Title
Fixing Gauge Redundancies in Quantum Gravity

Author
Weinberg, Sean Jason

Publication Date
2016

Peer reviewed|Thesis/dissertation

Fixing Gauge Redundancies in Quantum Gravity

By

Sean Jason Weinberg

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yasunori Nomura, Chair
Professor Ori J Ganor
Professor Alexander B Givental

Spring 2016

Abstract

Fixing Gauge Redundancies in Quantum Gravity

By

Sean Jason Weinberg

Doctor of Philosophy in Physics

University of California, Berkeley

Yasunori Nomura, Chair

Evidence has accumulated that descriptions of systems in quantum gravity depend strongly on various choices of gauge-fixing including a choice of "reference frame." We discuss several explicit examples of this reference frame dependence and, in doing so, clarify a number of general features of quantum gravity including the thermodynamics of spacetime, the holographic principle, and black hole complementarity.

Our discussion focuses on two superficially independent subjects. The first of these is that of holographic screens. These are codimension-one surfaces that are preferred from the perspective of the holographic principle. They are generated by a choice of null foliation and, in particular, can be fixed by the light cones of a worldline. We will study a class of holographic screens called past and future holographic screens and strengthen a recently proven area law for these surfaces. We then introduce a definition of holographic entanglement entropy associated with past and future holographic screens and, in doing so, provide new evidence for the importance of screens in quantum gravity. Our second major emphasis is on the black hole information paradox and the firewall paradox. We give a set of hypotheses for the microscopic structure of black holes that appears to be self-consistent and admit a smooth horizon despite the AMPS arguments. Our model relies on the principle that the quantum information associated with spacetime is both delocalized and reference frame dependent.

Contents

Acknowledgements

My research would not have been possible without the many opportunities I had during my time at Berkeley to interact with others. First, I would like to thank my collaborators: Yasunori Nomura, Fabio Sanches, and Jaime Varela. Each chapter of this thesis is essentially a modified version of scientific work completed with at least one of these three. I would like to extend special gratitude to Yasunori Nomura, my doctoral advisor, both for the knowledge of physics that he imparted to me, and for the professional mentorship and support he provided over my time as a graduate student.

While is not possible to give a complete list of all of those with whom I have had discussions that impacted this text, it is my pleasure to thank to C. Akers, R. Bousso, Z. Fisher, B. Freivogel, D. Harlow, J. Koeller, D. Marolf, M. Moosa, J. Polchinski, M. Van Raamsdonk, and A. Wall. My work was supported by the Berkeley Center for Theoretical Physics and by the BCTP Brantley-Tuttle Fellowship, generously provided by L. Brantley and D. Tuttle.

At a more personal level, I want to thank my mother, LaWanda Walters, my sister, Tess Weinberg, and my stepfather, John Drury. Their illogical degree of confidence in me is the only reason I have made it to this point. Finally, thank you to my dad, David Weinberg. I hope he would be proud that I am skipping a generation and following in his father's path.

Chapter 1

Introduction

The research discussed in this thesis is primarily focused on quantum gravity. The study of the quantum mechanics of spacetime has been a persistent challenge for theoretical physicists for half of a century. Fortunately, the past few decades have been been cause for real optimism. String theory offers a productive framework that may correctly describe all of physics. From string theory, the AdS/CFT correspondence [1, 2] emerged as an explicit example of a quantum theory of gravity. Despite all of these developments, major conceptual gaps in our knowledge of quantum gravity remain, and these gaps become apparent when one considers black holes.

The vital idea that black hole physics has suggested is that naïve descriptions of systems in quantum gravity arise only after fixing significant gauge redundancy. This is a familiar development in physics. Before relativity, it was believed that the time interval between two events was a physical invariant. General relativity tells a different story: to evolve a spacetime in a Hamiltonian formalism, the (coordinate) gauge freedom must first be fixed. The particular surface that defines a particular coordinate time depends strongly on the choice of gauge.

Now consider the question "is the subsystem A inside or outside of a black hole?" Black hole complementarity [3] declares that the answer is *not* gauge invariant, despite the fact that it certainly is an invariant according to quantum field theory on a classical spacetime background. The correct answer to this question appears to be that "it depends on the reference frame." The requirement that A is not tethered to a point on a spacetime manifold, but instead has a gauge-dependent location and description, is one of the most significant departures from quantum field theory and general relativity that appears necessary in a theory of quantum gravity.

Consider a state corresponding to a black hole viewed from a reference frame outside the horizon. If the black hole has area $A \gg l_{\rm P}^2$, there are many states that roughly look like the same black hole. This degeneracy is counted by the Bekenstein-Hawking entropy: if the quantum state of the black hole is denoted by $|\psi_k\rangle$, the index k takes value in $\{1, \ldots, e^S\}$ where $S = \frac{A}{4l_{\rm P}^2}$. My collaborators and I have investigated the nature of this degeneracy. One point that has guided us is that it has two distinct interpretations. First, it is often assumed that the stretched horizon of

a black hole has "intrinsically stringy" degrees of freedom. In this case, $|\psi_k\rangle$ is the state of the stretched horizon. The second interpretation is related to the observation that every black hole has uncertainty in its macroscopic parameters. The mass M, for instance, is only specified to a precision ΔM (this is $O(1/Ml_{\rm P}^2)$ for a Schwarzschild black hole). Uncertainty in a black hole's macrostate means that the spacetime background is not fixed, and k is then the index that lists states that are interpreted as black holes viewed from a distant frame that have M, J, and Q consistent with the uncertainties.

Unfortunately, the first of these interpretations leads to great difficulty. A unitarily evolving black hole that stores information on its stretched horizon, which in turn interacts with a local quantum field theory in the exterior of the black hole, succumbs to the AMPS arguments [4]: a smooth horizon is impossible in such a framework. However my collaborators and I suggested [5, 6, 7] that the degrees of freedom associated with the black hole (referred to as vacuum degrees of freedom) are thermally distributed both on *and outside* the stretched horizon of a black hole. By delocalizing black hole information, Hawking emission need not involve modes close to the horizon. Our proposal drops an assumption of complementarity: we demote field theory outside the horizon to something that only arises after coarse-graining the index k. This proposal is discussed in detail in chapter 6.

Before laying out the black hole proposal described above, we will focus a somewhat different line of work which may, fundamentally, be closely related to the idea of gauge fixing in quantum gravity. In [10], Bousso integrated the ideas of [11, 12] with his covariant entropy conjecture [9] and proposed a *covariant holographic principle*. The critical idea of the holographic principle is that a quantum states describing a spacetime should be defined on a a surface with one less dimension (as is the case in the AdS/CFT correspondence). Bousso suggested that the boundary of AdS spacetimes can be reasonably extended to general spacetimes by considering a special class of codimension one surfaces called *holographic screens* which are preferred from the perspective of the covariant entropy bound.

Recently [13, 14], a relation has been found between screens and surfaces studied by others [15, 16, 17, 18] in an attempt to develop a quasilocal definition of a black hole. This insight led to a refinement of the concept of holographic screens to that of *past and future holographic screens* and proved that these objects have monotonic area. This area law, analogous to the second law of black hole mechanics, is highly attractive from the perspective of quantum gravity. Unlike globally defined event horizons, holographic screens are highly non-unique objects, and can even be associated to observers in a very specific way (such a construction was discussed in [8]). They are thus well in-line with the ideas of gauge fixing. Moreover, holographic screens arise in a vast array of spacetimes including cosmological spacetimes where no asymptotic region exists.

In this thesis, we will discuss two critical developments in the theory of holographic screens:

- Past holographic screens are foliated by marginally anti-trapped surfaces called leaves. It was shown in [13, 14] that leaves have monotonic area. In chapter 2, we prove a stronger area law that shows that subregions of leaves also have monotonic area. This means suggests the validity of a "local second law of thermodynamics" that is valid in arbitrary spacetimes.

- Given that holographic screens are a natural extension of the AdS boundary to arbitrary spacetimes, in chapter 3, we present a generalization of holographic entanglement entropy proposals beyond the scope of AdS/CFT by anchoring extremal surfaces to past holographic screens. We will show that the properties of holographic screens are sufficient to prove that the areas of anchored extremal surfaces satisfy, for nontrivial reasons, expected properties of entanglement entropy like strong subadditivity.

The results and ideas we will discuss below are not intended to form an entirely precise theory. They must be treated as pieces of "data" that may eventually put physicists on the path to a coherent understanding of quantum gravity. A picture of spacetime that involves the concepts of holography and thermodynamics has been emerging ever since Bekenstein's work [20]. With such data constantly accumulating, we can perhaps begin to feel optimism that this picture will be completed soon.

Chapter 2

Refinement of the Holographic Screen Area Law

Black hole thermodynamics [19, 20, 22, 23, 21, 24, 25] is a critical principle that has guided the development of quantum gravity over the past few decades. In particular, Hawking's area theorem displayed parallels between the area of the event horizon of a black hole and entropy. This identification of entropy with area is the heart of the holographic behavior [11, 12] exhibited by gravity.

As discussed above, Bousso and Engelhardt [13, 14] recently proved an area law for surfaces called past and future holographic screens that arise in a more general setting than the spacetimes of black holes. These objects are not defined by the global notion of an event horizon and thus provide an example of "quasi-locally" defined surfaces with thermodynamic behavior.

Holographic screens are well-motivated from considerations in quantum gravity. The covariant entropy bound suggests [9, 10] that holographic screens play a role in general spacetimes that is analogous to the AdS boundary[1] in the context of the AdS/CFT correspondence [1, 2]. This hypothesis is supported by the recent demonstration that holographic entanglement entropy [26, 27, 28] can be defined for regions on past and future holographic screens in a way that is consistent with many known properties of entanglement entropy [29] (this idea will be presented in detail in chapter 3).

Holographic screens are generated by null foliations. Because null foliations are highly non-unique, holographic screens are also non-unique. For example, a past holographic screen can often be obtained by considering the surfaces of maximal area on the past light-cones of an observer's worldline. This procedure is only valid if the maximal area surfaces are anti-trapped (see equation 2.1 below) and compact which we assume. In this case, performing a modification to the worldline will modify the holographic screen.[2] From this point of view, holographic screens appear to be

[1] Ref. [8] studies a related construction.

[2] Note that the non-uniqueness of holographic screens for a given spacetime fits well with the ideas of [8, 6, 7] where

"pro-complementarity" objects.

Below we show that the Bousso-Engelhardt area law can be refined into a more local form. The original area law of [13, 14] states that preferred codimension-2 surfaces called leaves have monotonic area. We show that arbitrary subregions of leaves also have monotonic area. From the point of view of the holographic principle, this provides evidence that degrees of freedom of a holographic description for arbitrary spacetimes are locally distributed and satisfy a local version of the second law of thermodynamics.

Geometrical Preliminaries

Fix a globally hyperbolic spacetime M of dimension D satisfying the null curvature condition and the genericity conditions stated in [14].

Suppose that σ is a compact orientable codimension 2 spacelike submanifold of M. At any point on σ, up to normalization by positive real numbers, there exist exactly 4 null directions that are orthogonal to σ. These are often referred to as the future and past ingoing and outgoing light rays. Two of these four are future-directed. It is thus possible to find a pair of vector fields on σ, l and k, that are are null, orthogonal to σ, and future-directed. These vector fields are unique up to normalization by positive functions on σ.

Associated with these vector fields is a pair of scalar functions on σ called the *expansion* scalars, denoted by θ^l and θ^k. These scalars can be defined as functional derivatives of an area functional as follows. For any $p \in \sigma$, let $\gamma_p : \mathbf{R} \to M$ be the null geodesic intersecting p with tangent vector l. Let L denote the null surface generated by the collection of all such geodesics. Given a function $\lambda : \sigma \to \mathbf{R}$, we can obtain a corresponding codimension 2 surface $\Gamma(\lambda)$ called a *cross section* of L:

$$\Gamma(\lambda) = \{\gamma_p(\lambda(p)) \mid p \in \sigma\}.$$

If $\{x\}$ is a coordinate system on σ, it naturally generates coordinates on $\Gamma(\lambda)$ by following geodesics. We can compute the area of $\Gamma(\lambda)$, denoted by $\|\Gamma(\lambda)\|$, as

$$\|\Gamma(\lambda)\| = \int d^{D-2}x \sqrt{g^\lambda}$$

where g^λ is the determinant of the induced metric on the cross-section in these coordinates. Now, under an infinitesimal deformation $\lambda(x) \to \lambda(x) + \delta\lambda(x)$, there exists a function θ^l such that the area of the cross-section changes by

$$\delta\|\Gamma(\lambda)\| = \int d^{D-2}x \sqrt{g^\lambda}\, \theta^l(x, \lambda(x))\, \delta\lambda(x)$$

a strong emphasis is placed on the importance of "fixing the gauge" in quantum gravity. This is clearly discussed in [8] and in chapter 4 in which the role of a gauge-fixed apparent horizon (essentially a holographic screen though not a past or future screen) was discussed. In this chapter we do not commit to the pictures described in these papers.

which determines $\theta^l(x, \lambda)$ uniquely. The expansion scalar $\theta^l(x)$ on σ is defined by setting $\lambda(x) = 0$. θ^k is defined analogously. Intuitively, the expansion considers an infinitesimal patch of a spacelike surface and measures the rate of change of its area per unit area as it is deformed in a null direction.

We now introduce some terminology.

Definition 1. *Suppose that σ is a compact spacelike orientable codimension 2 submanifold of a spacetime. σ is said to be marginally anti-trapped if one of its two future-directed null expansion scalars is zero and the other is strictly positive. σ is said to be marginally trapped if one of its two future-directed null expansion scalars is zero and the other is strictly negative.*

Without loss of generality, when dealing with marginal surfaces we will usually take l to be the "marginal direction" i.e. the direction with $\theta^l = 0$. In this case, the above definition is

$$\begin{array}{cc} \underline{\text{Marginally Anti-Trapped}} & \underline{\text{Marginally Trapped}} \\ \theta^l = 0 & \theta^l = 0 \\ \theta^k > 0 & \theta^k < 0 \end{array} \tag{2.1}$$

The condition that $\theta^l = 0$ means that σ is the area-maximizing surface on the geodesic congruence generated by l and $-l$.

Holographic Screens and Area Laws

We now move to the most important definition in this chapter.

Definition 2. *A past holographic screen is a codimension-1 submanifold of the spacetime that is foliated by marginally anti-trapped surfaces called leaves. A future holographic screen is instead foliated by marginally trapped surfaces, also called leaves.*

For both past and future holographic screens, the foliation into leaves is unique: other splittings of a screen cannot satisfy the marginally anti-trapped or trapped condition.

The area law of [13, 14] is a statement about the evolution of leaves comprising a past or future holographic screen H. We denote the leaves of H by σ_r where r is a smooth parameter. In our notation, we can express the Bousso-Engelhardt area law as the statement that $\|\sigma_r\|$ is monotonic where $\|\cdot\|$ denotes the area functional. By convention, we will always choose the parameter r so that $\|\sigma_r\|$ is increasing. As before, on a particular leaf σ, let l and k denote the two future-directed null vector fields orthogonal to σ.

We define a vector field h on H by requiring that h is orthogonal to every leaf and by the normalization condition $dr(h) = 1$. The integral curves of h are called the *fibration*[3] of H. If we

[3]Note that h need not have definite signature. This is the key distinguishing feature between past (and future) holographic screens and related objects including future outer trapping horizons and dynamical horizons [15, 16, 17, 18]. Past and future holographic screens can be regarded as a synthesis such ideas with those of [9].

extend the definition of k and l to all of H, then $h = \alpha k + \beta l$ where α and β are smooth functions on H. The Bousso-Engelhardt area law was proven by showing that α never changes sign from which equation 2.1 implies that leaves have increasing area.

Our area law extends this result as follows. Suppose that A_0 is a region in σ_0. We can translate A_0 to a region A_r in σ_r by following the fibration from points in A_0 to σ_r. We will prove that the area of A_r is increasing. This conclusion relies on the fact that the area increase associated with zig-zagging along k and l is a first order effect in r, while the failure of such a zig-zag procedure to follow the fibration is at most a second-order effect.

Relation to the Screen Entanglement Conjecture

Holographic entanglement entropy proposals [26, 28] have recently been conjecturally generalized beyond the context of AdS/CFT by employing past or future holographic screens in arbitrary spacetimes (see chapter 3 and [29]). The proposed construction is to anchor extremal surfaces to the boundaries of subregions of leaves. The properties of past and future holographic screens are sufficient to ensure that the areas of these extremal surfaces satisfy expected properties of entanglement entropy like strong subadditivity. The statement that one fourth of the area of such extremal surfaces is in fact the entanglement entropy of a subsystem in a quantum theory holographically defining the spacetime in which the screen lies is called the "screen entanglement conjecture."

The area law proven here applies to subregions of leaves, the same objects to which an entanglement entropy-like quantity was assigned in [29]. Suppose that A_0 is a region in the leaf σ_0 and A_r is the result of translating A_0 along the fibration to σ_r. Let $S(A_r)$ denote the screen entanglement entropy of A_r as defined above via the extremal surface anchored to ∂A_r. With the exception of cases that are topologically nontrivial, $S(A_r)$ satisfies a "Page bound": $S(A_r) \leq \min(\|A_r\|, \|\sigma_r \setminus A_r\|)$. Our area law applies to the evolution of the subregions A_r and A_r^C and thus causes the Page bound to become less restrictive whenever r is increased. This does not prove that $S(A_r)$ increases monotonically.

2.1 Proof of the Area Law for Subregions

From here on we will assume that H is a past holographic screen. Our argument can be modified to the case of a future holographic screen in an obvious way. Because H is a past screen,

$$\begin{aligned} \theta^l &= 0 \\ \theta^k &> 0. \end{aligned} \tag{2.2}$$

Moreover, we now have $\alpha > 0$ on all of H.

To carefully study the evolution of areas of regions in leaves, it is convenient to consider the null surfaces passing through a leaf σ_r. First, extend l and k to a tubular neighborhood of H by following along the geodesics generated by l and k. Now let N_r denote the null surface obtained by starting from points on σ_r and following the integral curves of l in both the $+l$ and $-l$ directions. Let L_r^+ denote the null surface obtained by starting at σ_r and following the integral curves of k only in the $+k$ direction.

We now fix an (arbitrarily chosen) reference leaf σ_0. There exists an $r_0 > 0$ such that if $0 < r < r_0$, it is possible to define a "zig-zag" map $f_r : \sigma_0 \to \sigma_r$ as follows. If $p \in \sigma_0$, follow L_0^+ from p along a generator of L_0^+ (i.e. along the integral curve of k that p lies on) until L_0^+ intersects a generator of N_r. Then, follow the N_r generator to σ_r. Bousso and Engelhardt established that f_r is well-defined for sufficiently small r (this is why we restrict to $r < r_0$). f_r is, in fact, a diffeomorphism between σ_0 and σ_r.

Considering equation 2.2 and the fact that $\alpha > 0$, the zig-zag construction of f_r implies that if A_0 is a $D-2$ dimensional submanifold of σ_0,

$$\frac{d}{dr}\Big|_{r=0} \|f_r(A_0)\| = \int_{A_0} \sqrt{g^{\sigma_0}}\, \alpha\, \theta^k \; > 0. \tag{2.3}$$

The area law of Bousso and Engelhardt is obtained in the case where $A_0 = \sigma_0$ because f_r is surjective.

Aside from the case where $A_0 = \sigma_0$, the fact that $\|f_r(A_0)\|$ is an increasing function of r is an unattractive area law. One issue is that the definition of the function f_r involves the choice of the reference leaf (i.e. the choice of $r = 0$). Moreover, the family of regions $\{f_r(A_0)\,|\,r \in [0, r_0)\}$ cannot necessarily be extended to all r.

Fortunately, as described above, there is a simpler way to carry subregions from one leaf to the next. Let A_{r_1} be a $D-2$ dimensional submanifold of the leaf σ_{r_1}. Define $A_{r_2} \subset \sigma_{r_2}$ by starting from points in A_{r_1} and following along the fibration of H (i.e. the integral curves of h) by parameter $r_2 - r_1$. Note that this procedure gives a well-defined region $A_r \subset \sigma_r$ for the entire range of r. We now prove that $\|A_r\|$ is an increasing function.

First, the following Lemma shows that f_r behaves similarly to h-translation for small r:

Lemma 1. *If $p_0 \in \sigma_0$, let $\gamma : [0, r_0) \to H$ be the curve on H defined by $\gamma(r) = f_r(p_0)$. Then, the tangent vector of γ at $r = 0$ is $h(p_0)$.*

Proof. We will begin by introducing a set of convenient coordinates. Fix a coordinate chart on σ_0 for a neighborhood of p_0. We denote these coordinates by x^i, $i \in \{1, \ldots D-2\}$ and require that p_0 corresponds to the origin of $\mathbf{R}^{D-2}$. Extend to coordinates $\{(x^i, r)\}$ on H by following the integral curves of h from x^i by parameter r to reach the point labeled by (x^i, r). Note that this point will lie in σ_r. Finally, extend to coordinates $\{(x^i, r, z)\}$ by starting from the point (x^i, r) and following the integral curves of l by affine parameter z. Note that H is the $z = 0$ hypersurface.

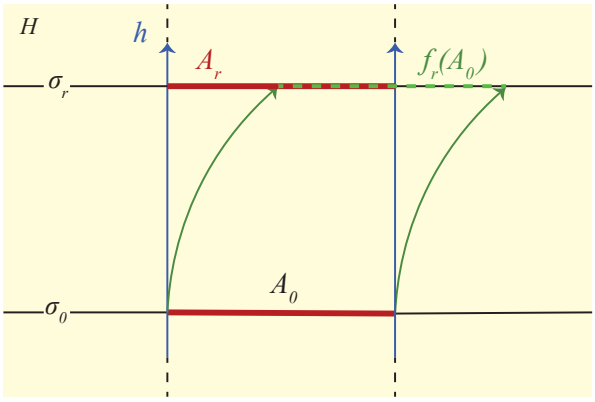


Figure 2.1: We show that A_r has monotonic area by comparing A_r with the region $f_r(A_0)$. As depicted here, A_r and $f_r(A_0)$ are identical at linear order in r.

Because $\alpha \neq 0$, we can put $k\big|_H = \frac{1}{\alpha}h - \frac{\beta}{\alpha}l$. Thus, in the coordinates (x^i, r, z) constructed above, we have

$$\begin{aligned} h &= (\mathbf{0}, 1, 0) \\ k\big|_{z=0} &= (\mathbf{0}, \frac{1}{\alpha}, -\frac{\beta}{\alpha}) \\ l\big|_{z=0} &= (\mathbf{0}, 0, 1) \end{aligned} \tag{2.4}$$

where $\mathbf{0}$ denotes $D-2$ zeros. The curve $\gamma(r)$ also takes a simple form in our coordinates: because f_r maps points in σ_0 to points in σ_r, we have

$$\gamma(r) = (x^i(r), r, 0) \tag{2.5}$$

where $x^i(r)$ is a curve in $\mathbf{R}^{D-2}$. Our Lemma will be proven by showing that $\dot{x}^i(0) = 0$.

Let $\xi_0(\lambda)$ and $\zeta_r(\lambda)$ denote, respectively, the geodesics generated by k and l from the points $\gamma(0) = (\mathbf{0}, 0, 0)$ and $\gamma(r) = (x^i(r), r, 0)$. The zig-zag definition of f_r implies that ξ_0 and ζ_r have an intersection: there exist functions $\lambda_1(r)$ and $\lambda_2(r)$ such that

$$\xi_0(\lambda_1(r)) = \zeta_r(\lambda_2(r)). \tag{2.6}$$

Meanwhile, equation 2.4 implies that

$$\begin{aligned} \xi_0(\lambda_1(r)) &= \Big(\mathbf{0}, \frac{1}{\alpha_0}\lambda_1(r), -\frac{\beta_0}{\alpha_0}\lambda_1(r)\Big) + O\big(\lambda_1(r)^2\big) \\ \zeta_r(\lambda_2(r)) &= \Big(x^i(r), r, \lambda_2(r)\Big) + O\big(\lambda_2(r)^2\big) \end{aligned} \tag{2.7}$$

where $\alpha_0 = \alpha(r=0)$ and $\beta_0 = \beta(r=0)$. Comparing the r and z components of equation 2.7 now gives

$$\begin{aligned} \lambda_1(r) &= \alpha_0\, r + O\big(\lambda_1(r)^2, \lambda_2(r)^2\big) \\ \lambda_2(r) &= -\beta_0\, r + O\big(\lambda_1(r)^2, \lambda_2(r)^2\big) \end{aligned} \tag{2.8}$$

which then implies that

$$x^i(r) = O\big(\lambda_1(r)^2, \lambda_2(r)^2\big) = O\big(r^2\big). \tag{2.9}$$

We conclude that $\dot{x}^i(r=0) = 0$. □

Theorem 1. *Let H be a past holographic screen in a D dimensional spacetime satisfying the genericity conditions of [14]. Let $\{\sigma_r\}$ be the foliation of H into marginally anti-trapped surfaces (i.e. leaves). Let h be the leaf orthogonal vector field in H normalized so that $dr(h) = 1$. Suppose that $A_0 \subset \sigma_0$ is a $D-2$ dimensional submanifold of σ_0 and define A_r as the result of translating A_0 along the integral curves of h by parameter r. Then, A_r has strictly increasing area.*

Proof. Take $r \in [0, r_0)$. We have

$$\Big| \|f_r(A_0)\| - \|A_r\| \Big| \leq \|f_r(A_0) \,\Delta\, A_r\| \tag{2.10}$$

where Δ denotes the symmetric difference of sets: $A \Delta B = (A \setminus B) \cup (B \setminus A)$. Now Lemma 1 and the compactness of σ_r implies that

$$\frac{d}{dr}\Big|_{r=0} \Big(\|f_r(A_0) \,\Delta\, A_r\| \Big) = 0.$$

Noting that both sides of equation 2.10 are nonnegative for all r and are zero at $r = 0$, we conclude that

$$\frac{d}{dr}\Big|_{r=0} \left(\Big| \|f_r(A_0)\| - \|A_r\| \Big| \right) = 0. \tag{2.11}$$

But equation 2.3 implies that $\|f_r(A_0)\|$ is increasing at $r = 0$ so we must have that $\|A_r\|$ is also increasing at $r = 0$.

While we have only proven that A_r has increasing area at $r = 0$, we can define a zig-zag function analogous to f from any reference leaf and repeat all arguments above for any r. Thus, we conclude that A_r has strictly increasing area. In fact, equations 2.3 and 2.11 show that

$$\frac{d}{dr}\|A_r\| = \int_{A_r} \sqrt{g^{\sigma_r}}\, \alpha\, \theta^k \; > 0.$$

□

Chapter 3

The Screen Entanglement Conjecture

We now turn to a major recent development in the study of past and future holographic screens. Holographic entanglement entropy, proposed by Ryu and Takayanagi (RT) [26], proved by Lewkowycz and Maldacena [27], and made covariant by Hubeny, Rangamani, and Takayanagi (HRT) [28], is a beautiful property (or, in the covariant case, conjecture) of AdS/CFT. In this chapter we will introduce a way to promote holographic entanglement entropy beyond the scope of AdS/CFT that applies just as well to cosmological spacetimes as it does to asymptotically AdS spacetimes. In the case of the latter, it reduces to the HRT proposal. Moreover, the promoted holographic entanglement entropy satisfies, for nontrivial reasons, expected properties of entanglement entropy like strong subadditivity.

The HRT prescription provides a way to compute entanglement entropy of a spatial region A in a quantum state dual to an AlAdS spacetime. The procedure is to consider ∂A, the boundary of the spatial region, and to find the area of a codimension 2 extremal surface that is anchored to ∂A. A naïve extension of this idea to general spacetimes would be to take A to be a region in the conformal boundary of an arbitrary spacetime. This approach fails: what is the boundary of a closed FRW universe with past and future singularities?

In our proposal, we anchor extremal surfaces to a past or future holographic screen. The definition of these codimension surfaces was given above in chapter 2. Note that holographic screens, in a less refined form, were proposed by Bousso [10] in an attempt to find the analogue of the AdS boundary when extending holography to general spacetimes. If one believes the covariant entropy bound [9], then there is essentially no other reasonable class of surfaces for this purpose.

3.1 Holographic Screen Entanglement Entropy

Throughout this chapter, we will work in a spacetime M that is globally hyperbolic (or, in the AlAdS case, satisfies a generalization of global hyperbolicity) and that satisfies the null energy condition. Put $d = \dim M$. Let $\mathcal{H}$ denote a past holographic screen. Everything below can be

modified to the case of a future holographic screen without subtlety.

We will now assume some *genericity conditions*.

- *Strict Focusing.* If B is a codimension 2 spacelike surface, the four surface-orthogonal null congruences have strictly decreasing expansion as they move away from B.
- *Strict Second Law of Holographic Screens.* If leaves of $\mathcal{H}$ are smoothly parameterized as $\sigma(r)$ with $h^a = \partial_r^a$ nonzero, then area$(\sigma(r))$ has a nonzero derivative for all r.

The sense in which the second condition is generic was reviewed in chapter 2. If our spacetime fails to satisfy these conditions, it can be made to do so by sprinkling a very small amount of classical matter everywhere.

As discussed in the previous chapter, there is a unique foliation of $\mathcal{H}$ into anti-trapped leaves. Let σ be a particular leaf in this foliation and let k and l denote the vector fields on σ that satisfy equation 2.2. Because M is globally hyperbolic, there exists a Cauchy surface S_0 containing σ such that $S_0 \setminus \sigma$ consists of a disconnected interior and exterior. The interior of S_0 is defined so that a vector on σ pointing toward the interior takes the form $-c_1 k + c_2 l$ with $c_1, c_2 > 0$. Let S denote the union of the interior of S_0 with σ. We will assume that S is compact and that it has the topology of a solid ball. Now let D_σ be the domain of dependence of S, $D_\sigma = D(S)$, with the convention that D_σ includes orthogonal null surfaces generated by l and $-k$.

Suppose that A is a $d-2$ dimensional submanifold of σ with a boundary. Consider the set of extremal codimension 2 surfaces that are anchored to and terminating at ∂A, and contained entirely in D_σ (see figure 3.1). In section 3.2 we will give conditions on D_σ that ensure that this set is not empty. Taking the existence of such a surface for granted, let the one of minimal area be denoted by ext(A) and define the *holographic screen entanglement entropy* (or *screen entanglement entropy* for brevity) of A as

$$S(A) = \frac{\text{area}(\text{ext}\,(A))}{4}. \tag{3.1}$$

The quantity $S(A)$ is the most natural generalization of the HRT proposal to general spacetimes. We emphasize that we have defined screen entanglement entropy geometrically without reference to a quantum theory. The term "entanglement entropy" is only meant suggestively. Nonetheless, below we state a *screen entanglement conjecture*: that $S(A)$ is in fact the von Neumann entropy of a subsystem of a holographic quantum state for general spacetimes. Regardless of the validity of this conjecture, we are free to study $S(A)$ as we have defined it. As we will see, the properties of holographic screens ensure that screen entanglement entropy possesses numerous properties reminiscent of von Neumann entropy which we now discuss.

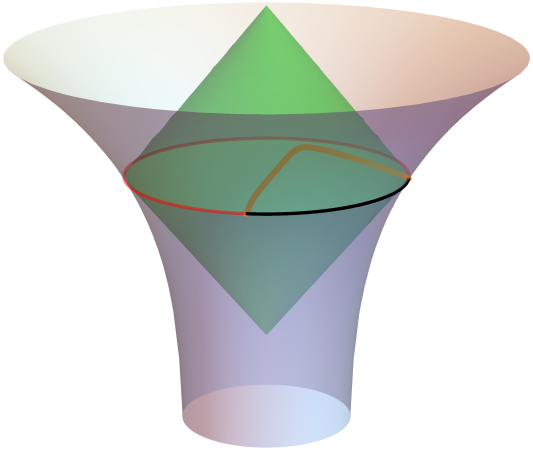


Figure 3.1: This figure depicts our construction of holographic entanglement entropy in general spacetimes. The horn-shaped surface is a past holographic screen $\mathcal{H}$. The black and red codimension 2 regions together form a single leaf σ. The black segment represents a region A and the extremal surface ext (A) (orange) is anchored to its boundary. The causal region D_σ is shown in green. Note that ext $A \subset D_\sigma$.

Properties of Holographic Screen Entanglement Entropy and Extremal Surfaces

- *Existence and Containment.* In section 3.2 we provide conditions for ext (A) to exist. This is a nontrivial issue because of the "containment condition" that ext $A \subset D_\sigma$. Arguments that D_σ contains an extremal surface rely critically on the assumption that A is in a leaf of a holographic screen. Moreover, the condition that ext $(A) \subset D_\sigma$ gives rise to properties of holographic screen entanglement entropy like strong subadditivity (see below) and will allow us to reasonably define an entanglement wedge for A. For an example of the importance of the containment condition, see equation 3.3 below and the paragraphs around it.

- *(Strong) Subadditivity.* Suppose that A and B are regions in σ. Then,

$$S(A) + S(B) \geq S(A \cup B) + S(A \cap B)$$

  where S is the function defined in 3.1. This result holds regardless of whether or not A and B intersect as long as we take the convention that $S(\emptyset) = 0$. As we will see in section 3.2, the proof of this is a modified version of Wall's [30] "maximim" proof for the HRT case. This does not mean that strong subadditivity is an obvious result: most of the work in section

3.2 is to show that the properties of leaves of holographic screens are sufficient to generalize Wall's arguments to our context.

- *Page Bounded.* Define the extensive entropy of A as $S_{\text{extensive}}(A) = \text{area}(A)/4$. Then, the holographic screen entanglement entropy satisfies the following *Page bound*:[1]

$$S(A) \leq \min\{S_{\text{extensive}}(A), S_{\text{extensive}}(\sigma \setminus A)\}. \tag{3.2}$$

  This is a simple consequence of the maximin construction we give in section 3.2. Note that the strict second law for holographic screens implies that this inequality becomes a weaker constraint if we transport A along the fibration vector field defined in chapter 2. In certain cases, the inequality saturates and $S(A)$ approaches a "random entanglement limit." (See section 3.3 for examples of this in cosmology.)

- *Reduction to the HRT Proposal.* As explained in detail in [10], the AdS boundary can be regarded as a holographic screen. In this case, surfaces of constant time in the dual field theory correspond to leaves, and our proposal becomes identical to the covariant holographic entanglement entropy conjecture of [28].

The Screen Entanglement Conjecture

The content of this work does not rely on any unverified statement. Nonetheless, for the purposes of better motivating the definition of holographic screen entanglement entropy, we allow ourselves to make a speculative conjecture about the role of $S(A)$ in quantum gravity.

Our proposal can be regarded as an extension of a covariant holographic principle due to Bousso which we now review. In [10], Bousso integrated the ideas of [11, 12] with his covariant entropy conjecture [9] and proposed that each marginal surface B foliating a holographic screen is associated with a Hilbert space $\mathcal{H}_B$ of dimension $\exp(\text{area}(B)/4)$ and that states in $\mathcal{H}_B$ holographically define the state on a null surface N passing through B in the marginal direction. For our purposes, this holographic principle takes the following form. To each leaf σ of a holographic screen we assign a density matrix ρ_σ. The density matrix acts on a Hilbert space of dimension $\exp(\text{area}(\sigma)/4)$ which may be a subspace of a "complete" Hilbert space.[2] The covariant entropy bound suggests that ρ_σ encodes the quantum information on the null slice generated by l and $-l$ in the notation of chapter 2.

