and Maldacena, 2013; Ryu and Takayanagi, 2006; Nishioka, Ryu, and Takayanagi, 2009) in the boundary theory. I will briefly discuss the latter in Sec. VI.J.

Returning now to question (2), recall that in the original example of AdS/CFT the number N of $D3$ branes is what set the AdS radius in Planck units. It was only in the large N limit that bulk gravity was approximately classical. We now extend this to a general statement about AdS/CFT. A general CFT has a large N limit if there exists a parameter N, possibly discrete, such that the set of primary operators whose dimensions do not scale with N have the following properties:[66]

- There is a finite set[67] of "single-trace" primary operators $\mathcal{O}_i$, each of which has spin ≤ 2. There is only one, the stress tensor, with spin exactly 2 and $\Delta = d$. If we normalize their two-point functions so that Eq. (149) (or its higher-spin generalization) holds with no extra prefactor, then the three-point function of any three of them is suppressed by powers of $1/N$.
- For any collection of single-trace operators $\{\mathcal{O}_{i_1}, \ldots, \mathcal{O}_{i_n}\}$ there exists a "multitrace" operator $\mathcal{O}_{i_1}\cdots\mathcal{O}_{i_n}$ with dimension $\Delta = \Delta_{i_1} + \cdots + \Delta_{i_n} + O(1/N)$.
- If we normalize the two-point functions of multitrace operators to $O(N^0)$, then their correlation functions with each other and with other single-trace operators are $O(1/N)$ unless their components can be matched in pairs. So, for example, $\langle\mathcal{O}_i(x)\mathcal{O}_j(y)\mathcal{O}_i\mathcal{O}_j(z)\rangle$ is $O(N^0)$, but $\langle\mathcal{O}_i(x)\mathcal{O}_j(y)\mathcal{O}_i\mathcal{O}_k(z)\rangle$ with $k \neq j$ is $O(1/N)$. Moreover if they can be matched in pairs, then to leading order in $1/N$ the correlation function is the sum over all such matchings of the product of the two-point functions of the matched pairs. This property is called large-N factorization.
- All operators whose dimensions are $O(N^0)$ are single-trace primary operators, multitrace primary operators, or their descendants.

These properties may seem somewhat *ad hoc*, but they can be easily remembered by considering bulk Feynman diagrams in a theory where all interactions are proportional to $1/N$. The parameter N is always proportional to some power of the AdS radius in Planck units. The last requirement is crucial; via the state-operator correspondence it says that the low-energy spectrum of the CFT is consistent with weakly coupled low-energy effective field theory in AdS. More explicitly the states corresponding to single-trace operators are single-particle states in the bulk, while the states corresponding to multitrace operators are multiparticle states. We now study this in a bit more detail.

C. Perturbations of the AdS vacuum

Consider a free massive scalar field in AdS_{d+1}, written in coordinates where the metric is Eq. (133). As usual we begin by looking for positive-frequency normalizable modes for use in Eq. (24), which we can expand as

$$f_{\omega\ell\vec{m}}(t,r,\Omega) \equiv r^{-(d-1)/2}Y_{\ell\vec{m}}(\Omega)\psi_{\omega\ell}(r). \tag{150}$$

[65]This can be determined from Eqs. (139) and (143).

[66]This list is not necessarily exhaustive. I am not sure if a completely sharp definition of what we mean by having a large-N limit exists. Also note that I have intentionally excluded theories with higher-spin symmetry in the bulk; although these exist (Vasiliev, 1990) and sometimes have simple gravity duals (Giombi and Yin, 2010; Klebanov and Polyakov, 2002; Gaberdiel and Gopakumar, 2011), they are sufficiently opaque that it is unclear to what extent they are good models for conventional gravity.

[67]Or a discrete tower with $O(1)$ spacing if there is a nontrivial compact manifold M.

As usual $\psi_{\omega\ell}$ obeys a Schrödinger equation

$$-\frac{d^2}{dr_*^2}\psi_{\omega\ell}+V(r)\psi_{\omega\ell}=\omega^2\psi_{\omega\ell}, \tag{151}$$

with the tortoise coordinate now being

$$r_*=\arctan r \tag{152}$$

and the potential being

$$V(r)=\left(1+\frac{1}{r^2}\right)\left[\left(m^2+\frac{(d-1)(d+1)}{4}\right)r^2 + \ell(\ell+d-2)+\frac{(d-1)(d-3)}{4}\right]. \tag{153}$$

This potential diverges at the boundary $r_*=\pi/2$, so $\psi_{\omega\ell}$ is confined to lie in the region $r_*\in(0,\pi/2)$. The basic features of the solutions are thus the same as for the infinite square well system. In particular, the frequency ω will be quantized with a constant high-energy density of states (at fixed ℓ). This is consistent with the gravity in a box intuition for physics in AdS; single-particle states created by the creation operators for these modes have no continuous quantum numbers. The "ground state" of this Schrödinger problem with $\ell=0$ corresponds to a particle localized "at rest" in the center of AdS, and excited states and/or states with $\ell>0$ correspond to the particle moving around.

In fact in this case the modes can be found analytically (Breitenlohner and Freedman, 1982), and the quantization condition is[68]

$$\omega_{n\ell}=\Delta+\ell+2n, \qquad n=0,1,2,\ldots, \tag{154}$$

with Δ given by Eq. (147). As expected, the lowest energy state for the particle has $\ell=n=0$.

Although we will not really need it, for posterity a properly normalized (in the KG norm) expression for the mode is

$$r^{-(d-1)/2}\psi_{n\ell}(r)=\frac{1}{\Gamma(\Delta-d/2+1)}\sqrt{\frac{\Gamma(n+\Delta-d/2+1)\Gamma(n+\Delta+\ell)}{\Gamma(n+1)\Gamma(n+\ell+d/2)}} \times r^{-\Delta}\left(1+\frac{1}{r^2}\right)^{-(\ell+\Delta+2n)/2} \times F\left(-n,-n+1-\ell-d/2,\Delta-d/2+1,-\frac{1}{r^2}\right), \tag{155}$$

where F is a hypergeometric function that goes to 1 as $r\to\infty$ and is actually equivalent to some polynomial for $n=0,1,2,\ldots$. This quantization condition is necessary for this mode to also be normalizable near $r=0$.

Equation (154) has a nice CFT interpretation via the state-operator correspondence. The state created by the $\ell=n=0$ creation operator is the CFT state produced by inserting the single-trace primary $\mathcal{O}$ dual to ϕ into the center of the Euclidean path integral as in Fig. 22, and the various excited states come from inserting its descendants. As a simple check consider the states with energy $\omega=\Delta+2$ in AdS_4. There are two types: a single state with $\ell=0$, $n=1$ and five states with $\ell=2$, $n=0$. In the CFT we get descendants of this dimension by acting on $\mathcal{O}$ with two derivatives. There is an angular momentum singlet where we contract the two derivatives and a traceless symmetric tensor where we do not. The latter has five linearly independent components so the degeneracies match. It is not hard to generalize this counting to arbitrary d, n, and ℓ, and needless to say it works.

We can also understand multiparticle states along these lines. The ground state with no particles is just the CFT ground state, produced by inserting the identity in the Euclidean path integral. The rest of the multiparticle Fock space can be built in the CFT by inserting multitrace operators. In fact this discussion can be condensed into the single statement that to leading order in $1/N$ the operator Fourier transform $\mathcal{O}_{n\ell\vec{m}}$ of the single-trace primary operator $\mathcal{O}$ has an algebra consistent with it being proportional to the lowering operator $a_{n\ell\vec{m}}$ for the mode $f_{n\ell\vec{m}}$ (Banks *et al.*, 1998). We can determine the constant of proportionality by comparing Eqs. (24), (144), and (155) in the limit $r\to\infty$, to find

$$\mathcal{O}_{n\ell\vec{m}}=\frac{1}{\Gamma(\Delta-d/2+1)} \times\sqrt{\frac{\Gamma(n+\Delta-d/2+1)\Gamma(n+\Delta+\ell)}{\Gamma(n+1)\Gamma(n+\ell+d/2)}}a_{n\ell\vec{m}}. \tag{156}$$

Since the left-hand side of this equation is a CFT operator, together with Eq. (24) this then allows us to write a CFT expression for a local bulk field at any bulk point. This may

[68]For people who are familiar with it, this is the "standard quantization" of the field (Breitenlohner and Freedman, 1982; Klebanov and Witten, 1999), where the boundary conditions require the modes to behave like $r^{-\Delta}$ at infinity. Generically normalizability requires this choice, but if $d/2<\Delta<(d+2)/2$ then we can instead choose modes which behave like $r^{\Delta-d}$ at infinity. The operator $\mathcal{O}$ will then have dimension $\Delta_-\equiv d-\Delta$, and formulas for that case can be obtained from those here by replacing $\Delta\to\Delta_-$.

seem disturbing, given our general arguments that this should not be possible, but remember that this construction is valid only to leading order in $1/N$, and only in states close to the vacuum. It can be "fixed up" perturbatively in $1/N$ (Heemskerk *et al.*, 2012; Kabat, Lifschytz, and Lowe, 2011), but there is no reason to expect a generalization that holds nonperturbatively and good reasons not to.[69] The regime of validity of this construction of local bulk fields has recently been reinterpreted in the language of quantum error correction, a subject that is unfortunately beyond the scope of this paper (Almheiri, Dong, and Harlow, 2015).

D. One-sided AdS black holes at fixed energy

The main new feature of black holes in AdS is that their Hawking radiation is reflected back by the boundary in finite time. For small enough black holes this is not really important, since after all the entire black hole could evaporate before the radiation gets to the boundary. But as the Schwarzschild radius of the black hole approaches the AdS radius we eventually reach a point where the radiation is being reflected back into the black hole as fast as it is being emitted. At this point the black hole never evaporates, so large enough black holes in AdS are eternal. One thus often hears discussion of "big" and "small" black holes in AdS, with the distinction almost always meaning that the big ones are stable and the small ones are not.[70]

The crossover point between stability and instability can be estimated by a simple statistical argument (Horowitz, 2000). A typical state of energy E in the CFT will have some fraction x of its energy in a black hole and the rest in the radiation field. We are interested in finding the x which maximizes the total entropy, which [in AdS_4, ignoring $O(1)$ factors, and temporarily restoring r_{AdS}] is

$$S \approx (E\ell_p)^2 x^2 + (E r_{AdS})^{3/4}(1-x)^{3/4}. \tag{157}$$

Here the first term is the black hole entropy and the second term is the entropy of the radiation field, which we can think of being in a box of linear size r_{AdS}.[71] At low energies the second term dominates and the function decreases monotonically. We maximize the entropy by taking $x = 0$ so typically there is no black hole. This matches our Minkowski intuition that black holes should evaporate. At sufficiently large E, however, the first term dominates so we win by taking x to be very close to 1; in fact there is a local maximum that is near but not quite at $x = 1$. Thus almost all of the energy (and entropy) are contained in a single black hole that never evaporates. The crossover apparently happens when the two terms are of comparable size, which happens when

$$E r_{AdS} = \left(\frac{r_{AdS}}{\ell_p}\right)^2 \left(\frac{r_{AdS}}{\ell_p}\right)^{-2/5}. \tag{158}$$

I have written it this way because the first term is the energy in AdS units of a black hole whose Schwarzschild radius is of the order of r_{AdS}, so we see that the crossover happens when the black hole is parametrically smaller in r_{AdS}/ℓ_p than the AdS radius (Horowitz, 2000).[72]

This crude argument cannot really tell us what happens when r_s of the order of the size of the AdS radius or larger, so to proceed further we need to address this. The AdS_{d+1} version of the Schwarzschild geometry, which is the unique spherically symmetric solution of Einstein's equation with negative vacuum energy, has (for $d \geq 3$) metric (again setting $r_{AdS} = 1$)

$$ds^2 = -f(r)dt^2 + \frac{dr^2}{f(r)} + r^2 d\Omega_{d-1}^2. \tag{159}$$

Here

$$f(r) = \left(r^2 + 1 - \frac{\alpha}{r^{d-2}}\right), \tag{160}$$

and α is related to the AdS version of the ADM mass M as

$$\alpha = \frac{16\pi G M}{(d-1)\Omega_{d-1}}. \tag{161}$$

The Schwarzschild radius r_s is the unique positive root of f. By demanding that the Euclidean version of this geometry be smooth at $r = r_s$ as we did for the Minkowski black hole in Sec. IV.H, one finds the temperature is

$$T = \frac{d - 2 + d r_s^2}{4\pi r_s}. \tag{162}$$

For $r_s \ll 1$ and $d = 3$ you can check that this agrees with Eq. (56). Combining this with

$$M = \frac{(d-1)\Omega_{d-1}}{16\pi G} r_s^{d-2}[1 + r_s^2] \tag{163}$$

[69] You might also worry about the gauge invariance of "local" bulk operators, but this can be dealt with by first fixing a gauge and then defining these operators (Heemskerk, 2012; Kabat and Lifschytz, 2013). This is somewhat similar to the situation with computing primordial density perturbations produced during inflation in cosmology (Maldacena, 2003b). In both cases, however, at higher orders in perturbation theory it is not completely clear whether or not the gauges used are "physical" in the sense that the quantities which appear simple are actually quantities that we observe.

[70] This distinction is a bit subtle, since as we will see the transition happens at different values of the energy in the microcanonical and canonical ensembles. Some usually seem to have the canonical transition in mind, and since as we will see it happens at higher energy it is always safe to assume this when a big black hole is being discussed.

[71] I assume here that the Schwarzschild radius of the black hole is small enough compared to r_{AdS} that we can use the usual Minkowski formula for the entropy, we will see momentarily that this is self-consistent.

[72] This holds up in AdS_{d+1}, with the suppression being $(r_{AdS}/\ell_p)^{-(d-1)(d-2)/(2d-1)}$. If there is a large compact manifold, such as for the $AdS_5 \times \mathbb{S}^5$ example, then $d+1$ is the total number of large dimensions.

for the energy we can integrate to find the entropy

$$S = \frac{\Omega_{d-1} r_s^{d-1}}{4G} = \frac{A}{4G}. \tag{164}$$

Thus as we keep increasing the energy, r_s and the entropy both continue to grow. There is less and less room for the radiation gas, so the black hole continues to win entropically. Through the AdS/CFT dictionary we thus arrive at the following statement: At sufficiently large energy, almost all states in the CFT have a bulk description as a single gigantic black hole. More carefully, black hole states dominate the microcanonical ensemble of the CFT at sufficiently large energy. This provides a concrete realization of Bekenstein's proposal that black hole entropy should actually count microstates.

In the special case of $d = 2$ we can actually quantitatively confirm this result. For $d = 2$ we replace the AdS-Schwarzschild geometry (159) by the BTZ black hole (Banados, Teitelboim, and Zanelli, 1992)

$$ds^2 = -(r^2 - r_s^2)dt^2 + \frac{dr^2}{r^2 - r_s^2} + r^2 d\theta^2, \tag{165}$$

but the entropy and temperature are still obtained by the $d \to 2$ limits of Eqs. (162) and (164). The energy differs from the $d \to 2$ limit of Eq. (163) by an r_s-independent shift, such that we still have $M \to 0$ as $r_s \to 0$.[73] The thermal partition function of a general unitary $1+1$ CFT with a discrete spectrum of primary operator dimensions quantized on a circle of radius L was shown by Cardy (1986) to scale at high temperature as

$$Z[\beta] = e^{\pi^2 cL/3\beta}[1 + O(e^{-4\pi^2 \Delta L/\beta})], \tag{166}$$

where Δ is the dimension of the lowest-dimension nontrivial primary and c is the "Virasoro central charge" of the CFT. This central charge can also be computed in the bulk (Brown and Henneaux, 1986), giving

$$c = \frac{3 r_{\mathrm{AdS}}}{2G}. \tag{167}$$

c thus plays the role of the parameter we have been calling N in higher dimensions, so it should be large in a CFT with a semiclassical bulk dual. The energy and entropy we compute from Eq. (166) using Eq. (91) are

$$E = \frac{\pi^2 Lc}{3\beta^2},$$
$$S = \frac{2\pi^2 cL}{3\beta} = 2\pi\sqrt{\frac{cLE}{3}}. \tag{168}$$

Replacing $L \to r_{\mathrm{AdS}} = 1$ (as the $r \to \infty$ behavior of the metric requires) and comparing with Eqs. (162) and (164), we find precise agreement between the entropy of the CFT and the entropy of the black hole. The energy also agrees with Eq. (163) up to the above-mentioned shift. Moreover from Eq. (166) this calculation stops being correct when $\beta \sim L$, which we will now see is exactly the order of the temperature where the black hole stops dominating the canonical ensemble.[74]

E. One-sided AdS black holes at fixed temperature and the Hawking-Page transition

The transition from unstable to stable black holes can also be studied at finite temperature instead of finite energy, where it is possible to be more rigorous along the lines of Sec. IV.H (Hawking and Page, 1983; Witten, 1998b). We should expect the transition to happen at a temperature corresponding to a higher energy than Eq. (158), since at finite temperature it is possible for energy to be absorbed by the heat bath instead of being reflected back into the black hole, which makes it harder for the black hole to win in the canonical ensemble. As in Sec. IV.H, one proceeds by evaluating the Euclidean gravitational action

$$I_E = -\frac{1}{16\pi G}\int_{\mathcal{M}} d^{d+1}x\sqrt{-g}[R + d(d-1)] - \frac{1}{8\pi G}\int_{\partial\mathcal{M}} d^d x\sqrt{\gamma}K, \tag{169}$$

of the AdS-Schwarzschild geometry (159) and comparing it to the Euclidean gravitational action of pure AdS space with compactified Euclidean time. In doing this one needs to cut off the geometry at some large r_c and ensure that the physical radius of the temporal $\mathbb{S}^1$ at this cutoff matches for the two geometries. I will not work out the details, but the result is that

$$I_E[g_{\mathrm{AdS}}] = -\frac{(d-1)\beta\Omega_{d-1}}{8\pi G}\frac{1}{\sqrt{1+1/r_c^2}} r_c^{d-1}(1 + r_c^2),$$
$$I_E[g_{\mathrm{Sch}}] = I_E[g_{\mathrm{AdS}}] + \frac{\beta\Omega_{d-1}}{16\pi G} r_s^{d-2}(1 - r_s^2) + O(1/r_c^2). \tag{170}$$

[73] Note that with this convention the energy of empty AdS, which we obtain from taking $r_s^2 \to -1$, is actually negative. Solutions with $-1 < r_s^2 < 0$ have a naked singularity and are usually considered to be unphysical, so there is thus a nontrivial energy gap between the vacuum and the "lightest BTZ black hole." It might seem more natural to define the energy so that the vacuum energy is zero, but for various reasons in $(1+1)$-dimensional CFTs the convention I use here is standard.

[74] An important subtlety here is that the asymptotics of Eq. (166) are only really rigorous in the "high temperature limit" where we keep c fixed and take $\beta \ll L$, whereas for the Bekenstein-Hawking entropy formula to be valid we also want the "semiclassical limit" $c \gg 1$ but only need $\beta \lesssim L$ to be above the Hawking-Page transition. Taking c to be large may in principle interfere with Eq. (166) if the density of states grows too rapidly with c. For an analysis see Hartman, Keller, and Stoica (2014), who confirm the validity of the Cardy formula in the semiclassical limit.

Here r_s is related to β by Eq. (162). The divergent parts of $I_E[g_{\text{AdS}}]$ can all be canceled by adding a series of boundary terms to Eq. (169) that depend only on the induced metric γ at the boundary. These terms correspond to possible counterterms in the CFT, and we are always free to add them without destroying the variational interpretation of Eq. (169). In fact it is necessary to add them if we wish the ground state energy to be zero.

In any event the main point is that, unlike what we found in asymptotically flat space, $I_E[g_{\text{Sch}}] - I_E[g_{\text{AdS}}]$ changes sign at $r_s = 1$ (Hawking and Page, 1983). The two saddle points exchange dominance in Eq. (96) for the partition function. This is the transition we are interested in; at sufficiently large temperatures the black hole wins, while at lower temperatures the thermal gas in AdS wins. As expected, the transition happens at a higher temperature than what we found in the previous section for the microcanonical ensemble.

This discussion has the great advantage over our discussion in Sec. IV.H that we now actually know what we are computing: the thermal partition function in the CFT. We saw already in the previous section that in $1+1$ dimensions this can be checked explicitly, and in higher dimensions the CFT interpretation of this transition is still fairly well understood (Witten, 1998b; Aharony *et al.*, 2004). The basic idea is that for a CFT quantized on a spatial $\mathbb{S}^{d-1}$, when the temperature is considerably larger than the inverse sphere radius the system basically behaves like a gas of $\sim N^\alpha$ free particles in $d-1$ spatial dimensions, where α is some $O(1)$ constant. This is consistent with Eqs. (163) and (164) for the black hole energy and entropy, which at high temperature scale as T^d and T^{d-1}, as they must if these quantities are to be extensive. This will no longer be true when the temperature is less than the inverse sphere radius, which must be the case since we now have a gas of free particles in d spatial dimensions in the bulk. On the CFT side the thermodynamics are now dominated by the constant modes of the fields, which in the special case of large N gauge theories like the $\mathcal{N}=4$ super Yang-Mills theory in $3+1$ dimensions are dominated by the holonomies of the gauge fields about the Euclidean thermal circle. It is reassuring to see the bulk and boundary descriptions of the physics line up in this manner.

F. Fields in the AdS-Schwarzschild background

We now briefly study fields propagating in the exterior of the AdS-Schwarzschild background. We can decompose the modes as Eq. (150), as we did for pure AdS, but the tortoise coordinate is now defined by

$$\frac{dr_*}{dr} = \frac{1}{f(r)}, \tag{171}$$

with $f(r)$ defined by Eq. (160). For simplicity I will take $r=\infty$ to be $r_*=0$, in which case we have

$$r_* = -\int_r^\infty \frac{dr'}{f(r')}. \tag{172}$$

The potential appearing in the effective Schrödinger equation (151) is now

$$V(r) = \frac{f(r)}{r^2}\left[\left(m^2 + \frac{(d+1)(d-1)}{4}\right)r^2 + \left(\ell(\ell+d-2) + \frac{(d-1)(d-3)}{4}\right) + \frac{(d-1)^2}{4}\cdot\frac{\alpha}{r^{d-2}}\right]. \tag{173}$$

The details here are not important; the main point is that, since $f(r)$ has a simple root at r_*, the tortoise coordinate r_* now runs from $-\infty$ to 0 as r runs from r_s to ∞. Moreover the effective potential now vanishes as $r_* \to -\infty$, so the effective Schrödinger problem is no longer confined to a box; the modes will now have a continuous frequency spectrum. As in our discussion of the brick wall model, however, this continuum is presumably discretized by Planckian physics near the horizon.[75]

It is worthwhile to note that the decomposition into "modes in the atmosphere" and "modes in the radiation" that we found for Minkowski black holes is not really valid once $r_s \gtrsim 1$. Once this is the case then the black hole potential barrier and the AdS barrier at $r_*=0$ essentially merge, so "the zone" fills the whole AdS space.[76]

It is interesting to understand to what extent we can extend our discussion of constructing local bulk field operators from Sec. VI.C to the AdS-Schwarzschild geometry. The situation is more subtle than it was around the vacuum, since there we had a clear picture of the structure of the Hilbert space from the state-operator correspondence. Acting on the vacuum with low-dimension primaries we were able to reproduce the detailed Fock space of low-energy field theory in AdS. A big black hole in a pure state by contrast is dual to a high-energy pure state in the CFT, which we can realize via the state-operator correspondence as the insertion of a very high dimension operator in radial quantization. This state is part of a densely spaced ensemble in the CFT. The subspace of states in some energy width of the order of the temperature has a dimensionality of the order of e^S. Understanding the details of this set of states is a hopeless task, so it is not immediately clear to what extent we can "find" the effective field theory Hilbert space buried within. The continuum spacing of the classical modes is certainly consistent with this dense spacing and introducing something like the brick wall will produce some set of states which have about the right density of states, but we want to have a prescription which does not require introduction of an arbitrary cutoff. To proceed further it is necessary to make some sort of *typicality* assumption about the state of the black hole. At least to leading order in $1/N$ this allows us to replace the detailed choice of pure state with a thermal density matrix, about which much more is known. In particular, the Fourier modes of CFT primary operators continue to behave in thermal expectation values as if they were creation and annihilation operators, but now for the

[75] We see CFT evidence for this in Sec. VI.I.

[76] A shallow barrier will still exist for $\ell \gg r_s^{2d/(d-2)}$, but the valley outside of it will be at very high energy so any mode localized there would be highly Boltzmann suppressed.

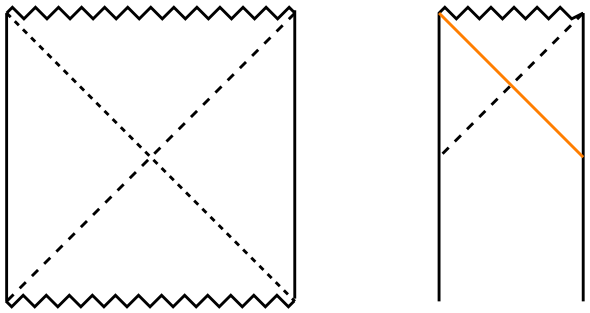


FIG. 23. The two-sided AdS-Schwarzschild wormhole and a one-sided big AdS black hole formed from collapse.

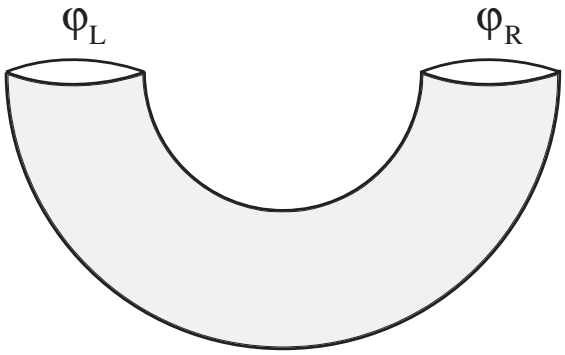


FIG. 24. The CFT construction of the Hartle-Hawking state. If the CFT lives on the boundary of the tube, so the field configurations at either end describe CFT field configurations on $\mathbb{S}^{d-1}$.

Schwarzschild modes we have been discussing (Papadodimas and Raju, 2013). This then seems to allow Eq. (24) for the bulk field, at least for fields located *outside* of the horizon. One can then attempt to extend the definition behind the horizon (Fidkowski *et al.*, 2004; Kabat and Lifschytz, 2014; Papadodimas and Raju, 2013), at least for the two-sided Schwarzschild geometry I will discuss in more detail, but there is considerably more to be said about this than what is currently known, especially about the regime of validity of the perturbative expansion in $1/N$.[77]

G. Collapsing shells and the two-sided AdS wormhole

So far I have not been particularly specific about the global structure of the geometry of AdS black holes. As in the asymptotically flat case, the full AdS-Schwarzschild geometry describes a wormhole connecting two asymptotic regions. Here each "exterior" region is asymptotically AdS. We can see this explicitly by introducing an AdS version of Kruskal coordinates[78]:

$$U \equiv -e^{(r_*-t)/2f'(r_s)},$$
$$V \equiv e^{(r_*+t)/2f'(r_s)}. \tag{174}$$

As in Sec. II.B there are two singularities at some positive value of UV, but now there are also two AdS boundaries at $UV=-1$. Defining $U=T-X$ and $V=T+X$ we can write the metric as

$$ds^2 = 4\frac{f(r)}{f'(r_s)}e^{-r_* f'(r_s)}(-dT^2+dX^2)+r^2 d\Omega_{d-1}^2. \tag{175}$$

Although not completely obvious, using Eq. (172) one can see that this geometry is smooth at the horizon at $UV=0$. As usual the Penrose diagram is more illuminating, however, it is shown on the left part of Fig. 23. I also show the one-sided geometry for a stable AdS black hole, created by some sort of collapsing shell. Note that in order to avoid the shell reflecting off of the boundary and going back in as we go back in time, we need to have it enter the system from the outside at $t=0$. Such a one-sided state will not be typical, after all its time reverse would be a black hole which spontaneously spits out all of its mass into a single narrow shell. A typical big black hole would be one that we assemble over a time that is exponentially long in N.

As with the two-sided asymptotically Minkowski black hole we discussed earlier, there is a natural choice of bulk ground state for the AdS wormhole: the Hartle-Hawking equation (88). As the geometry has two asymptotically AdS boundaries, the extrapolate dictionary strongly suggests that it should be realized as a state in the Hilbert space of *two copies* of the CFT (Maldacena, 2003a). Indeed by comparison with Eq. (88) there is a natural candidate for the CFT version of the Hartle-Hawking state:

$$|\psi_{\mathrm{HH}}\rangle \equiv \frac{1}{Z}\sum_i e^{-\beta H/2}|i^*\rangle_L |i\rangle_R, \tag{176}$$

where the states $|i\rangle_R$ are now understood as being energy eigenstates of a single copy of the CFT, and $|i^*\rangle_L = \Theta|i\rangle_R$, where Θ is an antiunitary operator that exchanges the two CFTs and reverses the direction of time in each.[79] This state is generated by the CFT Euclidean path integral on an interval times a sphere, as shown in Fig. 24. As in the asymptotically

[77] It was pointed out that, even if we stay outside the horizon, this construction cannot immediately be written in position space with the bulk field realized as an integral of some kernel times the position space CFT operator, contrary to the situation near the vacuum where it can (Leichenauer and Rosenhaus, 2013). This does not seem to be an insurmountable obstruction for two reasons: first, we can work in terms of the modes and not ask for such a formula. Second, if we allow ourselves to smear the bulk operator over a small region, then there is an expression of the desired form. In this sense the non-convergence discussed by Leichenauer and Rosenhaus (2013) is similar to the observation that formally the standard expression $\int[d^4k/(2\pi)^4]e^{ikx}/(k^2+m^2-i\epsilon)$ for a free field propagator is divergent at large k, which can also be resolved by smearing [see Morrison (2014) for a similar perspective].

[78] The nontrivial factors of $f'(r_s)$ are needed to ensure that U and V stay real under analytic continuation. In Sec. II.B we had $f'(r_s)=1/r_s$, so in the $r_s=1$ convention we were using we did not need them.

[79] Θ should not be confused with the natural CPT operation on a single copy of the CFT on a sphere, which reverses time but also reverses a longitudinal direction within the sphere.

Minkowski case discussed earlier, the Hartle-Hawking (or Hartle-Hawking-Israel) state is also often called the thermofield double state.

This proposal for the CFT description of the two-sided wormhole may seem obvious, but it has a rather surprising consequence. The Hamiltonian of the joint system is just the sum of the two CFT Hamiltonians; there are no interactions between the two CFTs. This is consistent with the bulk picture, where the ADM Hamiltonian is a sum of two boundary terms, one localized at each of the two boundaries, but it leads to the rather striking conclusion that two completely noninteracting systems can nonetheless have an alternate description where there is a single connected geometry where observers from the right and the left can jump in and meet each other in the middle. This is made possible by the diffeomorphism invariance of gravity. The bulk interactions that enable such a meeting are buried in the Hamiltonian constraint of canonical gravity and are invisible in the gauge-invariant CFT description of the system. This has led to a more general proposal that "entanglement generates geometry" (Van Raamsdonk, 2010), which recently has been given the name of "ER = EPR" (Maldacena and Susskind, 2013).[80]

The two-sided AdS-Schwarzschild wormhole is perhaps the best understood of all black hole-type systems, and it has justly taken a central role in many recent analyses of black hole physics. We will meet it again frequently in the remainder of this article.

H. The information problem in AdS/CFT

We now have all the pieces on the table, so we can return to the black hole information problem. I discuss it for small black holes in this section and big black holes in the following one. The thing to do is embed Hawking's original thought experiment of forming a black hole and watching it evaporate into AdS/CFT and see what happens. We need the black hole to be large enough to be semiclassical:

$$Er_{\mathrm{AdS}} \gg \left(\frac{r_{\mathrm{AdS}}}{\ell_p}\right), \tag{177}$$

but small enough to evaporate:

$$Er_{\mathrm{AdS}} \ll \left(\frac{r_{\mathrm{AdS}}}{\ell_p}\right)^{(d^2-1)/(2d-1)}. \tag{178}$$

Here Eq. (178) is the $(d+1)$-dimensional version of Eq. (158). Remember that it is stability in the microcanonical ensemble that decides whether or not a black hole of fixed energy evaporates. To create the black hole, we can act with the CFT creation operators we defined in Sec. VI.C to create an infalling spherical shell of matter that from the bulk point of view is expected to collapse into a black hole. We can then evolve this state forward in the CFT and see what it looks like after a time which is greater than the bulk evaporation time. This evolution is unitary, so to the extent that AdS/CFT is a definition of the bulk theory, this resolves the information problem in the sense of telling us the answer: information is preserved.

Because of the strongly coupled nature of the CFT it is difficult to actually compute the result of this evolution. But after we have evolved long enough for the CFT to thermalize there is a fairly simple argument that the state we get should typically have a bulk interpretation as a cloud of radiation in a pure quantum state with no significant projection onto any state whose semiclassical interpretation is not clear. Recall from Sec. VI.C that in the CFT we can produce multiparticle states by acting repeatedly on the vacuum with the Fourier modes of single-trace operators. Once we act with enough operators to get to energies of the order of Eq. (177) there will occasionally be states where the bulk wave packets are so close together that they collapse to form small black holes, but as long as we stay below the energy (178) we expect from the bulk discussion around Eq. (158) that states where the bulk radiation cloud is weakly coupled should be the most entropic.[81] I refer to such states as "radiation states." Classically if the set of radiation states has a larger entropy than its complement, we should expect to find ourselves in such a state after thermalization, but quantum mechanically it is a little less straightforward. The set of radiation states is closed under quantum superposition, so we can then write the CFT Hilbert space within some energy band as a direct sum of a "radiation" subspace whose states can be produced by acting with CFT creation operators from Sec. VI.C that are separated enough in the bulk to avoid a breakdown of bulk effective field theory, and its complement, which I will call the "black hole" subspace. We can then use our unitary integration technology to show that a typical state in this energy band has almost no projection onto the black hole subspace, or more carefully that a typical state is exponentially close in the trace norm to its projection onto the radiation subspace. As in our discussion of Page's theorem, we can write the typical state as

$$|\psi(U)\rangle = U|\psi_0\rangle, \tag{179}$$

where $|\psi_0\rangle$ is some reference state and U is chosen randomly from the Haar measure. The average trace norm distance (see Sec. V.C) between $|\psi(U)\rangle$ and its projection onto the radiation subspace then obeys

[80] This refers to the classic work of Einstein and Rosen on wormholes and Einstein, Podolsky, and Rosen on entanglement (Einstein, Podolsky, and Rosen, 1935; Einstein and Rosen, 1935).