[1] The term "Page bound" is motivated by Page's considerations of the entanglement entropies of subsystems [31].

[2] The concept that the states corresponding to any particular approximately fixed geometry form a subspace of a complete Hilbert space is due to Nomura [32, 33]. In his formulation, a larger Hilbert space for arbitrary geometries is a direct sum over subspaces for each geometry. This direct sum itself is only a subspace of the complete Hilbert space which may include an "intrinsically stringy" subspace with no geometrical interpretation. This construction may provide insight into how quantum mechanics can be unitary despite the fact that screens have non-constant area.

We now assume Bousso's holographic principle and state our new conjecture. We propose that every region A of σ (up to string scale resolution) corresponds to a subsystem of the Hilbert space that ρ_σ acts on. We conjecture that the von Neumann entropy of that subsystem in the density matrix ρ_σ is given, at leading order, by $S(A)$ as we have defined it in equation 3.1.

We refer to this statement as the *screen entanglement conjecture*. Because a holographic quantum theory dual to arbitrary spacetimes is not known, the screen entanglement conjecture is not a mathematical statement about the relation between two known theories (as in the case of HRT). Instead, our conjecture suggests a way to compute properties of quantum states in an unknown theory. It is our hope that this will, in fact, be a step toward developing a quantum theory for arbitrary spacetimes.

3.2 Proofs of Strong Subadditivity and Other Relations

In this section we prove key technical results about holographic screen entanglement entropy including many of the properties advertised above. The notation and conventions we will use are the same as those given in chapter 2 and section 3.1. In particular, $\mathcal{H}$ is a past holographic screen in a globally hyperbolic spacetime of dimension d that satisfies the genericity conditions of section 3.1. σ is a compact leaf of $\mathcal{H}$ which we assume to have the topology of S^{d-2}. k and l are null orthogonal vector fields on σ satisfying equation 2.2. S_0 is a Cauchy slice containing σ and S is the portion of S_0 that is enclosed by σ including σ itself (the enclosed side is defined in section 3.1). S is assumed to have the topology of a compact $d-1$ ball. D_σ is the domain of dependence of S.

As always, the case of a future holographic screen is omitted because it presents no additional subtlety.

Existence and Containment of Extremal Surfaces

As discussed in section 3.1, it is nontrivial and critical to show the existence of an extremal surface anchored to ∂A that lies entirely in D_σ. We now prove that such a surface exists under very generic conditions. Our first step is to show that ext (A) exists in the case that D_σ is compact. This is a common situation[3] although it is not the case if the ingoing light sheets of σ encounter a singularity.

[3]Suppose that the future and past ingoing light-sheets of σ terminate at caustics rather than singularities. Let C_+ and C_- denote the set of the first caustics encountered (local or nonlocal) by null geodesics in the future and past light sheets respectively. Then, if $D_\sigma = J_-(C_+) \cap J_+(C_-)$, we can conclude that D_σ is compact. This follows from the fact that $C_\pm$ inherits the compactness of σ and from the fact that global hyperbolicity implies that $J_-(K_1) \cap J_+(K_2)$ is compact if K_1 and K_2 are compact.

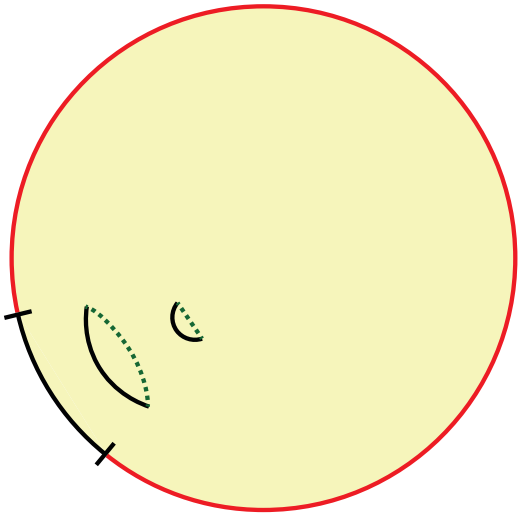


Figure 3.2: The proof of lemma 2 involves a continuous family of surfaces A_s along with their extremal surfaces (dotted curves).

Lemma 2. *If D_σ is compact, then there exists a codimenson 2 extremal surface anchored and terminating at ∂A that lies entirely in D_σ and that intersects ∂D_σ only at ∂A.*

Proof. Let Σ_+ and Σ_- denote the future and past ingoing light-sheets of σ. We now extend Σ_- to a slightly larger light-sheet, $\tilde{\Sigma}_-$, by following the future directed null congruence of k. Because $\theta^k > 0$ on σ, we can make this extension so that $\tilde{\Sigma}_-$ has $\theta^k > 0$ everywhere and so that there exists an open set in $\tilde{\Sigma}_-$ containing σ.

In the language of [34], both $\Sigma_+ \setminus \sigma$ and $\tilde{\Sigma}_-$ are extremal surface barriers because they have negative expansion in the l and $-k$ directions respectively. Moreover, $\partial D_\sigma \subset (\Sigma_+ \setminus \sigma) \cup \tilde{\Sigma}_-$. It follows that ∂D_σ is itself an extremal surface barrier for extremal surfaces in the interior[4] of D_σ.

Now consider the region A. The spherical topology[5] of σ ensures that it is possible to introduce a continuous one-parameter family of submanifolds of D_σ, A_s, such that

- A_0 consists of a single point in the interior of D_σ
- $A_1 = A$
- for $0 < s < 1$, A_s is a codimension 2 submanifold of the interior of D_σ that is diffeomorphic to A.

[4] In [34], extremal surfaces are confined to regions referred to as the "exterior" of an extremal surface barrier. The interior of D_σ, i.e. $D_\sigma \setminus \partial D_\sigma$, is analogous to exterior regions studied by Wall and Engelhardt.

[5] We remind the reader that our conventions are those laid out in the first paragraph of section 3.1. In particular, we are making simplifying topological assumptions about σ and S. We will leave it to future work to investigate the consequences of relaxing these assumptions.

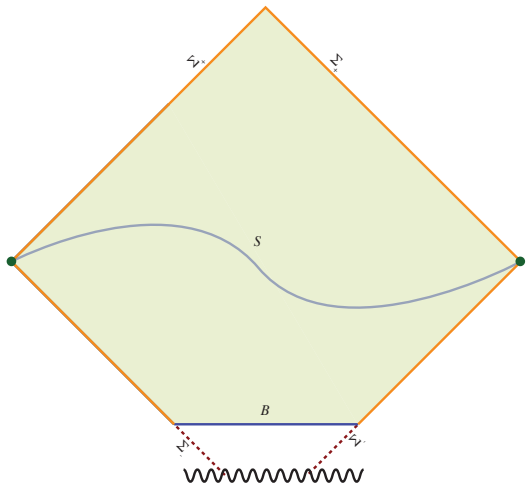


Figure 3.3: The idea of a compact restriction is shown here. The restriction R is the shaded region along with its boundary, the blue and orange lines. ∂R consists of two parts: an extremal surface barrier B (blue) and a portion of ∂D_σ (orange). In this figure, the barrier B protects extremal surfaces in R from a singularity. Not shown are extremal surfaces in R, none of which contact ∂R except at their anchor on the leaf σ.

This is shown in figure 3.2. Note, in particular, that if $s < 1$, $A_s \cap \partial D_\sigma = \emptyset$.

If $\epsilon > 0$ is sufficiently small, then the extremal surface of minimal area that is anchored to ∂A_ϵ lies entirely in the interior of D_σ. Denote this extremal surface by $\Gamma(\epsilon)$. Consider increasing the value of the parameter s from ϵ to 1. For each value of s, construct an extremal surface $\Gamma(s)$ (not necessarily the one of minimal area) anchored to ∂A_s. The compactness of D_σ (which ensures that it is bounded and has no singularities) together with the fact that, as discussed above, ∂D_σ is an extremal surface barrier, allows us to take $\Gamma(s)$ to not jump discontinuously and to be contained in the interior of D_σ for all $s < 1$. When we take the limit sending s to 1, the extremal surface anchored to ∂A must intersect ∂D_σ at ∂A and nowhere else: if it did intersect ∂D_σ outside of ∂A, the extremal surface would be locally tangent to an extremal surface barrier with strictly nonzero null extrinsic curvature.

□

The unwanted assumption that D_σ is compact (which fails in the event that Σ_+ or Σ_- encounter a singularity) can be dropped if there exists a codimension 0 submanifold (with boundary) of D_σ, R, which "restricts" extremal surfaces (see figure 3.3). By this we mean that

1. R is compact,

2. There exists an open set U containing S with $D_\sigma \cap U = R \cap U$, and

3. $\partial R = (\partial D_\sigma \cap R) \cup B$ where B is an extremal surface barrier for codimension 2 extremal surfaces inside in R.

These conditions for R are designed to ensure that R can be used in lemma 2 in place of D_σ without difficulty. The existence of such regions R relies on the existence of the barrier B. The arguments in theorem 11 of [30] show that Kasner singularities are always protected by such barriers. Hartman and Maldacena [35] encountered a barrier protecting black hole singularities from codimension 2 extremal surfaces. Constant time slices in FRW spacetimes are another example of suitable barriers.[6]

Any region $R \subset D_\sigma$ satisfying the conditions will be called a *compact restriction* of D_σ. Note that, in particular, if D_σ is compact then D_σ is a compact restriction of itself. Our findings can now be summarized by the following improvement upon lemma 2:

Theorem 2. *If D_σ possesses a compact restriction, then there exists a codimenson 2 extremal surface anchored and terminating at ∂A that lies entirely in D_σ and that intersects ∂D_σ only at ∂A.*

To better appreciate this theorem, it is helpful show that the statement is false if σ is not a leaf of a holographic screen. Consider $2+1$ dimensional Minkowski space with inertial coordinates (t, x, y) and let $\mathcal{C}$ denote the large cylinder satisfying $x^2 + y^2 = R^2$ with $R \gg 1$. Consider the two line segments on $\mathcal{C}$ that are approximately given by

$$\begin{aligned} A &= \{(t = \frac{1}{2}|x|, -1 > x \geq 0, y = R)\} \\ B &= \{(t = \frac{1}{2}|x|, 0 \leq x < 1, y = R)\} \end{aligned} \tag{3.3}$$

and construct any spacelike "time slice" on $\mathcal{C}$, σ, that includes AB. It is easy to see that the extremal surface anchored to $\partial(AB)$ is a straight line that is timelike related to AB and thus fails to lie within the domain of dependence D_σ. To see how severe this problem is, note that the segments A and B fail to satisfy subadditivity of entanglement entropy. That is, the inequality $S_A + S_B \geq S_{AB}$ is false. Note that in this example σ fails to satisfy equation 2.2 because of the kink at $A \cap B$.

A Maximin Construction for Holographic Screens

Theorem 2 ensures that holographic screen entanglement entropy is a well-defined quantity in a broad set of cases. We will now demonstrate that this quantity satisfies expected properties of

[6] Many extremal surfaces are anchored at singularities and thus pass through barriers. This is irrelevant because the barriers we are discussing here play the of ∂D_σ in the proof of lemma 2. As a region A_s is deformed from a point inside R into $A \subset \sigma$, extremal surfaces anchored to ∂A_s cannot smoothly pass B or ∂D_σ.

entanglement entropy. To do this, it is very useful to closely follow [30] and introduce a maximin construction of ext A. Our construction will be slightly modified from that used for HRT surfaces anchored to the AdS boundary. Wall's maximin prescription involves considering a collection of Cauchy slices that are anchored only to ∂A. Because we already know that ext A lies inside of D_σ, we will introduce a stronger constraint requiring that we only consider achronal slices that are anchored to all of σ.

Definition and Existence of Mm(A)

Our setup remains unchanged. Fix a past (or future) holographic screen $\mathcal{H}$ in a globally hyperbolic spacetime and let σ be a leaf. We take a Cauchy surface S_0 containing σ and define S as the closure of the portion of S_0 inside of σ. As before, we require that S is compact and that it has the topology of a solid $d-1$ ball. Let $D_\sigma = D(S)$. We also fix a region A in σ with a boundary. Now define $\mathcal{C}_\sigma$ as the collection of codimension 1 compact achronal surfaces that are anchored to σ and that have domain of dependence D_σ. Note, in particular, that $S \in \mathcal{C}_\sigma$. Moreover, note that the global hyperbolicity of D_σ ensures that every element of $\mathcal{C}_\sigma$ has the same topology as S: that of a compact $d-1$ ball.

Take any $\Sigma \in \mathcal{C}_\sigma$. Let $\min(\partial A, \Sigma)$ denote the codimension 2 surface of minimal area[7] on Σ that is anchored to ∂A. The existence of $\min(\partial A, \Sigma)$ is guaranteed by the compactness of Σ and theorem 9 of [30]. Define a function $F : \mathcal{C}_\sigma \to [0, \text{area}(A)]$ by $F(\Sigma) = \text{area}(\min(\partial A, \Sigma))$. Now assume that there exists a Σ_0 in $\mathcal{C}_\sigma$ that maximizes F (globally). We now define $\min(\partial A, \Sigma_0)$ as the maximin surface of A, and we will denote it by $\text{Mm}(A)$. If there are several maximin surfaces, $\text{Mm}(A)$ can refer to any of them.

The existence of $\text{Mm}(A)$ can be proven in many cases by appropriately importing the arguments of theorems 10 and 11 in [30] which we only briefly describe here. Consider the Cauchy surface S_0 which can be identified as a slice in a foliation of Cauchy surfaces $\{S_t\}$. Using this definition of time, we can identify a surface $\Sigma \in \mathcal{C}_\sigma$ with a function $t_\Sigma : S_0 \to \mathbf{R}$ in a natural way: if I_x denotes the integral curve of ∂_t that passes through a point $x \in S$, then $\Sigma = \{I_x \cap S_{t_\Sigma(x)} | x \in S\}$. From this viewpoint, F can be regarded as a real-valued functional on $\{t_\Sigma\}$. Now if D_σ is compact, we can find the maximum and minimum values of t for the set D_σ to obtain an upper and lower bound on t_Σ that applies for all Σ. Moreover, the condition that Σ be compact and achronal ensures that $\{t_\Sigma\}$ is equicontinuous. These facts imply that $\mathcal{C}_\sigma$ is compact (with the uniform topology) and that the extreme value theorem applies to the function F.

In the case where D_σ is not compact (for instance, due to a singularity terminating a light sheet of σ), we can still argue that F has a maximum as long as D_σ satisfies a condition similar to but

[7] Wall [30] added the condition that $\min(\partial A, \Sigma)$ be homologous to A. While this condition ought to be included in our discussion as well, the assumption that S (and thus every element of $\mathcal{C}_\sigma$) has the topology of a compact $d-1$ ball makes a homology condition trivial. We leave the task of investigating more general topologies to future work.

slightly stronger than the "compact restriction" idea discussed above. Suppose that B_+ is a surface in $\mathcal{C}_\sigma$ which is identical to Σ_+ in some neighborhood of S. For any $\Sigma \in \mathcal{C}_\sigma$, define another surface $\bar{\Sigma}$ by $t_{\bar{\Sigma}} = \min\{t_\Sigma, t_{B_+}\}$. If B_+ has the property that for any Σ we have $F(\Sigma) \leq F(\bar{\Sigma})$, then we will say that B_+ is a *future maximin barrier*. A past maximin barrier is defined analogously as a surface $B_- \in \mathcal{C}_\sigma$, identical to Σ_- in a neighborhood of S, such that for any Σ we have $F(\Sigma) \leq F(\bar{\Sigma})$ where $\bar{\Sigma}$ is defined by $t_{\bar{\Sigma}} = \max\{t_\Sigma, t_{B_-}\}$.

Now if D_σ possesses both a past and future maximin barrier, then we can restrict our attention to the subset of surfaces in $\mathcal{C}_\sigma$ that satisfy $t_{B_-} \leq t_\Sigma \leq t_{B_+}$. Let $\mathcal{C}_\sigma(B_-, B_+)$ denote this restricted set. Because B_- and B_+ are compact, $J_+(B_-) \cap J_-(B_+)$ is compact and so the set $\mathcal{C}_\sigma(B_-, B_+)$ is compact in the uniform topology and F has a maximum $\Sigma_0 \in \mathcal{C}_\sigma(B_-, B_+)$. The definition of past and future maximin barriers ensures us that if $\Sigma \in \mathcal{C}_\sigma$, then $F(\Sigma_0) \geq F(\Sigma)$. Thus, Σ_0 is a global maximum for F and we can safely define $\min(\partial A, \Sigma_0)$ as the maximin surface of A, $\mathrm{Mm}(A)$.

As in the case of the compact restriction of D_σ used in theorem 2, it is difficult to find examples where D_σ does not possess a past and future barrier. Wall [30] argued that such barriers protect maximin surfaces from a wide range of singularities: approximately Kasner singularities, BKL singularities, and FRW big bangs all lead to past or future maximin barriers. If $\Sigma_\pm$ simply terminate at caustics rather than singularities, then $B_\pm = \Sigma_\pm$ are barriers. In any event, if $B_\pm$ exist, then the region $J_+(B_-) \cap J_-(B_+)$ provides a compact restriction of D_σ in the sense of theorem 2. Thus, the existence of $B_\pm$ ensures both the existence of $\mathrm{Mm}(A)$ as well as the existence of ext (A). From here on, we will simply take for granted that a past and future maximin barrier exist.

Equivalence of Mm(A) and ext (A)

Below we will argue that $\mathrm{Mm}(A) = \text{ext}\ (A)$. However, it is first very useful to introduce two additional definitions first.

1. Take $\Sigma \in \mathcal{C}_\sigma$ and let Γ be a codimension 2 surface anchored to ∂A that lies in D_σ. Consider the intersection between Σ and the future and past-directed orthogonal null surfaces of Γ that are directed toward A. This intersection is called the representative of Γ on Σ and will be denoted by $\mathrm{rep}(\Gamma, \Sigma)$.

2. The domain of dependence of codimension 1 achronal surfaces anchored to $A \cup \text{ext}\ A$ lying in D_σ will be called the entanglement wedge of A.

Note that $\mathrm{rep}(\Gamma, \Sigma)$ is itself a codimension 2 surface anchored to ∂A that lies on Σ. Moreover, if Γ is extremal then, by the focusing theorem, $\mathrm{area}(\mathrm{rep}(\Gamma, \Sigma)) \leq \mathrm{area}(\Gamma)$.

We now demonstrate that our maximin procedure always finds ext A, the extremal surface of minimal area that is anchored to ∂A and which lies in D_σ. While much of the proof is similar to

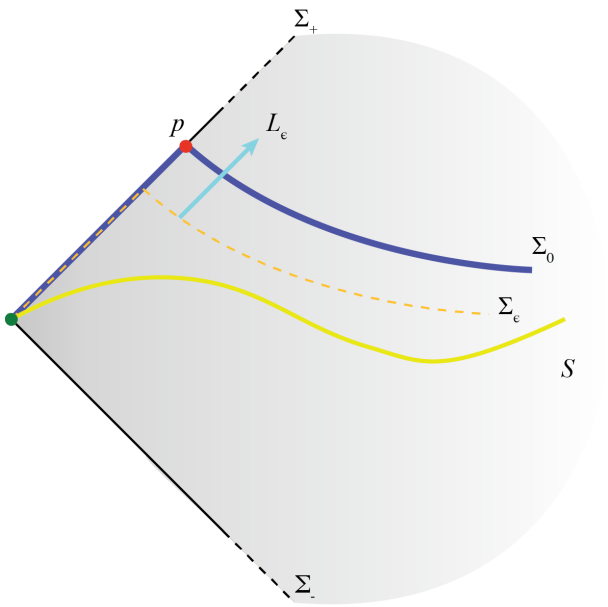


Figure 3.4: This figure depicts the argument of case 1 of the proof of theorem 3. Note that the surface S is shown here for reference and that it does not play a critical role in the proof. The shaded region is $D_\sigma = D(S)$ and the green dot is (a cross-section of) the leaf σ.

the arguments in [30], we will have to pay special attention to the possibility that the maximin surface could run into the boundary of D_σ.

Theorem 3. *$Mm(A) = ext\,(A)$.*

Proof. The argument of theorem 15 in [30] immediately shows that if a point $p \in \mathrm{Mm}(A)$ is also in the interior of D_σ (i.e. $D_\sigma \setminus \partial D_\sigma$), then $\mathrm{Mm}(A)$ is extremal at p. In particular, if $\mathrm{Mm}(A) \cap \partial D_\sigma = \partial A$, then $\mathrm{Mm}(A)$ is an extremal surface everywhere. We now argue that $\mathrm{Mm}(A)$ in fact cannot ever intersect ∂D_σ outside of ∂A.

Suppose there exists $p \in \mathrm{Mm}(A) \cap (\partial D_\sigma \setminus \partial A)$. There must be an open neighborhood of p in $\mathrm{Mm}(A)$ (open in the d-2 dimensional manifold $\mathrm{Mm}(A)$) that is entirely contained in ∂D_σ. If this were not the case, $\mathrm{Mm}(A)$ would be extremal at points arbitrarily close to p and would thus be extremal at p. Moreover, $\mathrm{Mm}(A)$ would be tangent to ∂D_σ at p. However, ∂D_σ is an extremal surface barrier (see lemma 2) so this is not possible. There are now two cases to consider.

- *Case 1*: $p \in \partial D_\sigma \setminus \sigma$.
  Figure 3.4 illustrates a construction that we will use for this case. Take $p \in \Sigma_+$ (the case of $p \in \Sigma_-$ is no different). By construction, $\mathrm{Mm}(A)$ is minimal on a surface Σ_0. There exists a (dimension d-1) open subset U of Σ_0 containing p such that $U \cap \mathrm{Mm}(A) \subset \Sigma_+$. Moreover, we can require that U is "split" by $\mathrm{Mm}(A)$ into two disconnected sets, N and V, such that N is the side of U closer to σ. Since Σ_0 is anchored to σ, we must have that $N \subset \Sigma_+$ and, in

particular, N is null. On the other hand, V cannot be a subset of Σ_+. If it were, then $\mathrm{Mm}(A)$ could decrease its area by being deformed up Σ_+ (by the focusing theorem). In particular, we can take U small enough to ensure that V is nowhere null in the direction of l^a.

We now consider the process of slightly sliding Σ_0 down Σ_+. More precisely, take a small parameter $\epsilon > 0$ and a corresponding one-parameter family of slices $\{\Sigma_\epsilon\}$ that are slightly deformed from Σ_0 in a way we now describe (an example of Σ_ϵ is depicted in figure 3.4 by an orange dashed line). The surface $\mathrm{Mm}(A) \cap U$ is described by a function $\lambda_0(x)$ giving the affine distance from σ up to $\mathrm{Mm}(A)$ at a point $x \in \sigma$. Now put $\lambda_\epsilon(x) = \lambda_0(x) - \epsilon f(x)$. Here, $f : \sigma \to [0, 1]$ is a smooth weighting function which equals 1 at the null generator x_p that p lies on. We take f to go to zero smoothly as x moves away from x_p, equaling zero exactly when x corresponds to a point outside of $U \cap \mathrm{Mm}(A)$. For $\lambda < \lambda_\epsilon(x)$, we require that the surface Σ_ϵ is identical to Σ_+. We extend Σ_ϵ beyond λ_ϵ by parallel transporting tangent vectors on $\mathrm{Mm}(A)$ directed toward V down to λ_ϵ. This prescription does not uniquely fix Σ_ϵ, but it is sufficient for our purposes.

Consider the one-parameter family of codimension 2 curves $\min(\partial A, \Sigma_\epsilon)$. For any $\epsilon > 0$, let L_ϵ denote the future-directed null congruence of $\min(\partial A, \Sigma_\epsilon)$ that points toward the interior of D_σ (see figure 3.4). The continuity of $\min(\partial A, \Sigma_\epsilon)$ as ϵ varies and the fact that Σ_+ is a light sheet ensures that there exists and $\epsilon_0 > 0$ such that for $\epsilon < \epsilon_0$,

- L_ϵ intersects Σ_0 to form a codimension 2 surface on Σ_0 anchored to ∂A and
- L_ϵ has negative future-directed expansion in the region between $\min(\partial A, \Sigma_\epsilon)$ and its intersection with Σ_0.

Denote this intersection by C_ϵ and observe that $C_0 = \mathrm{Mm}(A)$. But $\mathrm{Mm}(A)$ is minimal on Σ_0 so for sufficiently small ϵ,

$$\mathrm{area}(\mathrm{Mm}(A)) < \mathrm{area}(C_\epsilon) \leq \mathrm{area}(\min(\partial A, \Sigma_\epsilon))$$

which contradicts the assumption that $\mathrm{Mm}(A)$ has area greater than or equal to the minimal area surface on any slice. Note that the last inequality above follows from the focusing theorem applied to L_ϵ.

- *case 2*: $p \in \sigma$.
  Assume that there exists a (dimension d-2) open subset of $\mathrm{Mm}(A)$ that is contained in σ. (If not, there must be such an open set in $\partial D_\sigma \setminus \sigma$ which just leads to case 1 above.) Now consider the null vector field l^a on σ and the geodesics generated by it. Follow these geodesics from σ up along Σ_+ by a short affine distance $\epsilon > 0$ to generate a new codimension 2 surface,

σ_ϵ, which limits to σ when $\epsilon \to 0$. The focusing theorem now gives rise to a modified version of equation 2.2 at σ_ϵ:

$$\begin{aligned} \theta^k_\epsilon &> 0 \\ \theta^l_\epsilon &< 0. \end{aligned} \tag{3.4}$$

Along with moving σ up the light-sheet, we also translate A up the sheet to a one-parameter family of surfaces A_ϵ that limit to A. Consider the maximin construction applied to codimension 2 surfaces anchored to ∂A_ϵ that lie on codimension 1 surfaces anchored to σ_ϵ. We denote the result by $\mathrm{Mm}(A_\epsilon)$. We also define D_{σ_ϵ} in the obvious way. Now this maximin procedure leads to the same two cases that we are now studying. The first case, where $\mathrm{Mm}(A_\epsilon)$ intersects $\partial D_{\sigma_\epsilon} \setminus \sigma_\epsilon$ proceeds exactly as it did with $\epsilon = 0$. Now suppose that $\mathrm{Mm}(A_\epsilon)$ has an open set contained in σ_ϵ. $\mathrm{Mm}(A_\epsilon)$ must be minimal on some slice Σ_ϵ. However, equation 3.4 implies that σ_ϵ has negative (inward) extrinsic curvature on Σ_ϵ. It is thus impossible for $\mathrm{Mm}(A_\epsilon)$ to be minimal on Σ_ϵ since its area could be decreased by "cutting corners."

We can thus conclude that $\mathrm{Mm}(A_\epsilon) \cap \partial D_{\sigma_\epsilon} = \partial A_\epsilon$. This implies that $\mathrm{Mm}(A_\epsilon)$ is extremal. Taking the limit as $\epsilon \to 0$, we conclude that $\mathrm{Mm}(A)$ is extremal. But, given our assumption that part of $\mathrm{Mm}(A)$ lies on σ, equation 2.2 shows that $\mathrm{Mm}(A)$ cannot be extremal since extremal surfaces have zero null expansion in all directions.

At this point it is proven that $\mathrm{Mm}(A)$ is extremal. All that is left is to show that, of all the extremal surfaces in D_σ that are anchored to ∂A, $\mathrm{Mm}(A)$ is the smallest. Let $\Sigma_0 \in \mathcal{C}_\sigma$ be a slice on which $\mathrm{Mm}(A)$ is minimal. If Γ is another extremal surface anchored to ∂A then, as a result of the focusing theorem, we find that

$$\mathrm{area}(\mathrm{Mm}(A)) \leq \mathrm{area}(\mathrm{rep}(\Gamma, \Sigma_0)) \leq \mathrm{area}(\Gamma).$$

□

We are now in a position to prove a variety of properties of screen entanglement entropy. We begin with the "Page bound" advertised in section 3.1.

Corollary 1. *If A is a region in the leaf σ, then*

$$S(A) \leq \min\{S_{extensive}(A), S_{extensive}(\sigma \setminus A)\}$$

where S deonotes the holographic screen entanglement entropy of A and $S_{extensive}(X)$ denotes the area of a region $X \subset \sigma$ divided by 4.

Proof. $S(A) = \text{area}(\text{Mm(A)})/4$ but $\text{Mm}(A) = \min(\partial A, \Sigma_0)$ for some $\Sigma_0 \in \mathcal{C}_\sigma$. Both A and $\sigma \setminus A$ are codimension $d-2$ dimensional surfaces on Σ_0 anchored to ∂A so the area of $\text{Mm}(A)$ is less than or equal to the areas of both A and $\sigma \setminus A$. □

Next we turn to the proof of strong subadditivity for holographic screen entanglement entropy (other properties of entanglement entropy that admit covariant geometrical bulk proofs can be imported here as well). Unlike the case of theorems 2 and 3, the arguments below are essentially identical to those of [30] with little additional subtlety. We start with our version of theorem 17 in [30] which states that if $B \subset A$, then ext A lies "outside" of ext B.

Theorem 4. *Suppose that A and B are regions in the leaf σ with $B \subset A$. Then,*

1. *the entanglement wedge of A contains the entanglement wedge of B,*
2. *there exists a surface in $\mathcal{C}_\sigma$ on which both ext A and ext B are minimal.*

Sketch of Proof: The proof is the same as that of theorem 17 of [30] so we only sketch it here. For any surface in $\Sigma \in \mathcal{C}_\sigma$, consider a pair of codimension 2 surfaces constrained to lie on Σ, Γ_A and Γ_B, such that Γ_A is anchored to ∂A and Γ_B is anchored to ∂B. Then let $Z = \text{area}(\Gamma_A) + \text{area}(\Gamma_B)$. We now minimize the value of Z by varying over all possible choices of Γ_A and Γ_B. After that, we maximize the minimal values of Z by varying over all possible Σ.

This new maximin procedure gives a well-defined answer for the maximinimal value of Z. Moreover, a slice Σ_0 results on which both Γ_A and Γ_B are minimal. On this slice, it is impossible for Γ_A to cross Γ_B as this would necessarily give rise to a surface on Σ_0 anchored to ∂A with smaller area than Γ_A. A further observation is that if a connected component of A is distinct from a component of B, the corresponding connected components of Γ_A and Γ_B cannot come into contact even tangentially. The argument for this is that the component of Γ_B would necessarily have a different trace of its spatial extrinsic curvature than Γ_A at points close to the contact point. This would mean that either Γ_A or Γ_B is not minimal on Σ_0.

At this point it is known that components of Γ_A or Γ_B that are distinct have neighborhoods in Σ_0 that do not intersect the other surface. Within such neighborhoods, small deviations Σ_0 and the minimal surfaces can be made that prove that such surfaces are extremal.

The only remaining step is to show that, in fact, Γ_A and Γ_B are the extremal surfaces in D_σ of minimal area. If Γ'_A is an extremal surface in D_σ anchored to ∂A, then its representation on Σ_0 must have larger area than that of Γ_A but smaller area than that of Γ'_A. Thus, $\Gamma_A = \text{ext}\, A$. Similarly, $\Gamma_B = \text{ext}\, B$. By construction, both are minimal on the same surface $\Sigma_0 \in \mathcal{C}_\sigma$. Moreover, because Σ_0 is achronal, we must have that the entanglement wedge of A contains that of B. □

Corollary 2. *Suppose that A, B, and C are nonintersecting regions in σ. Then,*

$$S(AB) + S(BC) \geq S(ABC) + S(B)$$

where XY denotes $X \cup Y$ and where the function S is defined in equation 3.1.

Proof. By theorem 4, we can find a surface $\Sigma_0 \in \mathcal{C}_\sigma$ such that ext B and ext ABC are both minimal on Σ_0. Let $\tilde{S}(AB)$ and $\tilde{S}(BC)$ denote the areas of the representations of ext AB and ext BC on Σ_0. Then,

$$S(AB) + S(BC) \geq \tilde{S}(AB) + \tilde{S}(BC) \geq S(ABC) + S(B)$$

where the first inequality follows from the focusing theorem and the second inequality follows from the standard geometric proof of strong subadditivity [36]. □

Note that the inequality $S(A) + S(B) \geq S(AB)$ follows as a special case of this result.

3.3 Extremal Surfaces in FRW Cosmology

The conventional holographic entanglement entropy prescription, with its limitation to asymptotically locally AdS spacetimes, provides very little information about entanglement structure in cosmology. One of the most intriguing applications of our proposal, therefore, is to calculate holographic screen entanglement entropy in FRW universes. Assuming the screen entanglement conjecture, the calculations below give the entanglement entropy of subsystems in quantum states that are dual to cosmological spacetimes.

Holographic Screens in FRW Cosmology

First we review the holographic screen structure of FRW spacetimes. Consider a homogeneous and isotropic spacetime with the metric

$$ds^2 = -d\tau^2 + a(\tau)^2 \left(d\chi^2 + f(\chi)^2 d\Omega_2^2\right) \tag{3.5}$$

where $f(\chi) = \sinh(\chi), \chi$, or $\sin(\chi)$ in the open, flat, and closed cases respectively. Before computing extremal surfaces we must decide upon a null foliation for the spacetime and then identify the corresponding holographic screen. Null foliations (and thus holographic screens) are highly nonunique. The foliation we will consider here is that of past light cones from a worldline at $\chi = 0$.

To find the holographic screen for this foliation, it is convenient to introduce a conformal time coordinate η such that $d\tau/d\eta = a$. Then, the past light cone of the point $(\eta = \eta_0, \chi = 0)$ satisfies $\chi = \eta_0 - \eta$. Spheres along the past light cone can be parameterized by the coordinate η, and their area is given by

$$\mathcal{A}(\eta) = 4\pi\, a\Big(\tau(\eta_0 - \eta)\Big)^2\, f(\eta_0 - \eta)^2 \tag{3.6}$$

Assuming that $a = 0$ is not merely a coordinate singularity, the condition that $\mathcal{A}$ is maximized is equivalent to the condition that $d\mathcal{A}/d\eta = 0$. Thus, equation 3.6 gives the condition that fixes the holographic screen:

$$\frac{f(\chi)}{f'(\chi)} - \frac{1}{\dot{a}(\tau)} = 0. \tag{3.7}$$

The codimension 1 surface defined by this constraint may be timelike, spacelike, or null, depending on the particular choice of FRW spacetime. The foliating leaves of this holographic screen are spheres of constant τ and comoving radius χ satisfying equation 3.7. The covariant entropy bound implies that each leaf has sufficient area to holographically encode the information on one past light cone from the worldline at $\chi = 0$ [9, 10].