[81] For $E\ell_p \ll 1$ this follows from our large N assumptions about the spectrum of low-dimension operators in the theory. In the intermediate energy regime we are considering here there is perhaps an additional assumption that the states we produce this way continue to dominate the CFT spectrum as suggested by the bulk. In some special cases we can confirm it directly [see, for example, Shenker and Yin (2011)], but for this argument to really be airtight we would need to be able to show this in general in the strongly coupled CFT. For the conclusion to fail however we would need to be totally wrong in this estimate, and missing a few states would not change anything.

$$\int dU \left\| |\psi(U)\rangle\langle\psi(U)| - \frac{1}{\langle\psi(U)|\Pi_{\rm rad}|\psi(U)\rangle}\Pi_{\rm rad}|\psi(U)\rangle\langle\psi(U)|\Pi_{\rm rad}\right\|_1$$
$$= 2\int dU \sqrt{\langle\psi(U)|\Pi_{\rm bh}|\psi(U)\rangle} \le 2\sqrt{\int dU \langle\psi(U)|\Pi_{\rm bh}|\psi(U)\rangle} = \frac{2e^{-(S_{\rm rad}-S_{\rm bh})/2}}{\sqrt{1+e^{-(S_{\rm rad}-S_{\rm bh})}}}.$$

Here $\Pi_{\rm rad}$ is the projection onto the radiation subspace, $\Pi_{\rm bh}$ is the projection onto the black hole subspace, and as in the proof of Page's theorem I used Jensen's inequality and the unitary matrix technology of Sec. IV of the Supplemental Material [193]. Since we have $S_{\rm rad} > S_{\rm bh}$ by assumption, and since if we are not right up against the energy scale (178), their difference will be fairly large. We see that for all practical purposes the pure quantum state in the CFT that results from collapsing a shell and then waiting for the system to equilibrate will almost always be "all radiation."

What this "before" and "after" analysis leaves unclear is what happened in the middle; we have not yet achieved Strominger's success criterion of actually computing the Page curve. Nonetheless using assumptions about the structure of the CFT Hilbert space, we have ruled out remnants and information loss.[82] This is most assuredly progress.

I. Unitarity for big AdS black holes

The discussion of the information loss problem in the previous section required a treatment of the CFT realization of the bulk effective field theory Hilbert space of the radiation gas. In particular, we needed an assumption about the CFT spectrum for $E\ell_p \gg 1$ which, although well motivated, we would prefer to do without. In fact, Maldacena pointed out that in the context of big nonevaporating AdS black holes there is still an issue analogous to the information loss problem, but which has a cleaner resolution in AdS/CFT (Maldacena, 2003a) [see also Barbon and Rabinovici (2003, 2014)].

Consider two CFTs entangled in the Hartle-Hawking state (176). The essence of the information problem is that according to bulk quantum field theory the correlation between an object that we throw in early and the radiation that comes out late vanishes as the radiation comes out later and later. What Maldacena pointed out is that this assertion can also be tested simply by considering the two-point function of two primary operators in one of the CFTs in the limit in which we take their time separation to be large. The idea is that in the naive bulk theory this two-point function, interpreted via the extrapolate dictionary (144) as the boundary limit of a two-point function of bulk fields, will decay exponentially for arbitrarily long times. By contrast in the CFT it will decay only until it is of the order of e^{-S}, after which it will undergo chaotic quasiperiodic behavior. The latter is characteristic of unitary evolution in a system with finite entropy, so the former is inconsistent with unitarity. Thus the bulk will reproduce the "coarse-grained" behavior of this CFT two-point function, but will not get the detailed late-time structure right; this lends considerable support to option (3) of Sec. IV.F. The goal of this section is to explain these statements in some more detail.

We first get some idea of what sort of time dependence is expected for quantum field theory correlation functions in the thermal ensemble. A full treatment of this subject requires analytic continuations along the lines of Sec. IV.H, but we can get to the main point from our existing results on the Rindler decomposition. Recall from Secs. III.C and III.D that the reduced density matrix in the right Rindler wedge is thermal, so if we study the correlation functions in the Minkowski vacuum restricted to the wedge then we can reinterpret them as thermal QFT correlators with respect to the Rindler time τ_R. In particular, consider the free massless scalar field in $3+1$ dimensions; the time-ordered two-point function is given by the $m \to 0$ limit of Eq. (27):

$$\langle\Omega|T\phi(t,x)\phi(t',x')|\Omega\rangle = \frac{1}{4\pi^2}\frac{1}{|x-x'|^2-(t-t')^2}. \quad (180)$$

If we take both points to lie in the right Rindler wedge, with $\tau_R = \tau$ and $\tau_R' = \xi_R = \xi_R' = \vec{y} = \vec{y}' = 0$, then using Eq. (40) we can rewrite this as

$$\langle\Omega|T\phi(\tau,0)\phi(0,0)|\Omega\rangle = \frac{1}{8\pi^2}\frac{1}{1-\cosh\tau}. \quad (181)$$

Thus we see that at large time separation in the thermal ensemble the correlation decays exponentially in time. Recall that in Eq. (40) we suppressed a length scale that sets the effective temperature for an observer at $\xi_R = 0$, so it is the temperature that sets the exponential decay constant. This is quite intuitive; whatever perturbation we put in at $\tau = 0$ will be rapidly thermalized, so it will be difficult to detect the perturbation with a single local operator at much later times.[83] Moreover we saw in Sec. VI.F the fields on the AdS-Schwarzschild background effectively all live in the thermal atmosphere, where as discussed in Sec. IV.A the Rindler wedge is a good model. So we should expect that correlation functions of boundary operators in the right exterior should have the same qualitative behavior of exponential decay at large timelike separation. This can be confirmed explicitly for

[82] The argument against remnants can be strengthened if we allow large black holes in AdS to evaporate by locally coupling the CFT to an external system (Rocha, 2008). In this case to preserve unitarity using remnants we would need arbitrarily entropic low-energy states in the CFT, which certainly do not exist.

[83] Note that conceptually this is a different exponential decay than the exponential decay with distance that we found for massive fields in the vacuum expectation value (26).

the special case of AdS_3 and with somewhat more difficulty in higher dimensions (Maldacena, 2003a).

Now consider what sort of behavior is expected for unitary systems with finite entropy. We are interested in "thermodynamic" systems, where the thermal entropy $S = -\mathrm{tr}\rho(\beta)\log\rho(\beta)$ is very large. Moreover we assume that the energy spectrum is densely spaced in the vicinity of the energy that dominates the canonical ensemble, with typical energy spacing of the order of e^{-S} times the temperature. The energy levels are distributed chaotically, with the only degeneracies arising from a small number of compact symmetries such as rotations, parities, etc. In such a system we want to understand the long-time behavior of quantities like

$$G(t) \equiv \frac{1}{Z}\mathrm{tr}[e^{-\beta H}\mathcal{O}(t)\mathcal{O}(0)] \tag{182}$$

$$= \frac{1}{Z}\sum_{ij} e^{-\beta E_i + i(E_i - E_j)t}|\mathcal{O}_{ij}|^2, \tag{183}$$

where $\mathcal{O}(t)$ is a Heisenberg picture operator and $\mathcal{O}_{ij} \equiv \langle i|\mathcal{O}(0)|j\rangle$, with $|i\rangle$ a complete basis of energy eigenstates. One simple way to do this is to compute the time average of $|G(t)|^2$ over some very long time T:

$$\frac{1}{T}\int_0^T dt|G(t)|^2 = \frac{1}{Z^2}\sum_{ij,i'j'} e^{-\beta(E_i+E_{i'})}|\mathcal{O}_{ij}|^2|\mathcal{O}_{i'j'}|^2 \times \left[\frac{1}{T}\int_0^T dt e^{i(E_i - E_j + E_{j'} - E_{i'})t}\right]. \tag{184}$$

As $T \to \infty$, the quantity in square brackets is equal to 1 when $E_i - E_j + E_{j'} - E_{i'} = 0$ and is 0 otherwise. Given our assumptions about the spectrum, it will be nonzero only if $E_i - E_j = E_{i'} - E_{j'} = 0$ or $E_i - E_{i'} = E_j - E_{j'} = 0$. Thus we see that the average is finite in the limit $T \to \infty$, which means that, unlike for the black hole correlation function, $G(t)$ cannot decrease monotonically to zero at late times.

If we do not know anything about $\mathcal{O}$ we cannot say much more about the late-time behavior of $G(t)$, but using assumptions we can estimate its typical size. First I assume that there is a symmetry which acts on $\mathcal{O}$ as $\mathcal{O} \to -\mathcal{O}$. This is more of a convenience than a necessity, but it is consistent with $\mathcal{O}$ being a CFT operator dual to a bulk field ϕ which has such a symmetry. What this buys us is that, perhaps after some reshuffling of the $|i\rangle$'s within degenerate eigenspaces of H, we have $\mathcal{O}_{ii} = 0$; this implies that the thermal one-point function of $\mathcal{O}$ is zero. We also want to ensure that $G(t)$ is "stable" under small changes, either of β or of the high-energy structure of the theory. Concretely consider

$$G(0) = \frac{1}{Z}\sum_{ij} e^{-\beta E_i}|\mathcal{O}_{ij}|^2. \tag{185}$$

We want this quantity to be of the order of 1 and to be a reasonably continuous function of β. For the sum over j to even converge we need $\mathcal{O}_{ij}$ to fall off sufficiently fast at fixed i and large E_j, and since we think of the operator as being a "probe" we can further demand that it falls off fast enough that the sum is dominated by the region where $E_j - E_i \ll \langle H\rangle = (1/Z)\mathrm{tr}He^{-\beta H}$.[84] Since even in this region of j there are of the order of e^S states, we need the individual $\mathcal{O}_{ij}$'s to be $O(e^{-S/2})$. Moreover to get a reasonable function of β we need the dependence of $|\mathcal{O}_{ij}|$ on i, j to be a reasonably smooth function of E_i and E_j; we make no similar restriction on the i, j dependence of its phase.[85] With these assumptions we then see that the time average (184) is of the order of e^{-2S}, and thus that the typical late-time behavior of $G(t)$ is of the order of e^{-S}. Over long time scales the correlator will fluctuate erratically, and over even longer time scales it will sometimes come back up to an $O(1)$ value.

How then are we to reconcile this with the endless exponential decay of Eq. (182)? The resolution of course is that, as we found in the brick wall model of Sec. IV.G, the entropy of the Rindler wedge is infinite due to the infinite collection of modes near the horizon in tortoise coordinates. By studying the correlation at very large τ separation, we take advantage of these degrees of freedom near the horizon. If the black hole entropy is actually finite, as it certainly is in AdS/CFT, then the exponential decay must eventually stop. That the CFT agrees with the black hole result for this correlation function at short times [this is shown in more detail in Papadodimas and Raju (2013)], but disagrees at long times as required by unitarity, is compelling evidence for the unitarity of black hole evaporation.[86]

[84] In fact these statements will not be true for local field operators, which do actually have significant matrix elements between low- and high-energy states. This is the origin of the short-distance singularities in their correlation functions, which here would say that $G(t) \to \infty$ as $t \to 0$. We can fix this by "smearing" the operators against wave packets whose size in space and time is of the order of β, which will not affect the late-time behavior that we are interested in.

[85] These properties of $\mathcal{O}_{ij}$ are sometimes referred to as the "eigenstate thermalization hypothesis," with the hypothesis being that for local Hamiltonians H the local operators $\mathcal{O}$ that we are interested in obey them (Deutsch, 1991; Srednicki, 1996). More generally if we do not assume there is an $\mathcal{O} \to -\mathcal{O}$ symmetry the eigenstate thermalization hypothesis says that

$$\mathcal{O}_{ij} = \mathcal{O}(E_i)\delta_{ij} + e^{-S[(E_i+E_j)/2]/2}f(E_i, E_j)R_{ij}, \tag{186}$$

where $\mathcal{O}(E)$ and $f(E, E')$ are real smooth functions of their arguments, but R_{ij} is a complex (both in the sense of "not real" and in the sense of "complicated") $O(1)$ function of i and j.

[86] Maldacena (2003a) also pointed out that subleading saddles in the path integral can give rise to long-time corrections which are of the order of e^{-S} as required by unitarity. This is sometimes misunderstood as an argument that including such saddle points is sufficient to resolve the information problem and reproduce the chaotic late-time CFT behavior, but this is most likely not the case (Barbon and Rabinovici, 2003). There are presumably other nonperturbative effects in quantum gravity beyond those suggested by Euclidean gravity, and it would be unreasonable to expect a semiclassical interpretation for all of them.

J. von Neumann entropy and the Ryu-Takayanagi formula

This concludes the background in AdS/CFT needed for the remainder of this paper, but there is one more aspect which deserves to be mentioned. This is the proposal of Ryu and Takayanagi (RT), later generalized by Hubeny, Rangamani, and Takayanagi (HRT), for a holographic expression for the von Neumann entropy of a subregion in the boundary theory (Ryu and Takayanagi, 2006; Hubeny, Rangamani, and Takayanagi, 2007). A vast literature understanding and using this conjecture has appeared in the last few years, so here I will state only the conjecture and mention two applications of interest to black hole physics.

The basic idea is as follows: take the CFT to live on $\mathbb{S}^{d-1} \times \mathbb{R}$, and in the Schrödinger picture pick out any particular Cauchy slice with topology $\mathbb{S}^{d-1}$ and study a quantum state of the CFT on that slice. A natural thing to do with the state is decompose the slice into a region A and its complement B and then compute the von Neumann entropy

$$S_A = -\mathrm{tr}\rho_A \log \rho_A. \tag{187}$$

The RT-HRT proposal says that to compute S_A, using the bulk gravity theory, look for the codimension two extremal-area surface Σ in the bulk with the property in which $\partial\Sigma = \partial A$.[87] If there is more than one such Σ, take the one of smallest area. The proposal is then that to leading order in $1/N$ we have

$$S_A = \frac{A(\Sigma)}{4G}, \tag{188}$$

where $A(\Sigma)$ is the area of Σ in the bulk geometry. The basic idea is illustrated in Fig. 25.

The RT-HRT proposal has passed many nontrivial checks, especially for $\mathrm{AdS}_3/\mathrm{CFT}_2$, where both sides can be computed explicitly in many cases (Faulkner, 2013; Hartman, 2013). It also obeys nontrivial properties of entropy like strong subadditivity (Headrick and Takayanagi, 2007; Wall, 2012b). Indeed a general "heuristic proof" was given by Lewkowycz and Maldacena (2013), although the range of validity of this argument is still being actively explored by the community. The RT-HRT conjecture has also recently been extended to bulk theories more general than Einstein gravity (Dong, 2014), and it also has been used to derive formulas for the areas of more general nonextremal bulk surfaces (Balasubramanian *et al.*, 2014; Czech, Dong, and Sully, 2014).

A word of warning: the RT-HRT proposal is often described as computing entanglement entropy, but this is misleading since it is supposed to work even if the total state is not pure.

From the point of view of this article, however, the most important applications of the RT-HRT proposal involve its use in spacetimes with black holes. Two especially interesting examples of this are its use to study the two-sided AdS-Schwarzschild geometry by Hartman and Maldacena (2013), and perturbations thereof, by Shenker and Stanford (2013, 2014). The calculation of Hartman and Maldacena builds on an observation of Morrison and Roberts (2012), who emphasized that at $t = 0$ in the Hartle-Hawking state (176) of two CFTs there is considerable "local entanglement." What this means is that if I take a suitable region A_L in the left CFT and its mirror region A_R in the right CFT, their mutual information $I(A_L, A_R) \equiv S_{A_L} + S_{A_R} - S_{A_L A_R}$ is nonvanishing, and in fact is proportional to the black hole entropy. This statement can be confirmed in the bulk by using the RT-HRT proposal (Morrison and Roberts, 2012). What Hartman and Maldacena studied is how this statement evolves as we evolve time forward simultaneously on the two sides.[88] Say that the linear size of the region is L, and that the temperature is large enough compared to the Hawking-Page transition that we can have $\beta \ll L \ll 1$ (this is what "suitable" means). Hartman and Maldacena then argued that the mutual information $I(A_L, A_R)$ starts out positive and decreases linearly with time, changing nontrivially over a time of order β, until at a time of the order of L it drops to zero. In $1+1$ dimensions they were able to also confirm this behavior directly in the CFT. Moreover they found that the extremal surface used in computing $S_{A_L A_R}$ over this time scale extends through the wormhole, with the details of the result depending on the metric in the interior. The successful matching with the CFT calculation provides good evidence that, at least in this special case, the CFT knows about the interior geometry, which should thus be taken seriously despite the paradoxes I describe in the following section.

The decrease of the mutual information $I(A_L, A_R)$ found by Hartman and Maldacena is reminiscent of thermalization, since it represents the "dilution" of a special property of the state at $t = 0$, the local entanglement, into the rest of the system as time evolves. Soon after Shenker and Stanford introduced a modification of this setup in which the connection to thermalization in the CFT can be made even more explicit. Their idea was to instead study how this special property of the state is affected by introducing a small perturbation to the system at an earlier time $t = -t_w$ (with

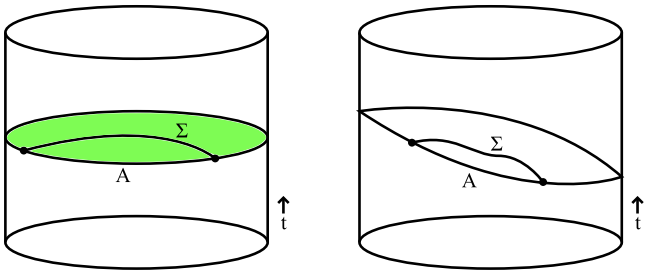


FIG. 25. Two examples of the RT-HRT proposal, shown in $\mathrm{AdS}_3/\mathrm{CFT}_2$ for simplicity. On the left there is a time-translation symmetry (the geometry is "stationary"), and I have chosen the spatial slice in the boundary to respect this. The extremal surface Σ is a geodesic of minimal area that lies in the green bulk Cauchy slice that also respects the symmetry. This was the setup originally studied by Ryu and Takayanagi. On the right I show a more arbitrary slice in a possibly nonstationary geometry. The extremal surface no longer lies in any preferred bulk slice.