Let $\sigma(\tau)$ be the leaf of the holographic screen at time τ and let $\rho(\tau)$ denote the energy density in the universe (measured by comoving observers) at time τ. Then, one can write a simple expression for the area of a leaf of the holographic screen at time τ, valid for any f:

$$\text{area}(\sigma(\tau)) = \frac{3}{2\rho(\tau)}. \tag{3.8}$$

In particular, this expression shows that holographic screens grow in area as the universe expands.

Extremal Surfaces in de Sitter Space

Consider 3+1 dimensional de Sitter space of radius α. This spacetime is $S^3 \times \mathbf{R}$ with the metric

$$ds^2 = -dT^2 + \alpha^2 \cosh^2\left(\frac{T}{\alpha}\right) d\Omega_3^2$$

where $d\Omega_3^2$ is the metric on a unit 3-sphere. Despite the fact that this spacetime has the form of equation 3.5 (with $f(\chi) = \sin\chi$), it is an awkward setting for the consideration of holographic screens: the null expansion on the past or future light cones of any point in de Sitter space goes to zero only at infinite affine parameter. This suggests that the appropriate "boundary" of de Sitter space is past or future infinity. Even if we do attempt to anchor extremal surfaces to spheres at infinity, the analysis in section 3.2 fails to apply because of the assumption made there that leaves are compact.

Fortunately these difficulties can be averted completely by considering an FRW spacetime that asymptotically approaches de Sitter space at late times. Specifically, we will consider a spacetime of the form of equation 3.5 with vacuum energy density ρ_Λ and, in addition, some matter content $\rho_{\text{matter}}(\tau)$ with the property the matter content gives rise to a big bang at $\tau = 0$ and dilutes completely[8] as $\tau \to \infty$.

[8] In particular, we are not considering spacetimes with a big crunch in this section.

Equation 3.8 immediately implies that

$$\lim_{\tau\to\infty} \text{area}(\sigma(\tau)) = \frac{3}{2\rho_\Lambda} = 4\pi\alpha^2 \tag{3.9}$$

where $\alpha = \sqrt{3/8\pi\rho_\Lambda}$. Because of the big bang singularity, we must have that $\text{area}(\sigma_{\tau=0}) = 0$. Thus, by the area law for holographic screens [13, 14], we can conclude that the leaves of our screen are spheres that monotonically increase in area, starting with 0 area at the big bang, and expanding to approach the de Sitter horizon of area $4\pi\alpha^2$ at late τ.

Now focus on a late time leaf $\sigma(\tau)$. As discussed in section 3.2, given a region $A \subset \sigma(\tau)$ with a boundary, we can determine the holographic screen entanglement entropy of A, $S(A)$, by considering an extremal surface anchored to and terminating at ∂A. In the notation of section 3.2, $D_{\sigma(\tau)}$ is compact so theorem 2 implies that an extremal surface anchored to ∂A exists and lies inside of $D_{\sigma(\tau)}$.

For any time τ, define

$$S^\tau_{\text{Page}}(A) = \begin{cases} \frac{1}{4}\text{area}(A) & \text{area}(A) \leq \frac{1}{2}\text{area}(\sigma(\tau)) \\ \frac{1}{4}\left(\text{area}(\sigma(\tau)) - \text{area}(A)\right) & \text{area}(A) > \frac{1}{2}\text{area}(\sigma(\tau)). \end{cases} \tag{3.10}$$

We allow this definition to extend to a function $S^\infty_{\text{Page}}(A)$ where A is a region on a 2-sphere of radius α. This $\tau = \infty$ case is defined exactly as in equation 3.10 if we take $\text{area}(\sigma(\infty)) = 4\pi\alpha^2$.

Below we will present an argument that if $A \subset \sigma(\tau)$, then

$$\lim_{\tau\to\infty} S(A) = S^\infty_{\text{Page}}(A). \tag{3.11}$$

(Note that in this limit, it is implied that A is transported to later and later leaves.) Thus, we will find that as $\tau \to \infty$, $S(A)$ approaches the random entanglement limit discussed in section 3.1.

Any interpretation of this result is necessarily speculative. Nevertheless, if one assumes the screen entanglement conjecture, then equation 3.11 implies that the the quantum state of an FRW universe asymptotically approaching de Sitter space has the property that its $O(\alpha^2)$ degrees of freedom are almost randomly entangled with one-another. At earlier times, the degrees of freedom are not randomly entangled because $S(A) < S^\infty_{\text{Page}}(A)$.

Random Entanglement and the Static Sphere Approximation

We now present a combination of rigorous arguments, numerical data, and analytic approximations suggesting that the approximate de Sitter cosmological spacetimes discussed above saturate the random entanglement bound in the $\tau \to \infty$ limit. As before, $\sigma(\tau)$ denotes a leaf at time τ in an FRW universe with vacuum energy as well as matter energy that dilutes at late time.

The entire region $D_{\sigma(\tau)}$ has a metric that can be made arbitrarily similar to that of a patch of empty de Sitter space by making τ large. To see this, first note that points in $D_{\sigma(\tau)}$ have

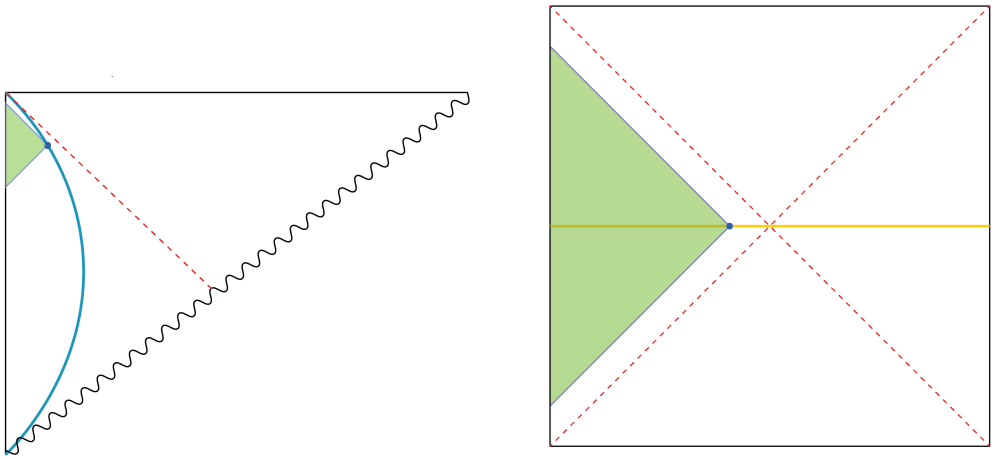


Figure 3.5: The domain of dependence $D_{\sigma(\tau)}$ for a late time leaf in the flat FRW universe (the small green triangle in the upper diagram) can be approximately mapped to a domain of dependence $D_{\tilde{\sigma}(\tau)}$ in empty de Sitter space (lower diagram). The mapping becomes increasingly accurate as τ becomes larger. The effect of increasing τ is to move the green triangle in the upper diagram into the top-left corner (along the blue curve), while the green triangle in the lower diagram moves to the right and approaches the entire left static wedge.

$\chi < \chi_{\text{screen}}(\tau)$ and $\chi_{\text{screen}}(\tau)$ can be made arbitrarily small by making τ large. (This follows from equation 3.9 and the fact that $\lim_{\tau\to\infty} a(\tau) = \infty$.) Meanwhile, the conformal diagram for our spacetime immediately shows that the minimal value of τ in $D_{\sigma(\tau)}$ can be made arbitrarily large by making τ large. Thus $D_{\sigma(\tau)}$ can be made to only cover arbitrarily large τ and arbitrarily small χ, in which case our metric of equation 3.5 takes the form

$$ds^2 \approx -d\tau^2 + c\, e^{2\tau/\alpha}(d\chi^2 + \chi^2 d\Omega_2^2) \tag{3.12}$$

where c is a constant and α is the same constant as before. Here we have made use of the Friedmann equations. The right-hand side of this equation is precisely the metric of de Sitter space in flat slicing. De Sitter space can also be described in static coordinates that make a time-translation Killing vector field manifest:

$$ds^2 \approx -\left(1 - \frac{r^2}{\alpha^2}\right) dt^2 + \left(1 - \frac{r^2}{\alpha^2}\right)^{-1} dr^2 + r^2 d\Omega_2^2. \tag{3.13}$$

Fortunately, $D_{\sigma(\tau)}$ lies in a region that is well-described by either the flat or static slicing of equations 3.12 and 3.13 respectively.

We can now identify $D_{\sigma(\tau)}$ with a region $D_{\tilde{\sigma}(\tau)}$ where $D_{\tilde{\sigma}(\tau)}$ denotes a corresponding region in exact de Sitter space obtained by finding a sphere $\tilde{\sigma}(\tau)$ in the static patch with area matching that of $\sigma(\tau)$. While it may seem natural to put $\tilde{\sigma}(\tau)$ at large static time, we can use the t translational

symmetry of de Sitter space to place $\tilde{\sigma}(\tau)$ at $t = 0$ for all τ. The effect of increasing τ is simply to bring $\tilde{\sigma}(\tau)$ closer to the bifurcation sphere on the de Sitter horizon. This identification is illustrated in figure 3.5. Note that as $\tau \to \infty$, the geometry of $D_{\sigma(\tau)}$ and $D_{\tilde{\sigma}(\tau)}$ become arbitrarily similar.

Consider the region $A \subset \sigma(\tau)$ which can be identified with a region $\tilde{A} \subset \tilde{\sigma}(\tau)$. At large τ, $\tilde{\sigma}(\tau)$ approaches the equator of a 3-sphere of radius α. The equator itself is an extremal surface so with $\tau < \infty$ but still large, there must be an extremal surface that is close to $\tilde{A}$ but not exactly on it. Its area will be slightly less than that of $\tilde{A}$. Note, moreover, that if the area of $\tilde{A}$ exceeds half the area of the equator, then a smaller extremal surface can be obtained by considering the complement of A.

This suggests but does not yet prove that at large τ, the holographic screen entanglement entropy of A is almost equal to a fourth of its own area in Planck units if A has less area than half of the de Sitter horizon. What we have proven so far is that an extremal surface exists with area almost equal to that of A (or $4\pi\alpha^2 - \text{area}(A)$).

What if there is another extremal surface with smaller area than the one we have found? It is easy to see that this is impossible. Following the notation in section 3.2, consider the spacelike surface Σ_0 that, after mapping to $D_{\tilde{\sigma}(\tau)}$, lies at static time $t = 0$, and that and terminates at $\tilde{\sigma}$. (Σ_0 is most of a hemisphere of the 3-sphere.) The Riemannian geometry of S^3 shows that the surface of minimal area anchored to ∂A is the one we have already found. If Γ is another extremal surface (not necessarily lying on Σ_0), then its representation on Σ_0, $\text{rep}(\Gamma, \Sigma_0)$, necessarily has larger area than the extremal surface close to the horizon. But $\text{area}(\Gamma) \geq \text{area}(\text{rep}(\Gamma, \Sigma_0))$ so we conclude that Γ does not have minimal area.

The arguments above show that the random entanglement limit is saturated at large τ. Taking $0 \ll \tau < \infty$ and $A \subset \sigma(\tau)$, we now explain a way to obtain a more accurate estimate for $S(A)$ than $S^{\tau}_{\text{Page}}(A)$. Calculating $S(A)$ without taking the large τ limit is more involved than what was done above. Nonetheless, it is worthwhile to investigate this case to better understand how the Page bound limit is approached. In particular, it is of interest to understand how the discontinuity of the derivative of S^{∞}_{Page} arises.

We begin by further discussing the role of the 3-sphere in de Sitter space. Figure 3.6 depicts a hemisphere of an S^3 of radius α which is precisely half of a static slice of de Sitter space (which we can freely take to be $t = 0$). Define a parameter z as $z = \sqrt{\alpha^2 - r^2}$ where r is the static radius appearing in equation 3.13. Note that a surface of constant z (and static time) is an S^2 of area $4\pi(\alpha^2 - z^2)$. This suggests a way to obtain an approximation for $S(A)$ if A is a region in the leaf $\sigma(\tau)$. Rather than taking A to be a region in $\sigma(\tau)$, we take figure 3.5 seriously and map A to a region in the S^2 of constant

$$z = \sqrt{\frac{4\pi\alpha^2 - \text{area}(\sigma(\tau))}{4\pi}} \tag{3.14}$$

which ensures that this S^2 has the same area as $\sigma(\tau)$. After this mapping is made, one computes

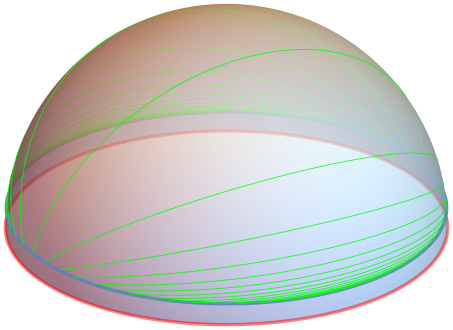


Figure 3.6: The upper hemisphere of a 3-sphere of radius α is half of a static slice in empty de Sitter space and serves as a good approximation for $D_{\sigma(\tau)}$ at large τ. The blue 2-sphere (appearing as a circle here) lies at constant z (equivalently, constant r where r is the radial coordinate in equation 3.13). This 2-sphere is an approximation for the leaf $\sigma(\tau)$. Green surfaces depict extremal spherical caps on S^3 that approximate ext A_ψ for various values of ψ. The many samples of extremal surfaces shown here have evenly spaced values of ψ. Figure 3.7 provides evidence that this static sphere approximation is accurate at late τ.

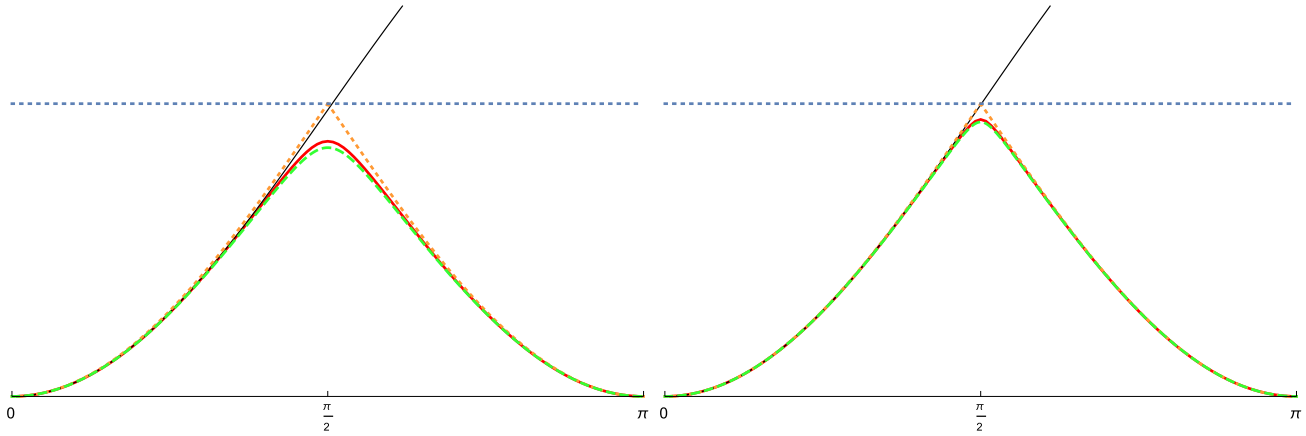
$S(A)$ by finding the extremal surface on the S^3 that is anchored to ∂A (which we take to lie at constant z). Below we will refer to this procedure as the "static sphere approximation."

Consider regions in $\sigma(\tau)$ that are spherical caps. Such a cap can be fixed (up to $SO(3)$ rotation) by a zenith opening angle angle ψ, so we will denote our region of $\sigma(\tau)$ by A_ψ. (With this notation, $A_{\pi/2}$ is a hemisphere and A_π is the entire leaf.) The static sphere approximation makes it is clear that for $0 < \psi \ll \pi/2$, ext A_ψ is close to A_ψ itself and that for $\pi/2 \ll \psi < \pi$, ext A_ψ approaches $\sigma(\tau) \setminus A_\psi$. As ψ passes the transition angle $\pi/2$, ext A_ψ quickly passes over the top of the 3-sphere of radius α. The closer area$(\sigma(\tau))$ is to $4\pi\alpha^2$, the faster ext A_ψ passes over the top of the sphere. This explains how the discontinuity in the derivative of $S^\infty_{\text{Page}}(A)$ arises in the large τ limit.[9]

Because the geometry of S^3 is simple, it is not difficult to obtain an explicit (if cumbersome) expression for $S(A_\psi)$ in the static sphere approximation:

$$S(A_\psi) \approx \pi \sin^2\left(\frac{1}{4}\cos^{-1}\left[\frac{z^2}{\alpha^2} + \left(1 - \frac{z^2}{\alpha^2}\right)\cos 2\psi\right]\right) \tag{3.15}$$

[9]For finite τ, there is always another extremal surface on the 3-sphere which goes around the sphere the wrong way. This surface always has area greater than ext A_ψ and, in any case, fails to lie in $D_{\sigma(\tau)}$. However, if we consider the $\tau = \infty$ limit, then ext A_ψ does not smoothly pass over the hemisphere of the 3-sphere, and in this case, the discontinuity in the derivative of $S^\infty_{\text{Page}}(A)$ is explained by the fact that the surface that wraps around the sphere the "wrong way" is now precisely the complement of A_ψ in the equator. If ψ exceeds $\pi/2$ in this case, then the complement of A_ψ has smaller area than A_ψ. We see that a phase transition occurs only in the exact $\tau = \infty$ limit.



(a) $z/\alpha \approx .05$ (b) $z/\alpha \approx .02$

Figure 3.7: Plots of $S(A_\psi)$ and other quantities for two leaves at different times in a universe with dust and vacuum energy. In both plots, the red curve is the numerically computed holographic screen entanglement entropy of A_ψ. The dashed green curve is the static sphere approximation for $S(A\psi)$ which becomes more accurate at later τ (smaller z). The orange curve with a sharp peak is $S_{\text{Page}}(A_\psi)$ as defined by equation 3.10 and the black curve is $S_{\text{extensive}}(A_\psi)$. The horizontal line, provided for scale, marks the value of $\pi\alpha^2/2$ which is precisely one fourth of the extensive entropy of the de Sitter horizon.

where z is given by equation 3.14 and, as before, $\alpha = \sqrt{3/8\pi\rho_\Lambda}$. This expression can be thought of as giving a correction to the "zeroth order" expression $S(A_\psi) \approx S^{\infty}_{\text{Page}}(A_\psi)$. Taking $\tau < \infty$ will lead to corrections in $1/\tau$ that are not described by the static sphere method. It is an open question as to whether or not such corrections can, in principle, be of the same (or greater) order in $1/\tau$ as the one we have studied here. However, numerical data that suggests that the static sphere approximation is accurate at large τ as we will now see.

As explained above, the cosmological spacetimes we have been discussing have vacuum energy ρ_Λ as well as some matter content that dilutes at late time. The simplest case of this is when the universe is flat ($f(\chi) = \chi$) and when the additional matter content consists of only one species with density ρ_{matter} and pressure $p_{\text{m}} = w\rho_{\text{m}}$. The scale factor for this case is

$$a(\tau) = C \sinh\left[\frac{3(1+w)\tau}{2\alpha}\right]^{\frac{2}{3(1+w)}} \tag{3.16}$$

where the normalization factor C is independent of τ.

This setting is very useful to test the theoretical apparatus developed in this section. In the case of $w = 0$, figure 3.7 shows a variety of quantities we have discussed. Figure 3.7 (a) and (b) depict the case of an earlier and later time leaf with $z/\alpha \approx .05$ and $z/\alpha \approx .02$ respectively. The solid red curves show $S(A_\psi)$ (computed numerically) while the green curves give the static sphere approximation of equation 3.15. The dotted horizontal line marks half of the de Sitter entropy: $S_{1/2} = \pi\alpha^2/2$. As expected, $S(A_\psi) < S_{1/2}$. The orange curve with a discontinuity in its derivative

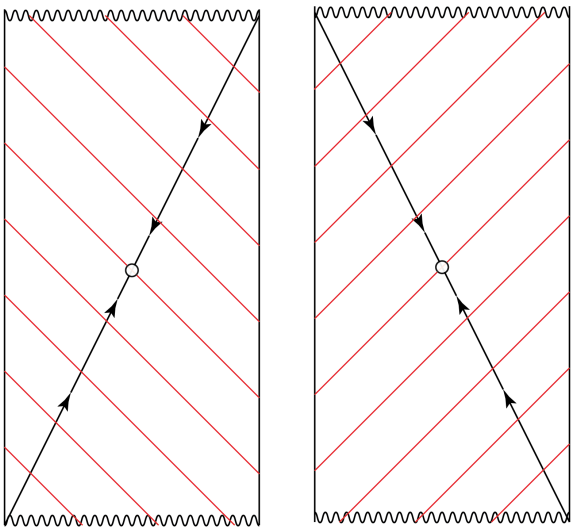


Figure 3.8: Both Penrose diagrams here are for the same spacetime: a closed FRW universe with dust. The red lines denote a null foliation and the black diagonals are the past and future holographic screens corresponding to the foliation. The two figures demonstrate that different foliations give rise to different screens. In both figures, the lower half of the diagonal is a past holographic screen and the upper half is a future holographic screen. Arrows show the direction of increasing area.

is $S^{\infty}_{\text{Page}}(A_\psi)$. Comparing figures 3.7 (a) and (b), one can see that $S(A_\psi)$ is approaching $S^{\infty}_{\text{Page}}(A_\psi)$ as $\tau \to \infty$. Finally, the black curves shows extensive entropy: $S_{\text{extensive}}(A_\psi) = (1/4)\text{area}(A_\psi)$. Note that $S(A_\psi) < S_{\text{extensive}}(A_\psi)$ for all ψ as required by corollary 1.

Closed Universe with a Big Crunch

The holographic screen entanglement entropy structure of a closed universe with a past and future singularity is similar to that of approximate de Sitter space. The spacetimes we consider have the metric of equation 3.5 with $f(\chi) = \sin(\chi)$. In this case the coordinate χ takes values from 0 to π. We put one species of matter content in the spacetime that satisfies $p = w\rho$ which gives rise to a big bang at $\tau = 0$ as well as a big crunch. As before, we introduce a conformal time coordinate η in terms of which the scale factor is

$$a(\eta) = c\left(\sin\frac{\eta}{q}\right)^q$$

where $q = 2/(1+3w)$ and c is constant. This shows that the Penrose diagram for this spacetime is a rectangle with a time-to-space aspect ratio of q.

Figure 3.8 shows the holographic screen structure of this spacetime for two examples of null foliations. We focus on the diagram to the left in which case the null foliation (partially) consists of

past light cones of a comoving worldline at the $\chi = 0$. As suggested by the figure, the holographic screen is given by

$$\chi_{\text{screen}} = \frac{1}{q}\eta.$$

However, a subtlety arises because the screen is a past holographic screen for $\eta < q\pi/2$ and a future screen for $\eta > q\pi/2$. The sphere that connects the past and future screen is extremal (this was called an "optimal" surface in [10]) and has area $4\pi c^2$. Let $\sigma(\eta)$ denote the leaf at conformal time η. We put $\sigma_0 = \sigma(\eta = q\pi/2)$.

Just as in the de Sitter case, this example leads to a saturation of the Page bound of equation 3.2 as leaves are maximized in area. More precisely, if $A \subset \sigma(\eta)$, then $\lim_{\eta \to q\pi/2} S(A) = S^{\infty}_{\text{Page}}(A)$ where in this case

$$S^{\infty}_{\text{Page}}(A) = \begin{cases} \frac{1}{4}\text{area}(A) & \text{area}(A) \leq \frac{1}{2}\pi c^2 \\ \frac{1}{4}\left(4\pi c^2 - \text{area}(A)\right) & \text{area}(A) > \frac{1}{2}\pi c^2. \end{cases}$$

It appears that $S(A)$ saturates the Page bound in a great variety of cases where the areas of leaves are bounded above.

3.4 Discussion

The proposal we have given above may open the door to a new research program: the study of the entanglement structure of general spacetimes. In light of this, and for the sake of clarity, we now summarize the recipe for computing von Neumann entropy under the assumption of the screen entanglement conjecture discussed in section 3.1:

1. Select a particular null foliation $\{N_r\}$ of a spacetime with dimension d.
2. Find the codimension 2 surfaces $\{\sigma_r\}$ with $\sigma_r \subset N_r$ that have maximal area on each N_r.
3. Take a $d-2$ dimensional subregion $A \subset \sigma_r$ with a boundary ∂A.
4. Of all extremal surfaces anchored to ∂A and lying in the causal region D_σ (see section 3.1), select the one of minimal area. The conjectured entropy $S(A)$ is then one fourth the area of the minimal extremal surface in Planck units.

Potential applications of our conjecture are numerous. One example not considered above is case of a spacetime with a black hole. Black holes formed through the collapse of matter possess future holographic screens in their interiors that approach their horizons at late times. It is of potential significance to investigate the entanglement structure of such spacetimes. Perhaps such an analysis will shed light on the firewall paradox [4].

If the screen entanglement conjecture is correct, it should still only be regarded as a leading order prescription for the computation of von Neumann entropies. A version of the analysis of [37] may be extendible to the context of holographic screens. It is not completely obvious how this should be done. If A is a region in a leaf σ lying on a Cauchy slice S_0, one may consider the region on S_0 bounded by A and its extremal surface ext (A) and compute the entanglement entropy of this region in a quantum field theory on the spacetime background. On the other hand, it may be necessary to modify the spacetime position of the holographic screen itself as was done in [38].

Chapter 4

Fixing the Reference Frame in Quantum Gravity

Counterintuitively, holographic screens are appealing because of their non-uniqueness. That descriptions in quantum gravity should depend strongly on the choice of reference frame is perhaps the most important concept in modern quantum gravity. In this chapter, we will turn our attention away from holographic screens and attempt to address such redundancies by "fixing the gauge" in an observer-dependent fashion.

As discussed in chapter 1, the standard local formulation of quantum general relativity leads to inconsistency when describing a process in which an object falls into a black hole that eventually evaporates, since it may employ a class of equal time hypersurfaces (called nice slices) on which quantum information is duplicated [60]. In the early 90's, a remarkable suggestion to avoid this difficulty—called the complementarity picture—has been made [3, 41]: the apparent cloning of the information occurring in black hole physics implies that the internal spacetime and horizon/Hawking radiation degrees of freedom appearing in different, i.e. infalling and distant, descriptions are not independent. This clearly signals a breakdown of the naive global spacetime picture of general relativity, and forces us to develop a new low energy theory of quantum gravity in which locality is preserved (if there exists such a formulation).

In this chapter, building on earlier suggestions in Refs. [33, 32], we propose an explicit framework in which low energy dynamics of quantum gravity is described preserving locality, and yet taking into account the effects that are not captured by the naive global spacetime picture. We introduce an explicit coordinate system associated with a freely falling reference frame, which we call the observer-centric coordinate system, that allows for a "special relativistic" description of gravity, i.e. treating gravity *as a force* measured by the observer tied to this coordinate system. This allows us to identify, in simple cases, boundaries of spacetime where the local description of the system breaks down, which we call the observer horizon. We work with a specific Hilbert space, which we refer to as the covariant Hilbert space for quantum gravity, in which the proposed scheme is

realized in a simple manner.

4.1 Covariant Hilbert Space for Quantum Gravity

Our construction is based on a series of hypotheses which we list here without much elaboration.

We postulate that

(i) A Hamiltonian formalism exists that describes a quantum mechanical system with gravity. Since a system with gravity in general has constraints, we consider the constrained Hamiltonian formalism developed by Dirac [42].

(ii) There is a way to restrict Hilbert space (e.g. fix intrinsically stringy gauge redundancies) in such a way that dynamics defined on it is local in spacetime at length scales larger than l_*. In other words, there is a way to formulate a theory such that "intrinsically quantum gravitational" (stringy) effects decouple at distances larger than l_* (the string scale).

(iii) The desired local description is obtained by restricting the Hilbert space such that an element represents either an appropriately restricted region of a spacetime hypersurface (when it allows for a spacetime interpretation) or an intrinsically quantum gravitational state (when it does not). In particular, the former can be taken to represent a state of physical degrees of freedom on *a portion of* the past light cone of a fixed reference point p_0.

A main motivation for the last hypothesis is that it seems to constitute the minimal deviation from the standard general relativistic view of spacetime, needed to address the issue of information cloning in the existence of a horizon. The use of a light cone is also motivated to make the causal structure manifest; the hypersurface represented by a state corresponds to the spacetime region from which a hypothetical observer at p_0 can obtain light ray signals. (The possibility of using a spacelike hypersurface will be mentioned later.)

We argue that the desired description is obtained by suitably dropping some of the constraints needed to reduce the Hilbert space to that of the physical states:

$$\mathcal{P}^\mu(\mathbf{x})|\Psi\rangle = 0, \tag{4.1}$$

where $\mu = 0, \cdots, 3$, and $\mathbf{x}$ are the coordinates parameterizing a hypersurface on which the states are defined. Note that $|\Psi\rangle$ represents a quantum state for the *entire* system, including possible degrees of freedom associated with the boundaries of space, which may be located at infinity. Now, a natural way to define locality is through the structure of the Hamiltonian. However, if we define the Hamiltonian operator (which is a linear combination of $\mathcal{P}^\mu(\mathbf{x})$'s) on Hilbert space $\mathcal{H}_{\rm phys}$ spanned by the independent physical states $|\Psi\rangle$, then it is simply zero. Furthermore, it is

in general not possible to take a basis in $\mathcal{H}_{\rm phys}$ such that all of its elements represent well-defined semi-classical spacetimes as they are generically in superposition states.[1] This precludes us from labeling the elements of $\mathcal{H}_{\rm phys}$ according to physical configurations in spacetime, since they do not even have well-defined spacetimes. In particular, in the Hilbert space $\mathcal{H}_{\rm phys}$ spanned by physical (gauge invariant) states $|\Psi\rangle$, local operators—or the concept of locality itself—cannot be defined in general.

These issues can be addressed if we consider a Hilbert space larger than $\mathcal{H}_{\rm phys}$ by appropriately dropping some of the constraints (which then must be imposed later as the "dynamics" of the system). Specifically, consider a (hypothetical) reference point p_0 at some $\mathbf{x}$ and a local Lorentz frame elected there. We may then change the basis of constraint operators $\mathcal{P}^\mu(\mathbf{x})$ (by taking their linear combinations) so that it minimizes the number of constraints corresponding to transformations affecting the local Lorentz frame. This leads to 10 constraint operators, H, P_i, $J_{[ij]}$, and K_i $(i = 1, 2, 3)$, associated with the change of the local Lorentz frame. These operators obey the standard Poincaré algebra. (The set of operators determined in this way is not unique, and each choice corresponds to adopting different, e.g. null or spacelike, quantization.)

We now postulate

(iv) By dropping the constraints related to the changes of the local Lorentz frame

$$H|\Psi\rangle = P_i|\Psi\rangle = J_{[ij]}|\Psi\rangle = K_i|\Psi\rangle = 0, \tag{4.2}$$

we obtain a Hilbert space $\mathcal{H}_{\rm QG}$ larger than $\mathcal{H}_{\rm phys}$. The elements of $\mathcal{H}$—the subspace of $\mathcal{H}_{\rm QG}$ allowing for a spacetime interpretation—can then be labeled by physical configurations in spacetime hypersurfaces (together with possible other labels such as spins of particles); in other words, we can take a basis of $\mathcal{H}$ such that all the basis states have well-defined semi-classical spacetimes. Physics defined on this space is local in the bulk of spacetime.

In particular, we assume that we can take specific linear combinations of the constraint operators $\mathcal{P}^\mu(\mathbf{x})$ such that the appropriate basis states of $\mathcal{H}$ represent the configurations of physical degrees of freedom on (portions of) the past light cone of p_0. We then call the corresponding enlarged Hilbert space $\mathcal{H}_{\rm QG}$ the *covariant Hilbert space for quantum gravity*.[2]

The Hamiltonian defined on $\mathcal{H}_{\rm QG}$ represents local physics on the past light cone of p_0 within a boundary, which we will explicitly determine in simple cases below. (Here we are considering

[1] Because the quantum state we consider here, $|\Psi\rangle$, is the state representing the entire system including clock degrees of freedom (as opposed to relative states $|\psi_i\rangle$ which may evolve in time), it satisfies all the constraints in Eq. (1), including the Hamiltonian constraint. This makes $|\Psi\rangle$ a superposition of terms representing semi-classical spacetimes because it takes the form of $|\Psi\rangle = \sum_i |i\rangle|\psi_i\rangle$, where $|i\rangle$ and $|\psi_i\rangle$ represent the clock degrees of freedom and the rest of the system, respectively.

[2] The physical Hilbert space, $\mathcal{H}_{\rm phys}$, is a subspace of $\mathcal{H}_{\rm QG}$. As such, any gauge-invariant (constrained) state, i.e. an element of $\mathcal{H}_{\rm phys}$, can be expanded as a superposition of elements in $\mathcal{H}_{\rm QG}$ in the "locality basis" that can be determined by the structure of the Hamiltonian defined in this enlarged Hilbert space.

each component state, i.e. a basis state in $\mathcal{H}$ in the basis given in (iv). The full quantum state is in general a superposition of these and other states.) This Hamiltonian is not *manifestly* local, since the constraints associated with the coordinate transformations on the past light cone of p_0 are still imposed on $\mathcal{H} \subset \mathcal{H}_{\rm QG}$. In other words, the elements of $\mathcal{H}$ represent physical states obtained after solving Einstein's equation on the light cone. To recover a manifest locality of the Hamiltonian, we need to introduce appropriate metric degrees of freedom on the light cone and drop the corresponding constraints from the definition of the Hilbert space. We assert that the resulting Hamiltonian is then manifestly local in the bulk of spacetime (but not at the boundary). In the rest of the letter, we do not bother with this last step and focus our attention on $\mathcal{H}_{\rm QG}$, which is enough to make *physics* local in the bulk (in the sense that there exists an equivalent, though more redundant, description in which the Hamiltonian takes a manifestly local form).[3]

The Hilbert space $\mathcal{H}_{\rm QG}$ is the relevant Hilbert space when we discuss "evolution" of a system with gravity. It is true that a physical state of the *entire* system must obey all the constraints, including those in Eq. (4.2), and thus satisfies

$$\frac{d}{dt}|\Psi\rangle = 0, \tag{4.3}$$

i.e. $|\Psi\rangle$ is static. However, in $|\Psi\rangle$ we can identify a (small) subsystem as the "clock" degrees of freedom, and rewrite the entanglement of these degrees of freedom—represented e.g. by a set of states $|i\rangle$—with the rest of the degrees of freedom—represented e.g. by a set of states $|\psi_i\rangle$—in the standard form of Schrödinger time evolution of a state $|\psi_i\rangle$, where i plays the role of time [43]. (In the Minkowski space, we are doing this operation implicitly by identifying boundary degrees of freedom at infinity as the clock degrees of freedom; this is why we can consider time evolution, or S matrix, in Minkowski space without explicitly being bothered by the clock degrees of freedom.) We may then view $\mathcal{H}_{\rm QG}$ as the Hilbert space in which $|\psi(t)\rangle \equiv |\psi_i\rangle$ evolves unitarily according to the "derived" Hamiltonian, which in general depends on the choice of the clock degrees of freedom.[4] (Note that $|\psi_i\rangle$ are no longer zero eigenvalue eigenstates of H, P_i, $J_{[ij]}$, or K_i, in general.) Furthermore, complementarity can be viewed as a relation between different low energy descriptions corresponding to different choices of clocks separated beyond each other's horizon, which are obtained after a suitable action of H, P_i, $J_{[ij]}$, or K_i to put the clock in the bulk of spacetime in each description. From this perspective, $|\Psi\rangle$ serves the role of a generating function from which physical predictions can be derived by identifying the clock degrees of freedom and extracting their entanglement with the rest.

[3]The commutation relations among field operators may contain apparent non-local terms associated with null quantization, which arise from the fact that massless particles can propagate along the light cone.

[4]In order for this operation to give well-defined time evolution of $|\psi_i\rangle$ by an ordered Hamiltonian at a macroscopic level, the state $|\Psi\rangle$ must be in a special low coarse-grained entropy state, at least in branches relevant for the clock degrees of freedom. In a real cosmological situation, when $|\Psi\rangle$ represents the entire "multiverse state," this leads to a set of conditions which the Hamiltonian H defined on $\mathcal{H}_{\rm QG}$ must satisfy [44].

We note that while our framework allows for formally writing down the Hamiltonian applicable at length scales larger than l_*, this is not directly useful in calculating the effect of dynamical spacetime or the result of a reference frame change, since they depend on unknown dynamics of degrees of freedom at the boundaries of space. This problem may be largely bypassed if we are interested only in a coarse-grained description of the system, by employing a certain correspondence principle which we may call the complementarity hypothesis [45]—we can then use a combination of quantum theory and classical general relativity to obtain a coarse-grained description of the evolution of the system. An advantage of our framework in doing this is that it clearly separates between the "low-energy" local physics and "trans-Planckian" intrinsically quantum gravitational (stringy) physics, so it allows for developing clear physical pictures of the origins of various effects. To obtain a complete dynamical theory, however, we would need to formulate the theory applicable above M_*—presumably string theory—along the lines described here. This is beyond the scope of the present work.

The structure of the covariant Hilbert space takes the form

$$\mathcal{H}_{\rm QG} = \mathcal{H} \oplus \mathcal{H}_{\rm sing}. \tag{4.4}$$

Here, $\mathcal{H}$ is spanned by all the possible physical configurations realized on (portions of) the past light cone of p_0 *as viewed from a local Lorentz frame at* p_0, while $\mathcal{H}_{\rm sing}$ contains intrinsically quantum mechanical states that do not allow for a spacetime interpretation (the states relevant when p_0 hits a spacetime singularity), where $\dim \mathcal{H}_{\rm sing} = \infty$ [33]. How do we define physical configurations "as viewed from a local Lorentz frame at p_0"? Where is the boundary of space that determines the relevant portion of the light cone for each element of $\mathcal{H}$? In the next section, we address these questions and provide an explicit prescription to specify elements of $\mathcal{H}$ which is applicable in simple cases. We also discuss a global structure of $\mathcal{H}$, based on a certain classification scheme for the elements.

4.2 Defining Boundaries and Classifying the States

We now focus on $\mathcal{H}$ and identify a spacetime region (in particular, a region on an "equal-time" hypersurface) represented by its element. We discuss how independent quantum states comprising $\mathcal{H}$ are specified, and classify them into elements of $\mathcal{H}_{\partial\mathcal{M}}$'s, the subsets of $\mathcal{H}$ labeled by "horizons" possessed by the states.

4.2.1 Observer-centric coordinates

We first introduce a useful coordinate system to describe our construction. Let us choose a fixed spacetime point p_0 in a fixed spacetime background. We consider that an element of $\mathcal{H}$ represents

physical configurations of dynamical degrees of freedom and their conjugate momenta on the past light cone of p_0, which we call L_{p_0}. In general, the elements of $\mathcal{H}$ are labeled by a set of quantum numbers (i.e. the response to a set of quantum operators), and in Section 4.2.5 we will discuss how many independent such quantum states exist in full quantum gravity. For now, however, it is sufficient to keep in mind that the state is specified by the response to the operators defined on L_{p_0}.

Now, consider a timelike geodesic $p(\tau)$ which passes through p_0 at $\tau = 0$: $p(0) = p_0$. We take τ to be the proper time measured at p. A set of local Lorentz frames elected along $p(\tau)$ corresponding to a freely falling frame can then be uniquely determined by specifying spacetime location q^μ and proper velocity v^i of p at $\tau = 0$

$$x^\mu_{p(0)} = q^\mu, \qquad \left.\frac{dx^i_{p(\tau)}}{d\tau}\right|_{\tau=0} = v^i, \tag{4.5}$$

as well as 3 Euler angles $\alpha^{[ij]}$ that determine the orientation of the coordinate axes, where $i = 1, 2, 3$. This is because all the axes of the local Lorentz frames along $p(\tau)$ can be obtained by parallel transporting the axes at $p(0)$.

We now introduce angular coordinates (θ, ϕ) at each τ which coincide with the angular variables of the spherical coordinate system of the local Lorentz frame in an infinitesimally small neighborhood of $p(\tau)$. We then define the "radial" coordinate λ for fixed τ, θ, ϕ as the affine parameter associated with the light ray emitted toward the past from $p(\tau)$ in the direction of (θ, ϕ). The origin and normalization of λ are taken so that the values of λ agree with those of the radial coordinate of the local Lorentz frame in an infinitesimally small neighborhood of $p(\tau)$. We perform this procedure in an inextendible spacetime; for example, we do not terminate the light ray at a coordinate singularity. This process allows us to introduce the coordinate system, which we call the *observer-centric coordinate system*. It has 4 coordinates τ, λ, θ, and ϕ, depicted schematically in Fig. 4.1, and provides a reference frame from which physics is described. Note that a hypersurface with constant τ corresponds to the past light cone of $p(\tau)$, which is a null, rather than spacelike, hypersurface. To describe a state, we need this coordinate system only in an infinitesimally small neighborhood of the $\tau = 0$ hypersurface. The reason why we need the neighborhood is that some phase space variables involve the τ derivative of quantum fields at $\tau = 0$.

We describe a quantum state, e.g. the configuration of matter on the "equal time" (null) hypersurface, using the observer-centric coordinate system throughout the evolution of the system. The introduction of this "absolute coordinate system" allows us to view gravity as a force measured in these coordinates—the motion of a particle of mass m under the influence of gravity can be expressed as $m\ddot{\chi} = F$, where $\chi = (\lambda, \theta, \phi)$ and the dot represents a τ derivative.

For a given spacetime, we may convert a coordinate system x^μ to the observer-centric one once a local Lorentz frame is elected. For this purpose, we regard x^μ to be functions of the observer-

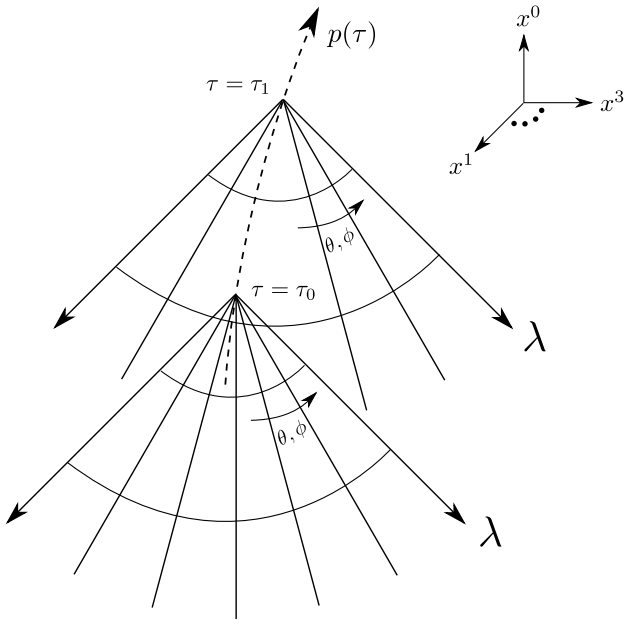


Figure 4.1: A schematic depiction of the observer-centric coordinate system.

centric coordinates, $x^\mu(\tau, \lambda, \theta, \phi)$, and derive equations that allow us to solve these functions. Note that the form of these functions depends on the choice of the local Lorentz frame, $(q^\mu, v^i, \alpha^{[ij]})$.

4.2.2 Gravitational observer horizon

In general, an element of $\mathcal{H}$ represents only a portion of L_{p_0}. Specifically, a past-directed light ray emitted from p_0 will hit a point beyond which the semi-classical description of spacetime is not applicable. The collection of these points forms a two-dimensional surface

$$\lambda = \lambda_{\rm obs}(\theta, \phi), \tag{4.6}$$

which we call the *gravitational observer horizon*, or the observer horizon for short. In general, we expect that this surface is determined by some condition which indicates that the intrinsically quantum gravitational physics becomes important there. In some simple cases, however, we may be able to state the condition more explicitly.

Consider a spacetime trajectory of a point with constant (λ, θ, ϕ) in the infinitesimal vicinity of L_{p_0}. Its proper velocity is given by

$$u^\mu = \frac{\frac{\partial x^\mu}{\partial \tau}}{\sqrt{-g_{\mu\nu}\frac{\partial x^\mu}{\partial \tau}\frac{\partial x^\nu}{\partial \tau}}}, \tag{4.7}$$

while the local proper acceleration by

$$a^\mu = u^\nu \nabla_\nu u^\mu. \tag{4.8}$$

Here, x^μ is an arbitrary coordinate system. $a^\mu(\tau, \lambda = 0, \theta, \phi) = 0$ since $p(\tau)$ is a geodesic, but if $\lambda > 0$, a trajectory of constant (λ, θ, ϕ) need not be a geodesic so we may have $a^\mu(\tau, \lambda, \theta, \phi) \neq 0$. a^μ has dimensions of energy in natural units $\hbar = c = 1$. Note that u^μ is timelike while a^μ is spacelike (or zero) within a (coordinate) horizon $g_{\tau\tau} = g_{\mu\nu}(\partial x^\mu/\partial\tau)(\partial x^\nu/\partial\tau) = 0$, where these vectors diverge.

In general, special behaviors of these quantities, e.g. $g_{\tau\tau} \to 0$ and $a^\mu \to \infty$, may be merely coordinate artifacts. We claim, however, that when the system under consideration is static, i.e. when the spacetime admits a timelike Killing vector k^μ and when the geodesic, $p(\tau)$, is approximately along this vector ($dp^\mu(\tau)/d\tau \propto k^\mu$), then the surface on which the magnitude of the local proper acceleration vector a^μ becomes the cutoff scale M_* signals the breakdown of the semi-classical description, giving the surface $\lambda = \lambda_{\rm obs}(\theta, \phi)$. Namely, in a static situation, the semi-classical picture is applicable only on a portion of L_{p_0} in which

$$A \equiv \sqrt{a^\mu a_\mu} \lesssim M_*. \tag{4.9}$$

This is a natural criterion given that a^μ measures acceleration relative to a free-fall. It can be interpreted as the condition that the gravitational acceleration measured from the reference frame—i.e. using the observer-centric coordinates—must be smaller than M_*.

In simple spacetimes, we can explicitly see that the local Hawking temperatures on surfaces $\lambda = \lambda_{\rm obs}$ determined by the condition in Eq. (4.9) actually become of order M_*, so the semi-classical picture is indeed expected to break down there. In these spacetimes, the observer horizons are reduced to the stretched horizons defined in Ref. [3]. In de Sitter space, for example, the observer horizon is located at $r = 1/H - O(H/M_*^2)$ in the static coordinates when calculated from $p(\tau)$ staying at the origin, where H is the Hubble constant. An important point, however, is that unlike the stretched horizon, the definition of the observer horizon does not require knowledge of spacetime outside of L_{p_0}. This is a desirable feature, as it allows us to construct a state without relying on the information in the spacetime region outside the one represented by the state. We also note that the spacetime location of the observer horizon, as well as the functional form of $\lambda_{\rm obs}(\theta, \phi)$, depends in general on the choice of the reference frame $(v^i, \alpha^{[ij]})$. This is another, important difference of the observer horizon from the stretched horizon defined in a conventional manner.

We consider that each region of the observer horizon holds $\mathcal{A}/4l_{\rm Pl}^2$ quantum degrees of freedom at the leading order in $l_{\rm Pl}^2/\mathcal{A}$, where $\mathcal{A}$ is the area of the region.[5] This comes from the requirement that

[5]The number of degrees of freedom is defined as the natural logarithm of the dimension of the corresponding

the spacetime region "outside" the observer horizon in the global spacetime picture is reproduced by an appropriate reference frame change (complementarity). (See Ref. [46] for recent discussions on how this may actually work.) Our picture is such that the degrees of freedom associated with the "outside spacetime" are entirely in the boundary degrees of freedom on the observer horizon. In fact, the number of the boundary degrees of freedom postulated here is sufficient for this purpose because of the holographic principle [11, 12, 9]. (In the case of a back hole viewed from a distant frame, these degrees of freedom are the stretched horizon degrees of freedom.) An element of $\mathcal{H}$, therefore, may be said to represent a physical state of the degrees of freedom in *and* on the observer horizon:

$$0 \leq \lambda \leq \lambda_{\rm obs}(\theta, \phi). \tag{4.10}$$

Note that the bulk and boundary degrees of freedom will in general be entangled since the horizon forms by a dynamical process. Entanglement between the two will also be necessary to reconstruct the outside region when a relevant reference frame change is made [47].

4.2.3 Other "ends" of spacetime on L_{p_0}

We now discuss other ways in which semi-classical spacetime ceases to exist on L_{p_0} along a light ray generating it. For this purpose, we assume that the observer horizon is located sufficiently far away, $\lambda_{\rm obs}(\theta, \phi) \to \infty$. We argue that there are two ways that the light ray may encounter the "end" of spacetime on L_{p_0} even in this case.

The first possibility is for a light ray to hit a spacetime singularity. Consider a null geodesic representing a light ray emitted from p_0 toward the past in the direction of (θ, ϕ). Suppose that the geodesic encounters a spacetime singularity in the sense that it is inextendible beyond some finite value of the affine parameter $\lambda_{\rm sing}(\theta, \phi)$ in an inextendible spacetime. In this case, semi-classical spacetime exists only in the region $\lambda < \lambda_{\rm sing}(\theta, \phi)$, and we consider that an element of $\mathcal{H}$ represents the physical state of the degrees of freedom only in that region.

The other possibility has to do with the behavior of the congruence of past-directed light rays emitted from p_0. Assuming the null energy condition, $T_{\mu\nu} v^\mu v^\nu \geq 0$ for all null vectors v^μ, the expansion of the light rays Θ satisfies [48]

$$\frac{\partial \Theta}{\partial \lambda} + \frac{1}{2}\Theta^2 \leq 0. \tag{4.11}$$

This implies that the light rays emitted from p_0 converge toward the past, starting from $\Theta = +\infty$ at $\lambda = 0_+$.

Suppose that a light ray reaches a point where $\Theta = -\infty$ at some finite value of the affine parameter $\lambda_{\rm conj}(\theta, \phi)$ (before it hits a spacetime singularity). Such a point is said to be conjugate to

Hilbert space factor. By the leading order, we mean that the number of degrees of freedom is $(\mathcal{A}/4l_{\rm Pl}^2)\{1 + O(l_{\rm Pl}^{2n}/\mathcal{A}^n)\}$ with $n > 1$.

p_0, and signals the failure of the light ray being on the boundary of the past of p_0 [48]. Specifically, there exists a family of timelike causal curves connecting p_0 and a point q on L_{p_0} with $\lambda > \lambda_{\rm conj}(\theta,\phi)$. Now, suppose semi-classical spacetime exits beyond $\lambda_{\rm conj}(\theta,\phi)$ in our framework. This would contradict the validity of null quantization, which we are assuming throughout. In particular, it would mean that a massive particle sent from q—which, being on L_{p_0}, is at an "equal time" as p_0—can travel backward in time and reach p_0 from the past (as there exits a timelike causal curve connecting q and p_0). We therefore consider that $\Theta = -\infty$ signals the end of spacetime, and that an element of $\mathcal{H}$ only represents the region $\lambda < \lambda_{\rm conj}(\theta,\phi)$.

Combining with the possibility of hitting a spacetime singularity discussed above, we conclude that an element of $\mathcal{H}$ represents a physical state of the degrees of freedom in the region

$$0 \le \lambda < \lambda_{\rm end}(\theta,\phi) \equiv \min\left\{\lambda_{\rm sing}(\theta,\phi), \lambda_{\rm conj}(\theta,\phi)\right\}, \tag{4.12}$$

where we have assumed that $\lambda_{\rm obs}(\theta,\phi) > \lambda_{\rm end}(\theta,\phi)$. If a light ray hits the observer horizon before it reaches a singularity or a conjugate point, i.e. $\lambda_{\rm obs}(\theta,\phi) < \lambda_{\rm end}(\theta,\phi)$, then spacetime must be terminated there and the boundary degrees of freedom must be attached, according to the discussion in the previous subsection.

We assume that, unlike the observer horizon, the two-dimensional surface determined by $\lambda = \lambda_{\rm end}(\theta,\phi)$ does *not* hold boundary degrees of freedom. This corresponds to the hypothesis that the evolution of a state can be determined without any information from the singularity or the region beyond $\lambda_{\rm conj}(\theta,\phi)$, in addition to what is already in the Hamiltonian. For example, the evolution of a big-bang universe is not affected by the "details" of the big-bang singularity that must be specified beyond the Einstein equation.

4.2.4 Apparent horizon "pull-back"

We have seen that spacetime on the past light cone of p_0 is extended only until λ reaches $\lambda_{\rm obs}$ of Section 4.2.2 or $\lambda_{\rm end}$ of Section 4.2.3. (Here and below, until Eq. (4.15), we omit the arguments from the boundary locations, but it should be remembered that they are functions of θ and ϕ.) In the former case, the boundary degrees of freedom are attached with the number $\mathcal{A}/4l_{\rm Pl}^2$ per area $\mathcal{A}$, while in the latter case, none are attached. Here we discuss a description in which this asymmetry of boundary degrees of freedom is dissolved and all the boundaries are treated on equal footing for the purpose of counting degrees of freedom. This description is available if the following condition is satisfied:

$$\lambda_{\rm obs} \le \lambda_{\rm sing} \qquad \text{or} \qquad \lambda_{\rm conj} \le \lambda_{\rm sing}, \tag{4.13}$$

i.e. a singularity is screened either by the observer horizon or conjugate point. Indeed, in example spacetimes we have investigated, this condition is always satisfied, although we do not have a proof of it. Below, we assume that Eq. (4.13) is valid, and disregard a singularity surface.

Let us define the apparent horizon as a surface on which the expansion of the past-directed light rays emitted from p_0 first crosses zero:[6]

$$\Theta = 0 \quad \text{at} \quad \lambda = \lambda_{\rm app}. \tag{4.14}$$

This implies that $\lambda_{\rm app} < \lambda_{\rm conj}$, since Θ is a monotonically decreasing function of λ. Now, if $\lambda_{\rm obs} < \lambda_{\rm app}$ for a range of (θ, ϕ), then in these directions spacetime ceases to exist at $\lambda = \lambda_{\rm obs}$, where a boundary degree of freedom is located per area $4l_{\rm Pl}^2$. On the other hand, if $\lambda_{\rm app} < \lambda_{\rm obs}$ for a range of (θ, ϕ), then there are two cases to consider:

1. $\lambda_{\rm app} < \lambda_{\rm conj} < \lambda_{\rm obs}$ — In this case, spacetime exists only for $\lambda < \lambda_{\rm conj}$. The covariant entropy bound then implies that the number of physical degrees of freedom in the region $\lambda > \lambda_{\rm app}$ is bounded by $\mathcal{A}/4l_{\rm Pl}^2$, where $\mathcal{A}$ is the area of the relevant portion of the apparent horizon [9, 49]. This suggests that these degrees of freedom may be replaced by $\mathcal{A}/4l_{\rm Pl}^2$ boundary degrees of freedom located on the apparent horizon.

2. $\lambda_{\rm app} < \lambda_{\rm obs} < \lambda_{\rm conj}$ — In this case, physical degrees of freedom outside the apparent horizon consist of the bulk degrees of freedom in $\lambda_{\rm app} < \lambda < \lambda_{\rm obs}$ and the boundary degrees of freedom at $\lambda = \lambda_{\rm obs}$. If the strengthened covariant entropy bound of Ref. [50] applies, then the number of the former is bounded by $(\mathcal{A} - \mathcal{A}_{\rm obs})/4l_{\rm Pl}^2$, while that of the latter is $\mathcal{A}_{\rm obs}/4l_{\rm Pl}^2$, where $\mathcal{A}$ and $\mathcal{A}_{\rm obs}$ are the areas of the relevant portions of the apparent and observer horizons, respectively. This suggests that physical degrees of freedom in the region $\lambda > \lambda_{\rm app}$ may be replaced by $\mathcal{A}/4l_{\rm Pl}^2$ boundary degrees of freedom on the apparent horizon. While the strengthened covariant entropy bound is known to be violated in some extreme cases, we assume that this replacement can always be done in our context.

We thus find that both cases allow for replacing physical degrees of freedom in the region $\lambda > \lambda_{\rm app}$ by a quantum degree of freedom per area $4l_{\rm Pl}^2$ on the apparent horizon. We call this replacement procedure *apparent horizon pull-back*.

With the apparent horizon pull-back, the structure of the physical region represented by an element of $\mathcal{H}$ can be stated in the following simple way. Spacetime on L_{p_0} exists only for

$$0 \le \lambda \le \lambda_B(\theta, \phi) \equiv \min\left\{\lambda_{\rm obs}(\theta, \phi), \lambda_{\rm app}(\theta, \phi)\right\}. \tag{4.15}$$

In addition to the degrees of freedom in the bulk of spacetime, the boundary at $\lambda = \lambda_B(\theta, \phi)$ also holds $\mathcal{A}/4l_{\rm Pl}^2$ quantum degrees of freedom (at the leading order in $l_{\rm Pl}^2/\mathcal{A}$), where $\mathcal{A}$ is the area of the boundary.

[6] This definition is different from that in Ref. [49], where the apparent horizon is defined as a surface on which at least one pair among *four* orthogonal null congruences have zero expansion. Here we only consider two directions along L_{p_0}.

4.2.5 Horizon decomposition of $\mathcal{H}$

So far, we have been discussing the structure of spacetime represented by *an* element of $\mathcal{H}$. The full Hilbert space $\mathcal{H}$ consists of the elements representing "all possible" physical configurations in "all possible" spacetimes, as viewed from the reference frame. What do we really mean by that? In other words, what is the structure of $\mathcal{H}$ concretely?

To address this question, let us adopt the apparent-horizon pulled-back description, discussed in the previous subsection. We now group the elements that have the same boundary $\partial\mathcal{M}$, and denote the Hilbert space spanned by these elements by $\mathcal{H}_{\partial\mathcal{M}}$.[7] The general definition of the boundary being the same is not obvious to give explicitly. One possible definition, which seems to work if the boundary is within the coordinate horizon $g_{\tau\tau} = 0$, is given as follows. Consider the induced metric on the boundary $\lambda = \lambda_B(\theta, \phi)$ with the arguments being the observer-centric angular variables:

$$h_{XY}(\theta, \phi) = \frac{\partial\lambda_B}{\partial X}\frac{\partial\lambda_B}{\partial Y}g_{\lambda\lambda} + \frac{\partial\lambda_B}{\partial X}g_{\lambda Y} + \frac{\partial\lambda_B}{\partial Y}g_{\lambda X} + g_{XY}, \tag{4.16}$$

where $X, Y = \theta, \phi$, and $g_{\lambda\lambda}$, $g_{\lambda X}$, and g_{XY} are spacetime metric components in the observer-centric coordinate system, evaluated at $\tau = 0$ and $\lambda = \lambda_B(\theta, \phi)$. We regard two boundaries as the same if the induced metrics on them are *explicitly* identical, i.e. all the h_{XY}'s $(X, Y = \theta, \phi)$ take the identical functional forms with respect to θ, ϕ.[8]

This definition reflects the fact that our description of physics is "special relativistic" or "as viewed from the reference frame." For example, a spacetime 2-surface is regarded as different boundaries when described from two different reference frames which are rotated with respect to with each other (unless the surface is spherically symmetric around p_0). This implies that depending on the choice of the reference frame, the identical physical configuration in spacetime can belong to different Hilbert subspaces $\mathcal{H}_{\partial\mathcal{M}}$. An operator corresponding to rotating the reference frame then transforms an element of a subspace into that of another. Note that here we are talking about a state $|\psi_i\rangle$ in $\mathcal{H} \subset \mathcal{H}_{\rm QG}$, which may be viewed as representing a physical state relative to clock degrees of freedom. The "full" quantum state (i.e. the multiverse state) $|\Psi\rangle \subset \mathcal{H}_{\rm phys}$ obtained after imposing the constraints in Eq. (4.2) is, of course, invariant under such a rotation (guaranteeing that there is no absolute frame in the universe).

Now, the elements of $\mathcal{H}_{\partial\mathcal{M}}$ represent all possible physical configurations in all possible spacetimes (or null slices of spacetimes) that share the same boundary $\partial\mathcal{M}$ as defined above. Let us denote the Hilbert space factors of $\mathcal{H}_{\partial\mathcal{M}}$ corresponding to the bulk and boundary degrees of

[7] The $\mathcal{H}_{\partial\mathcal{M}}$ here is the same as what is denoted by $\mathcal{H}_{\mathcal{M}}$ in earlier work Refs. [33, 44, 45].

[8] It is not entirely clear if there is no additional condition for the boundaries being the same; for example, we might have to require $\lambda_B(\theta, \phi)$ to be the same in addition to $h_{XY}(\theta, \phi)$. Here we postulate that the identity of $h_{XY}(\theta, \phi)$ is sufficient, and proceed with it.

freedom by $\mathcal{H}_{\partial\mathcal{M},\,\text{bulk}}$ and $\mathcal{H}_{\partial\mathcal{M},\,B}$, respectively:

$$\mathcal{H}_{\partial\mathcal{M}} = \mathcal{H}_{\partial\mathcal{M},\,\text{bulk}} \otimes \mathcal{H}_{\partial\mathcal{M},\,B}, \tag{4.17}$$

where the direct product structure arises from the locality hypothesis in our framework. According to the covariant entropy bound [9], the dimension of the Hilbert space factor $\mathcal{H}_{\partial\mathcal{M},\,\text{bulk}}$ is bounded by the area of the boundary $\mathcal{A}_{\partial\mathcal{M}}$ as $\dim\mathcal{H}_{\partial\mathcal{M},\,\text{bulk}} \leq \exp(\mathcal{A}_{\partial\mathcal{M}}/4l_{\text{Pl}}^2)$. On the other hand, by construction the dimension of the boundary factor is $\dim\mathcal{H}_{\partial\mathcal{M},\,B} = \exp(\mathcal{A}_{\partial\mathcal{M}}/4l_{\text{Pl}}^2)$. Therefore, we find

$$\dim\mathcal{H}_{\partial\mathcal{M}} = \dim\mathcal{H}_{\partial\mathcal{M},\,\text{bulk}} \times \dim\mathcal{H}_{\partial\mathcal{M},\,B} \leq \exp\left(\frac{\mathcal{A}_{\partial\mathcal{M}}}{2l_{\text{Pl}}^2}\right). \tag{4.18}$$

Note that this includes arbitrary fluctuations of spacetimes as well as arbitrary configurations of matter (which are related by Einstein's equation with each other) that keep the boundary fixed, namely with $h_{XY}(\theta,\phi)$ held fixed.[9]

The complete spacetime part of the Hilbert space $\mathcal{H}$ is then given by the direct sum of the Hilbert subspaces $\mathcal{H}_{\partial\mathcal{M}}$ for different $\partial\mathcal{M}$'s:

$$\mathcal{H} = \bigoplus_{\partial\mathcal{M}} \mathcal{H}_{\partial\mathcal{M}}, \tag{4.19}$$

where the direct sum runs over $\partial\mathcal{M} = \{h_{XY}(\theta,\phi)\}$. We call the expression of this form the *horizon decomposition* of $\mathcal{H}$. In general, what $\partial\mathcal{M}$'s are included in the decomposition of the complete Hilbert space $\mathcal{H}$ cannot be determined by the low energy consideration alone. For instance, some spacetimes such as stable (not cosmological) de Sitter space may be unrealistic mathematical idealizations and may not appear in the underlying full quantum theory of gravity. In practice, however, we may include only $\partial\mathcal{M}$'s that are relevant to the problem under consideration (the ones relevant for the clock degrees of freedom), and that is sufficient. For discussions of this issue in cosmology, especially in the eternally inflating multiverse, see Ref. [44].

4.2.6 Spacelike quantization

Finally, we discuss briefly if there is a way to use spacelike hypersurfaces, rather than null hypersurfaces, to quantize the system. Such a spacelike quantization would avoid technical subtleties

[9]Recently, the analysis above has been significantly refined in Ref. [46], which claims that for physical states the relevant space is given by $\mathcal{H}_{\partial\mathcal{M}}$ with $\dim\mathcal{H}_{\partial\mathcal{M}} = \exp(\mathcal{A}_{\partial\mathcal{M}}/4l_{\text{Pl}}^2)$ (at least at leading order in an $l_{\text{Pl}}^2/\mathcal{A}_{\partial\mathcal{M}}$ expansion in the exponent), which is much smaller than $\exp(\mathcal{A}_{\partial\mathcal{M}}/2l_{\text{Pl}}^2)$ appearing in the last expression in Eq. (4.18). This is possible because the contribution from the bulk region is in general tiny $\approx O(\mathcal{A}^n/l_{\text{Pl}}^{2n})$ $(n<1)$ [11] for physically realizable states, and hence can be neglected at the leading order. In fact, when $\partial\mathcal{M}$ is the observer horizon, we find that $\ln\dim\mathcal{H}_{\partial\mathcal{M}}$ for physical states is saturated (at the leading order in $l_{\text{Pl}}^2/\mathcal{A}_{\partial\mathcal{M}}$) by the *entropy of a vacuum*—the logarithm of the number of possible independent ways in which quantum field theory on a fixed classical spacetime background can emerge in a full quantum theory of gravity [46].

associated with null quantization, for example, non-commutativity of field operators at different points in a same angular direction (see footnote 3).

One possibility is simply to "round" the light cone L_{p_0} slightly to make an equal-time hypersurface spacelike. We can do this while keeping the boundary $\partial\mathcal{M}$ fixed. An advantage of this procedure is that the structure of the Hilbert space is unchanged from that in Eqs. (4.17 – 4.19). This is because the future-directed ingoing light sheet of $\partial\mathcal{M}$ (a portion of L_{p_0} bounded by $\partial\mathcal{M}$) is complete (ending at the caustic at p_0), so that the spacelike projection theorem of Ref. [9] applies. In a sense, our null quantization may be viewed as a limit of the spacelike quantization discussed here (although the limit is not completely smooth).

Another possibility is to adopt an "intrinsically spacelike" construction. Specifically, we may follow a similar construction to our covariant Hilbert space using spacelike geodesics attached to the local Lorentz frame at p_0 (e.g. with the affine parameters taken to agree with the radial coordinate in the infinitesimal vicinity of p_0), instead of null geodesics (light rays). In particular, we may define acceleration parameter A and the observer horizon similarly in a static situation. This construction corresponds to taking different linear combinations of the constraint operators $\mathcal{P}^\mu(\mathbf{x})$ as H, P_i, $J_{[ij]}$, and K_i (see discussion in Section 4.1). The validity of this approach or its relation to the null quantization presented here is not clear.

Chapter 5

Macroscopic Superpositions of Spacetimes

The last chapter emphasized the importance of fixing the reference frame in a quantum description of spacetime. In this chapter, we will begin to focus our attention on the case of a reference frame outside of or inside of a black hole. In the following chapter we will address detailed microscopic aspects of black holes which will include a discussion of firewalls [4]. However, it will first be necessary to consider quantum processes that lead to physics occurring at very large distance scales.

5.1 Black Holes and Unitarity—A Distant View

In this section we discuss the process in which a black hole unitarily forms and evaporates, as viewed from a distant reference frame. We clarify the meaning of the information in this context, and argue that it (partly) lies in relative coefficients—especially phases—of terms representing *macroscopically distinct* configurations in a full quantum state. We also elucidate the fact that a physical observer can never extract complete (quantum) information of the initial state forming the black hole; i.e., observing final-state Hawking radiation does not allow for him/her to infer the initial state, despite the fact that the evolution of the entire quantum state is fully unitary.

5.1.1 Black Hole Information

In his famous 1976 paper, Hawking argued, based on semi-classical considerations, that a black hole loses information [55]. Consider two objects having the same energy-momentum, represented by pure quantum states $|A\rangle$ and $|B\rangle$, which later collapse into black holes with the same mass $M(0)$. According to the semi-classical picture, the evolutions of the two states after forming the black holes are identical, leading to the same mixed state ρ_H, obtained by integrating the thermal

Hawking radiation states:

$$\begin{array}{rcl} |A\rangle & \to & \rho_H, \\ |B\rangle & \to & \rho_H. \end{array} \tag{5.1}$$

This phenomenon is referred to as the information loss in black holes.

What is the problem of this picture? The problem is that since the final states are identical, we cannot recover the initial state of the evolution just by knowing the final state, even in principle. This contradicts unitarity of quantum mechanical evolution, which says that time evolution of a state is reversible, i.e. we can always recover the initial state if we know the final state exactly by applying the inverse time evolution operator e^{+iHt}.

Based on various circumstantial evidence, especially AdS/CFT duality [1], we now do not think the above picture is correct. We think that the final states obtained from different initial states differ, and a state obtained by evolving any pure state is always pure even if the evolution involves formation and evaporation of a black hole. Namely, instead of Eq. (5.1), we have

$$\begin{array}{rcl} |A\rangle & \to & |\psi_A\rangle, \\ |B\rangle & \to & |\psi_B\rangle, \end{array} \tag{5.2}$$

where $|\psi_A\rangle \neq |\psi_B\rangle$ iff $|A\rangle \neq |B\rangle$. In this picture, quantum states representing black holes formed by different initial states are different, even if they have the same mass. (The dimension of the Hilbert space corresponding to a classical black hole of a fixed mass M is $\exp(\mathcal{A}_{\rm BH}/4)$ according to the Bekenstein-Hawking entropy, where $\mathcal{A}_{\rm BH} = 16\pi M^2$ is the area of the black hole horizon.) These states then evolve into different final states $|\psi_A\rangle$ and $|\psi_B\rangle$, representing states for emitted Hawking radiation quanta.

A question is in what form the information is encoded in the final state. On one hand, possible final states of evaporation of a black hole must have a sufficient variety to encode complete information about the initial state forming the black hole. This requires that the dimension of the Hilbert space corresponding to these states must be of order $\exp(\mathcal{A}_{\rm BH}(0)/4)$, where $\mathcal{A}_{\rm BH}(0) = 16\pi M(0)^2$ is the area of the black hole horizon right after the formation. On the other hand, Hawking radiation quanta emitted from the black hole must have the thermal spectrum (with temperature $T_H = 1/8\pi M$ when the black hole mass is M) in the regime where the semi-classical analysis is valid, $M \gg 1$. It is not clear how the state actually realizes these two features [66], although the generalized second law of thermodynamics guarantees that it can be done. Below, we argue that a part of the information that is necessary to recover the initial state is contained in relative coefficients of terms representing different macroscopic worlds, even if the initial state has a well-defined classical configuration.