[87]Technically we also require the fact that the surface Σ is homologous to the region A on the boundary (Headrick and Takayanagi, 2007).

[88]Note that evolving time forward on one side and backward on the other is a symmetry of the HH state (176), but evolving both sides forward is nontrivial.

$t_w > 0$). They modeled this in the bulk theory by sending in a spherical shell of matter on one side of the wormhole at this earlier time, whose energy was of the order of the temperature β^{-1}. They found that the local effect of this perturbation on the extremal surface at $t = 0$ grows exponentially in t_w, for essentially the same reason that the center of mass collision energy grew exponentially in our discussion of the trans-Planckian problem in Sec. IV.F, and thus that the backreaction of the shell becomes significant for the RT-HRT calculation at a time of the order of

$$t_w \approx \beta \log S. \tag{189}$$

This is nothing other than the scrambling time (129), which we argued in a rather different way in Sec. V.F is the relevant time scale for a black hole to absorb information. The effect of this backreaction is to make the wormhole "longer," which causes the mutual information to vanish since the minimal-area extremal surface used in computing $S_{A_L A_R}$ switches to a pair of surfaces that do not extend through the wormhole (this is also what causes it to drop to zero in the Hartman-Maldacena calculation). This is essentially a bulk illustration of the "butterfly effect" in the CFT evolution. We perturb a state at $t = -t_w$ which was carefully tuned to produce local entanglement at $t = 0$, but our perturbation prevents it from doing so. The black hole entropy S appears in the calculation since, unlike in the Hartman-Maldacena setup, we need to wait for the perturbation to be mixed throughout the system. It is gratifying to see such a chaotic effect in the CFTs emerging from a straightforward bulk calculation.

VII. PARADOXES FOR THE INFALLING OBSERVER

Having at least provisionally settled the information paradox, it is now time to return to the description of the black hole interior. In Sec. V.G we saw that unitarity of the evaporation process leads to a possible inconsistency in the description of the interior: a violation of the no-cloning theorem (Susskind and Thorlacius, 1994). We saw however that this cloning seemed to be unobservable (Susskind and Thorlacius, 1994; Hayden and Preskill, 2007), and thus considered the possibility that the principle of "black hole complementarity" could allow us to formulate a theory in which no observer sees a violation of quantum mechanics, avoiding both nonunitarity outside and cloning inside.

Many were reasonably satisfied with this state of affairs, but there were always some lingering doubts and, in particular, there was a sizable contingent who were never convinced (Unruh and Wald, 1995; Mathur, 2009; Giddings, 2012). If black hole complementarity is consistent and correct, should we not be able to find a real theory of the interior that realizes it? As we will now see, it seems to be the case that black hole complementarity as originally formulated cannot be consistent. We will also see in the following sections that there are a number of problems with any naive attempt to "reconstruct" the interior in AdS/CFT using the same machinery as we did for perturbations of the vacuum in Sec. VI.C. As of now the status of these arguments is somewhat controversial. One will notice a definite decrease in the precision of the arguments as we move in from the boundary, but they raise serious obstructions that one would need to address before claiming to possess a satisfactory theory of black holes.

A. The entanglement-monogamy problem

The argument against the consistency of complementarity I presented is due to Almheiri, Marolf, Polchinski, and Sully (2013), who I will refer to as AMPS, but its basic building blocks have a long history. Throughout the discussion of the information problem, there was some concern that changing the state of the evaporating modes from Hawking's result would be dangerous to the infalling observer (Giddings, 1994, 2012; Polchinski, 1995; Mathur, 2009; Avery, 2013). In particular, the main quantitative piece of the AMPS argument, based on the strong subadditivity of von Neumann entropy, is due to Mathur (2009). Many aspects of the argument were also independently realized by Braunstein, Pirandola, and Życzkowski (2013), who suggested the term "energetic curtain" for what AMPS later called a "firewall." The contribution of AMPS was to assemble these pieces and use them to attack complementarity in a concrete way. The argument presented in this section differs in detail from their argument, in particular, neither strong subadditivity nor a discussion of black hole mining (Brown, 2013) is needed, but the basic idea is the same.

The goal of the AMPS argument is to put all of the "moving parts" of the black hole information problem into the past light cone of a single observer, preventing any use of complementarity to avoid an observable violation of effective field theory or quantum mechanics. The basic setup is shown in Fig. 26. Rather than assuming that there is some well-defined state of the quantum fields on an entire nice slice such as the blue one in Fig. 20, we make the weaker assumption that there exists such a state only on the red slice that stays within the past light cone of an infalling observer Alice in Fig. 26. This is still a strong assumption, one that goes well beyond the asymptotic S matrix and boundary correlator assumptions used in this article, and it almost certainly will not survive in the correct theory of the interior. Nonetheless it has proven quite difficult

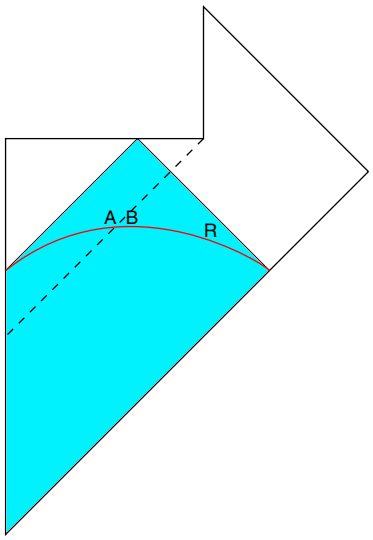


FIG. 26. A spatial slice in the diamond of an infalling observer.

to come up with consistent alternatives that are able to reproduce quantum field theory in the expected limits, so for now we make this assumption; it might be called "taking the quantum mechanics of effective field theory seriously." We will not need to assume anything about the dynamics of how different such time slices are related. We can get into trouble just by trying to find any quantum field theory state for this slice that is consistent both with the horizon being smooth for an infalling observer and with unitarity.

To proceed, in Fig. 26 I have indicated three important sets of modes. The mode B is one of the thermally occupied Schwarzschild modes (59) in the atmosphere, confined between the horizon and the barrier of the effective potential (61) (I have assumed here that we have waited long enough after the black hole formation that the geometry in the vicinity of B is well approximated by the upper corner of the Schwarzschild geometry). The mode A is the "mirror" of B behind the horizon. It is analytically continued up from right-moving modes near the horizon of the left exterior region of the two-sided Schwarzschild geometry in Fig. 1. If the horizon is smooth, we expect A and B to be highly entangled as in Eq. (45). If not then at least in quantum field theory we would expect something singular to happen at the horizon, as explained in Sec. III.E.[89]

You might think that unitarity does not impose any significant constraint on the colored slice of Fig. 26; after all this observer never gets to measure the S matrix, so what would unitarity mean operationally? One answer is provided by Page's theorem. Once the black hole is sufficiently old, we saw in Sec. V.C that we should expect a considerable degree of entanglement between the black hole BH and its early radiation R. In fact we saw that, ignoring energy conservation, they should be maximally entangled. Once we include energy conservation what should be true instead is that BH will typically be thermally entangled with R, in the sense that $\rho_{\mathrm{BH}} = (1/Z_{\mathrm{BH}})e^{-\beta H_{\mathrm{BH}}}$. Moreover as discussed in Sec. III.B.3 of the Supplemental Material this means that, provided that B is sufficiently weakly interacting with the rest of the black hole, it can be purified by some tensor factor R_B in the early radiation in the sense that ρ_{BR_B} is close to a pure quantum state. This is a statement that lies within the infalling observer's diamond; it must then be a property of the state on the colored slice of Fig. 26. This is what we need to form a contradiction. If B is strongly entangled with R_B, then it cannot also be significantly entangled with A without violating a principle: called monogamy of entanglement (Koashi and Winter, 2004).

We can illustrate the contradiction more precisely using some entropy inequalities from Sec. III of the Supplemental Material [193]. We begin by assuming that B and R_B are thermally entangled, so that $S_{BR_B} = 0$. From the non-negativity of the mutual information $I_{A,BR_B} \equiv S_A + S_{BR_B} - S_{ABR_B}$ and the triangle inequality (40), we see that

$$S_A = S_{ABR_B}, \tag{190}$$

and thus that

$$I_{A,BR_B} \equiv S_A + S_{BR_B} - S_{ABR_B} = 0. \tag{191}$$

We saw in Sec. III of the Supplemental Material that the mutual information between two tensor factors vanishes only if the state is a product state, so

$$\rho_{ABR_B} = \rho_A \otimes \rho_{BR_B}, \tag{192}$$

which is incompatible with any entanglement, or indeed correlation of any kind, between A and B. Conversely we could assume that A and B are thermally entangled, in which case there is no correlation between B and R. We thus seem to be led to the conclusion that either unitarity is violated or sufficiently old black holes have singular horizons; this is the firewall paradox.

One issue with this argument is whether or not B really is "sufficiently weakly interacting" with the rest of the black hole to allow us to approximate the thermal density matrix ρ_{BH} as a product between a thermal density matrix for B and a thermal density matrix for the rest of the black hole. This is not a trivial point. We saw in Sec. III.E that gradient interaction between the two Rindler wedges leads to a high "energy cost" for firewalls, which causes them to be Boltzmann suppressed in a thermal distribution defined with respect to the Minkowski Hamiltonian H. The difference here however is that the asymptotic time-translation symmetry of Minkowski space, which is conjugate to the conserved energy, behaves in the vicinity of the Schwarzschild horizon as a boost K rather than H. From our Eq. (29) for the quantum field theory boost operator, we see that the gradient interactions between the two sides are suppressed by an extra factor of x at the origin, which prevents them from being as restrictive. Alternatively we saw in Sec. IV.G that modes with Schwarzschild energy of the order of the temperature can be thought of as living in a box of size $\log(r_s/\ell\ell_p)$ in the tortoise coordinate r_*. This allows us to form wave packets of narrow frequency that are nonetheless well localized away from the horizon, so any contribution from the interactions near the horizon is suppressed appropriately. Nonetheless I think it is worthwhile to understand these energetics better, especially in the context of a careful treatment of the Hamiltonian formulation of canonical gravity, but I leave this to future work.[90]

[89] One might worry that we have evolved the modes up from the bifurcate horizon in Fig. 1, which is not part of the diagram in Fig. 26, but the Hamiltonian is conserved by this evolution so we can recast the discussion of Sec. III.E entirely as a calculation at some later time where the two relevant modes are given by A and B in Fig. 26.

[90] Even if there is some subtlety in general with the energetics of B, it seems unlikely that it would be so radical as to impose entanglement across the horizon for the lowest ℓ modes. These are the ones which directly carry out the information, and it is difficult to see how the evaporation could be unitary if they do not come out entangled with R. For this reason some sometimes contemplate an "s-wave firewall," which corrupts only those B modes with $O(1)$ angular momentum. This has some resonance with the second energetic argument just given, since for sufficiently large ℓ we cannot think of $\log(r_s/\ell\ell_p)$ as being large; perhaps energetics prevent a full firewall but allow an s wave one?

B. Firewall typicality

The AMPS argument of the previous section does not apply to big asymptotically AdS black holes; they do not evaporate. One way to deal with this is to enable them to evaporate, either by coupling the CFT to an external system or by "mining" them, but it would be better to have a version of the paradox that does not require evaporation. Indeed the only thing we really needed the evaporation for in the previous section was to argue that the black hole was in a thermally mixed state. At least if we are playing by the usual rules of quantum mechanics, and assuming some crude form of locality, then any local experiment in the vicinity of the black hole should not care whether this mixed state is purified by the early Hawking radiation.[91] If we instead interpret the thermal density matrix as a classical ensemble of pure states, then the energetic argument of the previous section suggests that firewalls are "typical" in this ensemble. Unlike in ordinary Minkowski space they are not energetically suppressed (Bousso, 2013b).

One can make this "typicality" argument more precisely for big AdS black holes (Marolf and Polchinski, 2013). Consider again our Schwarzschild mode B, which has a frequency of the order of the temperature and is localized away from the horizon. Since we are now considering big AdS black holes, meaning for simplicity black holes that are stable in the canonical ensemble and thus have a Schwarzschild radius at least of the order of the AdS radius, the mode B will automatically be in the thermal atmosphere ("the zone"), since as we saw in Sec. VI.F this fills the entire space. The AdS/CFT description of such black holes is as excited states of a conformal field theory quantized on a spatial sphere cross time. We also saw in Sec. VI.F that to leading order in $1/N$ we can plausibly interpret the operator Fourier transform $\mathcal{O}_{\omega\ell m}$ of a single-trace primary $\mathcal{O}$ dual to a bulk field ϕ as the annihilation operator for a bulk Schwarzschild mode such as B, and, in particular, we can define a number operator

$$N_B \equiv \mathcal{O}^\dagger_{\omega\ell m}\mathcal{O}_{\omega\ell m} + O(1/N). \tag{193}$$

The operator $\mathcal{O}^\dagger_{\omega\ell m}\mathcal{O}_{\omega\ell m}$ exactly commutes with the CFT Hamiltonian, so to leading order in $1/N$ we can simultaneously diagonalize N_B and H. This is rather problematic, since it implies that we can find a complete basis for the microcanonical ensemble (the subspace of all states built from energy eigenstates within a narrow energy width centered at some large energy E_0) with the property that each basis element is an eigenstate of the occupation number for B. Such states are very far from the expected entanglement of Eq. (45), so they cannot be expected to locally resemble the Minkowski vacuum.

We now argue that the existence of a complete basis of "bad" states implies that almost all pure states sampled from the microcanonical ensemble are bad, but to really make this argument we need to make one final assumption based on the linearity of quantum mechanics. Namely, we assume that in addition to the existence of a CFT operator $\mathcal{O}_{\omega\ell m}$ which annihilates B, there is also a CFT operator $\tilde{\mathcal{O}}_{\omega\ell m}$ which annihilates A. Unlike $\mathcal{O}_{\omega\ell m}$ we do not have an explicit expression for this operator, but it is quite natural to assume it exists. After all the occupation number for A is an observable and thus it should correspond to a self-adjoint operator according to the general principles of quantum mechanics. Given this operator we can then define the annihilation operator

$$c_{1\omega\ell m} = \frac{1}{\sqrt{1-e^{-\beta\omega}}}(\mathcal{O}_{\omega\ell m} - e^{-\beta\omega/2}\tilde{\mathcal{O}}^\dagger_{\omega\ell(-m)}), \tag{194}$$

which we saw in Secs. III.D and III.E can be used as "diagnostic" for firewalls. Indeed we saw that its number operator $c^\dagger c$ will annihilate any state with a smooth horizon, whereas it will have $O(1)$ expectation value in a disentangled state. If we now compute the expectation value of $c^\dagger c$ in the microcanonical ensemble, however, we are free to use the basis of N_B eigenstates. In this case the average will be $O(1)$. Moreover it is a general fact [see Sec. II.D of Harlow (2014)] that the microcanonical expectation value of some operator is exponentially close to the expectation value of the same operator in a randomly chosen pure state from the ensemble. Thus $c^\dagger c$ will have an $O(1)$ expectation value in almost all pure states. Since this argument can be applied to any mode, firewalls that corrupt all B modes are typical (Marolf and Polchinski, 2013).

This argument has several technical points which have not been completely explicated in the literature. In particular, I am not really sure that the $1/N$ corrections are under control, and I am also not totally sure about the bulk interpretation of $\mathcal{O}_\omega$. Nonetheless so far no decisive objection has been raised, and even if this argument does not end up being correct it will be illuminating to understand why it fails.

One aspect of the argument of this section which is less satisfying than the original AMPS argument of the previous section is that it is less operational. There is no low-energy experiment that illustrates the paradox. I will return to this in Sec. VIII.A.

C. The creation operator problem

The paradoxes of the previous two sections arose from taking the quantum mechanics of bulk effective field theory seriously. In this section and the next, I will describe two additional arguments that suggest that one might not want to do this.

I first focus on the mode A, which lives just behind the horizon. One way to think about our construction of the CFT representation of creation and annihilation operators for the mode B is that it proceeds by solving the bulk equations of motion in from the boundary, with the boundary conditions given by the extrapolate dictionary (144) (Heemskerk *et al.*, 2012). A similar evolution for the mode A however would involve propagating either forward to the singularity or backward through the trans-Planckian bulk collision with the infalling shell, as described in Sec. IV.F. This means that

[91] We will see some tension with this statement in Sec. VII.D, as well as in Sec. VIII.B.

any attempt to produce a CFT expression for creation or annihilation operators for A via the same method will require an understanding of trans-Planckian bulk physics (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013).

In fact there is a more basic problem with finding a CFT expression for this mode. Consider a candidate raising operator $\tilde{\mathcal{O}}^\dagger$ for A. Within effective field theory we expect this operator to lower the energy of the quantum state. Remember that the Schwarzschild isometry acts within the vicinity of the horizon as a boost, and the A mode has negative boost energy. Moreover within effective field theory there are no states that this operator can annihilate. If we assume that these two properties are true for the CFT operator $\tilde{\mathcal{O}}^\dagger$, we reach a contradiction. The density of states of the CFT decreases as we go to lower energy, so it is impossible for an operator that lowers the energy on all states to not have any states it annihilates. Thus we seem to have an obstruction to a naive representation of creation operators for the A modes within the CFT (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013).