Our analysis does not prove unitarity of the black hole formation/evaporation process, or address the question of how the complete information of the initial state is encoded in the emitted Hawking quanta at the microscopic level. Rather, we assume that unitarity is preserved at the

microscopic level, and study manifestations of this assumption when we describe the process at a semi-classical level. This will provide implications on how such a description must be constructed. For example, in order to preserve all the information in the initial state, the description must not be given on a fixed black hole background in an intermediate stage of the evaporation, since it would correspond to ignoring a part of the information contained in the full quantum state manifested as macroscopic properties of the remaining black hole. Note that we do not claim that these macroscopic properties contain independent information beyond what is in the emitted Hawking quanta—the two are certainly correlated by energy-momentum conservation. The analysis presented here also has implications on the complementarity picture, which will be discussed in Section 5.2.

5.1.2 Where is the information in the black hole state?

Let us consider a process in which a black hole is formed from a pure state $|A\rangle$ and then evaporates. For simplicity, we assume that the black hole formed does not have a spin or charge. We describe this process in a distant reference frame, i.e. a freely falling (local Lorentz) frame whose origin p is outside the black hole horizon all the time; see the left panel of Fig. 5.1. In its minimal implementation, the framework of Ref. [33] says that quantum states represent physical configurations on the past light cone of p in and on the stretched/apparent horizon.[1] This description, therefore, represents evolution of the system in the shaded spacetime region in the left panel of Fig. 5.1.

An important point is that this provides a *complete* description of the *entire* system [62, 3]—it is not that we describe only a part of the system corresponding to the shaded region; physics is complete in that spacetime region. The picture describing the infalling matter inside the horizon can be obtained only after performing a unitary transformation on the state corresponding to a change of the reference frame to an infalling one [33] (which in general leads to a superposition of infalling and distant views, as will be explained in Section 5.2). In this sense, the *entire* spacetime is better represented by a Penrose diagram in the right panel of Fig. 5.1 when the system is described in a distant reference frame. As is clear from the figure, this allows for an S-matrix description of the process in Hilbert space representing Minkowski space $\mathcal{H}_{\rm Minkowski}$, which is a subspace of the whole covariant Hilbert space for quantum gravity: $\mathcal{H}_{\rm Minkowski} \subset \mathcal{H}_{\rm QG}$. This is the case despite the fact that in general quantum mechanics requires only that the evolution of a state is unitary in the whole Hilbert space $\mathcal{H}_{\rm QG}$; see Section 5.2 for more discussions on this point.

What does the evolution of a quantum state look like in this description? Let us denote the black hole state right after the collapse of the matter by $|{\rm BH}^0_A\rangle$. Since the subsequent evolution

[1] The stretched horizon is defined as a time-like hypersurface on which the local Hawking temperature becomes of order the Planck scale and thus short-distance quantum gravity effects become important (where we have not discriminated between the string and Planck scales). In the Schwarzschild coordinates, it is located at $r-2M \approx 1/M$.

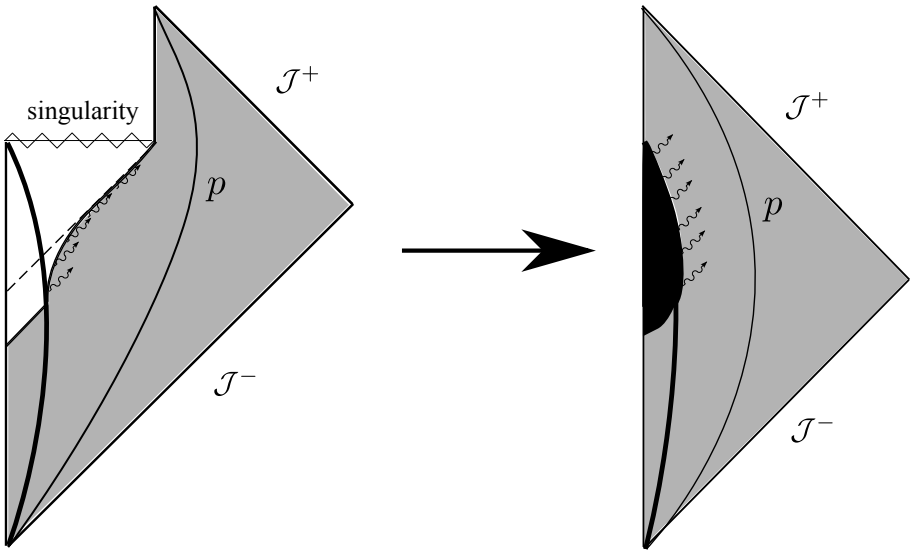


Figure 5.1: The Penrose diagram representing a black hole formed from a collapsing shell of matter (represented by the thick solid curve) which then evaporates. The left panel shows the standard "global spacetime" picture, in which Hawking radiation (denoted by wavy arrows) comes from the stretched horizon. To obtain a consistent quantum mechanical description, we must fix a reference frame (freely falling frame) and then describe the system from that viewpoint [33]. Quantum states then corresponds to physical configurations in the past light cone of the origin p of that reference frame. Here we choose a "distant" reference frame; the trajectory of its origin p is depicted by a thin solid curve. With this choice, a *complete* description of the evolution of the system is obtained in the shaded region in the panel. In other words, the conformal structure of the *entire* spacetime is as in the right panel, when the system is described in this reference frame.

is unitary, the state can be written in the form $\sum_i a_i^t |\mathrm{BH}_i^t\rangle \otimes |\psi_i^t\rangle$. Here, $|\mathrm{BH}_i^t\rangle$ represent states of the black hole (i.e. the horizon degrees of freedom) when time t is passed since the formation, while $|\psi_i^t\rangle$ those of the rest of the world at the same time, where t is the proper time measured at the origin of the reference frame p. (The dimension of the Hilbert space for $|\mathrm{BH}_i^t\rangle$, $\mathcal{H}_{\mathrm{BH}}^t$, is $\exp(\mathcal{A}_{\mathrm{BH}}(t)/4)$ with $\mathcal{A}_{\mathrm{BH}}(t) = 16\pi M(t)^2$, where $M(t)$ is the mass of the black hole at time t; the state $|\mathrm{BH}_A^0\rangle$ is an element of $\mathcal{H}_{\mathrm{BH}}^0$.) The entire state then evolves into a state representing the final Hawking radiation quanta, which can be written as $\sum_i a_i^\infty |\psi_i^\infty\rangle$. Summarizing, the evolution of the system is described as

$$|A\rangle \longrightarrow |\mathrm{BH}_A^0\rangle \longrightarrow \sum_i a_i^t |\mathrm{BH}_i^t\rangle \otimes |\psi_i^t\rangle \longrightarrow \sum_i a_i^\infty |\psi_i^\infty\rangle. \tag{5.3}$$

The complete information about the initial state is contained in the state at any time t in the set of complex coefficients when the state is expanded in fixed basis states. In particular, after the evaporation it is contained in $\{a_i^\infty\}$ showing how the radiation states are superposed.

5.1.3 Black hole drifting: a macroscopic uncertainty of the black hole location after a long time

What actually are the states $|\psi_i^t\rangle$? Namely, what does the intermediate stage of the evaporation look like when it is described from a distant reference frame? Here we argue that $|\psi_i^t\rangle$ for different i span macroscopically different worlds. In particular, the state of the black hole becomes a superposition of macroscopically different geometries (in the sense that they represent different spacetimes as viewed from the reference frame) throughout the course of the evaporation. The analysis here builds upon an earlier suggestion by Page, who noted a large backreaction of Hawking emissions to the location of an evaporating black hole [52].

To analyze the issue, let us take a semi-classical picture of the evaporation but in which the backreaction of the Hawking emission to the black hole energy-momentum is explicitly taken into account. Specifically, we model it by a process in which the black hole emits a massless quantum with energy $\sim 1/M$ in a random direction in each time interval $\sim M$, in the rest frame of the black hole. Here, M is the mass of the black hole at the time of the emission. Suppose that the velocity of the black hole is $\mathbf{v}$ before an emission; then the emission of a Hawking quantum will change the four-momentum of the black hole as

$$p^\mu_{\rm BH} = \begin{pmatrix} M\gamma \\ M\gamma\mathbf{v} \end{pmatrix} \longrightarrow \begin{pmatrix} M\gamma - \frac{\gamma}{M}(1-\mathbf{n}\cdot\mathbf{v}) \\ M\gamma\mathbf{v} + \frac{1}{M}\mathbf{n} - \frac{1-\gamma}{M}\frac{\mathbf{n}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v} - \frac{\gamma}{M}\mathbf{v} \end{pmatrix}, \tag{5.4}$$

where $\gamma \equiv 1/\sqrt{1-|\mathbf{v}|^2}$ and $\mathbf{n}$ is a unit vector pointing to a random direction. The mass and the velocity of the black hole, therefore, change by

$$\Delta M \;=\; \sqrt{M^2-2} - M \;\approx\; -\frac{1}{M}, \tag{5.5}$$

$$\Delta\mathbf{v} \;=\; \frac{1}{\gamma\{M^2-(1-\mathbf{n}\cdot\mathbf{v})\}}\left\{\mathbf{n} - \left(1-\frac{1}{\gamma}\right)\frac{\mathbf{n}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v}\right\} \;\approx\; \frac{1}{M^2}\mathbf{n} - \frac{\mathbf{n}\cdot\mathbf{v}}{2M^2}\mathbf{v}, \tag{5.6}$$

in each time interval

$$\Delta t \;=\; M\gamma \;\approx\; M, \tag{5.7}$$

where we have taken the approximation that $M \gg 1$ and $|\mathbf{v}| \ll 1$ in the rightmost expressions. In general, the emission of a Hawking quantum can also change the black hole angular momentum $\mathbf{J}$. We consider this effect in section 5.1.4, where we find that the black hole accumulates macroscopic angular momentum, $|\mathbf{J}| \gg 1$, after long time. This, however, does not affect the essential part of the discussion below, so we will suppress it in most part.

Now, suppose that a (non-spinning) black hole is formed at $t = 0$ with the initial mass $M_0 \equiv M(0)$. Then, in timescales of order M_0^3 or shorter, the black hole mass is still of order M_0 until the very last moment of the evaporation. (For example, at the Page time $t_{\rm Page} \sim M_0^3$, at which the black hole loses a half of its initial entropy, the black hole mass is still $M \approx M_0/\sqrt{2}$.) The

above process, therefore, can be well approximated by a process in which the black hole receives a velocity kick of $|\Delta\mathbf{v}| \approx 1/M_0^2$ in each time interval $\Delta t \approx M_0$, which after time t leads to black hole velocity

$$|\mathbf{v}_{\rm BH}| \approx |\Delta\mathbf{v}|\sqrt{\frac{t}{\Delta t}} \sim \frac{1}{M_0^{5/2}}\sqrt{t}, \tag{5.8}$$

whose direction does not change appreciably in each kick (and so is almost constant throughout the process). This implies that after time t ($M_0 \ll t \lesssim M_0^3$), the location of the black hole drifts in a random direction by an amount

$$|\mathbf{x}_{\rm BH}| \approx |\mathbf{v}_{\rm BH}|t \sim \frac{1}{M_0^{5/2}}t^{3/2}. \tag{5.9}$$

For $t \sim M_0^3$, this gives $|\mathbf{v}_{\rm BH}|_{t\sim M_0^3} \sim 1/M_0$ and

$$|\mathbf{x}_{\rm BH}|_{t\sim M_0^3} \sim M_0^2, \tag{5.10}$$

which is much larger than the Schwarzschild radius of the initial black hole, $R_S = 2M_0$. By the time of the final evaporation, the velocity is further accelerated to $|\mathbf{v}_{\rm BH}| \sim 1$, but the final displacement is still of the order of Eq. (5.10).

To appreciate how large the value of Eq. (5.10) is, consider a black hole whose lifetime is of the order of the current age of the universe, $t_{\rm evap} \sim 10^{10}$ years. It has the initial mass of $M_0 \sim 10^{12}$ kg, implying the initial Schwarzschild radius of $R_S \sim 1$ fm. The result in Eq. (5.10) says that the displacement of such a black hole is $|\mathbf{x}_{\rm BH}| \sim 100$ km at the time of evaporation! The origin of this surprisingly large effect is the longevity of the black hole lifetime, $t_{\rm evap} \sim M_0^3$. For example, for a black hole of the solar mass $M = M_\odot \sim 10^{30}$ kg (i.e. $R_S \sim 1$ km), the evaporation time is $t_{\rm evap} \sim 10^{62}$ years—52 orders of magnitude longer than the age of the universe.

In Fig. 5.2, we show the result of our simulations of the random process described above. In the left panel, we show the average value of $|\mathbf{x}_{\rm BH}|$ when the black hole mass is reduced to $M_0/\sqrt{2}$, i.e. at the Page time, as a function of M_0. We see the expected behavior of $\langle|\mathbf{x}_{\rm BH}|\rangle \sim M_0^2$. In the right panel, we show the distribution of $|\mathbf{x}_{\rm BH}|$ for a fixed M_0, which we take $M_0 = 5000$, with a large number of simulations, $N_{\rm total} = 10000$. We find that the probability distribution of $|\mathbf{x}_{\rm BH}|$ has the form

$$dP(|\mathbf{x}_{\rm BH}|) \propto |\mathbf{x}_{\rm BH}|^2 \exp\left(-c\frac{|\mathbf{x}_{\rm BH}|^2}{M_0^4}\right) d|\mathbf{x}_{\rm BH}|, \tag{5.11}$$

where c is a constant of $O(1)$, as implied by the central limit theorem, i.e. each component of $\mathbf{x}_{\rm BH}$ having the Gaussian distribution centered at zero with a width $\sim M_0^2$. We emphasize that the precise value of c obtained by the plot does not have a physical significance, since it reflects our particular modeling of evaporation and omission of various numerical coefficients such as 8π in the expression of Hawking temperature $T_H = 1/8\pi M$. Our point here is to show that the displacement

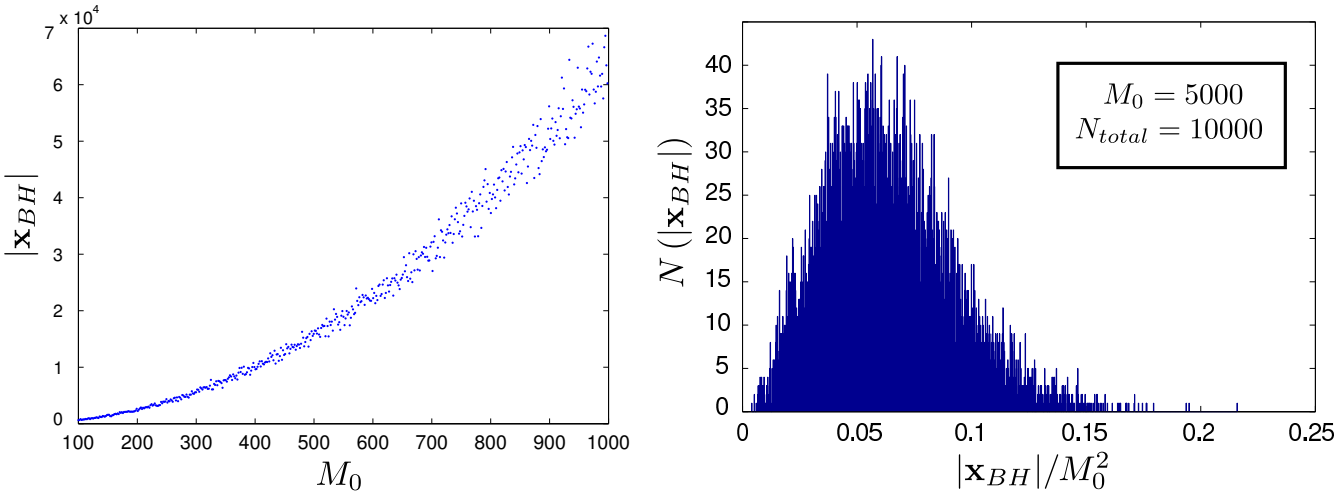


Figure 5.2: In the left panel, we show the result of simulating the black hole displacement $|\mathbf{x}_{\rm BH}|$ at the Page time, $t_{\rm Page} \sim M_0^3$, as a function of the initial black hole mass, M_0. We find the behavior expected from the general argument, $|\mathbf{x}_{\rm BH}| \sim M_0^2$. In the right panel, we show the probability distribution of the displacement $|\mathbf{x}_{\rm BH}|$ for a fixed $M_0 = 5000$, obtained by performing a larger number of simulations, $N_{\rm total} = 10000$. The distribution takes the form expected from the central limit theorem; see Eq. (5.11).

of the black hole is indeed of $O(M_0^2)$ and its distribution follows what is expected from the theory of statistics. Note also that the reason why the left plot appears to have smaller distributions in $|\mathbf{x}_{\rm BH}|$ is because it has a smaller sample size; $N_{\rm total}$ for each point is of $O(10)$ in that plot. In Fig. 5.3, we show typical paths of the black hole drift in three spatial dimensions. We see that the direction of the velocity stays nearly constant along a path, as suggested by the general analysis.

Quantum mechanically, the result described above implies that the state of the black hole becomes a superposition of terms in which the black hole exists in macroscopically different locations, even if the initial state forming the black hole is a classical object having a well-defined macroscopic configuration. At time $t \sim M_0^{7/3}$ after the formation (where t is the proper time measured at p), the uncertainty of the black hole location becomes of order M_0, comparable to the Schwarzschild radius of the original black hole. At the timescale of evaporation, $t \sim M_0^3$, the uncertainty is of order M_0^2, much larger than the initial Schwarzschild radius. This is illustrated schematically in Fig. 5.4. Note that each term in the figure still represents a superposition of terms having different phase space configurations of emitted Hawking quanta. Also, as shown in the section 5.1.4, each black hole at a fixed location is a superposition of black holes having macroscopically different angular momenta.

The evolution of the state depicted in Fig. 5.4 is obviously physical if we consider, for example, a super-Planckian scattering experiment. In this case, we will find that Hawking quanta emitted at the last stage of the evaporation will come from $\sim M_0^2$ away from the interaction point, according

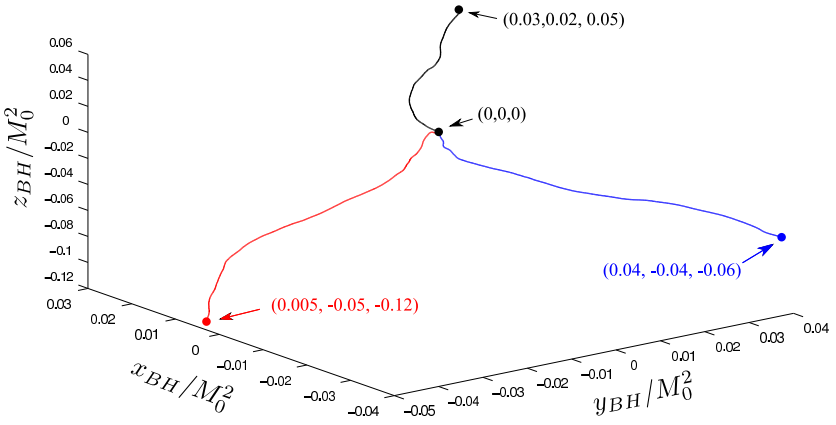


Figure 5.3: Typical paths of the black hole drifting in the three dimensional space $\mathbf{x}_{\rm BH} = (x_{\rm BH}, y_{\rm BH}, z_{\rm BH})$, normalized by M_0^2.

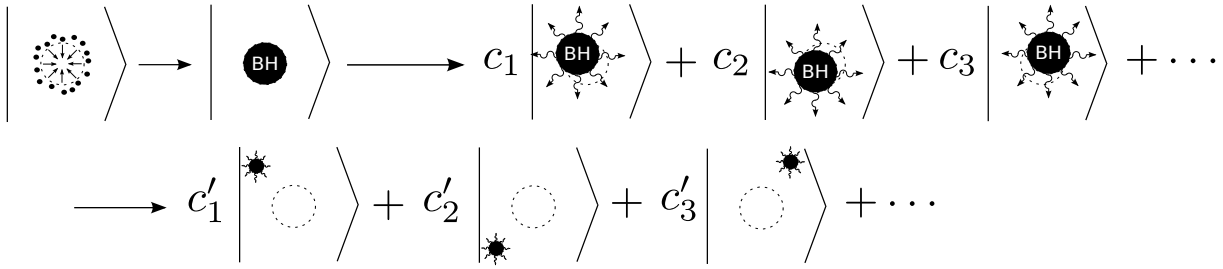


Figure 5.4: A schematic depiction of the evolution of a black hole state formed by a collapse of matter. After long time, the state will evolve into a superposition of terms representing the black hole to be in macroscopically different locations, even if the initial collapsing matter has a well-defined macroscopic configuration. The variation of the final locations in the evaporation timescale, $t \sim M_0^3$, is of order M_0^2, which is much larger than the Schwarzschild radius of the initial black hole, $R_S = 2M_0$.

to the distribution in Eq. (5.11); and we can certainly measure this because the wavelengths of these quanta are much smaller than M_0^2, and the interaction point is defined clearly with respect to, e.g., the beam pipe. An important point here, however, is that the superposition nature of the black hole state is physical *even if there is no physical object other than the black hole*, e.g., the beam pipe. This is because the location of an object with respect to the origin p of the reference frame is a physically meaningful quantity in the framework of Ref. [33]. In other words, the superposition nature discussed here is an *intrinsic* property of the black hole state, not one arising only in relation to other physical objects.

While relative values of the moduli of coefficients in front of terms representing different black hole locations, e.g. $|c_1/c_2|$ in Fig. 5.4, are determined by the statistical analysis leading to Eq. (5.11), their relative phases are unconstrained by the analysis. Moreover, it is possible that there are higher

order corrections to the moduli that are not determined by any semi-classical analysis. These quantities, therefore, can contain the information about the initial state; i.e., they can reflect the details of the initial configuration of matter that has collapsed into the black hole. (This actually *should* be the case because a particular initial state leads to particular values for the relative phases because the Schrödinger equation is deterministic.) Together with the relative coefficients of terms representing different phase space configurations of emitted Hawking quanta for each black hole location (more precisely, their parts that are not fixed by semi-classical analyses, e.g. the relative phases), these quantities must be able to reproduce the initial state of the evolution by solving the appropriate Schrödinger equation backward in time.

5.1.4 Spontaneous Spin-up of a Schwarzschild Black Hole

Just as a black hole accumulates momentum over its lifetime through randomly recoiling from Hawking emissions, we can ask if a black hole also accumulates angular momentum due to the spin and orbital angular momentum of emitted particles. In this section, we argue that the answer is yes: non-rotating black holes with initial mass M_0 spontaneously spin up to angular momentum $J \equiv |\mathbf{J}| \sim M_0$ at a time of order M_0^3. This implies that a Schwarzschild black hole evolves into a superposition of Kerr black holes with different values of $\mathbf{J}$, although the resulting angular momenta will be small enough, $J/M^2 \ll 1$, that the geometry of each term is still well approximated by the Schwarzschild one.

To begin with, let us consider how many Hawking quanta are emitted by the time at which an initial black hole loses some fixed fraction of its mass, e.g. the Page time at which the black hole mass becomes $M = M_0/\sqrt{2}$. The number of emitted quanta is

$$N \sim \frac{M_0}{T_H} \sim M_0^2, \tag{5.12}$$

where $T_H \sim 1/M_0$ is the Hawking temperature. If the emitted quanta consist of a particle with spin $s > 0$, then each emission changes the angular momentum of the black hole by $\Delta J \sim s$, depending on the direction of the spin. Assuming that the emission is unbiased in the direction of angular momentum (see below), we find that the black hole accumulates the angular momentum

$$J \sim s\sqrt{N} \sim M_0, \tag{5.13}$$

at a time of order M_0^3, where we have taken $s \sim O(1)$ in the last expression.

If the Hawking quanta consist of a scalar ($s = 0$), then most of the emissions do not affect the black hole angular momentum since the emissions are dominated by s-wave. However, there is a small probability that a quantum is emitted in a higher angular momentum mode. The probability is dominated by p-wave ($l = 1$), which can be calculated for small J/M^2 as $p \simeq 0.002 + O(J/M^2)$,

independent of M [71, 72]. Therefore, the number of Hawking quanta that affect the black hole angular momentum is pN, and the accumulated angular momentum of the black hole is

$$J \sim \sqrt{pN} \sim M_0, \tag{5.14}$$

which is parametrically the same as in the case of a particle with spin.

One might think that once the accumulated angular momentum becomes macroscopic, $J \gg 1$, the black hole becomes a Kerr black hole, so that there is a bias in the Hawking spectrum that preferentially selects emissions that reduce J [25], preventing a further accumulation of J. We now argue, however, that until the time $t \sim M_0^3$ when the mass of the black hole starts decreasing significantly, the evolution of $\mathbf{J}$ is well approximated by a random walk process as described above.

To see this, at a given time t, let us call the direction of $\mathbf{J}$ the z-axis. Suppose an emission of a particle with spin s changes $J = J_z$, which occurs with $O(p)$ and $O(1)$ probabilities for $s = 0$ and $s > 0$, respectively. For small J/M^2, the probability ρ_+ (ρ_-) that the emission increases (reduces) J is [72]:

$$\rho_\pm = \frac{1}{2} \mp c\frac{J}{M^2}, \tag{5.15}$$

where J and M are the magnitude of angular momentum and the mass before the emission takes place, and c is an $O(1)$ coefficient which depends on the type of a particle emitted and is independent of J and M to first order in J/M^2. Numerical simulations of this process indicate that this bias is not strong enough to prevent a black hole from spinning up to $J \sim M_0$ by the Page time, $t_{\rm Page}$. Results of these simulations are shown in Fig. 5.5, where we have assumed a change of J according to Eq. (5.15) in each time interval M_0. The results indicate that

$$J \sim f(c)M_0 \sim M_0 \tag{5.16}$$

at $t \sim t_{\rm Page}$, where f is a monotonically decreasing function of c; in fact, our simulations suggest that $f(c) \propto 1/\sqrt{c}$ for $c \gtrsim 1$.

The results obtained above can be understood by the following simple argument. Imagine that at some late time $t \gtrsim M_0^3$, the probability distribution of the black hole angular momentum reaches some "equilibrium" distribution $P(J)$, in which the random walk effect increasing J is balanced with the bias of the emission reducing J. According to Eq. (5.15), this implies

$$\rho_+ P(J) = \rho_- P(J+1), \tag{5.17}$$

leading to

$$\frac{P(J+1)}{P(J)} = \frac{1 - 2c\frac{J}{M^2}}{1 + 2c\frac{J+1}{M^2}} \approx 1 - 4c\frac{J}{M^2}. \tag{5.18}$$

Here, we have used $1 \ll J \ll M^2$ in the last expression. This has the solution

$$P(J) \sim e^{-2c\frac{J^2}{M^2}}. \tag{5.19}$$

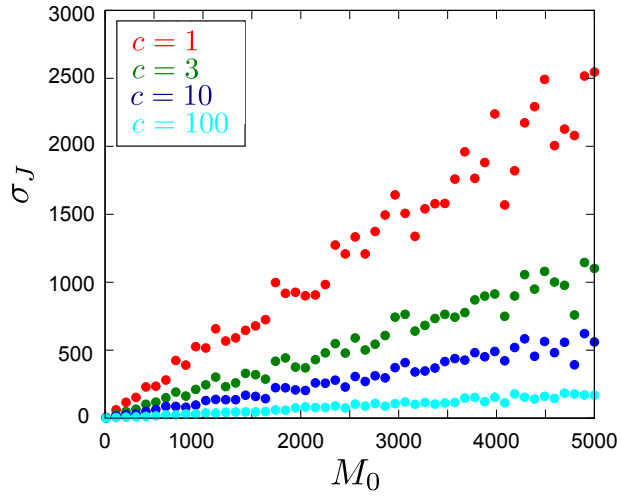


Figure 5.5: Plot of σ_J, the square root of the variance of J at the Page time, as a function of the initial mass M_0. Each data point represents σ_J obtained by 100 simulations. The different colors correspond to different values of coefficient c in Eq. (5.15), which measures the strength of the angular momentum emission bias from a Kerr black.

Namely, the black hole angular momentum has a characteristic size

$$J \sim \frac{1}{\sqrt{c}} M, \tag{5.20}$$

consistent with the result obtained in Eq. (5.16).

In summary, we conclude that a Schwarzschild black hole with initial mass M_0 will spontaneously spin up to $J \sim M_0$ by a timescale of order M_0^3. When the black hole mass starts decreasing significantly, its angular momentum will also start decreasing, following Eq. (5.20). The combination J/M^2 keeps increasing as $1/M$ but is still (much) smaller than 1, as long as $M \gg 1$ where our analysis is valid. What happens at the real end of the evaporation is unclear, but we can say that while the evolution of a Schwarzschild black hole leads to a superposition of Kerr black holes with distinct angular momenta, the probability of it becoming a macroscopic extremal black hole ($J = M^2 \gg 1$) is, most likely, exponentially suppressed.

5.1.5 Evolution in the covariant Hilbert space for quantum gravity

Let us now formulate more precisely how the black hole state, formed by a collapse of matter, evolves in the covariant Hilbert space for quantum gravity, Eq. (4.4). Recall that a Hilbert subspace $\mathcal{H}_\mathcal{M}$ in Eq. (4.4) corresponds to the states realized on a fixed semi-classical three-geometry $\mathcal{M}$ (more precisely, a set of three-geometries $\mathcal{M} = \{\mathcal{M}_i\}$ having the same boundary $\partial\mathcal{M}$). In our context, the relevant $\mathcal{M}$'s for spacetime with the black hole are specified by the location of the black hole

$\mathbf{x}_{\rm BH}$ (which can be parameterized, e.g., by the direction $\{\theta, \phi\}$ and the affine length λ of the past-directed light ray connecting reference point p to the closest point on the stretched horizon) and the size of the black hole (which can be parameterized, e.g., by its mass M or area $\mathcal{A} = 16\pi M^2$). Here and below, we ignore the angular momentum of the black hole, for simplicity. We also need to consider the Hilbert subspace corresponding spacetime without the black hole, $\mathcal{H}_0$.

The part of $\mathcal{H}_{\rm QG}$ relevant to our problem here is then

$$\mathcal{H} = \left(\bigoplus_{\mathbf{x}_{\rm BH},\, M} \mathcal{H}_{\mathbf{x}_{\rm BH},\, M} \right) \oplus \mathcal{H}_0, \tag{5.21}$$

where $0 < M \leq M_0$, and we have used the notation in which $\mathbf{x}_{\rm BH}$ and M are discretized. The Hilbert subspace $\mathcal{H}_{\mathbf{x}_{\rm BH},\, M}$ consists of the factor associated with the black hole horizon $\mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\, M}$ and that with the rest $\mathcal{H}^{\rm bulk}_{\mathbf{x}_{\rm BH},\, M}$ (which represents the region outside the horizon):

$$\mathcal{H}_{\mathbf{x}_{\rm BH},\, M} = \mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\, M} \otimes \mathcal{H}^{\rm bulk}_{\mathbf{x}_{\rm BH},\, M}. \tag{5.22}$$

According to the Bekenstein-Hawking entropy, the size of the horizon factor is given by

$$\dim \mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\, M} = e^{\frac{\mathcal{A}_{\rm BH}}{4}} = e^{4\pi M^2}, \tag{5.23}$$

regardless of $\mathbf{x}_{\rm BH}$. Because of this, Hilbert space factors $\mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\, M}$ for different $\mathbf{x}_{\rm BH}$ are all isomorphic with each other, which allows us to view $\mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\, M}$ for any fixed $\mathbf{x}_{\rm BH}$ as the intrinsic structure of the black hole.

Now, right after the formation of the black hole, which we assume to have happened at $\mathbf{x}_{\rm BH,0}$ at $t = 0$, the system is in a state that is an element of $\mathcal{H}_{\mathbf{x}_{\rm BH,0},\, M_0}$. In the case of Eq. (6.21)

$$|\Psi(0)\rangle \equiv |{\rm BH}^0_A\rangle \in \mathcal{H}_{\mathbf{x}_{\rm BH,0},\, M_0}. \tag{5.24}$$

This state then evolves into a superposition of states in different $\mathcal{H}_{\mathcal{M}}$'s.[2] At time t, the state of the system can be written as

$$|\Psi(t)\rangle = \sum_{\mathbf{x}_{\rm BH}} \alpha^t_{\mathbf{x}_{\rm BH}} |\phi^t_{\mathbf{x}_{\rm BH}}\rangle, \tag{5.25}$$

where $|\phi^t_{\mathbf{x}_{\rm BH}}\rangle \in \mathcal{H}_{\mathbf{x}_{\rm BH},\, M(t)}$, and we have ignored possible fluctuations of the black hole mass at a fixed time t, for simplicity. (Including this effect is straightforward; we simply have to add terms corresponding to $\mathcal{H}_{\mathbf{x}_{\rm BH},\, M}$ with $M \neq M(t)$.) The state $|\phi^t_{\mathbf{x}_{\rm BH}}\rangle$ contains the horizon *and* other degrees of freedom, according to Eq. (5.22). We can expand it in some basis in $\mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\, M(t)}$ (e.g. the one spanned by states having well-defined numbers of Hawking quanta emitted afterward) or in some

[2] This is precisely analogous to the case of e^+e^- scattering, in which the initial state $|e^+e^-\rangle \in \mathcal{H}_2$ evolves into a superposition of states in different $\mathcal{H}_n$'s, e.g. $|e^+e^-\rangle \to c_e|e^+e^-\rangle + \cdots + c_{ee}|e^+e^-e^+e^-\rangle + \cdots$, where $\mathcal{H}_n$ is the n-particle subspace of the entire Fock space: $\mathcal{H} = \bigoplus_n \mathcal{H}_n$.

basis in $\mathcal{H}^{\rm bulk}_{\mathbf{x}_{\rm BH},\,M(t)}$ (e.g. the one spanned by states having well-defined phase space configurations of already emitted Hawking quanta). In either case, it takes the form

$$|\phi^t_{\mathbf{x}_{\rm BH}}\rangle = \sum_n \beta^t_n |{\rm BH}^t_{\mathbf{x}_{\rm BH},n}\rangle \otimes |\psi^t_{\mathbf{x}_{\rm BH},n}\rangle, \tag{5.26}$$

where $|{\rm BH}^t_{\mathbf{x}_{\rm BH},n}\rangle \in \mathcal{H}^{\rm horizon}_{\mathbf{x}_{\rm BH},\,M(t)}$ and $|\psi^t_{\mathbf{x}_{\rm BH},n}\rangle \in \mathcal{H}^{\rm bulk}_{\mathbf{x}_{\rm BH},\,M(t)}$. Plugging this into Eq. (5.25) and defining

$$a^t_i \equiv \alpha^t_{\mathbf{x}_{\rm BH}} \beta^t_n, \tag{5.27}$$

where $i \equiv \{\mathbf{x}_{\rm BH}, n\}$, we reproduce the third expression in Eq. (6.21). In this formulation, the statement that the black hole state is a superposition of macroscopically different geometries refers to the fact that coefficients $|\alpha^t_{\mathbf{x}_{\rm BH}}|$ have a significant support in a wide range of $\mathbf{x}_{\rm BH}$ extending beyond the original Schwarzschild radius M_0.

5.1.6 What does a physical observer actually see?

We have found that a late black hole state is far from a semi-classical state in which the spacetime has a fixed geometry; rather, it involves a superposition of macroscopically different geometries. Does this mean that a physical observer sees something very different from what the usual picture based on general relativity predicts?

The answer is no. To understand this, let us consider a physical observer watching the evaporation process from a distance by measuring (all or parts of) the emitted Hawking quanta. For simplicity, we consider that he/she does that using usual measuring devices, e.g. by locating photomultipliers around the black hole from which he/she collects the data. This leads to an entanglement between the system and the observer (or his/her brain states). And because the interactions leading to it are local, the observer is entangled with the basis in $\mathcal{H}^{\rm bulk}_{\mathbf{x}_{\rm BH},\,M}$ spanned by the states that have well-defined phase space configurations of emitted Hawking quanta (within the errors dictated by the uncertainty principle) and well-defined locations for the black hole (since the black hole location can be inferred from the momenta of the Hawking quanta) [33]. Namely, the combined state of the black hole and the observer evolves as

$$|{\rm BH}^0_A\rangle \otimes \big|\ \dot{}\ \big\rangle \longrightarrow \sum_{\mathbf{x}_{\rm BH},\,n} a^t_{\mathbf{x}_{\rm BH},n} |{\rm BH}^t_{\mathbf{x}_{\rm BH},n}\rangle \otimes |\psi^t_{\mathbf{x}_{\rm BH},n}\rangle \otimes \big|\ \dot{}\ \ {}^{\mathbf{x}_{\rm BH}}_{n}\big\rangle, \tag{5.28}$$

where $|\psi^t_{\mathbf{x}_{\rm BH},n}\rangle$ represents the state in which the black hole is in a well-defined location $\mathbf{x}_{\rm BH}$ and Hawking quanta have a well-defined phase space configuration n. The last factor in the right-hand side implies that the observer recognized that the black hole is at $\mathbf{x}_{\rm BH}$ and the configuration of emitted Hawking quanta is n.

Since terms in the right-hand side of Eq. (5.28) have macroscopically different configurations, e.g. the brain state of the observer differs, their mutual overlaps are exponentially suppressed

(e.g. by $\sim \prod_{i=1}^{N} \epsilon_i$, where $\epsilon_i < 1$ is the overlap of each atom and N the total number of atoms). The observer in each term (or branch), therefore, sees his/her own universe; i.e., the interferences between different terms are negligible. For any of these observers, the behavior of the black hole is controlled by semi-classical physics (but with the backreaction of the emission taken into account). For instance, they all see that the black hole keeps emitting Hawking quanta consistent with the thermal spectrum with temperature $T_H(t) = 1/8\pi M(t)$, and that it drifts in a fixed direction as a result of backreactions, eventually evaporating at a location $\sim M_0^2$ away from that of the formation. A single observer cannot predict the direction to which the black hole will drift, reflecting the fact that the entire state is a superposition of terms having different $(\mathbf{x}_{\rm BH} - \mathbf{x}_{\rm BH,0})/|(\mathbf{x}_{\rm BH} - \mathbf{x}_{\rm BH,0})|$, but all these observers find a set of common properties for the black hole, including the relation between T_H and M.[3]

It is these "intrinsic properties" of the black hole that the semi-classical gravity on a fixed Schwarzschild geometry (in which the black hole is located at the "center") really describes. A physical observer watching the evolution does not see anything contradicting what is implied by the semi-classical analysis about these intrinsic properties. This is true despite the fact that the full quantum state obtained by evolving collapsing matter that initially had a well-defined configuration takes the form in Eqs. (5.25, 5.26), which involves a superposition of macroscopically different geometries and is very different at late times from a "semi-classical state" having a fixed geometry.

5.1.7 Can a physical observer recover the information?

The black hole evaporation process is often compared with burning a book in classical physics: if we measure all the details of the emitted Hawking quanta, we can recover the initial state from these data by solving the Schrödinger equation backward in time. Is this correct?

It is true that if we know the coefficients of all the terms in a state when it is expanded in a fixed basis, e.g. $\{a_i^\infty\}$ in Eq. (6.21), then unitarity must allow us to recover the initial state unambiguously. However, a physical observer measuring Hawking radiation from black hole evaporation can *never* obtain the complete information about these coefficients, even if he/she measures *all* the radiation quanta. In the state in Eq. (5.28), for example, a physical observer "lives" in *one* of the terms in the right-hand side and, therefore, cannot have the information about the coefficients of the other terms. The other terms are already decohered—or "decoupled"—so that they are other worlds/universes for the observer.

In fact, the situation is exactly the same in usual scattering experiments. Consider two initial

[3]More precisely, there are rare observers who find deviations from these relations, but the probability for that to happen is exponentially suppressed.

states $|e^+e^-\rangle$ and $|\mu^+\mu^-\rangle$ with the same $\sqrt{s}$ ($> 2m_\tau$) and angular momentum. They evolve as

$$|e^+e^-\rangle \longrightarrow a_1|e^+e^-\rangle + a_2|\mu^+\mu^-\rangle + a_3|\tau^+\tau^-\rangle + \cdots , \tag{5.29}$$

$$|\mu^+\mu^-\rangle \longrightarrow b_1|e^+e^-\rangle + b_2|\mu^+\mu^-\rangle + b_3|\tau^+\tau^-\rangle + \cdots , \tag{5.30}$$

where we have ignored the momenta and spins of the final particles. The information about an initial state is in the *complete* set of coefficients in the final superposition state; i.e., if we know the entire $\{a_i\}$ (or $\{b_i\}$), then we can recover the initial state by solving the evolution equation backward. However, if a physical observer measures a final state, e.g., as $\tau^+\tau^-$, how can he/she know that it has arisen from e^+e^- or $\mu^+\mu^-$ scattering? In general, if an observer measures the final outcome of a process, he/she will be entangled with one of the terms in the final state (in the above case, $|\tau^+\tau^-\rangle$), so there is no way that he/she can learn all the coefficients in the final state.

The situation does not change even if the observer uses a carefully-crafted quantum device which, upon interacting with the radiation, is entangled *not* with a well-defined phase space configuration of the radiation quanta but with a macroscopic superposition of those configurations. In this case, the basis of the final state to which the observer is entangled may be changed, but it still cannot change the fact that he/she will be entangled with *one* of the terms in the final state, i.e., he/she will measure *a* possible outcome among all the possibilities.

Therefore, in quantum mechanics, an observer can never recover the initial state by observing the final state. The statement that the final state of an evolution contains all the information about the initial state is *not* the same as the statement that a physical observer measuring the final state can recover the initial state if he/she measures a system with high enough (or even infinite) precision. The only way that an observer can test the relation between the initial and final states is to *create* the same initial state many times and perform multiple (including quantum) measurements on the final states. (Note that creating many initial states in this context differs from producing a copy of a generic unknown state, which is prohibited by the quantum no-cloning theorem [67].) A single system does not allow for doing this, no matter how high the precision of the measurement is, and no matter how clever the measurement device is.

5.2 Complementarity as a Reference Frame Change

So far, we have been describing the formation and evaporation of a black hole from a distant reference frame. In this reference frame, the complete description of the process is obtained in the spacetime region outside and on the (time-like) stretched horizon, where intrinsically quantum gravitational—presumably stringy (such as fuzzball [63])—effects become important. What then is the significance of the interior of the black hole horizon, where we expect to have regular low-curvature spacetime according to general relativity?

As discussed in Ref. [33], and implied by the original complementarity picture [3], a description of the internal spacetime is obtained (only) after changing the reference frame. An important point is that the reference frame change is represented as a unitary transformation acting on a quantum state, so if we want to discuss the precise mapping between the pictures based on different reference frames, then we need to keep all the terms in the state. In this section, we carefully study issues associated with the reference frame changes, especially in describing an old black hole.

5.2.1 Describing the black hole interior

Suppose collapsing matter, which initially had a well-defined classical configuration, forms a black hole, which then eventually evaporates. In a distant reference frame, this process is described as in Eq. (6.21), which we denote by $|\Psi(t)\rangle$. How does the process look from a different reference frame?

Since a reference frame can be any freely falling (locally Lorentz) frame, the new description can be obtained by performing a translation, rotation, or boost on a quantum state at fixed t [33]. In general, the state on which these transformations act, however, contains the horizon degrees of freedom as well as the bulk ones. How do they transform under the transformations?

We do not know the microscopic description of the horizon degrees of freedom or their explicit transformations under the reference frame changes. Nevertheless, we can know which spacetime regions are transformed to which horizon degrees of freedom, and vice versa, by *assuming* that the global spacetime picture in semi-classical gravity is consistent with the one obtained by a succession of these reference frame changes. Here we phrase this in the form of a hypothesis:

> **Complementarity Hypothesis:** The transformation laws of a quantum state under the reference frame changes are consistent with those obtained in the global spacetime picture based on general relativity. In particular, the transformation laws between the horizon and bulk degrees of freedom are constrained by this requirement.

As discussed in Refs. [33, 32], this hypothesis is fully consistent with the holographic principle formulated in the form of the covariant entropy conjecture [9]. Specifically, the dimension of the Hilbert space representing horizon degrees of freedom and that representing the corresponding spacetime region before (or after) a transformation are the same for general spacetimes, including the cosmological ones, as it should be. Alternatively, we can take a view that if we require that the above hypothesis is true in the covariant Hilbert space $\mathcal{H}_{\rm QG}$, then the covariant entropy conjecture is obtained as a consequence.

Let us now consider a reference frame change induced by a boost performed at some early time $t_{\rm boost} < 0$ (before the black hole forms at $t = 0$) in such a way that the origin p of the reference frame enters the black hole horizon at some late time $t_{\rm enter} > 0$. In this subsection, we focus on

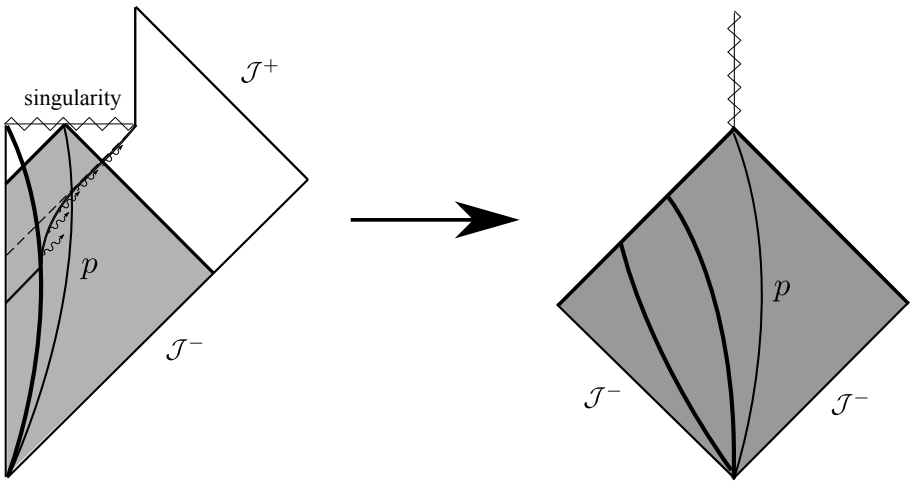


Figure 5.6: The left panel shows the standard global spacetime picture for the formation and evaporation of a black hole, with the shaded region representing the spacetime region described by an infalling reference frame. (The trajectory of the origin, p, of the reference frame is also depicted.) As discussed in the text, this is the *entire* spacetime when the system is described in this reference frame, so its conformal structure is in fact as in the right panel. Here, the wavy line with a solid core represents singularity states.

the case

$$t_{\rm enter} \ll M_0^{7/3}, \tag{5.31}$$

so that the uncertainty of the black hole location at the time when p enters the horizon is negligible, and we ignore the (exponentially) small probability that p misses the horizon. (The possibility of p missing the horizon becomes important when we discuss the description of the interior of an older black hole.)

Recall that quantum states in the present framework represent physical configurations on the past light cone of p (in and on the apparent horizon) when they allow for spacetime interpretation, i.e. when the curvature at p is smaller than the Planck scale. Therefore, the spacetime region represented by the state of the system after the reference frame change

$$|\Psi'(t)\rangle = e^{-iH(t-t_{\rm boost})} U_{\rm boost} e^{iH(t-t_{\rm boost})} |\Psi(t)\rangle, \tag{5.32}$$

where $U_{\rm boost}$ is the boost operator represented in $\mathcal{H}_{\rm QG}$, corresponds to the shaded region in the left panel of Fig. 5.6. Specifically, $|\Psi'(t)\rangle$ at

$$t < t_{\rm enter} + t_{\rm fall} \tag{5.33}$$

describes this region, with $t_{\rm fall} \approx O(M_0)$ being the time needed for p to reach the singularity after it passes the horizon. After $t = t_{\rm enter} + t_{\rm fall}$, the state evolves in the Hilbert subspace $\mathcal{H}_{\rm sing}$, which consists of states that are associated with spacetime singularities and thus do not allow for

spacetime interpretation. The detailed properties of these "intrinsically quantum gravitational" states are unknown, except that $\dim \mathcal{H}_{\rm sing} = \infty$, implying that generic singularity states do not evolve back to the usual spacetime states [33].

In the right panel of Fig. 5.6, we depict the causal structure of the spacetime as viewed from the new reference frame. Because of the lack of the spherical symmetry, we have depicted the region swept by two past-directed light rays emitted from p in the opposite directions (while in the left panel we have depicted only the region with fixed angular variables with respect to the center of mass of the system). The singularity states are represented by the wavy line with a solid core at the top. Note that, as in the case of the description in the distant reference frame (depicted in Fig. 5.1), this is the *entire* spacetime region when the system is described in this infalling reference frame—the non-shaded region in the left panel simply does not exist. (Including the non-shaded region, indeed, is overcounting as indicated by the standard argument of information cloning in black hole physics.) A part of the non-shaded region appears if we change the reference frame, but only at the cost of losing some of the shaded region. The global spacetime picture in the left panel appears only if we "patch" the views from different reference frames, which, however, grossly overcounts the correct quantum degrees of freedom.

There are two comments. First, the reference frame change considered here is (obviously) only *a* reference frame change among possible (continuously many) reference frame changes, all of which lead to different descriptions of the same physical process. Second, a unitary transformation representing this reference frame change,

$$U(t) = e^{-iH(t-t_{\rm boost})} U_{\rm boost} e^{iH(t-t_{\rm boost})}, \tag{5.34}$$

does not close in the Hilbert space $\mathcal{H}$ in Eq. (5.21), although it closes in the whole covariant Hilbert space $\mathcal{H}_{\rm QG}$. Before the reference frame change, the evolution of the state is given by a trajectory in $\mathcal{H} = (\oplus_{\mathbf{x}_{\rm BH},\, M} \mathcal{H}_{\mathbf{x}_{\rm BH},\, M}) \oplus \mathcal{H}_0$. The action of $U(t)$ maps this into a trajectory in

$$\mathcal{H}' = \mathcal{H}_0 \oplus \mathcal{H}_{\rm sing}, \tag{5.35}$$

with $|\Psi'(-\infty)\rangle \in \mathcal{H}_0$ and $|\Psi'(+\infty)\rangle \in \mathcal{H}_{\rm sing}$.[4] As a result, in this new reference frame, the evolution of the system does not allow for an S-matrix description in $\mathcal{H}_0$ (or $\mathcal{H}_{\rm Minkowski}$), although it still allows for an "S-matrix" description in the whole $\mathcal{H}_{\rm QG}$ (or in $\mathcal{H}_{\rm Minkowski} \oplus \mathcal{H}_{\rm sing}$), which contains the singularity states in $\mathcal{H}_{\rm sing}$.

5.2.2 Complementarity for an old black hole

Let us now try to describe the interior of an older black hole, specifically the spacetime inside the black hole horizon after a time $> O(M_0^{7/3})$ is passed since the formation. To do this, we can

[4] Note that $\mathcal{H}_0$ contains a set of states that represent three-geometries whose *boundary* (at an infinity) is that of the flat space, i.e. a two-dimensional section of $\mathcal{J}^-$.

consider performing a boost at time $t_{\rm boost} < 0$ on $|\Psi(t)\rangle$ in such a way that p enters the black hole horizon at time $t_{\rm enter} \gg M_0^{7/3}$. What does the resultant state $|\Psi'(t)\rangle$ look like?

As discussed in the previous subsection, this can be done by applying an operator of the form of Eq. (5.34) on $|\Psi(t)\rangle$, where $U_{\rm boost}$ now represents a different amount of boost than the one considered before. In general, the relation between the states before and after a reference frame change is highly nontrivial. For example, time t is measured by the proper time at p, but relations between the proper times of the two frames depend on the geometries as well as the paths of p therein. Therefore, various terms in $|\Psi'(t)\rangle$ for a fixed t may correspond to terms in $|\Psi(t)\rangle$ of different t's. Without knowing the explicit form of H and $U_{\rm boost}$ represented in the whole $\mathcal{H}_{\rm QG}$, which includes the horizon degrees of freedom, how can we know the form of the state after the transformation?

According to our complementarity hypothesis, the probability of finding a certain history for the evolution of geometry must agree in the two pictures before and after the reference frame change if the geometries are appropriately transformed, i.e. according to the global spacetime picture in general relativity. To elucidate this, let us consider the black hole evolution described in Eqs. (5.25, 5.26) in a distant reference frame, and ask what is the probability that the black hole follows a particular path $\mathbf{r}(t)$ in a time interval between t_I and t_F within the error $|\Delta\mathbf{r}| < \epsilon(t)$. For simplicity, we do this by requiring that the black hole satisfies the above conditions at discretized times t_i; $i = 0, \cdots, N$ ($\gg 1$), with $t_0 \equiv t_I$ and $t_N \equiv t_F$. The probability is then given by

$$P = \prod_{i=0}^{N} \left(\sum_{|\Delta\mathbf{r}|<\epsilon_i} |\alpha^{t_i}_{\mathbf{r}_i+\Delta\mathbf{r}}|^2 \right), \tag{5.36}$$

where $\mathbf{r}_i \equiv \mathbf{r}(t_i)$ and $\epsilon_i \equiv \epsilon(t_i)$. This provides the probability of a particular semi-classical history to appear, given the state $|\Psi(t)\rangle$. We can now ask a similar question for the state $|\Psi'(t)\rangle$: what is the probability of having the black hole to follow the trajectory $\mathbf{r}'(t)$ between t'_I and t'_F within the error $\epsilon'(t)$? The resulting probability is

$$P' = \prod_{i=0}^{N} \left(\sum_{|\Delta\mathbf{r}'|<\epsilon'_i} |\alpha^{t_i}_{\mathbf{r}'_i+\Delta\mathbf{r}'}|^2 \right), \tag{5.37}$$

where $t_0 = t'_I$ and $t_N = t'_F$. The complementarity hypothesis in the previous subsection asserts that the two probabilities are the same

$$P = P', \tag{5.38}$$

if the relation between $\{\mathbf{r}(t), \epsilon(t), t_I, t_F\}$ and $\{\mathbf{r}'(t), \epsilon'(t), t'_I, t'_F\}$ is the one obtained by performing the corresponding transformation in general relativity on the semi-classical background selected by Eq. (5.36).

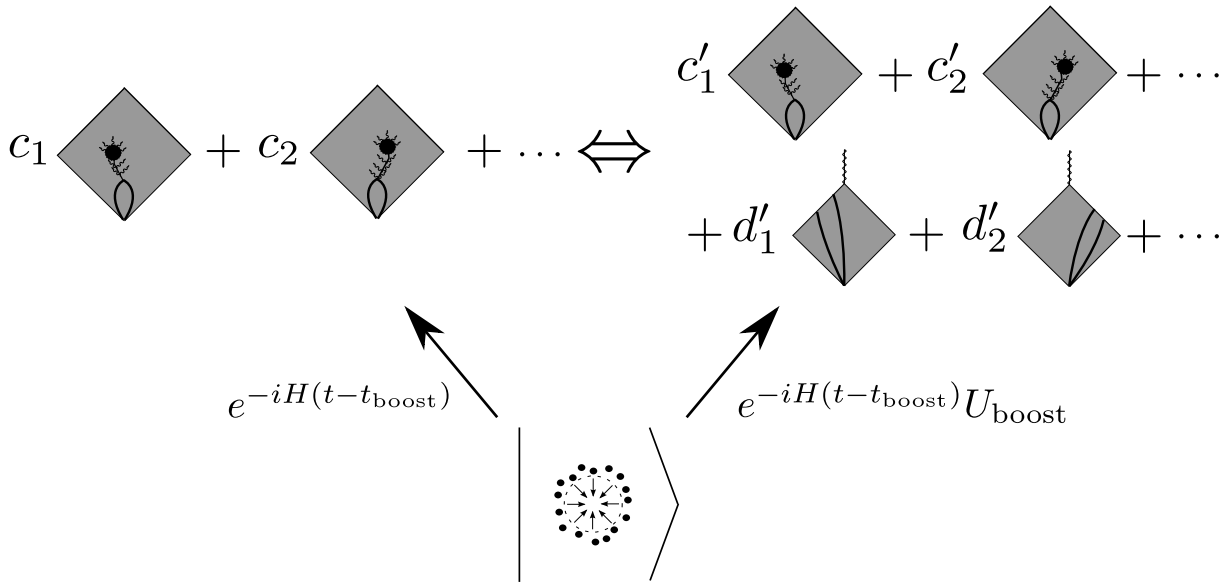


Figure 5.7: A schematic picture of a relation between the two descriptions based on two different reference frames of an old black hole formed by collapsing matter that initially had a well-defined classical configuration. In one reference frame, the black hole is viewed from outside, and the state becomes a superposition of black holes in different locations at late times (depicted schematically in the left-hand side). In the other reference frame, obtained by acting $U_{\rm boost}$ on the state at $t_{\rm boost}$, the reference point p enters the black hole horizon at late time $t_{\rm enter} \gg M_0^{7/3}$, allowing for a description of internal spacetime (in the right-hand side). This, however, happens only for some of the terms, depicted in the second line, since p misses the horizon in most of the terms because of the large uncertainty of the black hole location, i.e. $\sum_i |d_i'|^2 \ll \sum_i |c_i'|^2$.

The above analysis implies that when we perform a boost on $|\Psi(t)\rangle$ at an early time $t_{\rm boost} < 0$, trying to describe the interior of an old black hole with $t_{\rm enter} \gg M_0^{7/3}$, then the resultant state can only be a *superposition* of infalling and distant descriptions of the process, since in most of the semi-classical histories represented by $|\Psi(t)\rangle$, the trajectory of p obtained by the boost will miss the black hole horizon because of the large uncertainty of the black hole location. Namely, complementarity obtained by this reference frame change is the one between the distant description and the superposition of the infalling and distant descriptions specified by the state $|\Psi'(t)\rangle$. This is illustrated schematically in Fig. 5.7.

Is it possible to obtain a direct correspondence between the interior and exterior of an old black hole, without involving a superposition? This can be done if we focus only on a term in $|\Psi(t)\rangle$ in which p just misses the black hole horizon, with the smallest distance between p and the horizon achieved at some time $t_{\rm min} \gg M_0^{7/3}$. We can then evolve this term slightly backward in time, to $t_{\rm boost} = t_{\rm min} - \epsilon$ ($\epsilon \ll M_0^{7/3}$), and perform a boost there so that p enters into the horizon at some time after $t_{\rm boost}$. In this way, the correspondence between the terms representing the interior

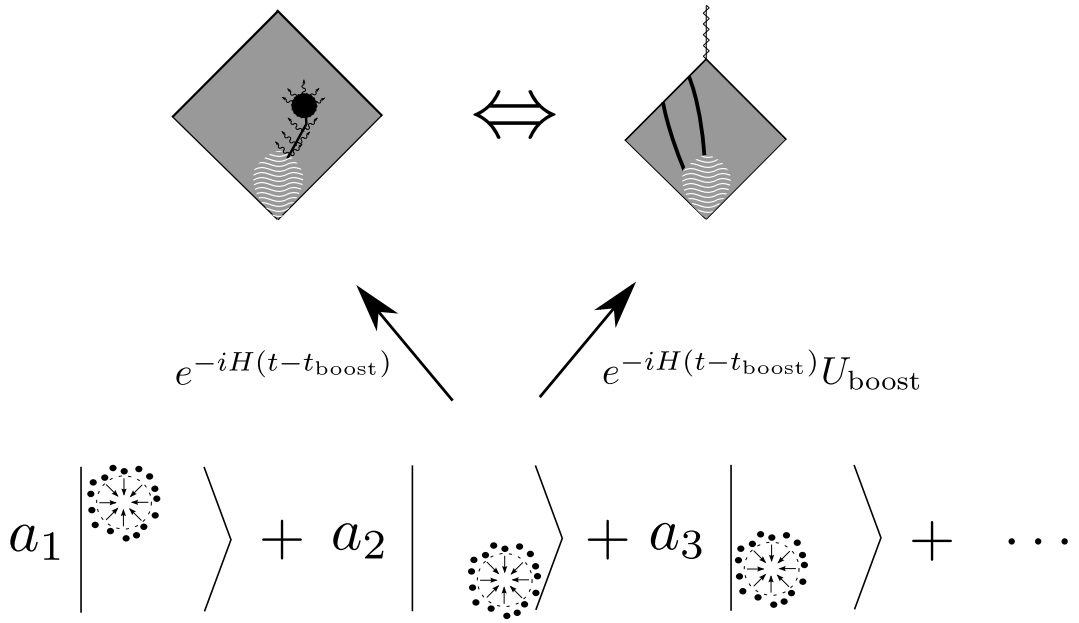


Figure 5.8: A complementarity relation between the internal and external descriptions of an old black hole can be obtained if we consider a state in which the black hole has a well-defined semi-classical configuration at late time $t_{\rm enter}$. Such a state, however, can arise through evolution only if we consider a special initial state in which coefficients a_i of the terms representing well-defined configurations of collapsing matter are finely-adjusted so that the state represents a black hole in a well-defined location at late time $\sim t_{\rm enter}$. (The regions with wavy white lines indicate superpositions of classical geometries.)

and exterior can be obtained. An important point, however, is that neither of these terms can be obtained by evolving initial collapsing matter that had a well-defined classical configuration (which would lead to a superposition of the black hole in vastly different locations). Rather, by evolving the state further back beyond $t_{\rm boost}$, we would obtain a superposition of states each of which represents collapsing matter with a well-defined classical configuration. (This state would have finely-adjusted coefficients so that after evolving to $t_{\rm boost} \sim t_{\rm min}$, the black hole is in a well-defined location with respect to p.) This situation is illustrated in Fig. 5.8.

The discussion above implies that there is no well-defined complementarity map between the interior and exterior of an old black hole throughout the course of the black hole evolution within the purely semi-classical picture. Such a map must involve a superposition of semi-classical geometries at some point in the evolution. We note that while the state in the intermediate stage of the evolution can be a superposition of elements in $\mathcal{H}_0$, $\mathcal{H}_{\mathbf{x}_{\rm BH},M}$, and $\mathcal{H}_{\rm sing}$, it becomes a superposition of elements in $\mathcal{H}_0$ and $\mathcal{H}_{\rm sing}$ at $t \to \infty$. Therefore, the "S-matrix" description discussed in the previous subsection is still available in this case in the Hilbert space of $\mathcal{H}_{\rm Minkowski} \oplus \mathcal{H}_{\rm sing}$.

Chapter 6

A Frame-Dependent Model for Microscopic Black Hole Evolution

We now turn to the hardest problem discussed in this thesis: the black hole information puzzle. As emphasized in previous chapters, it has become increasingly apparent that the concept of spacetime must receive substantial revisions when it is treated in a fully quantum mechanical manner. Consider describing a process in which an object falls into a black hole, which eventually evaporates, from the viewpoint of a distant observer. Unitarity of quantum mechanics suggests that the information content of the object will first be stored in the black hole system, and then emitted back to distant space in the form of Hawking radiation [62]. On the other hand, the equivalence principle implies that the object should not find anything special at the horizon, when the process is described by an observer falling with the object. These two pictures lead to inconsistency if we adopt the standard formulation of quantum field theory on curved spacetime, since it allows us to employ a class of equal time hypersurfaces (called nice slices) that pass through both the fallen object and late Hawking radiation, leading to violation of the no-cloning theorem of quantum mechanics [67].

Black hole complementarity [3] was a suggested to avoid this difficulty: the apparent cloning of the information occurring in black hole physics implies that the internal spacetime and horizon/Hawking radiation degrees of freedom appearing in different, i.e. infalling and distant, descriptions are not independent. This signals a breakdown of the naive global spacetime picture of general relativity at the quantum level, and it forces us to develop a new view of how classical spacetime arises in the full theory of quantum gravity. One of the main purposes of this thesis is to present a coherent picture of this issue. We discuss how a series of well-motivated hypotheses leads to an apparently consistent view of the effective emergence of global spacetime from a fundamental theory of quantum gravity. In particular, we elucidate how this picture avoids the recently raised firewall paradox [4, 73, 74], which can be viewed as a refined version of the old information paradox [55]. Our analysis provides a concrete answer to how the information can be preserved at

the quantum level in the black hole formation and evaporation processes.

A key element in developing our picture is to identify the origin and nature of the "entropy of spacetime," first discovered by Bekenstein and Hawking in studying black hole physics [21, 25]. In a previous work [5], it was argued that this entropy—the Bekenstein-Hawking entropy—is associated with the degrees of freedom that are coarse-grained *to obtain* the semiclassical description of the system: quantum theory of matter and radiation on a fixed spacetime background. This picture is consonant with the fact that in quantum mechanics, having a well-defined geometry of spacetime, e.g. a black hole in a well-defined spacetime location, requires taking a superposition of an enormous number of energy-momentum eigenstates, so we expect that there are many different ways to arrive at the same background for the semiclassical theory within the precision allowed by quantum mechanics. This implies that, when a system with a black hole is described in a distant reference frame, the information about the microstate of the black hole is *delocalized* over a large spatial region, since it is encoded globally in the way of taking the energy-momentum superposition to arrive at the geometry under consideration. In particular, we may naturally identify the spatial distribution of this information as that of the gravitational thermal entropy calculated using the semiclassical theory. This leads to a fascinating picture: the degrees of freedom represented by the Bekenstein-Hawking entropy play dual roles of spacetime and matter—they represent how the semiclassical geometry is obtained at the microscopic level and at the same time can be viewed as the origin of the thermal entropy, which is traditionally associated with thermal radiation in the semiclassical theory.

The delocalization of the microscopic information described above plays an important role in addressing the firewall/information paradox. As described in a distant reference frame, a general black hole state is specified by the following three classes of indices at the microscopic level:

- Indices labeling the (field or string theoretic) degrees of freedom in the exterior spacetime region, excited over the vacuum of the semiclassical theory;[1]
- Indices labeling the excitations of the stretched horizon;[2]
- Indices representing the degrees of freedom that are coarse-grained to obtain the semiclassical description, which we will collectively denote by k. The information in k represents how the black hole geometry is obtained at the microscopic level, and cannot be resolved by semiclassical operators. It is regarded as being delocalized following the spatial distribution of the gravitational thermal entropy, calculated using the semiclassical theory.

[1] Note that the concepts of the breakdown of a semiclassical description and that of semiclassical *field* theory are not the same—there can be phase space regions in which an object can be well described as a string (or brane) propagating in spacetime, but not as a particle.

[2] The stretched horizon is located at a microscopic distance outside of the mathematical horizon, and is regarded as a physical (timelike) membrane which may be physically excited [3].

In a distant reference frame, an object falling into the black hole is initially described by the first class of indices, and then by the second when it hits the stretched horizon. The information about the fallen object will then reside there for, at least, time of order $Ml_{\rm P}^2 \ln(Ml_{\rm P})$ (the scrambling time [75]), after which it will be transmitted to the index k. Here, M and $l_{\rm P}$ are the mass of the black hole and the Planck length, respectively. Finally, the information in k, which is delocalized in the whole zone region, will leave the black hole system through the Hawking emission, or black hole mining, process.

Since the microscopic information about the black hole is considered to be delocalized from the semiclassical standpoint, the Hawking emission, or black hole mining, process can be viewed as occurring at a *macroscopic* distance away from the stretched horizon without contradicting information conservation. In this region, degrees of freedom represented by the index k are converted into modes that have clear identities as semiclassical excitations, i.e. matter or radiation, above the spacetime background. This conversion process, i.e. the emission of Hawking quanta or the excitation of a mining apparatus, is accompanied by the appearance of negative energy excitations, which have *negative entropies* and propagate inward to the stretched horizon. As we will see, the microscopic dynamics of quantum gravity allows these processes to occur unitarily without violating causality among events described in low energy quantum field theory. This picture avoids firewalls as well as information cloning.

In the description based on a distant reference frame, a falling object can be described by the semiclassical theory only until it hits the stretched horizon, after which it goes outside the applicability domain of the theory. We may, however, describe the fate of the object using the semiclassical language somewhat longer by performing a *reference frame change*, specifically until the object hits a singularity, after which there is no reference frame that admits a semiclassical description of the object. This reference frame change is the heart of complementarity: the emergence of global spacetime in the classical limit. We argue that while descriptions in different reference frames (the descriptions before and after a complementarity transformation) apparently look very different, e.g. in locations of the degrees of freedom representing the microscopic information of the black hole, their predictions about the same physical question are consistent with each other. This consistency is ensured by an intricate interplay between the properties of microscopic information and the causal structure of spacetime.

It is striking that the concept of spacetime, e.g. the region in which a semiclassical description is applicable, depends on a reference frame. This extreme "relativeness" of the description is a result of nonzero Newton's constant $G_{\rm N}$. The situation is analogous to what happened when the speed of light, c, was realized to be finite [33]: in Galilean physics ($c = \infty$) a change of the reference frame leads only to a constant shift of all the velocities, while in special relativity ($c =$ finite) it also alters temporal and spatial lengths (time dilation and Lorentz contraction) and makes the concept

of simultaneity relative. With gravity ($G_{\rm N} \neq 0$), even the concept of spacetime becomes relative. The trend is consistent—as we "turn on" fundamental constants in nature ($c = \infty \to$ finite and $G_{\rm N} = 0 \to \neq 0$), physical descriptions become more and more relative: descriptions of the same physical system in different reference frames appear to differ more and more.

The organization of this chapter is the following. In Section 6.1, we discuss some basic aspects of the breakdown of global spacetime, setting up the stage for later discussions. In Sections 6.2 and 6.3, we describe how our picture addresses the problem of black hole formation and evaporation. We discuss the quantum structure of black hole microstates and the unitary flow of information as viewed from a distant reference frame (in Section 6.2), and how it can be consistent with the existence of interior spacetime (in Section 6.3). In particular, we elucidate how this picture addresses the arguments for firewalls and provides a consistent resolution to the black hole information paradox. In Section 6.4, we give our summary by presenting a grand picture of the structure of quantum gravity implied by our analysis of a system with a black hole.