It is not clear to me that constructing the interior really requires an operator with the assumed properties. As mentioned previously I want to see a more careful treatment of energetics and bulk diffeomorphism invariance, but it is interesting to note that this argument seems to oppose the firewall typicality argument of the previous section. It says that one of the assumptions of that argument, the existence of a $\tilde{\mathcal{O}}$ operator in the CFT with the expected properties from effective field theory, cannot be true. On some level this is encouraging; it means that rather than accepting firewalls we should look for a less naive description of the interior.[92]

D. The Marolf-Wall paradox

Another interesting obstruction to a naive "inclusion" of effective field theory into the CFT was pointed out by Marolf and Wall (2013) [see also Avery and Chowdhury (2014)]. The idea is as follows: consider a single large-N CFT in a thermal density matrix

$$\rho_{\mathrm{CFT}} = \frac{1}{Z} e^{-\beta H}, \tag{195}$$

with $\beta > 1$ so that black holes dominate the ensemble. In ordinary quantum mechanics, we are free to interpret this state either as a classical probability distribution for energy eigenstates or as being purified by an auxiliary system. But for the black hole interior it seems like there is a difference between the two cases.

Say that we view this density matrix as a probability distribution for one-CFT states. According to the rules of quantum mechanics there should be linear operators on the single CFT whose expectation values we can compute to see what is going on in the interior. But alternatively say we view this density matrix as being purified by a second copy of the same CFT, and moreover say that we choose the joint system to be in the thermofield double state (176). As described in Sec. VI.G, this system is usually interpreted as describing the two-sided AdS-Schwarzschild geometry, with the two sides connected by a wormhole. For definiteness we can interpret the left CFT as the "auxiliary" one and the right CFT as the one we started with. There is now something of a paradox; in the two-sided system there is nothing to stop us from acting on the left CFT with a unitary operator that sends a signal into the wormhole. An observer on the right side could jump into the wormhole and receive this signal. But if he or she uses the right-CFT operators we just motivated in the single-CFT case, their expectation values will be identical whether or not we send a signal from the left. Any dictionary for the interior which could detect this signal would need to involve operators from both CFTs.

What are we to make of this? Marolf and Wall suggested that we need additional degrees of freedom beyond the two CFTs in describing this setup. They wanted to use these extra degrees of freedom to distinguish between two black holes which just happen to be entangled with each other but do not share a common interior (interpretation one of the previous paragraph) and two entangled black holes connected by a wormhole (interpretation two). I find it more natural to say that there are just two different interpretations of the two-CFT system, one which assumes there is a bridge and the other which does not. The Hamiltonian and Hilbert space are the same in both cases but the dictionary for observables is different. It is interesting that we seem to have this choice in defining the dictionary. There does not seem to be an analogous ambiguity for observables that are not behind horizons, since they can always be unambiguously evolved back to the boundary and matched onto the extrapolate dictionary.

VIII. PROPOSALS FOR THE INTERIOR

We have now seen that there are several interesting obstructions to a quantum description of the black hole interior. In this section I will describe what I consider to be some of the more promising ideas that have been proposed to resolve the paradoxes of the previous section. Since none of these ideas are unambiguously successful, I will be brief.

A. Complementarity from computational complexity?

The AMPS argument of Sec. VII.A presented a thought experiment testing the unitarity of black hole evaporation that appears to be doable by a single observer without violating causality. Could it be however that there is some principle other than causality that prevents the experiment from being done? If so, then it may be that some version of black hole complementarity as described in Sec. V.G may yet provide an escape from the apparent inconsistency of unitary evaporation and smooth infall without requiring an observable breakdown of some principle of physics. Several reasons why the AMPS experiment might be

[92]This objection does not invalidate the argument that there is a complete basis of N_B eigenstates spanning the microcanonical ensemble.

impossible have been proposed (Harlow, 2012; Freivogel *et al.*, 2014; Hui and Yang, 2014; Ilgin and Yang, 2014), but the most robust is that the computational complexity of "distilling" the purification R_B from the Hawking radiation is so severe that it almost certainly requires an exponential amount of time in the entropy of the black hole (Harlow and Hayden, 2013). This far exceeds the evaporation time, which is only of the order of $S^{3/2}$ in Planck units, so the infalling observer of Fig. 26 will not succeed in extracting R_B until long after the black hole has evaporated and she can no longer jump in and see a contradiction. In the remainder of this section I explain why this is probably the case and then briefly comment on to what extent this is a satisfactory resolution of the paradoxes of the previous section.

To be concrete we can model the old black hole of Sec. VII.A as a qubit system, with the mode B taken to be a single qubit, its complement H in the black hole taken to be m qubits, and the radiation R being a further n qubits. The black hole will be old in the Page sense if $n \gg m$. We can take the state of the system to be a random pure state of BHR, which by Page's theorem will be maximally mixed on BH. By the Schmidt decomposition of Sec. III in the Supplemental Material [193], we can represent the state as

$$|\psi\rangle = \frac{1}{\sqrt{2|H|}} \sum_{bh} |b\rangle_B |h\rangle_H U_R |bh0\rangle_R, \tag{196}$$

where b and h label convenient bases for B and H, and we have chosen a local basis for the radiation in the sense that a state like $|10110\cdots\rangle_R$ is analogous to a state where all the radiation modes have definite occupation number. U_R is some unitary transformation on the radiation that relates this basis to the natural basis of the Schmidt decomposition where the entanglement of the state $|\psi\rangle$ is manifest. U_R is defined only up to an arbitrary unitary on the orthogonal complement of the subspace spanned by $|bh0\rangle$. In order to observe a violation of entanglement monogamy along the lines of the AMPS experiment, our infalling observer must first use a quantum computer to act with $U_R^\dagger$ to "distill" the purification of B in the radiation into an easily usable form. We want to assess the "difficulty" of doing this, as defined using the quantum circuit model described in Sec. V of the Supplemental Material [193].

The first obvious objection to any claim that implementing $U_R^\dagger$ requires a time that is exponential in n is that the black hole certainly does not require exponential time to produce the state (196). Indeed we should probably accept the fact that there exists a polynomial-sized circuit U_{dyn} with the property that acting on the state $|0\rangle_B |0\rangle_H |000\rangle_R$ it produces the state $|\psi\rangle$. This amounts to the quite plausible assumption that quantum gravity can be "efficiently simulated."[93] The problem however is that even if we have available a polynomial-sized circuit for U_{dyn}, we cannot use its inverse to decode the Hawking radiation, since that would only work if we are also able to act on the degrees of freedom H that remain inside the black hole [this is explained in more detail in Harlow and Hayden (2013)]. The argument above Eq. (196) does ensure the existence of a distilling $U_R^\dagger$ that acts only on the radiation, but it is nonconstructive and gives no information about the complexity of this distillation. Without a construction, we are left with the basic fact reviewed in Sec. V of the Supplemental Material that almost all unitaries on n qubits require a circuit whose size is of the order of 2^{2n} to implement.

Of course it could still be the case that we can somehow use the simplicity of U_{dyn} to argue in a more complicated way that there must be a polynomial-sized circuit for U_R, but we were able to give a complexity theoretic argument that this is probably not the case. We can phrase this as a general question about quantum circuits.

- *Hawking distillation problem:* Say we are given a product Hilbert space of three-qubit systems B, H, and R, along with a polynomial-sized circuit U_{dyn} that prepares a quantum state $|\psi\rangle = U_{\text{dyn}}|0\rangle$, where $|0\rangle$ is the product state that is all 0's for all three factors, and moreover say that $|\psi\rangle$ has the property that it is maximally mixed on BH. Also say that the dimensionality $|B|$ is some $O(1)$ number. Does there exist a small quantum circuit $U_R^\dagger$, meaning a circuit whose size is at most polynomial in $m = \log_2 |H|$ and $n = \log_2 |R|$, that distills the purification of B, in the sense that

$$U_R^\dagger |\psi\rangle \approx \frac{1}{\sqrt{|B||H|}} \left(\sum_b |b\rangle_B |b\rangle_{R_B} \right) \times \left(\sum_h |h\rangle_H U_R' |h0\rangle_{R_{\bar{B}}} \right). \tag{197}$$

  I have here allowed for some remaining scrambling U_R' of the purification of H, although this can be removed by a basis redefinition of H. The approximation should be understood as saying the two states are close in the trace norm.

We argued based on the assumed hardness of a quantum complexity class called QSZK (quantum statistical zero knowledge that the answer to this question is typically no, but I will instead give a more elegant pair of arguments due to Scott Aaronson that lead to the same conclusion.[94] As with most complexity theoretic arguments, it is too difficult to directly prove that no efficient distillation $U_R^\dagger$ exists. So what one does instead is show that if it did exist, this would enable us to do something which is widely expected to be difficult.

The difficult task Aaronson uses is the inversion of "one-way functions," meaning functions that are easy to evaluate but difficult to invert. It is not *a priori* clear that such functions should exist, but they are widely expected to; indeed almost all of modern cryptography is based on their

[93] In the cases where we really understand it, such as the BFSS matrix model or AdS/CFT, this seems likely to be the case (Feynman, 1982; Lloyd *et al.*, 1996; Jordan, Lee, and Preskill, 2011).

[94] The first argument is also explained at http://www.scottaaronson.com/talks/hawking.ppt, the second came out of a recent discussion with Aaronson.

assumed existence (Arora and Barak, 2009). In quantum language we can think of a one-way function as a map f from m-bit strings to n-bit strings, with $n \geq m$, such that there exists a polynomial-sized circuit U_f, where

$$U_f|x,0\rangle = |x,f(x)\rangle, \tag{198}$$

but no polynomial-sized circuit $U_{f^{-1}}$ such that

$$U_{f^{-1}}|0,f(x)\rangle = |x,0\rangle. \tag{199}$$

Note that although U_f is invertible, we cannot use its inverse to invert the function. If we give it $|0,f(x)\rangle$ it will not necessarily do anything useful.[95] What Aaronson was able to do was show that if our Hawking distillation problem has a positive answer, then one can use that answer to efficiently invert any candidate injective one-way function; in other words no injective one-way function could exist.[96] This may not sound so bad to physics readers, but in computer science and cryptography it would be a genuine catastrophe (or revelation) of almost a similar order of magnitude as a proof that $P = NP$ (Impagliazzo, 1995).

Indeed say we are given a candidate injective one-way function f on m-bit strings and an efficient circuit U_f that implements it, as in Eq. (198). It is not difficult using the Hadamard and CNOT gates of Sec. V of the Supplemental Material [193] to come up with a polynomial-sized quantum circuit which, given the all zero state of BHR, prepares the state

$$|\psi\rangle = \frac{1}{\sqrt{2|H|}}\sum_h |h\rangle_H(|0\rangle_B|0,h,0\rangle_R + |1\rangle_B|1,f(h)\rangle_R). \tag{200}$$

Now assume that the Hawking distillation problem has a positive answer: there then must exist an efficient distilling unitary transformation $U_R^\dagger$ which acts as

$$U_R^\dagger|0,h,0\rangle_R = |0,g(h)\rangle_R, \tag{201}$$

$$U_R^\dagger|1,f(h)\rangle_R = |1,g(h)\rangle_R, \tag{202}$$

where $|g(h)\rangle$ is some given set of 2^m states of the radiation minus a qubit. But now by combining U_R and $U_R^\dagger$ with the X operation that flips the first qubit we can directly construct an efficient circuit $U_{f^{-1}}$ which implements Eq. (199), contradicting the one-way nature of f.[97] Thus the Hawking distillation problem is tractable in general only if injective one-way functions do not exist. Since there is strong evidence that they do exist, Hawking distillation is most likely a hard problem.

A subtlety with this argument however is that, although we need to do something hard to distill manifest entanglement, the state (200) already exhibits *classical correlation* with the radiation in the sense that its mutual information with the first qubit of R is nonzero. This is in fact already sufficient to prevent maximal entanglement between B and something else, so a slight modification of the AMPS experiment is sufficient to argue that the state (200) cannot have a smooth horizon even if we cannot distill manifest entanglement. What we want to argue is that in the setup of the Hawking distillation problem, even distilling classical correlation between B and R is usually exponentially difficult. In fact a small modification of Aaronson's argument, again due to Aaronson, is able to resolve this. Namely, we consider instead the state

$$\frac{1}{\sqrt{2|H|}}\sum_{h_1,h_2}|h_1,h_2\rangle_H(|h_1\cdot h_2\rangle_B|f(h_1),h_2,0\rangle_R + |h_1\cdot h_2+1\rangle_B|f(h_1),h_2,1\rangle_R), \tag{203}$$

where we split H into two equal size pieces H_1 and H_2. Here f is a candidate injective one-way function on $m/2$ bits, and by $h_1\cdot h_2$ I mean the inner product of the two strings, computed mod 2. It is not difficult to see that this state satisfies the conditions of the Hawking distillation problem. It can be simply generated using a circuit U_f that computes f as in Eq. (198), and it is maximally mixed on BH. Now imagine that there exists an efficient circuit $U_R^\dagger$ that is able to distill a bit of R that is classically correlated with B in the sense that their mutual information is close to log 2 (or close to 1 if we define the entropy with a base two logarithm). We must then have

$$U_R^\dagger|f(h_1),h_2,0\rangle_R = |h_1\cdot h_2, g(h_1,h_2)\rangle_R,$$
$$U_R^\dagger|f(h_1),h_2,1\rangle_R = |h_1\cdot h_2+1, g'(h_1,h_2)\rangle_R, \tag{204}$$

which means that using $U_R^\dagger$ to determine $h_1\cdot h_2$ given $(f(h_1),h_2)$. This however again allows us to invert the function (Goldreich and Levin, 1989); for example, given $f(h_1)$ we can determine the first bit of the input by choosing h_2 to be 1 for the first qubit and 0 for the rest. Thus the assumed existence of one-way functions also ensures the exponential difficulty of distilling classical correlation.

Given these arguments, it seems highly likely that computational complexity prevents an AMPS experiment from being done. What then are we to conclude? One option would be to declare victory for black hole complementarity and move on; by the standards of the mid-1990s this would be what we would do. What the past few years have taught us however is that those

[95] In fact, the standard definition of one-way functions is slightly weaker than this. It demands only that no efficient algorithm exists which can invert the function on a fraction of inputs which at most are polynomially small in m. Since we will see that solving Hawking distillation cracks the stronger version this distinction is not relevant for our purposes, although it could be if we tried to define a version of Hawking distillation that allowed for "imperfect distillation."

[96] Injectivity is not a very strong restriction; if $n > 2m$ then we avoid the birthday paradox and the function f will typically be injective.

[97] More carefully it implements it with the assistance of a single extra ancillary qubit, which does not increase the complexity in any practical sense.

standards were too low. Without an actual theory of the interior we cannot be sure that complementarity actually provides the mechanism whereby unitarity is made consistent with a smooth experience for the infalling observer. One problem is that somebody might later come up with another thought experiment that cannot be resolved in the same way, and then we would be back in the soup.[98] Even if they do not, however, there are basic physical questions such as "what happens if we form and evaporate a little black hole inside of a big one?" that seem to really require a theory of the black hole interior that goes beyond effective field theory. It is only after we find this real theory of the interior that we will be able to see whether or not limitations from computational complexity play an important role in its consistency; at best they currently can be thought of as a tool to use in looking for that theory. In the historical analogy suggested in Sec. V.G, it could be that our computational complexity restrictions on experiments are analogous to the uncertainty principle, but if so then we have not yet discovered the theory analogous to quantum mechanics. Until we do, there will always be stodgy classical physicists saying "of course you can measure the position and momentum, you just need to think more about how to do it."[99] In particular, the firewall typicality arguments of Sec. VII.B are not expressed as operational contradictions accessible to a low-energy thought experimenter in the bulk and thus are not immediately addressed by complexity theoretic concerns. Perhaps these arguments are resolved by noting that they are based on assumptions that are analogous to asserting that a particle has both a position and a momentum, despite the unobservability of this notion. To confirm (or refute) this however, we need a theory.

B. Nonlinearity?

I now turn to proposals that, unlike complementarity, give some sort of positive prescription for what the theory of the interior might be.

One idea for evading the entanglement-monogamy paradox is to declare that in the theory of quantum gravity the interior mode A is simply defined to be "whatever is entangled with B." In the language of Sec. VII.A this is sometimes called $A = R_B$. Versions of this idea have been proposed in many places (Bousso, 2013a; Harlow and Hayden, 2013; Papadodimas and Raju, 2013; Verlinde and Verlinde, 2013b; Maldacena and Susskind, 2013). I have recently written a general discussion of what I consider to be the best defined of the options, that of Raju and Papadodimas (Verlinde and Verlinde, 2013a; Papadodimas and Raju, 2014a, 2014b); see there for more details (Harlow, 2014). In this section I make a few general points about the features of these proposals in the context of a simple qubit model.

Before beginning, however, it may seem crazy that we should think of the interior of an old black hole as having anything to do with the distant Hawking radiation that has already been emitted. We know that there is some type of nonlocality in holography, but this may seem to "just be too much." An argument in this regard was emphasized by Maldacena and Susskind, and by Van Raamsdonk, who suggested taking the Hawking radiation cloud of a black hole that has just evaporated halfway and collapsing it into a second black hole that is thermally entangled with the first in the thermofield double state (Maldacena and Susskind, 2013; Van Raamsdonk, 2013). From Sec. VI.G it then seems at least somewhat plausible that the two black holes will be connected by a wormhole. Assuming that this is the case, from the Penrose diagram in Fig. 4 (or Fig. 23) we can affect the interior of the original black hole by throwing things into the other black hole. It thus seems that at least there is one unitary operation we can do on the Hawking radiation which directly affects the interior.[100] Given that this is the case, it does not seem so unreasonable to try to look for a general theory of the interior that allows us to use whatever the black hole is entangled with as part of the theory.