Throughout the chapter, we adopt the Schrödinger picture for quantum evolution, and use natural units in which $\hbar = c = 1$ unless otherwise stated. We limit our discussions to 4-dimensional spacetime, although we do not expect difficulty in extending to other dimensions. The value of the Planck length in our universe is $l_{\rm P} = G_{\rm N}^{1/2} \simeq 1.62 \times 10^{-35}$ m. A concise summary of the implications of our framework for black hole physics can be found in Ref. [7].

6.1 Failure of Global Spacetime

As described in the introduction, semiclassical theory applied to an entire global spacetime leads to overcounting of the true degrees of freedom at the quantum level. This implies that in the full theory of quantum gravity, a semiclassical description of physics emerges only in some limited sense. Here we discuss basic aspects of this limitation, setting up the stage for later discussions.

The idea of complementarity [3] is that the overcounting inherent in the global spacetime picture may be avoided if we limit our description to what a single "observer"—represented by a single worldline in spacetime—can causally access. Depending on which observer we choose, we obtain different descriptions of the system, which are supposed to be equivalent. Since the events an observer can see lie within the causal patch associated with the worldline representing the observer, we may assume that this causal patch is the spacetime region a single such description may represent. In particular, one may postulate the following [33, 32]:

- For a single description allowing a semiclassical interpretation of the system, the spacetime region represented is restricted to the causal patch associated with a single worldline. With this restriction, the description can be local in the sense that any physical correlations between low energy field theoretic degrees of freedom respect causality in spacetime (beyond

some microscopic quantum gravitational distance l_*, meaning that possible nonlocal corrections are exponentially suppressed $\sim e^{-r/l_*}$).

Depending on the worldline we take, we may obtain different descriptions of the same system, which are all local in appropriate spacetime regions. A transformation between different descriptions is nothing but the complementarity transformation.

To implement Hamiltonian quantum mechanics, we must introduce a time variable. This corresponds to foliating the causal patch by equal-time hypersurfaces, with a state vector $|\Psi(t)\rangle$ representing the state of the system on each hypersurface.[3] Let $\mathbf{x}$ be spatial coordinates parameterizing each equal-time hypersurface. Physical quantities associated with field theoretic degrees of freedom can then be obtained using field theoretic operators $\phi(\mathbf{x})$ and the state $|\Psi(t)\rangle$. (Excited string degrees of freedom will require the corresponding operators.) In general, the *procedure* of electing coordinates $(t, \mathbf{x})$, which we need to *define* states and operators, must be given independently of the background spacetime, since we do not know it a priori (and states may even represent superpositions of very different semiclassical geometries); an example of such procedures is described in Ref. [8]. In our discussions in this chapter, however, we mostly consider issues addressed on a fixed background spacetime (at least approximately), so we need not be concerned with this problem too much—we may simply use any coordinate system adapted to a particular spacetime we consider, e.g. Schwarzschild-like coordinates for a black hole.

In the next two sections, we discuss how the complementarity picture described above works for a dynamical black hole. We discuss the semiclassical descriptions of the system in various reference frames, as well as their mutual consistency. In these discussions, we focus on a black hole that is well approximated by a Schwarzschild black hole in asymptotically flat spacetime. We do not expect difficulty in extending it to more general cases.

6.2 Black Hole—A Distant Description

Suppose we describe the formation and evaporation of a black hole in a distant reference frame. Following Ref. [62], we postulate that there exists a unitary description which involves only the degrees of freedom that can be viewed as being on and outside the (stretched) horizon. To describe quantum states with a black hole, we adopt Schwarzschild-like time slicings to define equal-time

[3] In general, the "time variable" of (constrained) Hamiltonian quantum mechanics may not be related directly with time we observe in nature [43]. Indeed, the whole "multiverse" may be represented by a state that does not depend on the time variable and is normalizable in an appropriate sense [44]. Even if this is the case, however, when we describe only a branch of the whole state, e.g. when we describe a system seen by a particular observer, the state *of the system* may depend on time. Here we discuss systems with black holes, which are *parts of* the multiverse so their states may depend on time.

hypersurfaces.[4] We argue that the origin of the Bekenstein-Hawking entropy may be viewed as a coarse-graining performed to obtain a semiclassical description of the evolving black hole. We then discuss implications of such a coarse-graining, in particular how it reconciles unitarity of the Hawking emission and black hole mining processes in the fundamental theory with the non-unitary (thermal) view in the semiclassical description.

6.2.1 Microscopic structure of a dynamical black hole

Consider a quantum state which represents a black hole of mass M located at some place at rest, where the position and velocity are measured with respect to some distant reference frame, e.g. an inertial frame elected at asymptotic infinity. Because of the uncertainty principle, such a state must involve a superposition of energy and momentum eigenstates. Let us first estimate the required size of the spread of energy ΔE, with E measured in the asymptotic region. According to the standard Hawking calculation, a state of a black hole of mass M will evolve after Schwarzschild time $\Delta t \approx O(M l_{\rm P}^2)$ into a state representing a Hawking quantum of energy $\approx O(1/M l_{\rm P}^2)$ and a black hole with the correspondingly smaller mass. The fact that these two states—before and after the emission—are nearly orthogonal implies that the original state must involve a superposition of energy eigenstates with

$$\Delta E \approx \frac{1}{\Delta t} \approx O\left(\frac{1}{M l_{\rm P}^2}\right). \tag{6.1}$$

Of course, this is nothing but the standard time-energy uncertainty relation, and here we have assumed that a state after time $t \ll M l_{\rm P}^2$ is not clearly distinguishable from the original one, so that the uncertainty relation is almost saturated.

Next, we consider the spread of momentum Δp, where p is again measured in the asymptotic region. Suppose we want to identify the spatial location of the black hole with precision comparable to the quantum stretching of the horizon $\Delta r \approx O(1/M)$, i.e. $\Delta d \approx O(l_{\rm P})$, where r and d are the Schwarzschild radial coordinate and the proper length, respectively. This implies that the superposition must involve momenta with spread $\Delta p \approx (1/M l_{\rm P})(1/\Delta d) \approx O(1/M l_{\rm P}^2)$, where the factor $1/M l_{\rm P}$ in the middle expression is the redshift factor. This value of Δp corresponds to an

[4] Strictly speaking, to describe a general gravitating system we need a procedure to foliate the relevant spacetime region in a background independent manner, as discussed in the previous section. For our present purposes, however, it suffices to employ any foliation that reduces to Schwarzschild-like time slicings when the black hole exists. Note that macroscopic uncertainties in the black hole mass, location, and spin caused by the stochastic nature of Hawking radiation [52, 45] require us to focus on appropriate branches in the full quantum state in which the black hole in a given time has well-defined values for these quantities at the classical level. The relation between the Schwarzschild-like foliation and a general background independent foliation is then given by the standard coordinate transformation, which does not introduce subtleties beyond those discussed in this chapter. The effect on unitarity by focusing on particular branches in this way is also minor, so we ignore it. The full unitarity, however, can be recovered by keeping all the branches in which the black hole has different classical properties at late times [45].

uncertainty of the kinetic energy $\Delta E_{\rm kin} \approx p\Delta p/M \approx O(1/M^3 l_{\rm P}^4)$, which is much smaller than ΔE in Eq. (6.1). The spread of energy thus comes mostly from a superposition of different rest masses: $\Delta E \approx \Delta M$.

How many different independent ways are there to superpose the energy eigenstates to arrive at the same black hole geometry, at a fixed position within the precision specified by Δr and of mass M within an uncertainty of ΔM? We assume that the Bekenstein-Hawking entropy, $\mathcal{A}/4l_{\rm P}^2$, gives the logarithm of this number (at the leading order in expansion in inverse powers of $\mathcal{A}/l_{\rm P}^2$), where $\mathcal{A} = 16\pi M^2 l_{\rm P}^4$ is the area of the horizon. While the definition of the Bekenstein-Hawking entropy does not depend on the precise values of ΔM or Δp, a natural choice for these quantities is

$$\Delta M \approx \Delta p \approx O\left(\frac{1}{M l_{\rm P}^2}\right), \tag{6.2}$$

which we will adopt. The nonzero Bekenstein-Hawking entropy thus implies that there are exponentially many independent states in a small energy interval of $\Delta E \approx O(1/M l_{\rm P}^2)$. We stress that it is not appropriate to interpret this to mean that quantum mechanics introduces exponentially large degeneracies that do not exist in classical black holes. In classical general relativity, a set of Schwarzschild black holes located at some place at rest are parameterized by a continuous mass parameter M; i.e., there are a continuously infinite number of black hole states in the energy interval between M and $M + \Delta M$ for any M and small ΔM. Quantum mechanics *reduces* this to a finite number $\approx e^{S_0}\Delta M/M$, with S_0 given by[5]

$$S_0 = \frac{\mathcal{A}}{4l_{\rm P}^2} + O\left(\frac{\mathcal{A}^q}{l_{\rm P}^{2q}};\, q < 1\right). \tag{6.3}$$

This can also be seen from the fact that S_0 is written as $\mathcal{A}c^3/4l_{\rm P}^2\hbar$ when $\hbar$ and c are restored, which becomes infinite for $\hbar \to 0$.

As is clear from the argument above, there are exponentially many independent microstates, corresponding to Eq. (6.3), which are all black hole *vacuum* states: the states that do not have a field or string theoretic excitation on the semiclassical black hole background and in which the stretched horizon, located at $r_{\rm s} = 2Ml_{\rm P}^2 + O(1/M)$, is not excited.[6] Denoting the indices representing these exponentially many states collectively by k, which we call the *vacuum index*, basis states for the general microstates of a black hole of mass M (within the uncertainty of ΔM) can be given by

$$|\Psi_{\bar{a}\,a\,a_{\rm far};k}(M)\rangle. \tag{6.4}$$

[5]Of course, quantum mechanics allows for a superposition of these finite number of independent states, so the number of possible (not necessarily independent) states is continuously infinite. The statement here applies to the number of independent states, regarding classical black holes with different M as independent states.

[6]These states can be defined, for example, as the states obtained by first forming a black hole of mass M and then waiting sufficiently long time after (artificially) switching off Hawking emission. Note that at the level of full quantum gravity, all the black hole states are obtained as excited states. Any semiclassical description, however, treats some of them as vacuum states on the black hole background.

Here, $\bar{a}$, a, and $a_{\rm far}$ represent the indices labeling the excitations of the stretched horizon, in the near exterior zone region (i.e. the region within the gravitational potential barrier defined, e.g., as $r \leq R_{\rm Z} \equiv 3Ml_{\rm P}^2$), and outside the zone ($r > R_{\rm Z}$), respectively.[7] As we have argued, the index k runs over $1, \cdots, e^{S_0}$ for the vacuum states $\bar{a} = a = a_{\rm far} = 0$. In general, the range for k may depend on $\bar{a}$ and a, but its dependence is higher order in $l_{\rm P}^2/\mathcal{A}$; i.e., for fixed $\bar{a}$ and a

$$k = 1, \cdots, e^{S_{\bar{a}a}}; \qquad S_{\bar{a}a} - S_0 \approx O\left(\frac{\mathcal{A}^q}{l_{\rm P}^{2q}}; \, q < 1\right). \tag{6.5}$$

We thus mostly ignore this small dependence of the range of k on $(\bar{a}, a)$, i.e. the non-factorizable nature of the Hilbert space factors spanned by these indices, except when we discuss negative energy excitations associated with Hawking emission later, where this aspect plays a relevant role in addressing one of the firewall arguments.

Since we are mostly interested in physics associated with the black hole region, we also introduce the notation in which the excitations in the far exterior region are separated. As we will see later, the degrees of freedom represented by k can be regarded as being mostly in the region $r \leq R_{\rm Z}$, so we may write the states of the entire system in Eq. (6.4) as

$$|\Psi_{\bar{a}\, a\, a_{\rm far};k}(M)\rangle \approx |\psi_{\bar{a}a;k}(M)\rangle |\phi_{a_{\rm far}}(M)\rangle, \tag{6.6}$$

and call $|\psi_{\bar{a}a;k}(M)\rangle$ and $|\phi_{a_{\rm far}}(M)\rangle$ as the black hole and exterior states, respectively. Note that by labeling the states in terms of localized excitations, we need not write explicitly the trivial vacuum entanglement between the black hole and exterior states that does not depend on k, which typically exist when they are specified in terms of the occupation numbers of modes spanning the entire space.

How many independent quantum states can the black hole region support? Let us label appropriately coarse-grained excitations in the region $r_{\rm s} \leq r \leq R_{\rm Z}$ by $i = 1, 2, \cdots$, each of which carries entropy S_i. Suppose there are n_i excitations of type i at some fixed locations. The entropy of such a configuration is given by the sum of the "entropy of vacuum" in Eq. (6.3) and the entropies associated with the excitations:

$$S_I = S_0 + \sum_i n_i S_i. \tag{6.7}$$

The energy of the system in the region $r \leq R_{\rm Z}$ is given by the sum of the mass M of the black hole, which we define as the energy the system would have in the absence of an excitation outside

[7] Strictly speaking, the states may also have the vacuum index associated with the ambient space in which the black hole exists. The information in this index, however, is not extracted in the Hawking evaporation or black hole mining process, so we ignore it here. (For more discussions, see, e.g., Section 5 of Ref. [5].) We will also treat excitations spreading both in the $r \leq R_{\rm Z}$ and $r > R_{\rm Z}$ regions only approximately by including them either in a or $a_{\rm far}$. The precise description of these excitations will require more elaborate expressions, e.g. than the one in Eq. (6.6), which we believe is an inessential technical subtlety in addressing our problem.

the stretched horizon, and the energies associated with the excitations in the zone. Note that excitations here are defined as fluctuations with respect to a fixed background, so their energies E_i as well as entropies S_i can be either positive or negative, although the signs of the energy and entropy must be the same: $E_i S_i > 0$. The meaning of negative entropies will be discussed in detail in Sections 6.2.4 and 6.2.5.

Since excitations in the zone affect geometry, spacetime outside the stretched horizon, when they exist, is not exactly that of a Schwarzschild black hole. We require that these excitations do not form a black hole by themselves or become a part of the black hole at the center; otherwise, the state must be viewed as being built on a different semiclassical vacuum.[8] The total entropy S of the region $r \leq R_{\rm Z}$, i.e. the number of independent microscopic quantum states representing this region, is then given by

$$S = \ln\Big(\sum_I e^{S_I}\Big), \tag{6.8}$$

where I represents possible configurations of excitations, specified by the set of numbers $\{n_i\}$ and the locations of excitations of each type i, that do not modify the semiclassical vacuum in the sense described above. As suggested by a representative estimate [11], and particularly emphasized in Ref. [46], the contribution of such excitations to the total entropy is subdominant in the expansion in inverse powers of $\mathcal{A}/l_{\rm P}^2$: $S = S_0 + O(\mathcal{A}^q/l_{\rm P}^{2q}; q < 1)$. The total entropy in the near black hole region, $r \leq R_{\rm Z}$, is thus given by

$$S = \frac{\mathcal{A}}{4l_{\rm P}^2}, \tag{6.9}$$

at the leading order in $l_{\rm P}^2/\mathcal{A}$.

6.2.2 Emergence of the semiclassical picture and coarse-graining

The fact that all the independent microstates with different values of k lead to the same geometry suggests that the semiclassical picture is obtained after coarse-graining the degrees of freedom represented by this index; namely, any result in semiclassical theory is a statement about the maximally mixed ensemble of microscopic quantum states consistent with the specified background within the precision allowed by quantum mechanics [5]. According to this picture, the black hole vacuum state in the semiclassical description is given by the density matrix

$$\rho_0(M) = \frac{1}{e^{S_0}} \sum_{k=1}^{e^{S_0}} |\Psi_{\bar{a}=a=a_{\rm far}=0;k}(M)\rangle\langle\Psi_{\bar{a}=a=a_{\rm far}=0;k}(M)|. \tag{6.10}$$

[8] More precisely, we regard two geometries as being built on different classes of semiclassical vacua when they have different horizon configurations as viewed from a fixed reference frame. On the other hand, if two geometries have the same horizon, they belong to the same "vacuum equivalence class" in the sense that one can be converted into the other with "excitations." For more discussions on this point, see Ref. [8] and Section 6.2.2.

Because of the coarse-graining of an enormous number of degrees of freedom, this density matrix has statistical characteristics.

In order to obtain the response of this state to the operators in the semiclassical theory, we may trace out the subsystem on which they do not act. As we will discuss more later, the operators in the semiclassical theory in general act on a part, but not all, of the degrees of freedom represented by the k index. Let us denote the subsystem on which semiclassical operators act nontrivially by C, and its complement by $\bar{C}$. The index k may then be viewed as labeling the states in the combined $C\bar{C}$ system which satisfy certain constraints, e.g. the total energy being M within ΔM. The density matrix representing the semiclassical vacuum state in the Hilbert space in which the semiclassical operators act nontrivially, C, is given by

$$\tilde{\rho}_0(M) = \mathrm{Tr}_{\bar{C}}\, \rho_0(M). \tag{6.11}$$

Consistently with our identification of the origin of the Bekenstein-Hawking entropy, we assume that this density matrix represents the thermal density matrix with temperature $T_{\mathrm{H}} = 1/8\pi M l_{\mathrm{P}}^2$ in the zone region (as measured at asymptotic infinity):

$$\tilde{\rho}_0(M) \approx \frac{1}{\mathrm{Tr}\, e^{-\beta H_{\mathrm{sc}}(M)}} e^{-\beta H_{\mathrm{sc}}(M)}; \qquad \beta = \begin{cases} \frac{1}{T_{\mathrm{H}}} & \text{for } r \leq R_{\mathrm{Z}}, \\ +\infty & \text{for } r > R_{\mathrm{Z}}, \end{cases} \tag{6.12}$$

where $H_{\mathrm{sc}}(M)$ is the Hamiltonian of the semiclassical theory in the distant reference frame, which is defined in the region $r \geq r_{\mathrm{s}}$ on the black hole background of mass M.[9] (The meaning of position-dependent β is that the expression $\beta H_{\mathrm{sc}}(M)$ should be interpreted as β times the Hamiltonian density integrated over space.) Note that this procedure of obtaining Eq. (6.12) from Eq. (6.10) can be viewed as an example of the standard procedure of obtaining the canonical ensemble of a system from the microcanonical ensemble of a larger (isolated) system that contains the system of interest. In fact, if the system traced out is larger than the system of interest, $\dim \bar{C} \gtrsim \dim C$, we expect to obtain the canonical ensemble in this manner (see Ref. [31] for a related discussion). Below, we drop the tilde from the density matrix in Eq. (6.12), as it represents the same state as the one in Eq. (6.10)—$\rho_0(M)$ must be interpreted to mean either the right-hand side of Eq. (6.10) or of Eq. (6.12), depending on the Hilbert space under consideration.

In semiclassical field theory, the density matrix of Eq. (6.12) is obtained as a reduced density matrix by tracing out the region within the horizon in the *unique* global black hole vacuum state. Our view is that this density matrix, in fact, is obtained from a mixed state of exponentially many pure states, arising from a coarse-graining performed in Eq. (6.10); the prescription in the

[9] The Hilbert space of the semiclassical theory for states which have a single black hole at a fixed location at rest may be decomposed as $\mathcal{H} = \oplus_M \mathcal{H}_M$, where $\mathcal{H}_M$ is the space spanned by the states in which there is a black hole of (appropriately coarse-grained) mass M. In this language, $H_{\mathrm{sc}}(M)$ is a part of the semiclassical Hamiltonian acting on the subspace $\mathcal{H}_M$.

semiclassical theory provides (merely) a useful way of obtaining the same density matrix, in a similar sense in which the thermofield double state was originally introduced [76]. We emphasize that the information in k is invisible in the semiclassical theory (despite the fact that it involves subsystem C) as it is already coarse-grained *to obtain* the theory; in particular, the dynamics of the degrees of freedom represented by k cannot be described in terms of the semiclassical Hamiltonian $H_{\rm sc}(M)$.[10] As we will see explicitly later, it is this inaccessibility of k that leads to the apparent violation of unitarity in the semiclassical calculation of the Hawking emission process [55]. Note that because $\rho_0(M)$ takes the form of the maximally mixed state in k, results in the semiclassical theory do not depend on the basis of the microscopic states chosen in this space.

A comment is in order. In connecting the expression in Eq. (6.10) to Eq. (6.12), we have (implicitly) assumed that $|\Psi_{\bar{a}=a=a_{\rm far}=0;k}(M)\rangle$ represent the black hole vacuum states in the limit that the effect from evaporation is (artificially) shut off.[11] With this definition of vacuum states, the evolution effect necessarily "excites" the states, making $a \neq 0$, as we will see more explicitly in Section 6.2.4. As a consequence, the density matrix for the semiclassical operators representing the evolving black hole deviates from Eq. (6.12) even without matter or radiation. (In the semiclassical picture, this is due to the fact that the effective gravitational potential is not truly confining, so that the state of the black hole is not completely stationary.) If one wants, one can redefine vacuum states to be these states: the states that do not have any matter or radiation excitation on the *evolving* black hole background—the original vacuum states are then obtained as excited states on the new vacuum states.[12] This redefinition is possible because the two semiclassical "vacua" represented by the two classes of microstates belong to the same "vacuum equivalence class" in the sense described in the last paragraph of Section 6.2.1; specifically, they possess the same horizon for the same black hole mass, as defined for the evaporating case in Ref. [79].

As was mentioned above, semiclassical operators, in particular those for modes in the zone, act nontrivially on both a and k indices of microstates $|\Psi_{\bar{a}\,a\,a_{\rm far};k}(M)\rangle$. This can be seen as follows. If the operators acted only on the a index, the maximal mixture in k space with $a = 0$, Eq. (6.10), would look like a pure state from the point of view of these operators, contradicting the thermal

[10] This does not mean that a device made out of semiclassical degrees of freedom cannot probe information in k. Since there are processes in the fundamental theory (i.e. Hawking evaporation and mining processes) in which information in k is transferred to that in semiclassical *excitations* (i.e. degrees of freedom represented by the a and $a_{\rm far}$ indices), information in k can be probed by degrees of freedom appearing in the semiclassical theory. It is simply that these information extraction processes cannot be described within the semiclassical theory, since it can make statements only about the ensemble in Eq. (6.10) and excitations built on it.

[11] This is analogous to the treatment of a meta-stable vacuum in usual quantum field theory. At the most fundamental level (or on a very long timescale), such a state must be viewed as a scattering state built on the true ground state of the system. In practice (or on a sufficiently short timescale), however, we regard it as a vacuum state, which is approximately the ground state of a theory in which the tunneling out of this state is artificially switched off, e.g. by making the relevant potential barriers infinitely high.

[12] In the standard language in semiclassical theory, the original vacuum states correspond essentially to the Hartle-Hawking vacuum [77], while the new ones (very roughly) to the Unruh vacuum [78].

nature in Eq. (6.12). On the other hand, if the operators acted only on the k index, they would commute with the maximally mixed state in k space, again contradicting the thermal state. Since the thermal nature of Eq. (6.12) is prominent only for modes whose energies as measured in the asymptotic region are of order the Hawking temperature or smaller

$$\omega \lesssim T_{\rm H}, \tag{6.13}$$

i.e. whose energies as measured by local (approximately) static observers are of order or smaller than the blueshifted Hawking temperature $T_{\rm H}/\sqrt{1-2Ml_{\rm P}^2/r}$, this feature is significant only for such infrared modes—operators representing modes with $\omega \gg T_{\rm H}$ act essentially only on the a index. For operators representing the modes with Eq. (6.13), their actions on microstates can be very complicated, although they act on the coarse-grained vacuum state of Eq. (6.10) as if it is the thermal state in Eq. (6.12), up to corrections suppressed by the exponential of the vacuum entropy S_0. The commutation relations of these operators defined on the coarse-grained states take the form as in the semiclassical theory, again up to exponentially suppressed corrections.

There is a simple physical picture for this phenomenon of "non-decoupling" of the a and k indices for the infrared modes. As viewed from a distant reference frame, these modes are "too soft" to be resolved clearly above the background—since the derivation of the semiclassical theory involves coarse-graining over microstates in which the energy stored in the region $r \lesssim R_{\rm Z}$ has spreads of order $\Delta E \approx 1/Ml_{\rm P}^2$, infrared modes with $\omega \lesssim T_{\rm H} \approx O(1/Ml_{\rm P}^2)$ are not necessarily distinguished from "spacetime fluctuations" of order ΔE. One might think that if a mode has nonzero angular momentum or charge, one can discriminate it from spacetime fluctuations. In this case, however, it cannot be clearly distinguished from vacuum fluctuations of a Kerr or Reissner-Nordström black hole having the corresponding (minuscule) angular momentum or charge. In fact, we may reverse the logic and view that this lack of a clear identity of the soft modes is the physical origin of the thermality of black holes (and thus of Hawking radiation).

Once the state for the vacuum of the semiclassical theory is obtained as in Eq. (6.10) (or Eq. (6.12) after partial tracing) and appropriate coarse-grained operators acting on it are identified, it is straightforward to construct the rest of the states in the theory—we simply have to act these operators (either field theoretic or of excited string states) on $\rho_0(M)$ to obtain the excited states. For example, to obtain a state which has a field theoretic excitation in the zone, one can apply the appropriate linear combination of creation and/or annihilation operators in the semiclassical theory, $a^\dagger_{\omega\ell m}$ and/or $a_{\omega\ell m}$:

$$\rho_{\bar{a}=0\,a\,a_{\rm far}=0}(M) = \left(\sum_{\ell,m}\int (c^a_{\omega\ell m}a_{\omega\ell m} + c'^a_{\omega\ell m}a^\dagger_{\omega\ell m})d\omega\right)\rho_0(M)\left(\sum_{\ell,m}\int (c^a_{\omega\ell m}a_{\omega\ell m} + c'^a_{\omega\ell m}a^\dagger_{\omega\ell m})d\omega\right)^\dagger, \tag{6.14}$$

where $c^a_{\omega\ell m}$ and $c'^a_{\omega\ell m}$ are coefficients. In the case that the applied operator is that for an infrared mode, this represents a state in which the thermal distribution for the infrared modes is "modulated" by an excitation over it. A construction similar to Eq. (6.14) also works for excitations in the far region. To obtain excitations of the stretched horizon, i.e. $\bar{a} \neq 0$, operators dedicated to describing them must be introduced. The detailed dynamics of these degrees of freedom, i.e. the $r = r_{\rm s}$ part of $H_{\rm sc}(M)$, is not yet fully known, however.

6.2.3 "Constituents of spacetime" and their distribution

While not visible in semiclassical theory, the black hole formation and evaporation (or mining) processes do involve the degrees of freedom represented by k, which we call *fine-grained vacuum degrees of freedom*, or vacuum degrees of freedom for short. The dynamics of these degrees of freedom as well as their interactions with the excitations in the semiclassical theory are determined by the fundamental theory of quantum gravity, which is not yet well known. We may, however, anticipate their basic properties based on some general considerations. In particular, motivated by the general idea of complementarity, we assume the following:

- Interactions with vacuum degrees of freedom do not introduce violation of causality among field theory degrees of freedom (except possibly for exponentially suppressed corrections, $\sim e^{-r/l_*}$ with l_* a short-distance quantum gravitational scale).

- Interactions between vacuum degrees of freedom and excitations in the semiclassical theory are such that unitarity is preserved at the microscopic level.

The first assumption is a special case of the postulate discussed in Section 6.1, applied to the distant reference frame description of a black hole. This implies that we cannot send superluminal signals among field theory degrees of freedom using interactions with vacuum degrees of freedom. The second assumption has an implication for how the vacuum degrees of freedom may appear from the semiclassical standpoint, which we now discuss.

In quantum mechanics, the information about a state is generally delocalized in space—locality is a property of dynamics, not that of states. In the case of black hole states, the information about k, which roughly represents slightly different "values" (superpositions) of M, is generally delocalized in a large spatial region, so that it can be accessed physically in a region away from the stretched horizon (e.g. around the edge of the zone $r \sim R_{\rm Z}$). This, however, does not mean that the complete information about the state can be recovered by a physical process occurring in a limited region in spacetime. For example, if we consider the set of e^{S_0} different black hole vacuum states, a physical detector occupying a finite spatial region can only partially discriminate these states in a given finite time.

To see how much information a physical detector in spatial region i can resolve, we can consider the reduced density matrix obtained after tracing out the subsystems that cannot be accessed by the semiclassical degrees of freedom associated with this region. In particular, we may consider the set of all field theory (and excited string state) operators that have support in i, and trace out the subsystems that do not respond to any of these operators (which we denote by $\bar{C}_i$):

$$\rho_0^{(i)} = \mathrm{Tr}_{\bar{C}_i}\, \rho_0(M), \tag{6.15}$$

where $\rho_0(M)$ is given by Eq. (6.10), and we have omitted the argument M for $\rho_0^{(i)}$. The von Neumann entropy of this density matrix, $S_0^{(i)} = -\mathrm{Tr}\, \rho_0^{(i)} \ln \rho_0^{(i)}$, then indicates the discriminatory power the region i possesses—a physical process occurring in region i can, at most, discriminate the e^{S_0} states into $e^{S_0^{(i)}}$ ($\ll e^{S_0}$) types in a characteristic timescale of the system, $1/\Delta E \approx O(M l_{\mathrm{P}}^2)$. According to the assumption in Eq. (6.12), this entropy is the gravitational thermal entropy contained in region i, calculated using the semiclassical theory.

We therefore arrive at the following picture. Let us divide the region $r \geq r_{\mathrm{s}}$ into N (arbitrary) subregions, each of which is assumed to have a sufficiently large number of degrees of freedom so that the thermodynamic limit can be applied. A basis state in the semiclassical theory can be written as

$$\rho_{\bar{a}\, a\, a_{\mathrm{far}}}(M) = \rho_{a_1}^{(1)} \otimes \rho_{a_2}^{(2)} \otimes \cdots \otimes \rho_{a_N}^{(N)}, \tag{6.16}$$

where $\rho_{a_i}^{(i)}$ are states defined in the i-th subregion, with a_i representing excitations contained in that region. (Following the convention in Section 6.2.2, we regard the vacuum states, $\bar{a} = a = a_{\mathrm{far}} = 0$, to be defined in the limit that the effect from evaporation is ignored.) Now, in the full Hilbert space of quantum gravity, there are e^{S_0} independent states that all reduce to the same $\rho_{\bar{a}\, a\, a_{\mathrm{far}}}(M)$ at the semiclassical level. These states can be written as

$$|\Psi_{\bar{a}\, a\, a_{\mathrm{far}};k=\{k_i\}}(M)\rangle = |\psi^{(1)}_{a_1;k_1}\rangle\, |\psi^{(2)}_{a_2;k_2}\rangle \cdots |\psi^{(N)}_{a_N;k_N}\rangle, \tag{6.17}$$

where $k_i = 1, \cdots, e^{S_0^{(i)}}$ with

$$S_0^{(i)} \approx \text{ gravitational thermal entropy contained in subregion } i, \tag{6.18}$$

calculated using the semiclassical theory for subregions that do not contain the stretched horizon. The $S_0^{(i)}$'s for the subregions involving the stretched horizon are determined by the condition

$$\sum_{i=1}^{N} S_0^{(i)} = S_0 \approx \frac{\mathcal{A}}{4 l_{\mathrm{P}}^2}, \tag{6.19}$$

which is valid in the thermodynamic limit. Assuming that the entropy on the stretched horizon is distributed uniformly on the surface, this condition determines the entropies contained in all the subregions.

The association of k_i's to each subregion, as in Eq. (6.17), corresponds to taking a specific basis in the space spanned by k. While the expressions above are strictly valid only in the thermodynamic limit, the corrections caused by deviating from it (e.g. due to correlations among subregions) do not affect our later discussions. In particular, it does not change the fact that the region around the edge of the zone, $r \leq R_{\rm Z}$ and $r - 2Ml_{\rm P}^2 \not\ll Ml_{\rm P}^2$, contains $O(1)$ bits of information about k (as it contains $O(1)$ bits of gravitational thermal entropy), which becomes important when we discuss the Hawking emission process in Section 6.2.4. Incidentally, the picture described here leads to the natural interpretation that the subsystem that is traced out when going from Eq. (6.10) to Eq. (6.12) corresponds to the stretched horizon; i.e. $\bar{C}$ lives on the stretched horizon, while C in the zone.[13]

We stress that by the gravitational thermal entropy in Eq. (6.18), we mean that associated with the equilibrium vacuum state. It counts the thermal entropy within the zone, since this region is regarded as being in equilibrium because of its boundedness due to the stretched horizon and the potential barrier; on the other hand, Eq. (6.18) does not count the thermal entropy associated with Hawking radiation emitted from the zone, which is (artificially) switched off in defining our vacuum microstates. In other words, when calculating $S_0^{(i)}$'s using Eq. (6.18) we should use the vacuum state in Eq. (6.12), implying that we should use the local temperature, i.e. the temperature as measured by local static observers, of

$$T(r) \simeq \begin{cases} \frac{T_{\rm H}}{\sqrt{1-\frac{2Ml_{\rm P}^2}{r}}} & \text{for } r \leq R_{\rm Z}, \\ 0 & \text{for } r > R_{\rm Z}. \end{cases} \tag{6.20}$$

When the evolution effect is turned on, which we will analyze in Section 6.2.4, the state of the zone is modified ($a \neq 0$) due to an ingoing negative energy flux, while the state outside the zone is excited ($a_{\rm far} \neq 0$) by Hawking quanta, which are emitted from the edge of the zone and propagate freely in the ambient space. The contribution of the negative energy flux to the entropy within the zone is small, as we will see in Section 6.2.4.

The distribution of vacuum degrees of freedom in Eqs. (6.17, 6.18) is exactly the one needed for the interactions between these degrees of freedom and semiclassical excitations to preserve unitarity [5]. Imagine we put a physical detector at constant r in the zone. The detector then sees the thermal bath for all the modes with blueshifted Hawking temperature, Eq. (6.20), including higher angular momentum modes. This allows for the detector(s) to extract energy from the black

[13] This in turn gives us a natural prescription to determine the location of the stretched horizon precisely. Since the semiclassical expression in Eq. (6.12) is expected to break down for $\ln \dim C > \ln \dim \bar{C}$, a natural place to locate the stretched horizon, i.e. the cutoff of the semiclassical spacetime, is where the gravitational thermal entropy outside the stretched horizon becomes $S_0/2 = \mathcal{A}/8l_{\rm P}^2$. For n low energy species, this yields $r_{\rm s} - 2Ml_{\rm P}^2 \sim n/M \sim l_*^2/Ml_{\rm P}^2$, where l_* is the string (cutoff) scale and we have used the relation $l_*^2 \sim nl_{\rm P}^2$, which is expected to apply in any consistent theory of quantum gravity (see, e.g., Ref. [80]). This scaling is indeed consistent, giving the local Hawking temperature at the stretched horizon $T(r_{\rm s}) \sim 1/l_*$, where $T(r)$ is given in Eq. (6.20).

hole at an accelerated rate compared with spontaneous Hawking emission: the mining process [81, 82]. In order for this process to preserve unitarity, the detector must also extract information at the correspondingly accelerated rate. This is possible if the information about the microstate of the black hole, specified by the index k, is distributed according to the gravitational thermal entropy, as in Eqs. (6.17, 6.18). A similar argument also applies to the spontaneous Hawking emission process, which is viewed as occurring around the edge of the zone, $r \sim R_{\rm Z}$, where the gravitational thermal entropy is small but not negligible. The microscopic and semiclassical descriptions of these processes will be discussed in detail in Sections 6.2.4 and 6.2.5.