In trying to turn this observation into a theory of the interior however one runs into serious problems that so far have resisted a clear resolution. For simplicity I will illustrate them in the simple qubit model of the previous section, where we think of the factors B, H, and R as qubit systems. I will again model B as a single qubit, so by the

[98] Indeed there have been several attempts to come up with thought experiments that evade our computational constraints. Some of these involve direct manipulation of the microscopic degrees of freedom from the outside of the system and thus cannot be performed by observers who are actually part of the system and are restricted to local operations allowed by perturbative semiclassical physics (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013), while others require the ability to precisely manipulate the degrees of freedom in the atmosphere without introducing any decoherence (Oppenheim and Unruh, 2014). At the moment I do not find either proposal convincing. It is not clear what we should expect to come out of direct manipulation of UV degrees of freedom, indeed because of the fundamental nonlocality in holography I expect that large scale action at a distance is possible in the bulk by such operations, so asking for a semiclassical description of such an experiment may well be futile. The second type of proposal requires large amounts of machinery in the vicinity of the black hole, and it is hard to see how this could be possible without introducing at least some decoherence. Oppenheim and Unruh (2014) suggested a mechanism to "correct" the decoherence, but implementing it again seems to require an exponentially long quantum computation.

[99] One direction for a more "positive" approach to complementarity is called "strong complementarity," and operates under the basic assumption that different observers have different Hilbert spaces and observables, with the only consistency condition being that they must agree on the results of experiments that they can both do (Banks and Fischler, 2012; Harlow, 2012; Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013; Bousso, 2013a; Harlow and Hayden, 2013). So far this idea is somewhat ambiguous and no precise version of it has appeared.

[100] We can make this more precise by starting with a large AdS black hole in a CFT and allowing it to evaporate into a second copy of the CFT by coupling the two through a simple local coupling at the boundary as in Rocha (2008). By then manipulating the second copy we can prepare the thermofield double state of the two CFTs, which is quite plausibly described in the bulk by a wormhole connecting the two asymptotically AdS regions.

Schmidt decomposition I can represent a typical state of the system as

$$|\psi_+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_B|\tilde{0}\rangle_{\rm HR} + |1\rangle_B|\tilde{1}\rangle_{\rm HR}). \tag{205}$$

I will also imagine that the "smooth" horizon state for A and B is $|00\rangle_{AB} + |11\rangle_{AB}$, and I have labeled the states of HR appearing in the Schmidt decomposition appropriately. The problems arise from the interpretation of the three other states

$$|\psi_-\rangle \equiv \frac{1}{\sqrt{2}}(|0\rangle_B|\tilde{0}\rangle_{\rm HR} - |1\rangle_B|\tilde{1}\rangle_{\rm HR}),$$
$$|\chi_\pm\rangle \equiv \frac{1}{\sqrt{2}}(|0\rangle_B|\tilde{1}\rangle_{\rm HR} \pm |1\rangle_B|\tilde{0}\rangle_{\rm HR}). \tag{206}$$

If we are trying to define the interior modes using whatever B is entangled with, then since all of these states have maximal entanglement it is tempting to look for a definition where they all "look smooth"; after all they are just as typical as the state $|\psi_+\rangle$ that we started with. We then find ourselves confronted by the following issues (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013; Bousso, 2014; Harlow, 2014) [see also Bousso (2013a) and Chowdhury (2013)].

- *Nonlinearity*: By taking simple superpositions of these four states we can produce states where B is pure and unentangled with anything; in fact there is a complete basis of such states. We saw in Sec. VII.B that it is impossible for such states to have unexcited horizons. This means that we cannot view "unexcitedness" as a conventional observable realized by a self-adjoint operator on the Hilbert space. If it were then since all four of these states are unexcited the "unexcitedness operator" is the identity on this subspace and any superposition must also be unexcited, contradicting the fact that there are excited states. There is thus a basic tension between wanting all four states to have a smooth horizon and the linearity of quantum mechanics. This situation is sometimes described as "state dependence": for a given observable, say the excitation number of the A qubit behind the horizon, one tries to use a different self-adjoint operator depending on which of the four states the system is in (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013).
- *Frozen vacuum*: Say we want to excite the horizon by acting with the X operator on B. From a semiclassical point of view this should produce an excitation at the horizon, but here this operator just permutes us among the four states which are all taken to be smooth. We then do not seem to be able to realize these excited states, even though they exist semiclassically (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013; Bousso, 2014; Harlow, 2014).
- *Nonunitary measurement*: Now say we want to design an apparatus which measures the Z operator on the interior qubit A which is entangled with B. A straightforward argument (Harlow, 2014) shows that, unlike in conventional quantum mechanics, the measurement process cannot be described as unitary evolution on the system together with the apparatus. This is essentially a consequence of the nonlinearity; it is a general problem for attempts to formulate "state-dependent" quantum mechanics. It is also a good concrete criterion for distinguishing "illegal" state dependence of the type needed here from more conventional experiments that naively seem to involve state-dependent observables (Harlow, 2014).
- *Nonuniqueness*: We see in Sec. III of the Supplemental Material [193] that the purification R_B is not uniquely defined. We are free to conjugate it by any unitary transformation which fixes the subspace of states appearing in the Schmidt decomposition, here the space spanned by $\{|\tilde{0}\rangle_{BH}, |\tilde{1}\rangle_{BH}\}$. This means that any attempt to define interior operators acting directly on the subfactor R_B will not be unique. How then are we to choose which operators to use?
- *Commutator problem*: Once we want to define interior operators out of operators that act nontrivially on H and R, it is no longer clear that these operators will commute with "simple" operators on H and R (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013). This raises the possibility of acausal signaling; we can in principle use this commutator to either create a firewall by doing something simple on the distant radiation or communicate from inside the black hole to the outside.

So far no version of $A = R_B$ has appeared which can give completely satisfying resolutions to these objections. The proposal that comes closest is that of Papadodimas and Raju (2014a, 2014b), which among other things is able to "postpone" the commutator problem to operations that involve some order one fraction of the radiation, but in my view it still is ultimately vulnerable to (somewhat more precise) versions of these problems (Harlow, 2014).[101] In the end if some version of $A = R_B$ is to work, it will need to come with a completely developed measurement theory that generalizes and replaces that of quantum mechanics, and it will need to make clear predictions for any thought experiment we can reasonably imagine doing.

C. Postselection?

Another interesting proposal for modifying quantum mechanics to obtain a consistent description of the black hole interior is the quantum postselection formalism of Horowitz and Maldacena (2004). Quantum postselection is a generalization of quantum mechanics to allow probabilities to depend not only on the initial state of the system, as is usual, but also on some additional "final state" that in general has no relation to the initial state (Aharonov, Bergmann, and Lebowitz, 1964). To concisely state the proposal, it is convenient to first introduce a condensed formalism for describing the predictions of quantum mechanics for successive experiments. Say that we have a quantum system that

[101] In particular, the operations for which the proposal fails are vastly simpler than the exponentially complex quantum computations considered in the previous section; there does not seem to be any serious obstruction to doing them.

begins life in a (possibly mixed) quantum state ρ_i. If we then measure a series of observables $A_1, A_2, \ldots, A_n$, it is not too hard to show that the joint probability distribution for the outcomes $a_1, a_2, \ldots, a_n$ for all the experiments is given by

$$P(a_1, a_2, \ldots, a_n) = \mathrm{tr}(\Pi_{a_n} \cdots \Pi_{a_1} \rho_i \Pi_{a_1} \cdots \Pi_{a_n}), \quad (207)$$

where Π_{a_j} is the projection operator onto the outcome a_j from measuring A_j. This remarkable formula combines the Born rule and the collapse of the wave function into a single equation. It elegantly captures some of the more surprising features of quantum measurement theory.

- If two sequential observables do not commute, then the probability distribution depends on which we measure first. If they commute then it does not.
- Regardless of commutators, the probability distribution for the outcomes of the first m measurements is independent of any measurements that come later. For example,

$$P(a_1) \equiv \sum_{a_2} P(a_1, a_2) = \mathrm{tr}(\Pi_{a_1} \rho_i \Pi_{a_1}). \quad (208)$$

  This is essential for quantum mechanics to respect causality; our results for measurements today should not depend on what we decide to do tomorrow.
- Averaging over results of a measurement in the *middle* of the chain affects the probability distribution for the results of the other measurements. For example, if $[A_1, A_2] \neq 0$ we usually have

$$P(a_2) \equiv \sum_{a_1} P(a_1, a_2) \neq \mathrm{tr}(\Pi_{a_2} \rho_i \Pi_{a_2}). \quad (209)$$

  "Averaging out" a measurement is not in general equivalent to not doing the measurement, unlike in classical physics where we allow measurements that acquire any information we like about the system without disturbing it.[102]

Now to generalize to including a "final state," the idea is simply to replace Eq. (207) by

$$P(a_1, \ldots, a_n) = \frac{1}{\mathcal{N}} \mathrm{tr}(\rho_f \Pi_{a_n} \cdots \Pi_{a_1} \rho_i \Pi_{a_1} \cdots \Pi_{a_n}), \quad (210)$$

where the final state is ρ_f (Aharonov, Bergmann, and Lebowitz, 1964) and

$$\mathcal{N} = \sum_{a_1' \cdots a_n'} \mathrm{tr}(\rho_f \Pi_{a_n'} \cdots \Pi_{a_1'} \rho_i \Pi_{a_1'} \cdots \Pi_{a_n'}). \quad (211)$$

This defines a normalized set of probabilities for the results of any experiment we can imagine doing. The first thing to note however is that we now lose our ability to average out later measurements as in Eq. (208); postselected quantum mechanics violates causality. It also immediately grants us the ability to solve nondeterministic polynomial (NP)-complete problems in polynomial time (Aaronson, 2005). These may already be sufficient grounds to reject it, but as we will now see it may provide a surprisingly simple way of getting information out of a black hole without disrupting the experience of an infalling observer.

To see what postselection has to do with black holes, following Horowitz and Maldacena we can again take seriously the effective field theory Hilbert space on a nice slice such as the blue one shown in Fig. 20. In Sec. V.G we saw that together with unitarity this led to quantum cloning, but now we are violating quantum mechanics anyway so let us press on. We can model the Hilbert space as a tensor product

$$\mathcal{H} = \mathcal{H}_M \otimes \mathcal{H}_A \otimes \mathcal{H}_B, \quad (212)$$

where now we are interpreting M as the full set of "left-moving" modes behind the horizon, A as the full set of "right-moving" modes behind the horizon, and B as the full set of "outgoing" modes in the atmosphere. We can think of them as all having dimensionality $|M| = e^{S_{BH}}$, since M describes "all things that we throw in," B describes "all radiation that comes out," and A is "all Hawking partners of the radiation." We can take the initial state to be

$$|\psi_i\rangle_{MAB} = |\psi\rangle_M \otimes |\phi\rangle_{AB}, \quad (213)$$

with

$$|\phi\rangle_{AB} \equiv \frac{1}{\sqrt{|M|}} \sum_a |a\rangle_A |a\rangle_B. \quad (214)$$

Here I have denoted the initial quantum state of the infalling matter as $|\psi\rangle_M$ and taken A and B to be maximally entangled as required for a smooth horizon. Were we to proceed as usual we would conclude that this state has information loss, since the state of B is mixed and has no memory of $|\psi\rangle_M$, but the insight of Horowitz and Maldacena was that if we now introduce a final state

$$\rho_f = |\chi\rangle\langle\chi|_{MA} \otimes \frac{I_B}{|M|}, \quad (215)$$

with

$$|\chi\rangle_{MA} = \frac{1}{\sqrt{|M|}} \sum_a S^\dagger |a\rangle_M |a\rangle_A, \quad (216)$$

the information actually gets out. More explicitly, say that we want to compute the probability distribution $P(b)$ for some set

[102] A sufficient condition for a sequence of measurements to allow such averaging is for the decoherence functional $D(a_1, \ldots, a_n, a_1', \ldots, a_n') \equiv \mathrm{tr}(\Pi_{a_n} \cdots \Pi_{a_1} \rho_i \Pi_{a_1'} \cdots \Pi_{a_n'})$ to be diagonal in the sense of vanishing unless $a_j = a_j'$ $\forall$ j. Such sequences of projection operators are sometimes called "consistent histories," although this term is rather misleading since there is nothing inconsistent about more general sequences of measurements. What "consistent" really means here is that when $D(a, a')$ is diagonal, we are allowed to think of the system as having a definite "classical history" in the sense that we can imagine that each observable has a well-defined value at each point in time whether or not we look at it. Of course this classical history is just us lying to ourselves about the fundamental indeterminacy of quantum mechanics, but it can be useful in understanding under what circumstances classical mechanics emerges from quantum mechanics (Gell-Mann and Hartle, 1993).

of measurements on the radiation B. We can represent the sequence of projectors for some set of outcomes as $C_b \equiv \Pi_{b_1} \cdots \Pi_{b_n}$, in which case it is not too hard to see that

$$\frac{\mathrm{tr}(\rho_f C_b^\dagger \rho_i C_b)}{\sum_{b'} \mathrm{tr}(\rho_f C_{b'}^\dagger \rho_i C_{b'})} = \langle \psi | S^\dagger C C^\dagger S \psi \rangle. \tag{217}$$

In other words the outcomes of all measurements on the outgoing radiation are consistent with taking it to be in the pure state $S|\psi\rangle_B$ and using ordinary quantum mechanics.

Of course to really address the AMPS paradox, we need to understand the implications of this proposal for experiments that involve the infalling observer (Bousso and Stanford, 2014; Lloyd and Preskill, 2014). Consider an experiment where an infalling observer attempts to confirm that A and B are indeed entangled as in Eq. (213). She can do this by measuring the projection operator $\Pi_\phi = I_M \otimes |\phi\rangle\langle\phi|_{AB}$. Using Eq. (210), a straightforward computation shows that the probability for observing $|\phi\rangle_{AB}$ is 1; she will always see the desired entanglement. Thus we see that postselected quantum mechanics allows us to both have our cake and eat it too; the Hawking radiation is unitary, but the horizon is smooth.

Unfortunately the simple story presented so far in this section becomes more complicated once we try to include more details. In the formalism I have presented here I have not taken into account interactions between M and A, and doing so requires a more delicate choice of final state (Gottesman and Preskill, 2004). Moreover once we consider experiments like the AMPS experiment that involves both attempting to verify the entanglement between A and B and confirms the purity of the radiation, the acausality intrinsic to the proposal rears its ugly head (Bousso and Stanford, 2014; Lloyd and Preskill, 2014). One point that is especially confusing is that the apparatus of the infalling observer must be included as part of the infalling system M and must interact with A. Since this is eventually postselected on it may be necessary to "undo" any measurement that happens by a redefinition of the final state. At the moment there does not seem to be completely satisfactory resolutions of these issues. One possibility that is perhaps worth exploring more is that the computational complexity restrictions we discussed in Sec. VIII.A may relieve some or all of the pressure on the Horowitz-Maldacena proposal (Bousso and Stanford, 2014; Lloyd and Preskill, 2014).

D. Firewalls?

The final possibility I consider is that some version of a firewall actually exists. I take this term to mean any observable violation of low-energy effective field theory for simple experiments in the vicinity of a black hole horizon. Six options, in particular, have been discussed in some detail.

- *Full-strength firewall:* AMPS originally argued that the most conservative resolution of the tension in the previous section is to simply imagine that the horizon becomes singular and the interior no longer exists, either for typical big AdS black holes or for old asymptotically flat black holes (Almheiri, Marolf, Polchinski, and Sully, 2013; Marolf and Polchinski, 2013). Given the proposals in the previous sections, one may be sympathetic to considering this option more seriously. It still has substantial drawbacks however: there is currently no dynamical explanation for why a singularity should form at the horizon "out of nothing," and if one does it seems rather unlikely that we should still take Hawking's calculation of the temperature and entropy seriously. Since this calculation is what justified black hole thermodynamics in the first place, the firewall proposal is somewhat self-defeating.[103]
- *Typical-states-only firewall:* Another possibility is to accept that black holes formed in typical states have firewalls, but then attempt to argue that the ones that form in nature never do, even if they are old evaporating black holes. This could be possible because, as argued in Sec. V.E, the black holes we make in short collapses are only a vanishingly small fraction of the total ensemble whose dimensionality is $e^{S_{BH}}$. As the black hole evaporates the state of the remaining black hole becomes more typical by Page's theorem, but the state of the joint black hole and radiation system stays atypical. If we are willing to allow the radiation to be used in constructing the interior according to Sec. VIII.B, then we can use this to avoid the genericity argument of Sec. VII.B. Recently Susskind has been exploring arguments based on computational complexity that may support this possibility (Susskind, 2014) [see also Stanford and Susskind (2014) and Susskind and Zhao (2014)], although this proposal still seems to be subject to the criticism of invalidating Hawking's calculation of the temperature and entropy, with the same caveats as before.
- *S-wave firewall:* We saw in Sec. VII.A that at least some of the paradoxes can be satisfied by a firewall that affects only low angular momentum modes. So far this proposal has not received too much attention in the literature, perhaps mostly because, to quote Raphael Bousso, one would like to "get rid of firewalls entirely or don't bother." Nonetheless this idea does have some things going for it, perhaps chief among them that the backreaction becomes small and thus the basic structure of Hawking's calculation should go through. One would

[103]Lenny Susskind emphasized to me however that one can attempt a "strictly exterior" calculation of the entropy and temperature by arguing that quantum fields outside the horizon have a large backreaction in the Schwarzschild geometry if we put them at a temperature other than T_{Hawking}. This is true, but if we are willing to allow large backreaction right at the horizon in the form of a firewall, why should we not also allow it further out in the atmosphere? Keeping the bad behavior quarantined at the horizon may be the "least objectionable" thing to do, but without an explanation for how the firewall forms we cannot be sure it does not extend further. We know from AdS/CFT that at least in some cases it does not, for example, in $\mathrm{AdS}_3/\mathrm{CFT}_2$, where we can compare the *BH* entropy and the Cardy formula as in Sec. VI.D. Logically we can only interpret this as constraining firewalls to lie strictly at the horizon and not outside, but it seems natural to interpret it as validating the whole coarse-grained semiclassical picture of the horizon, including a smooth experience for an infalling observer. Of course we still need to understand how to resolve the paradoxes of Sec. VII before really dismissing firewalls.

still need a mechanism for how the *S*-wave firewall develops however.

- *Nonviolent nonlocality:* Another possibility that Giddings has been exploring is that there is a more diffuse violation of effective field theory that is spread out nonlocally through the atmosphere rather than concentrated at the horizon (Giddings, 2013a; Giddings and Shi, 2014).[104] The models he has studied so far indeed have the property that black hole thermodynamics tends to be modified. Preventing this requires large modifications of the Schwarzschild geometry outside the horizon that may even be detectable in upcoming experiments (Almheiri, Marolf, Polchinski, Stanford, and Sully, 2013; Giddings, 2013b, 2014).[105] Once we allow such large modifications of effective field theory outside the horizon, however, it is difficult to see how they will not arise in other situations as well; small violations of causality tend to have a way of not staying small.
- *Fuzzballs:* For certain higher-dimensional black holes with large charges under various gauge form fields such as one finds in string theory, there exist so-called "fuzzball" solutions, which resemble the black hole solution near infinity but near the horizon cap off into higher dimensions in various ways, effectively excising the interior. There is a vast literature discussing these solutions; see Gibbons and Warner (2014) for an introduction and further references. It is sometimes argued that there might be enough of these solutions to account for the full Bekenstein-Hawking entropy of these black holes. Despite the capping off of fuzzball geometries at the horizon, there have been some attempts to argue that an infalling observer nonetheless experiences a smooth horizon (Mathur and Turton, 2014a, 2014b), but this certainly is not what the fuzzball geometries naively tell us (Bena, Puhm, and Vercnocke, 2012), so if it is true it will require some new idea. In any case fuzzball solutions exist only in these special cases, and so far there are no analogous solutions for uncharged black holes and in fact there are theorems forbidding their existence, which makes the relevance of these solutions for a general solution of the information problem unclear.
- *Shut up and calculate in the UV theory:* In string theory we already have available at least one candidate theory of quantum gravity; should we not see what it predicts at the horizon? If it predicts a firewall should we not then just accept it? Of course the problem with this correct philosophy is that we do not understand the theory well enough to decide what it predicts; the only case where it really seems well defined so far is in the AdS/CFT correspondence, which as discussed has not yet given us a clear answer. Nonetheless some attempts to study the black hole information problem have been made using *perturbative* string theory in the bulk. For some preliminary results that go in the direction of nonlocality, see Lowe *et al.* (1995) Amati, Ciafaloni, and Veneziano (2008), Giddings, Gross, and Maharana (2008), Dodelson and Silverstein (2014), and Silverstein (2014), although at this point it is not so clear whether these results actually lead us to expect observable violations of effective field theory. See also Horowitz, Lawrence, and Silverstein (2009) for another stringy attempt at behind the horizon physics.

Thus we find ourselves in the enviable situation of having an interesting problem with no really satisfying answer. If we are lucky this means that we will learn something deep. It is my hope that what we learn can then be applied to the other profound problem of quantum gravity—making sense of cosmology. Observers behind black hole horizons have many common features with observers in expanding universes, and what is in my view the best proposal for the global structure of spacetime, the landscape of string theory populated by eternal inflation, has perplexing difficulties which seem to be grown-up versions of the problems we are already confronting in black hole physics. If you have made it this far in this paper I hope it is clear to you that at least one major new idea is needed to understand the black hole interior, and it is exciting to imagine where it might lead us in the future.

ACKNOWLEDGMENTS

This paper would not have been possible without the help of many. This is a difficult subject, and I have benefited from many conversations with some of the world's experts. In particular, Raphael Bousso, Patrick Hayden, Juan Maldacena, Joe Polchinski, Douglas Stanford, Herman Verlinde, and my co-advisers Steve Shenker and Lenny Susskind have taught me many things over the years. It is impossible to list everyone else I am indebted to, but certainly the set includes Scott Aaronson, Ahmed Almheiri, Nima Arkani-Hamed, Tom Banks, Charlene Borsack, Borun Chowdhury, Bartek Czech, Xi Dong, Willy Fischler, Ben Freivogel, Steve Giddings, Tom Hartman, Matt Headrick, Idse Heemskerk, Simeon Hellerman, Veronica Hubeny, Dan Kabat, Igor Klebanov, Nima Lashkari, Albion Lawrence, Stefan Liechenauer, Don Marolf, Samir Mathur, Jonathan Oppenheim, Don Page, Kyriakos Papadodimas, John Preskill, Andrea Puhm, Eliezer Rabinovici, Suvrat Raju, Mukund Rangamani, Vladimir Rosenhaus, Julie Shih, Eva Silverstein, Jamie Sully, Mark Van Raamsdonk, Erik Verlinde, Aron Wall, and Edward Witten. You all have had to put up with me to varying degrees these past years, and I hope I have given something back. I thank two anonymous referees, and also Steve Giddings, for very detailed feedback on an earlier version of this paper, as well as Arash Ardehali for pointing out an impressive number of typos. I also thank the participants of the 2014 Jerusalem winter school for providing a stimulating audience for the lectures, the Israel Institute for Advanced Studies for

[104]Giddings objects to this proposal being included as a type of firewall, since for him a firewall is necessarily violent. For me however the real question is whether or not there is a simple experiment an infalling observer can do that detects a violation of effective field theory. I am less concerned with how traumatic the experience is, after all the *s*-wave firewall is also "nonviolent."

[105]At least somebody in this business is making experimental predictions.

the invitation to give them and hospitality during them, the Weizmann Institute and the Aspen Center for Physics for hospitality (twice) while the paper was being completed, as well as the Pacific Institute for Theoretical Physics at UBC, and Eliezer Rabinovici for a warm welcome to Israel during my first trip there several years ago. I am supported by the Princeton Center for Theoretical Science. Finally I would like to thank the late Professor Jacob Bekenstein, who was an active member of the audience when these lectures were delivered. His seminal realization that the entropy of black holes should be taken seriously lies at the foundation of everything discussed in these notes.

REFERENCES

Aaronson, Scott, 2005, "NP-complete problems and physical reality," arXiv:quant-ph/0502072.

Abramowitz, Milton, and Irene A. Stegun, 1965, *Handbook of Mathematical Functions* (Dover Publications, New York), p. 55.

Aharonov, Yakir, Peter G. Bergmann, and Joel L. Lebowitz, 1964, "Time symmetry in the quantum process of measurement," Phys. Rev. **134**, B1410.

Aharony, Ofer, Steven S. Gubser, Juan Martin Maldacena, Hirosi Ooguri, and Yaron Oz, 2000, "Large N field theories, string theory and gravity," Phys. Rep. **323**, 183.

Aharony, Ofer, Joseph Marsano, Shiraz Minwalla, Kyriakos Papadodimas, and Mark Van Raamsdonk, 2004, "The Hagedorn—deconfinement phase transition in weakly coupled large N gauge theories," Adv. Theor. Math. Phys. **8**, 603.

Almheiri, Ahmed, Xi Dong, and Daniel Harlow, **2015**, "Bulk Locality and Quantum Error Correction in AdS/CFT," J. High Energy Phys. 04, 163.

Almheiri, Ahmed, Donald Marolf, Joseph Polchinski, Douglas Stanford, and James Sully, 2013, "An Apologia for Firewalls," J. High Energy Phys. 09, 018.

Almheiri, Ahmed, Donald Marolf, Joseph Polchinski, and James Sully, 2013, "Black Holes: Complementarity or Firewalls?," J. High Energy Phys. 02, 062.

Amati, D., M. Ciafaloni, and G. Veneziano, 2008, "Towards an S-matrix description of gravitational collapse," J. High Energy Phys. 02, 049.

Arnowitt, Richard L., Stanley Deser, and Charles W. Misner, 2008, "The Dynamics of general relativity," Gen. Relativ. Gravit. **40**, 1997.

Arora, Sanjeev, and Boaz Barak, 2009, *Computational complexity: A modern approach* (Cambridge University Press, Cambridge, England).

Avery, Steven G., and Borun D. Chowdhury, **2014**, "Firewalls in AdS/CFT," J. High Energy Phys. 10, 174.

Avery, Steven G., 2013, "Qubit Models of Black Hole Evaporation," J. High Energy Phys. 01, 176.

Balasubramanian, Vijay, Borun D. Chowdhury, Bartlomiej Czech, Jan de Boer, and Michal P. Heller, 2014, "A hole-ographic spacetime," Phys. Rev. D **89**, 086004.

Banados, Maximo, Claudio Teitelboim, and Jorge Zanelli, 1992, "The Black hole in three-dimensional space-time," Phys. Rev. Lett. **69**, 1849.

Banks, Tom, Michael R. Douglas, Gary T. Horowitz, and Emil J. Martinec, 1998, "AdS dynamics from conformal field theory," arXiv:hep-th/9808016.

Banks, Tom, and W. Fischler, 2012, "Holographic Space-Time Does Not Predict Firewalls," arXiv:1208.4757.

Banks, Tom, W. Fischler, S. H. Shenker, and Leonard Susskind, 1997, "M theory as a matrix model: A Conjecture," Phys. Rev. D **55**, 5112.

Barbon, J. L. F., and E. Rabinovici, **2003**, "Very long time scales and black hole thermal equilibrium," J. High Energy Phys. 11, 047.

Barbon, Jose L. F., and Javier M. Magan, 2011, "Chaotic Fast Scrambling At Black Holes," Phys. Rev. D **84**, 106012.

Barbon, Jose L. F., and Javier M. Magan, **2012**, "Fast Scramblers, Horizons and Expander Graphs," J. High Energy Phys. 08, 016.

Barbon, Jose L. F., and Eliezer Rabinovici, 2014, "Geometry And Quantum Noise," arXiv:1404.7085.

Bardeen, James M., B. Carter, and S. W. Hawking, 1973, "The Four laws of black hole mechanics," Commun. Math. Phys. **31**, 161.

Bekenstein, Jacob D., 1973, "Black holes and entropy," Phys. Rev. D **7**, 2333.

Bekenstein, Jacob D., 1974, "Generalized second law of thermodynamics in black hole physics," Phys. Rev. D **9**, 3292.

Bekenstein, Jacob D., 1981, "A Universal Upper Bound on the Entropy to Energy Ratio for Bounded Systems," Phys. Rev. D **23**, 287.

Bena, Iosif, Andrea Puhm, and Bert Vercnocke, **2012**, "Non-extremal Black Hole Microstates: Fuzzballs of Fire or Fuzzballs of Fuzz?," J. High Energy Phys. 12, 014.

Bennett, Charles H., Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters, 1993, "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," Phys. Rev. Lett. **70**, 1895.

Bombaci, I., 1996, "The maximum mass of a neutron star," Astron. Astrophys. **305**, 871.

Bousso, Raphael, 2002, "The Holographic principle," Rev. Mod. Phys. **74**, 825.

Bousso, Raphael, 2013a, "Complementarity Is Not Enough," Phys. Rev. D **87**, 124023.

Bousso, Raphael, 2013b, "Firewalls From Double Purity," Phys. Rev. D **88**, 084035.

Bousso, Raphael, 2014, "Frozen Vacuum," Phys. Rev. Lett. **112**, 041102.

Bousso, Raphael, Horacio Casini, Zachary Fisher, and Juan Maldacena, 2014, "Proof of a Quantum Bousso Bound," arXiv:1404.5635.

Bousso, Raphael, **1999**a, "A Covariant entropy conjecture," J. High Energy Phys. 07, 004.

Bousso, Raphael, 1999b, "Holography in general space-times," J. High Energy Phys. 06, 028.

Bousso, Raphael, and Douglas Stanford, 2014, "Measurements without Probabilities in the Final State Proposal," Phys. Rev. D **89**, 044038.

Braunstein, Samuel L., Stefano Pirandola, and Karol Życzkowski, 2013, "Entangled black holes as ciphers of hidden information," Phys. Rev. Lett. **110**, 101301.

Breitenlohner, Peter, and Daniel Z. Freedman, 1982, "Stability in Gauged Extended Supergravity," Ann. Phys. (N.Y.) **144**, 249.

Brown, Adam R., 2013, "Tensile Strength and the Mining of Black Holes," Phys. Rev. Lett. **111**, 211301.

Brown, J. David, and M. Henneaux, 1986, "Central Charges in the Canonical Realization of Asymptotic Symmetries: An Example from Three-Dimensional Gravity," Commun. Math. Phys. **104**, 207.

Buhrman, Harry, Richard Cleve, John Watrous, and Ronald de Wolf, 2001, "Quantum fingerprinting," Phys. Rev. Lett. **87**, 167902.

Burgess, C. P., and C. A. Lutken, 1985, "Propagators and Effective Potentials in Anti-de Sitter Space," Phys. Lett. **153B**, 137.

Cardy, John L., 1986, "Operator Content of Two-Dimensional Conformally Invariant Theories," Nucl. Phys. **B270**, 186.

Casini, H., 2008, "Relative entropy and the Bekenstein bound," Classical Quantum Gravity **25**, 205021.
Chowdhury, Borun D., **2013**, "Cool horizons lead to information loss," J. High Energy Phys. 10, 034.
Corley, Steven, and Ted Jacobson, 1996, "Hawking spectrum and high frequency dispersion," Phys. Rev. D **54**, 1568.
Czech, Bartlomiej, Xi Dong, and James Sully, 2014, "Holographic Reconstruction of General Bulk Surfaces," arXiv:1406.4889.
Dankert, Christoph, Richard Cleve, Joseph Emerson, and Etera Livine, 2009, "Exact and approximate unitary 2-designs and their application to fidelity estimation," Phys. Rev. A **80**, 012304.
Demers, Jean-Guy, Rene Lafrance, and Robert C. Myers, 1995, "Black hole entropy without brick walls," Phys. Rev. D **52**, 2245.
Deutsch, J. M., 1991, "Quantum statistical mechanics in a closed system," Phys. Rev. A **43**, 2046.
Dieks, D., 1982, "Communication by EPR devices," Phys. Lett. **92A**, 271.
Dodelson, Matt, and Eva Silverstein, 2014, "String-theoretic breakdown of effective field theory near black hole horizons," arXiv:1504.05536.
Dong, Xi, **2014**, "Holographic Entanglement Entropy for General Higher Derivative Gravity," J. High Energy Phys. 01, 044.
Einstein, Albert, Boris Podolsky, and Nathan Rosen, 1935, "Can quantum mechanical description of physical reality be considered complete?," Phys. Rev. **47**, 777.
Einstein, Albert, and N. Rosen, 1935, "The Particle Problem in the General Theory of Relativity," Phys. Rev. **48**, 73.
Faulkner, Thomas, 2013, "The Entanglement Renyi Entropies of Disjoint Intervals in AdS/CFT," arXiv:1303.7221.
Feynman, R. P., 1982, "Simulating physics with computers," Int. J. Theor. Phys. **21**, 467.
Fidkowski, Lukasz, Veronika Hubeny, Matthew Kleban, and Stephen Shenker, **2004**, "The Black hole singularity in AdS/CFT," J. High Energy Phys. 02, 014.
Fitzpatrick, A. Liam, Jared Kaplan, and Matthew T. Walters, 2014, "Universality of Long-Distance AdS Physics from the CFT Bootstrap," arXiv:1403.6829.
Flanagan, Eanna E., Donald Marolf, and Robert M. Wald, 2000, "Proof of classical versions of the Bousso entropy bound and of the generalized second law," Phys. Rev. D **62**, 084035.
Freivogel, Ben, Robert A. Jefferson, Laurens Kabir, and I-Sheng Yang, 2014, "Geometry of the Infalling Causal Patch," arXiv:1406.6043.
Gaberdiel, Matthias R., and Rajesh Gopakumar, 2011, "An AdS_3 Dual for Minimal Model CFTs," Phys. Rev. D **83**, 066007.
Gell-Mann, Murray, and James B. Hartle, 1993, "Classical equations for quantum systems," Phys. Rev. D **47**, 3345.
Gibbons, G. W., and S. W. Hawking, 1977, "Action Integrals and Partition Functions in Quantum Gravity," Phys. Rev. D **15**, 2752.
Gibbons, G. W., and N. P. Warner, 2014, "Global structure of five-dimensional fuzzballs," Classical Quantum Gravity **31**, 025016.
Giddings, Steven B., 1994, "Quantum mechanics of black holes," arXiv:hep-th/9412138.
Giddings, Steven B., 1995, "Why aren't black holes infinitely produced?," Phys. Rev. D **51**, 6860.
Giddings, Steven B., 2006, "Black hole information, unitarity, and nonlocality," Phys. Rev. D **74**, 106005.
Giddings, Steven B., 2012, "Models for unitary black hole disintegration," Phys. Rev. D **85**, 044038.
Giddings, Steven B., 2013a, "Nonviolent nonlocality," Phys. Rev. D **88**, 064023.
Giddings, Steven B., 2013b, "Statistical physics of black holes as quantum-mechanical systems," Phys. Rev. D **88**, 104013.
Giddings, Steven B., 2014, "Possible observational windows for quantum effects from black holes," arXiv:1406.7001.
Giddings, Steven B., David J. Gross, and Anshuman Maharana, 2008, "Gravitational effects in ultrahigh-energy string scattering," Phys. Rev. D **77**, 046001.
Giddings, Steven B., and Yinbo Shi, 2014, "Effective field theory models for nonviolent information transfer from black holes," Phys. Rev. D **89**, 124032.
Giombi, Simone, and Xi Yin, **2010**, "Higher Spin Gauge Theory and Holography: The Three-Point Functions," J. High Energy Phys. 09, 115.
Goldreich, Oded, and Leonid A. Levin, 1989, "A hard-core predicate for all one-way functions," in *Proceedings of the twenty-first annual ACM symposium on Theory of computing* (ACM, Seattle, WA), pp. 25–32.
Gottesman, Daniel, and John Preskill, **2004**, "Comment on 'The Black hole final state'," J. High Energy Phys. 03, 026.
Gubser, S. S., Igor R. Klebanov, and Alexander M. Polyakov, 1998, "Gauge theory correlators from noncritical string theory," Phys. Lett. B **428**, 105.
Harlow, Daniel, 2012, "Complementarity, not Firewalls," arXiv:1207.6243v2.
Harlow, Daniel, 2014, "Aspects of the Papadodimas-Raju Proposal for the Black Hole Interior," arXiv:1405.1995.
Harlow, Daniel, and Patrick Hayden, **2013**, "Quantum Computation vs. Firewalls," J. High Energy Phys. 06, 085.
Harlow, Daniel, and Douglas Stanford, 2011, "Operator Dictionaries and Wave Functions in AdS/CFT and dS/CFT," arXiv:1104.2621.
Hartle, J. B., and S. W. Hawking, 1976, "Path Integral Derivation of Black Hole Radiance," Phys. Rev. D **13**, 2188.
Hartman, Thomas, 2013, "Entanglement Entropy at Large Central Charge," arXiv:1303.6955.
Hartman, Thomas, Christoph A. Keller, and Bogdan Stoica, 2014, "Universal Spectrum of 2d Conformal Field Theory in the Large c Limit," arXiv:1405.5137.
Hartman, Thomas, and Juan Maldacena, **2013**, "Time Evolution of Entanglement Entropy from Black Hole Interiors," J. High Energy Phys. 05, 014.
Hawking, S. W., 1971, "Gravitational radiation from colliding black holes," Phys. Rev. Lett. **26**, 1344.
Hawking, S. W., 1975, "Particle Creation by Black Holes," Commun. Math. Phys. **43**, 199.
Hawking, S. W., 1976, "Breakdown of Predictability in Gravitational Collapse," Phys. Rev. D **14**, 2460.
Hawking, S. W., 2005, "Information loss in black holes," Phys. Rev. D **72**, 084013.
Hawking, S. W., and Don N. Page, 1983, "Thermodynamics of Black Holes in anti-De Sitter Space," Commun. Math. Phys. **87**, 577.
Hayden, Patrick, and John Preskill, **2007**, "Black holes as mirrors: Quantum information in random subsystems," J. High Energy Phys. 09, 120.
Headrick, Matthew, and Tadashi Takayanagi, 2007, "A Holographic proof of the strong subadditivity of entanglement entropy," Phys. Rev. D **76**, 106013.
Heemskerk, Idse, **2012**, "Construction of Bulk Fields with Gauge Redundancy," J. High Energy Phys. 09, 106.
Heemskerk, Idse, Donald Marolf, Joseph Polchinski, and James Sully, 2012, "Bulk and Transhorizon Measurements in AdS/CFT," J. High Energy Phys. 10, 165.

Heemskerk, Idse, Joao Penedones, Joseph Polchinski, and James Sully, 2009, "Holography from Conformal Field Theory," J. High Energy Phys. 10, 079.
Henneaux, M., and C. Teitelboim, 1985, "Asymptotically anti-De Sitter Spaces," Commun. Math. Phys. **98**, 391.
Holzhey, Christoph, Finn Larsen, and Frank Wilczek, 1994, "Geometric and renormalized entropy in conformal field theory," Nucl. Phys. **B424**, 443.
Horowitz, Gary, Albion Lawrence, and Eva Silverstein, **2009**, "Insightful D-branes," J. High Energy Phys. 07, 057.
Horowitz, Gary T., 2000, "Comments on black holes in string theory," Classical Quantum Gravity **17**, 1107.
Horowitz, Gary T., and Juan Martin Maldacena, **2004**, "The Black hole final state," J. High Energy Phys. 02, 008.
Horowitz, Gary T., and Joseph Polchinski, 1997, "A Correspondence principle for black holes and strings," Phys. Rev. D **55**, 6189.
Hubeny, Veronika E., Mukund Rangamani, and Tadashi Takayanagi, **2007**, "A Covariant holographic entanglement entropy proposal," J. High Energy Phys. 07, 062.
Hui, Lam, and I-Sheng Yang, 2014, "Complementarity + Back-reaction is enough," Phys. Rev. D **89**, 084011.
Ilgin, Irfan, and I-Sheng Yang, 2014, "Causal Patch Complementarity: The Inside Story for Old Black Holes," Phys. Rev. D **89**, 044007.
Impagliazzo, Russell, 1995, "A personal view of average-case complexity," in *Structure in Complexity Theory Conference, 1995, Proceedings of Tenth Annual IEEE* (IEEE, New York), pp. 134–147.
Israel, W., 1976, "Thermo field dynamics of black holes," Phys. Lett. **57A**, 107.
Jordan, Stephen P., Keith S. M. Lee, and John Preskill, 2011, "Quantum Algorithms for Quantum Field Theories," arXiv:1111.3633.
Kabat, Daniel, and Gilad Lifschytz, 2013, "CFT representation of interacting bulk gauge fields in AdS," Phys. Rev. D **87**, 086004.
Kabat, Daniel, and Gilad Lifschytz, 2014, "Finite N and the failure of bulk locality: Black holes in AdS/CFT," arXiv:1405.6394.
Kabat, Daniel, Gilad Lifschytz, and David A. Lowe, 2011, "Constructing local bulk observables in interacting AdS/CFT," Phys. Rev. D **83**, 106009.
Kiem, Youngjai, Herman L. Verlinde, and Erik P. Verlinde, 1995, "Black hole horizons and complementarity," Phys. Rev. D **52**, 7053.
Klebanov, I. R., and A. M. Polyakov, 2002, "AdS dual of the critical O(N) vector model," Phys. Lett. B **550**, 213.
Klebanov, Igor R., and Edward Witten, 1999, "AdS/CFT correspondence and symmetry breaking," Nucl. Phys. **B556**, 89.
Koashi, Masato, and Andreas Winter, 2004, "Monogamy of quantum entanglement and other correlations," Phys. Rev. A **69**, 022309.
Lashkari, Nima, Douglas Stanford, Matthew Hastings, Tobias Osborne, and Patrick Hayden, **2013**, "Towards the Fast Scrambling Conjecture," J. High Energy Phys. 04, 022.
Leichenauer, Stefan, and Vladimir Rosenhaus, 2013, "AdS black holes, the bulk-boundary dictionary, and smearing functions," Phys. Rev. D **88**, 026003.
Lewkowycz, Aitor, and Juan Maldacena, **2013**, "Generalized gravitational entropy," J. High Energy Phys. 08, 090.
Lloyd, S., 1996, "Universal quantum simulators," Science **273**, 1073.
Lloyd, Seth, and John Preskill, **2014**, "Unitarity of black hole evaporation in final-state projection models," J. High Energy Phys. 08, 126.
Lowe, David A., Joseph Polchinski, Leonard Susskind, Larus Thorlacius, and John Uglum, 1995, "Black hole complementarity versus locality," Phys. Rev. D **52**, 6997.
Maldacena, Juan, and Leonard Susskind, 2013, "Cool horizons for entangled black holes," Fortschr. Phys. **61**, 781.
Maldacena, Juan Martin, 1998, "Wilson loops in large N field theories," Phys. Rev. Lett. **80**, 4859.
Maldacena, Juan Martin, 1999, "The Large-N limit of superconformal field theories and supergravity," Int. J. Theor. Phys. **38**, 1113.
Maldacena, Juan Martin, **2003**a, "Eternal black holes in anti-de Sitter," J. High Energy Phys. 04, 021.
Maldacena, Juan Martin, 2003b, "Non-Gaussian features of primordial fluctuations in single field inflationary models," J. High Energy Phys. 05, 013.
Marolf, Donald, Djordje Minic, and Simon F. Ross, 2004, "Notes on space-time thermodynamics and the observer dependence of entropy," Phys. Rev. D **69**, 064006.
Marolf, Donald, and Joseph Polchinski, 2013, "Gauge/Gravity Duality and the Black Hole Interior," Phys. Rev. Lett. **111**, 171301.
Marolf, Donald, and Rafael D. Sorkin, 2004, "On the status of highly entropic objects," Phys. Rev. D **69**, 024014.
Marolf, Donald, and Aron C. Wall, 2013, "Eternal Black Holes and Superselection in AdS/CFT," Classical Quantum Gravity **30**, 025001.
Mathur, Samir D., 2009, "The Information paradox: A Pedagogical introduction," Classical Quantum Gravity **26**, 224001.
Mathur, Samir D., and David Turton, **2014**a, "Comments on black holes I: The possibility of complementarity," J. High Energy Phys. 01, 034.
Mathur, Samir D., and David Turton, 2014b, "The flaw in the firewall argument," Nucl. Phys. **B884**, 566.
Morrison, Ian A., **2014**, "Boundary-to-bulk maps for AdS causal wedges and the Reeh-Schlieder property in holography," J. High Energy Phys. 05, 053.
Morrison, Ian A., and Matthew M. Roberts, 2012, "Mutual information between thermo-field doubles and disconnected holographic boundaries," arXiv:1211.2887.
Nishioka, Tatsuma, Shinsei Ryu, and Tadashi Takayanagi, 2009, "Holographic Entanglement Entropy: An Overview," J. Phys. A **42**, 504008.
Oppenheim, Jonathan, and William G. Unruh, **2014**, "Firewalls and flat mirrors: An alternative to the AMPS experiment which evades the Harlow-Hayden obstacle," J. High Energy Phys. 03, 120.
Page, Don N., 1976, "Particle Emission Rates from a Black Hole: Massless Particles from an Uncharged, Nonrotating Hole," Phys. Rev. D **13**, 198.
Page, Don N., 1993a, "Average entropy of a subsystem," Phys. Rev. Lett. **71**, 1291.
Page, Don N., 1993b, "Information in black hole radiation," Phys. Rev. Lett. **71**, 3743.
Page, Don N., **2013**, "Time Dependence of Hawking Radiation Entropy," J. Cosmol. Astropart. Phys. 09, 028.
Papadodimas, Kyriakos, and Suvrat Raju, 2013, "An Infalling Observer in AdS/CFT," J. High Energy Phys. 10, 212.
Papadodimas, Kyriakos, and Suvrat Raju, 2014a, "State-Dependent Bulk-Boundary Maps and Black Hole Complementarity," Phys. Rev. D **89**, 086010.
Papadodimas, Kyriakos, and Suvrat Raju, 2014b, "The Black Hole Interior in AdS/CFT and the Information Paradox," Phys. Rev. Lett. **112**, 051301.
Polchinski, J., 1998a, "String theory. Vol. 1: An introduction to the bosonic string" (Cambridge University Press, Cambridge, UK).
Polchinski, J., 1998b, "String theory. Vol. 2: Superstring theory and beyond" (Cambridge University Press, Cambridge, UK).
Polchinski, Joseph, 1995, "String theory and black hole complementarity," arXiv:hep-th/9507094.

Preskill, John, 1992, "Do black holes destroy information?," arXiv:hep-th/9209058.
Price, R. H., and K. S. Thorne, 1986, "Membrane Viewpoint on Black Holes: Properties and Evolution of the Stretched Horizon," Phys. Rev. D **33**, 915.
Regge, Tullio, and Claudio Teitelboim, 1974, "Role of Surface Integrals in the Hamiltonian Formulation of General Relativity," Ann. Phys. (N.Y.) **88**, 286.
Rocha, Jorge V., **2008**, "Evaporation of large black holes in AdS: Coupling to the evaporon," J. High Energy Phys. 08, 075.
Ryu, Shinsei, and Tadashi Takayanagi, 2006, "Holographic derivation of entanglement entropy from AdS/CFT," Phys. Rev. Lett. **96**, 181602.
Sekino, Yasuhiro, and Leonard Susskind, **2008**, "Fast Scramblers," J. High Energy Phys. 10, 065.
Shenker, Stephen H., and Douglas Stanford, 2013, "Multiple Shocks," arXiv:1312.3296.
Shenker, Stephen H., and Douglas Stanford, 2014, "Black holes and the butterfly effect," J. High Energy Phys. 03, 067.
Shenker, Stephen H., and Xi Yin, 2011, "Vector Models in the Singlet Sector at Finite Temperature," arXiv:1109.3519.
Silverstein, Eva, 2014, "Backdraft: String Creation in an Old Schwarzschild Black Hole," arXiv:1402.1486.
Srednicki, Mark, 1996, "Thermal fluctuations in quantized chaotic systems," J. Phys. A **29**, L75.
Stanford, Douglas, and Leonard Susskind, 2014, "Complexity and Shock Wave Geometries," arXiv:1406.2678.
Streater, R. F., and A. S. Wightman, 2000, "*PCT, Spin and Statistics, and All That*" (Princeton University Press, Princeton, NJ).
Strominger, Andrew, and Cumrun Vafa, 1996, "Microscopic origin of the Bekenstein-Hawking entropy," Phys. Lett. B **379**, 99.
Susskind, L., and J. Lindesay, 2005, "*An Introduction to Black Holes, Information and the String Theory Revolution: The Holographic Universe*" (World Scientific, Singapore).
Susskind, Leonard, 1993, "Some speculations about black hole entropy in string theory," arXiv:hep-th/9309145.
Susskind, Leonard, 1995a, "The World as a hologram," J. Math. Phys. (N.Y.) **36**, 6377.
Susskind, Leonard, 1995b, "Trouble for remnants," arXiv:hep-th/9501106.
Susskind, Leonard, 2014, "Computational Complexity and Black Hole Horizons," arXiv:1402.5674.
Susskind, Leonard, and Larus Thorlacius, 1994, "Gedanken experiments involving black holes," Phys. Rev. D **49**, 966.
Susskind, Leonard, Larus Thorlacius, and John Uglum, 1993, "The Stretched horizon and black hole complementarity," Phys. Rev. D **48**, 3743.
Susskind, Leonard, and John Uglum, 1994, "Black hole entropy in canonical quantum gravity and superstring theory," Phys. Rev. D **50**, 2700.
Susskind, Leonard, and Ying Zhao, 2014, "Switchbacks and the Bridge to Nowhere," arXiv:1408.2823.
Takahashi, Y., and H. Umezawa, 1996, "Thermo field dynamics," Int. J. Mod. Phys. B **10**, 1755.
't Hooft, Gerard, 1985, "On the Quantum Structure of a Black Hole," Nucl. Phys. **B256**, 727.
't Hooft, Gerard, 1993, "Dimensional reduction in quantum gravity," arXiv:gr-qc/9310026.
Unruh, W. G., 1994, "Dumb holes and the effects of high frequencies on black hole evaporation," arXiv:gr-qc/9409008.
Unruh, W. G., and Robert M. Wald, 1982, "Acceleration Radiation and Generalized Second Law of Thermodynamics," Phys. Rev. D **25**, 942.
Unruh, William G., and Robert M. Wald, 1984, "What happens when an accelerating observer detects a Rindler particle," Phys. Rev. D **29**, 1047.
Unruh, William G., and Robert M. Wald, 1995, "On evolution laws taking pure states to mixed states in quantum field theory," Phys. Rev. D **52**, 2176.
Van Raamsdonk, Mark, 2010, "Building up spacetime with quantum entanglement," Gen. Relativ. Gravit. **42**, 2323.
Van Raamsdonk, Mark, 2013, "Evaporating Firewalls," arXiv:1307.1796.
Vasiliev, Mikhail A., 1990, "Consistent equation for interacting gauge fields of all spins in $(3+1)$-dimensions," Phys. Lett. B **243**, 378.
Verlinde, Erik, and Herman Verlinde, 2013a, "Behind the Horizon in AdS/CFT," arXiv:1311.1137.
Verlinde, Erik, and Herman Verlinde, **2013**b, "Black Hole Entanglement and Quantum Error Correction," J. High Energy Phys. 10, 107.
Wald, Robert M., 1975, "On Particle Creation by Black Holes," Commun. Math. Phys. **45**, 9.
Wald, Robert M., 2010, *General relativity* (University of Chicago Press, Chicago).
Wall, Aron C., 2010, "A Proof of the generalized second law for rapidly-evolving Rindler horizons," Phys. Rev. D **82**, 124019.
Wall, Aron C., 2012a, "A proof of the generalized second law for rapidly changing fields and arbitrary horizon slices," Phys. Rev. D **85**, 104049.
Wall, Aron C., 2012b, "Maximin Surfaces, and the Strong Subadditivity of the Covariant Holographic Entanglement Entropy," arXiv:1211.3494.
Weinberg, Steven, 1995, "The Quantum theory of fields. Vol. 1: Foundations" (Cambridge University Press, Cambridge,UK).
Witten, Edward, 1998a, "Anti-de Sitter space and holography," Adv. Theor. Math. Phys. **2**, 253 [arXiv:hep-th/9802150].
Witten, Edward, 1998b, "Anti-de Sitter space, thermal phase transition, and confinement in gauge theories," Adv. Theor. Math. Phys. **2**, 505 [arXiv:hep-th/9803131].
Wootters, W. K., and W. H. Zurek, 1982, "A single quantum cannot be cloned," Nature (London) **299**, 802.
See Supplemental Material at http://link.aps.org/supplemental/10.1103/RevModPhys.88.015002 for many of the needed concepts from general relativity and quantum information theory. There are also exercises for students.