It is natural to interpret the expression in Eq. (6.17) to mean that k_i labels possible configurations of "physical soft quanta"—or the "constituents of spacetime"—that comprise the region i. In a certain sense, this interpretation is correct. The dimension of the relevant Hilbert space, $e^{S_0^{(i)}}$, controls possible interactions of the vacuum degrees of freedom with the excitations in the semiclassical theory in region i, e.g. how much information a detector located in region i can extract from the vacuum degrees of freedom. This simple picture, however, breaks down when we describe the same system from a different reference frame. As we will discuss in Section 6.3, the distribution of the vacuum degrees of freedom *depends on the reference frame*—they are not "anchored" to spacetime. Nevertheless, in a fixed reference frame, the concept of the spatial distribution of the degrees of freedom represented by the index k does make sense. In particular, in a distant reference frame the distribution is given by the gravitational thermal entropy calculated in the semiclassical theory, as we discussed here.

6.2.4 Hawking emission—"microscopic" and semiclassical descriptions

The formation and evaporation of a black hole involve processes in which the information about the initial collapsing matter is transferred into the vacuum index k, which will later be transferred back to the excitations in the semiclassical theory, i.e. the state of final Hawking radiation. Schematically, we may write these processes as

$$|m_{\rm init}\rangle \;\to\; \sum_{k=1}^{e^{S_0(M(t))}} \sum_{l} c_{kl}(t)\, |\psi_k(M(t))\rangle\, |r_l(t)\rangle \;\to\; |r_{\rm fin}\rangle, \tag{6.21}$$

where $|m_{\rm init}\rangle$, $|\psi_k(M(t))\rangle$, $|r_l(t)\rangle$, and $|r_{\rm fin}\rangle$ represent the states for the initial collapsing matter, the black hole of mass $M(t)$ (which includes the near exterior zone region; see Eq. (6.6)), the subsystem complement to the black hole at time t, and the final Hawking quanta after the black hole is completely evaporated, respectively. Here, we have suppressed the indices representing excitations for the black hole states. For generic initial states and microscopic emission dynamics, this evolution satisfies the behavior outlined in Ref. [53] on general grounds.

In this subsection, we discuss how the black hole evaporating process in Eq. (6.21) proceeds in details, elucidating how the arguments for firewalls in Refs. [4, 73, 74] are avoided. We also discuss how the semiclassical theory describes the same process, elucidating how the thermality of Hawking radiation arises despite the unitarity of the process at the fundamental level.

"Microscopic" (unitary) description

Let us first consider how the "elementary" Hawking emission process is described at the microscopic level,[14] i.e. how a "single" Hawking emission occurs in the absence of any excitations other than those directly associated with the emission. (As we will see later, this is not a very good approximation in general, but the treatment here is sufficient to illustrate the basic mechanism by which the information is transferred from the black hole to the ambient space.)

Suppose a black hole of mass M is in microstate k:

$$|\Psi_k(M)\rangle = |\psi_k(M)\rangle|\phi_I\rangle, \tag{6.22}$$

where $|\psi_k(M)\rangle$ is the black hole state, in which we have omitted indices representing excitations, while $|\phi_I\rangle$ is the exterior state, from which we have suppressed small M dependence (which, e.g., causes a small gravitational redshift of a factor of about 1.5 for the emitted Hawking quanta to reach the asymptotic region). As discussed in Sections 6.2.2 and 6.2.3, we consider $|\Psi_k(M)\rangle$ to be one of the black hole vacuum microstates in the limit that the evolution effect is shut off; see, e.g., Eqs. (6.12, 6.20). The effect of the evolution, which consists of successive elementary Hawking emission processes, will be discussed later.

After a timescale of $t \approx O(Ml_{\rm P}^2)$, the state in Eq. (6.22) evolves due to Hawking emission as

$$|\psi_k(M)\rangle|\phi_I\rangle \to \sum_{i,a,k'} c^k_{iak'}|\psi_{a;k'}(M)\rangle|\phi_{I+i}\rangle, \tag{6.23}$$

where $|\phi_{I+i}\rangle$ is the state in which newly emitted Hawking quanta, labeled by i and having total energy E_i, are added to the appropriately time evolved $|\phi_I\rangle$. The index a represents the fact that the black hole state has negative energy excitations of total energy $-E_a$ ($E_a > 0$) around the edge of the zone, created in connection with the emitted Hawking quanta; the coefficients $c^k_{iak'}$ are nonzero only if $E_i \approx E_a$ (within the uncertainty).[15] The negative energy excitations then propagate

[14] By the "microscopic" description, we mean a description in which the vacuum index k is kept (i.e. not coarse-grained as in the semiclassical description) so that the process is manifestly unitary at each stage of the evolution. A complete description of the microscopic dynamics of the vacuum degrees of freedom requires the fundamental theory of quantum gravity, which is beyond the scope of this work.

[15] To be precise, the sum in the right-hand side of Eq. (6.23) contains the "$i = 0$ terms" representing the branches in which no quantum is emitted: $|\phi_{I+0}\rangle = |\phi_I\rangle$. In these terms, there is no negative energy excitation: $c^k_{0ak'} \neq 0$ only for $a = 0$. The following expressions are valid including these terms with the definition $E_{i=0} = E_{a=0} = 0$.

inward, and after a time of order $Ml_{\rm P}^2 \ln(Ml_{\rm P})$ collide with the stretched horizon, making the black hole states relax as

$$|\psi_{a;k'}(M)\rangle \to \sum_{k_a} d_{k_a}^{ak'} |\psi_{k_a}(M - E_a)\rangle. \tag{6.24}$$

The combination of Eqs. (6.23, 6.24) yields

$$|\psi_k(M)\rangle|\phi_I\rangle \to \sum_{i,k_i} \alpha_{ik_i}^k |\psi_{k_i}(M - E_i)\rangle|\phi_{I+i}\rangle, \tag{6.25}$$

where $\alpha_{ik_i}^k = \sum_{a,k'} c_{iak'}^k d_{k_i}^{ak'}$, and we have used $E_i = E_a$; here, $M - E_i$ for different i may belong to the same mass within the precision ΔM, i.e. $M - E_i = M - E_{i'}$ for $i \neq i'$. This expression shows that information in the black hole can be transferred to the radiation state i.

It is important that the negative energy excitations generated in Eq. (6.23) come with *negative entropies*, so that each of the processes in Eqs. (6.23, 6.24) (as well as the propagation of the negative energy excitations in the zone) is separately unitary. This means that as k and i run over all the possible values with a being fixed, the index k' runs only over $1, \cdots, e^{S_0(M-E_a)}$, the dimension of the space spanned by k_a. In fact, this is an example of the non-factorizable nature of the Hilbert space factors spanned by k and a discussed in Eq. (6.5), which we assume to arise from the fundamental theory. This structure of the Hilbert space allows for avoiding the argument for firewalls in Ref. [73]—unlike what is imagined there, elements of the naive Fock space built on each k in a way isomorphic to that of quantum field theory are not all physical; the physical Hilbert space is smaller than such a (hypothetical) Fock space. This implies, in particular, that the Fock space structure of a semiclassical theory does not factor from the space spanned by the vacuum index k, as is also implied by the analysis in Section 6.2.2.

To further elucidate the point made above, we can consider the following simplified version of the relevant processes. Suppose a black hole in a superposition state of $|\psi_k(M)\rangle$'s ($k = 1, \cdots, e^{S_0(M)}$) releases 1 bit of information through Hawking emission of the form:

$$|\psi_k(M)\rangle|\phi_0\rangle \to \begin{cases} |\psi_{a;\frac{k+1}{2}}(M)\rangle|\phi_1\rangle & \text{if } k \text{ is odd,} \\ |\psi_{a;\frac{k}{2}}(M)\rangle|\phi_2\rangle & \text{if } k \text{ is even,} \end{cases} \tag{6.26}$$

where we have assumed $E_1 = E_2 = (\ln 2)/8\pi M l_{\rm P}^2 \simeq T_{\rm H}$, so that the entropy of the black hole after the emission is reduced by 1 bit: $S_0(M - E_1) = S_0(M) - \ln 2$. Note that the index representing the negative energy excitation (of energy $-E_1$) takes the same value a in the first and second lines. Namely, while the entire process in Eq. (6.26) is unitary, the initial states with $k = 2n - 1$ and $2n$ lead to the *same black hole state*. After the negative energy excitation reaches the stretched horizon, the black hole states relax into vacuum states for a smaller black hole:

$$|\psi_{a;k'}(M)\rangle \to |\psi_{k_1=k'}(M - E_1)\rangle. \tag{6.27}$$

While the resulting black hole has a smaller entropy than the original black hole, this relaxation process is unitary because k' in the left-hand side runs only over $1, \cdots, e^{S_0(M)}/2 = e^{S_0(M-E_1)}$. We note that the creation of a positive energy Hawking quantum and a negative energy excitation in Eq. (6.26) (and in Eq. (6.23)) takes a form very different from the standard "pair creation" of particles, which is often invoked to visualize the Hawking emission process. In the pair creation picture, the positive and negative energy excitations are maximally entangled with each other, which is not the case here. In fact, it is this lack of entanglement that allows the emission process to transfer the information from the black hole to radiation.

We emphasize that from the semiclassical spacetime viewpoint, the emission of Eq. (6.23) is viewed as occurring locally around the edge of the zone, which is possible because the information about the black hole microstate extends into the whole zone region according to Eqs. (6.17, 6.18). To elucidate this point, we may consider the tortoise coordinate

$$r^* = r + 2Ml_{\rm P}^2 \ln \frac{r - 2Ml_{\rm P}^2}{2Ml_{\rm P}^2}, \tag{6.28}$$

in which the region outside the Schwarzschild horizon $r \in (2Ml_{\rm P}^2, \infty)$ is mapped into $r^* \in (-\infty, \infty)$. This coordinate is useful in that the kinetic term of an appropriately redefined field takes the canonical form, so that its propagation can be analyzed as in flat space. In this coordinate, the stretched horizon, located at $r = 2Ml_{\rm P}^2 + O(l_*^2/Ml_{\rm P}^2)$ (see footnote 13), is at

$$r_{\rm s}^* \simeq -4Ml_{\rm P}^2 \ln \frac{Ml_{\rm P}^2}{l_*} \simeq -4Ml_{\rm P}^2 \ln(Ml_{\rm P}), \tag{6.29}$$

where l_* is the string (or gravitational cutoff) scale, which we take to be within a couple of orders of magnitude of $l_{\rm P}$. This implies that there is a large distance between the stretched horizon and the potential barrier region when measured in r^*: $\Delta r^* \approx 4Ml_{\rm P}^2 \ln(Ml_{\rm P}) \gg O(Ml_{\rm P}^2)$ for $\ln(Ml_{\rm P}) \gg 1$. On the other hand, a localized Hawking quantum is represented by a wavepacket with width of $O(Ml_{\rm P}^2)$ in r^*, since it has an energy of order $T_{\rm H} = 1/8\pi Ml_{\rm P}^2$ defined in the asymptotic region.

The point is that, given the state $|\Psi_k(M)\rangle = |\psi_k(M)\rangle|\phi_I\rangle$, the process in Eq. (6.23) occurs in the region $|r^*| \approx O(Ml_{\rm P}^2)$ (i.e. the region in which the effective gravitational potential starts shutting off toward large r^*) without involving deep interior of the zone $-r^* \gg Ml_{\rm P}^2$. In this region, information stored in the vacuum state is converted into that of a particle state outside the zone. More specifically, the information in the vacuum represented by the k index (which may also be viewed as a thermal bath of infrared modes, Eq. (6.13), though only in certain senses) is transferred into that in modes $a_{\rm far} \neq 0$, i.e. Hawking quanta, which have clear independent identities over the background spacetime. Due to energy conservation, this process is accompanied by the creation of ingoing negative energy excitations; however, they are not maximally entangled with the emitted Hawking quanta.

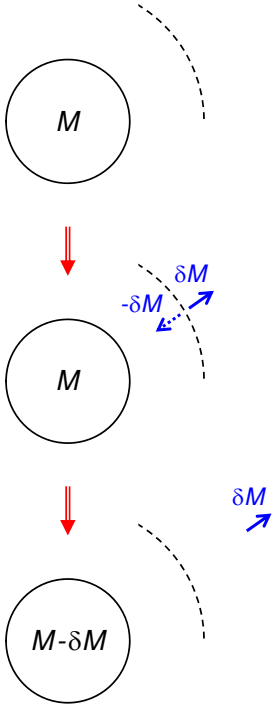


Figure 6.1: A schematic picture of the elementary Hawking emission process; time flows from the top to the bottom. The edge of the zone, i.e. the barrier region of the effective gravitational potential, is shown by a portion of a dashed circle at each moment in time. The emitted Hawking quanta as well as negative energy excitations are depicted by arrows (solid and dotted, respectively) although they are mostly s-waves.

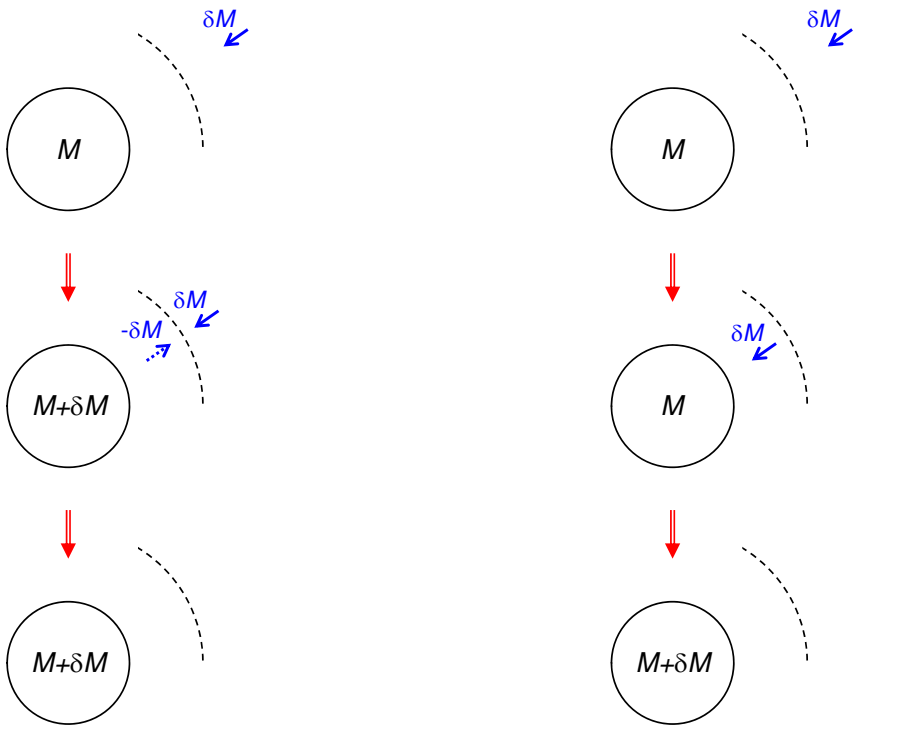
In Fig. 6.1, we depict schematically the elementary Hawking emission process described here. In the figure, we have denoted the emitted Hawking quanta as well as negative energy excitations by arrows, although they are mostly s-waves [72]. The discussion here makes it clear that the purifiers of the emitted Hawking quanta in the Hawking emission process are microstates which semiclassical theory describes as a vacuum. In particular, the emission process does *not* involve any excitation which, in the near horizon Rindler approximation, appears as a mode breaking entanglement between the two Rindler wedges necessary to keep the horizon smooth. Outgoing Hawking quanta emerge at the edge of the zone, living outside the applicability of the Rindler approximation. Ingoing negative energy excitations appear, in the Rindler approximation, as modes smooth in Minkowski space, which involve necessary entanglements between Rindler modes in the two wedges and have frequencies of order $1/Ml_{\rm P}^2$ in the Minkowski frame. Unlike what was considered in Ref. [4], and unlike what a "naive" interpretation of semiclassical theory might seem to suggest, Hawking quanta are not modes associated solely with one of the Rindler wedges (b

modes in the notation of Ref. [4]) nor outgoing Minkowski modes (a modes), which would appear to have high energies for observers who are freely falling into the black hole. This allows for avoiding the entropy argument for firewalls given in Ref. [4] as well as the typicality argument in Ref. [74].

In the discussion of the Hawking emission so far, we have assumed that a single emission of Hawking quanta as well as the associated creation of ingoing negative energy excitations occur in a black hole vacuum state consisting of $|\Psi_k(M)\rangle$'s, which are defined in the limit that the evolution effect is ignored. In reality, however, there are always of order $\ln(Ml_{\rm P})$ much of negative energy excitations in the zone, since the emission process occurs in every time interval of order $Ml_{\rm P}^2$ and the time it takes for a negative energy excitation to reach the stretched horizon is of order $Ml_{\rm P}^2 \ln(Ml_{\rm P})$ (both measured in the asymptotic region)—an evaporating black hole has an ingoing flux of negative energy excitations of entropy $\approx O(-\ln(Ml_{\rm P}))$ at all times. This flux of excitations modifies spacetime geometry from that of a Schwarzschild black hole; in particular, the geometry near the horizon is well described by the advanced/ingoing Vaidya metric [79]. Note that as discussed in Section 6.2.2, we may redefine our vacuum states to include these negative energy excitations, although we do not do it here.

Finally, it is instructive to consider the time reversal of the Hawking emission process. In this case, radiation coming from the far exterior region and outgoing negative energy excitations emitted from the stretched horizon meet around the edge of the zone; see Fig. 6.2(a). This results in a black hole state of mass given by the sum of the mass M of the original black hole (before emitting the negative energy excitations) and the energy δM of the incoming radiation. It is a "vacuum" state in the sense that there is no excitation in the zone except for those associated with a steady flux of outgoing negative energy excitations. We emphasize that this process is very different from what happens when generic incoming radiation of energy $\delta M \approx O(1/Ml_{\rm P}^2)$ is sent to a usual (i.e. evaporating, not anti-evaporating) black hole. In this case, the radiation enters into the zone without being "annihilated" by a negative energy excitation, which after hitting the stretched horizon will lead to a black hole state of mass $M + \delta M$; see Fig. 6.2(b). In fact, the process in Fig. 6.2(a) is a process which leads to a decrease of coarse-grained (or thermal) entropy, as implied by the fact that the coarse-grained entropy increases in the standard Hawking emission process [83]. In order for this to happen, therefore, the initial radiation and black hole state must be exponentially fine-tuned; otherwise, the radiation would simply propagate inward in the zone as depicted in Fig. 6.2(b) (although it can be subject to significant scattering by the effective gravitational potential at the time of the entrance). The origin of the conversion from radiation to vacuum degrees of freedom for such a fine-tuned initial state can be traced to the non-decoupling of the a and k indices discussed in Section 6.2.2.[16]

[16]If the black hole vacuum states are redefined as discussed in Section 6.2.2, the outgoing negative energy flux cannot be seen as excitations. The physics described here, however, will not change; in particular, only exponentially



(a) Time reversal of Fig. 6.1

(b) Generic incoming radiation

Figure 6.2: Time reversal of the Hawking emission process (a) as opposed to the process in which generic incoming radiation enters into the zone of a usual black hole (b). The former is an entropy decreasing process requiring an exponentially fine-tuned initial state, while the latter is a standard process respecting the (generalized) second law of thermodynamics.

Semiclassical (thermal) description

The expression in Eq. (6.21) implies that at an intermediate stage of the evolution, the information about the initial collapsing matter is encoded in the black hole microstates labeled by k and their entanglement with the rest of the system (which will later be transformed into the state of final-state Hawking radiation). Since semiclassical theory is incapable of describing the dynamics associated with the index k, it leads to apparent violation of unitarity *at all stages* of the black hole formation and evaporation processes. In particular, the state of the emitted Hawking quanta in each time interval of order $M(t)l_{\rm P}^2$ is given by the incoherent thermal superposition with temperature $1/8\pi M(t)l_{\rm P}^2$, making the final Hawking radiation state a mixed thermal state—this is an intrinsic limitation of the semiclassical description, which involves a coarse-graining.

To see in detail how thermal Hawking radiation in the semiclassical picture results from unitary

fine-tuned initial states allow for converting radiation to vacuum degrees of freedom around the edge of the zone.

evolution at the fundamental level, let us analyze the elementary Hawking emission process given in Eq. (6.25). Following Eq. (6.10), we consider the "semiclassical vacuum state" with a black hole of mass M, obtained after taking the maximally mixed ensemble of microstates:

$$\rho(M) = \frac{1}{e^{S_0(M)}} \sum_{k=1}^{e^{S_0(M)}} |\psi_k(M)\rangle|\phi_I\rangle\langle\psi_k(M)|\langle\phi_I|. \tag{6.30}$$

The evolution of this state under Eq. (6.25) is then given by

$$\rho(M) \to \frac{1}{e^{S_0(M)}} \sum_{k=1}^{e^{S_0(M)}} \sum_{i,i'} \sum_{k_i=1}^{e^{S_0(M-E_i)}} \sum_{k'_{i'}=1}^{e^{S_0(M-E_{i'})}} \alpha^k_{ik_i}\alpha^{k*}_{i'k'_{i'}} |\psi_{k_i}(M-E_i)\rangle|\phi_{I+i}\rangle\langle\psi_{k'_{i'}}(M-E_{i'})|\langle\phi_{I+i'}|. \tag{6.31}$$

Now, assuming that the microscopic dynamics of the vacuum degrees of freedom are generic, we expect using $S_0(M) = 4\pi M^2 l_{\rm P}^2$ that tracing out the black hole states leads to

$$\mathrm{Tr}\left[\frac{1}{e^{S_0(M)}} \sum_{k=1}^{e^{S_0(M)}} \sum_{k_i=1}^{e^{S_0(M-E_i)}} \sum_{k'_{i'}=1}^{e^{S_0(M-E_{i'})}} \alpha^k_{ik_i}\alpha^{k*}_{i'k'_{i'}} |\psi_{k_i}(M-E_i)\rangle\langle\psi_{k'_{i'}}(M-E_{i'})|\right] \approx \frac{1}{Z} g_i e^{-\frac{E_i}{T_{\rm H}}} \delta_{ii'}, \tag{6.32}$$

where $T_{\rm H} = 1/8\pi M l_{\rm P}^2$, $Z = \sum_i g_i e^{-E_i/T_{\rm H}}$, and g_i is a factor that depends on i. This allows us to write the reduced density matrix representing the exterior state after the evolution in Eq. (6.31) as

$$\rho_{\rm ext} \approx \frac{1}{Z} \sum_i g_i e^{-\frac{E_i}{T_{\rm H}}} |\phi_{I+i}\rangle\langle\phi_{I+i}|, \tag{6.33}$$

which is the result obtained in Hawking's original calculation, with g_i representing the gray-body factor calculable in the semiclassical theory [72].

The analysis given above elucidates why the semiclassical calculation sees apparent violation of unitarity in the Hawking emission process, i.e. why the final expression in Eq. (6.33) does not depend on microstates of the black hole, despite the fact that the elementary process in Eq. (6.25) is unitary, so that the coefficients $\alpha^k_{ik_i}$ depend on k. It is because the semiclassical calculation (secretly) deals with the mixed state, Eq. (6.30), from the beginning—states in semiclassical theory are maximal mixtures of black hole microstates labeled by vacuum indices, i.e. k's. By construction, the semiclassical theory cannot capture unitarity of detailed microscopic processes involving these indices, including the black hole formation and evaporation processes.

We finally discuss how the unitarity and thermal nature of the black hole evaporation process may appear in (thought) experiments, illuminating physical implications of the picture described here. Suppose we prepare an ensemble of a large number of black holes of mass M all of which are in an identical microstate k, and collect the Hawking quanta emitted from these black holes in a time interval of order $Ml_{\rm P}^2$. The quanta emitted from each black hole are then in the same quantum

state throughout the ensemble, so that a measurement of the spectrum of all the emitted quanta does *not* reveal the thermal property predicted by the semiclassical theory. On the other hand, if the members of the ensemble are in different microstates distributed randomly in k space, then the collection of the Hawking quanta emitted from all the black holes do exhibit the thermal nature consistent with the prediction of the semiclassical theory within the Hilbert space describing the quanta emitted from *each* black hole (which has dimension only of order unity).

What is the significance of the thermal nature for a single black hole, rather than an ensemble of a large number of black holes? If we form a black hole of mass M in a particular microstate k and collect all the Hawking quanta emitted throughout the evaporation process *without measuring them along the way*, then the state of the quanta contains the complete information about k, reflecting unitarity of the process at the fundamental level—the concept of thermality does not apply to this particular state *as a whole*. On the other hand, if an observer measures Hawking quanta emitted in each time interval of order $M(t)l_{\rm P}^2$, then the (incoherent) ensemble of measurement outcomes does exhibit the thermal nature as predicted by the semiclassical theory.[17] Since this is the kind of measurement that a realistic observer typically makes, the semiclassical theory can be said to provide a good prediction even for the outcome of (a series of) measurements a single observer performs on a single black hole.

6.2.5 Black hole mining—"microscopic" and semiclassical descriptions

It is known that one can accelerate the energy loss rate of a black hole faster than that of spontaneous Hawking emission by extracting its energy from the thermal atmosphere using a physical apparatus: the mining process. This acceleration occurs largely because the number of "channels" one can access increases by going into the zone—unlike the case of spontaneous Hawking emission, which is dominated by s-wave radiation, higher angular momentum modes can also contribute to the energy loss in this process [82]. Note that the rate of energy loss associated with *each* channel, however, is still the same order as that in the spontaneous Hawking emission process: energy of order $E \approx O(1/Ml_{\rm P}^2)$ is lost in each time interval of $t \approx O(Ml_{\rm P}^2)$, with E and t both defined in the asymptotic region. This fact will become important in Section 6.3 when we discuss the mining process as viewed from an infalling reference frame.

The information transfer associated with the mining process occurs in a similar way to that in the spontaneous Hawking emission process. An essential difference is that since the process involves

[17] In the more fundamental, many-world picture, this implies that the record of a physical observer who has "measured," or interacted with, emitted quanta in multiple moments shows a result consistent with the thermality predicted by the semiclassical theory. Note that a single branch in which such an observer lives does *not* in general contain the whole information about the initial black hole state k. The complete information about k (as well as that of the initial state of the observer) is contained only in a state given by a superposition of all possible branches resulting from interactions (and non-interactions) between the observer and quanta, representing all the possible "outcomes" the observer could have had (the probability distribution of which is consistent with thermality).

higher angular momentum modes, the negative energy excitations arising from backreactions can now be localized in angular directions. Specifically, consider a physical detector (or a system of detectors) located at a fixed Schwarzschild radial coordinate $r = r_{\rm d}$ within the zone, $r_{\rm s} < r_{\rm d} < R_{\rm Z}$. The detector then responds as if it is immersed in the thermal bath of blueshifted Hawking temperature $T(r_{\rm d})$, with $T(r)$ given by Eq. (6.20). Suppose the detector has the ground state $|d_0\rangle$ and excited states $|d_i\rangle$ $(i = 1, 2, \dots)$ playing the role of the "ready" state and pointer states, respectively, and that the proper energies needed to excite $|d_0\rangle$ to $|d_i\rangle$ are given by $E_{{\rm d},i}$. The mining process can then be written such that after a timescale of $t \approx O(Ml_{\rm P}^2)$ (as measured in the asymptotic region), the state of the combined black hole and detector system evolves as

$$|\psi_k(M)\rangle|d_0\rangle \to \sum_{i,a,k'} c^k_{iak'}|\psi_{a;k'}(M)\rangle|d_i\rangle, \tag{6.34}$$

where we have assumed, as in the discussion of "elementary" Hawking emission, that there are no excitations other than those directly associated with the process. The state $|\psi_{a;k'}(M)\rangle$ arises as a result of backreaction of the detector response; it contains a negative energy excitation a with energy $-E_a$, which is generally localized in angular directions. The coefficients $c^k_{iak'}$ are nonzero only if $E_a \approx E_{{\rm d},i}\sqrt{1 - 2Ml_{\rm P}^2/r_{\rm d}}$ within the uncertainty.

Once created, the negative energy excitations propagate inward, and after time of $t \approx r^*_{\rm d} - r^*_{\rm s}$ collide with the stretched horizon, where r^* is the tortoise coordinate in Eq. (6.28). This will make the black hole states relax as

$$|\psi_{a;k'}(M)\rangle \to \sum_{k_a} d^{ak'}_{k_a}|\psi_{k_a}(M - E_a)\rangle, \tag{6.35}$$

in the scrambling time of $t \approx O(Ml_{\rm P}^2 \ln(Ml_{\rm P}))$. As in the case of spontaneous Hawking emission, this relaxation process is unitary because the negative energy excitations carry negative entropies; i.e. for a fixed a, the index k' runs only over $1, \cdots, e^{S_0(M-E_a)} \ll e^{S_0(M)}$. The combination of Eqs. (6.34, 6.35) then yields

$$|\psi_k(M)\rangle|d_0\rangle \to \sum_{i,k_i} \alpha^k_{ik_i}|\psi_{k_i}(M - E_i)\rangle|d_i\rangle, \tag{6.36}$$

where $\alpha^k_{ik_i} = \sum_{a,k'} c^k_{iak'} d^{ak'}_{k_i}$ and $E_i = E_{{\rm d},i}\sqrt{1 - 2Ml_{\rm P}^2/r_{\rm d}}$. This represents a microscopic, unitary description of the elementary mining process.

In the description given above, we have separated the detector state from the state of the black hole, but in a treatment fully consistent with the notation in earlier sections, the detector itself must be viewed as excitations over $|\psi_k(M)\rangle$. After the detector response process in Eq. (6.34), these excitations can be entangled with Hawking quanta emitted earlier, reflecting the fact that the detector can extract information from the black hole. Since the detector can now be put deep in

the zone, in which the Rindler approximation is applicable, this implies that excitations localized within the Rindler wedge corresponding to the region $r > r_{\rm s}$ are entangled with early Hawking radiation. Does this lead to firewalls as discussed in Ref. [4]? The answer is no. The excitations describing the detector are, in the near horizon Rindler approximation, those of modes that are smooth in Minkowski space (a modes in the notation of Ref. [4]). Likewise, modes representing negative energy excitations arising from the backreactions are also ones smooth in Minkowski space. Excitations of these modes, of course, *do* perturb the black hole system, which can indeed be significant if the detector is held very close to the horizon. This effect, however, is caused by physical interactions between the detector and vacuum degrees of freedom, and is confined in the causal future of the interaction event. This is not the firewall phenomenon.

The semiclassical description of the mining process in Eq. (6.36) is obtained by taking maximal mixture for the vacuum indices. Specifically, the semiclassical state before the process starts is given by

$$\rho(M) = \frac{1}{e^{S_0(M)}} \sum_{k=1}^{e^{S_0(M)}} |\psi_k(M)\rangle|d_0\rangle\langle\psi_k(M)|\langle d_0|. \tag{6.37}$$

The evolution of this state under Eq. (6.36) is then

$$\rho(M) \to \frac{1}{e^{S_0(M)}} \sum_{k=1}^{e^{S_0(M)}} \sum_{i,i'} \sum_{k_i=1}^{e^{S_0(M-E_i)}} \sum_{k'_{i'}=1}^{e^{S_0(M-E_{i'})}} \alpha^k_{ik_i}\alpha^{k*}_{i'k'_{i'}} |\psi_{k_i}(M-E_i)\rangle|d_i\rangle\langle\psi_{k'_{i'}}(M-E_{i'})|\langle d_{i'}|. \tag{6.38}$$

This leads to the density matrix describing the detector state after the process

$$\rho_{\rm d} = \sum_{i,i'} \gamma_{ii'}\, |d_i\rangle\langle d_{i'}|, \tag{6.39}$$

where

$$\gamma_{ii'} = {\rm Tr}\left[\frac{1}{e^{S_0(M)}} \sum_{k=1}^{e^{S_0(M)}} \sum_{k_i=1}^{e^{S_0(M-E_i)}} \sum_{k'_{i'}=1}^{e^{S_0(M-E_{i'})}} \alpha^k_{ik_i}\alpha^{k*}_{i'k'_{i'}} |\psi_{k_i}(M-E_i)\rangle\langle\psi_{k'_{i'}}(M-E_{i'})|\right]. \tag{6.40}$$

Assuming that the microscopic dynamics of the vacuum degrees of freedom are generic, $\gamma_{ii'}$ is expected to take the form

$$\gamma_{ii'} \approx \frac{1}{Z} f_i e^{-\frac{E_{{\rm d},i}}{T(r_{\rm d})}}\, \delta_{ii'}, \tag{6.41}$$

where $Z = \sum_i f_i e^{-E_{{\rm d},i}/T(r_{\rm d})}$, and f_i is the detector response function reflecting intrinsic properties of the detector under consideration. This implies that in the semiclassical approximation, the final detector state does not have any information about the original black hole microstate, despite the fact that the fundamental process in Eq. (6.36) is, in fact, unitary.

6.2.6 The fate of an infalling object

We now discuss how an object falling into a black hole is described in a distant reference frame. As we have seen, having a well-defined black hole geometry requires a superposition of an enormous number of energy-momentum eigenstates. While the necessary spreads in energy and momentum are small when measured in the asymptotic region, the spreads of *local* energy and momentum (i.e. those measured by local approximately static observers) are large in the region close to the horizon, because of large gravitational blueshifts. This makes the local temperature $T(r)$ associated with the vacuum degrees of freedom, Eq. (6.20), very high near the horizon. We expect that the semiclassical description becomes invalid when this temperature exceeds the string (cutoff) scale, $T(r) \gtrsim 1/l_*$. Namely, semiclassical spacetime exists only in the region

$$r > r_{\rm s} = 2Ml_{\rm P}^2 + O\left(\frac{l_*^2}{Ml_{\rm P}^2}\right), \tag{6.42}$$

where $r_{\rm s}$ is identified as the location of the stretched horizon. The same conclusion can also be obtained by demanding that the gravitational thermal entropy stored in the region where the semiclassical spacetime picture is applicable is a half of the Bekenstein-Hawking entropy, $\mathcal{A}/8l_{\rm P}^2$, as discussed in footnote 13.

Let us consider that an object is dropped from $r = r_0$ with vanishing initial velocity, where $r_0 - 2Ml_{\rm P}^2 \approx O(Ml_{\rm P}^2) > 0$. It then freely falls toward the black hole and hits the stretched horizon at $r = r_{\rm s}$ in Schwarzschild time of about $4Ml_{\rm P}^2 \ln(Ml_{\rm P}^2/l_*)$. Before it hits the stretched horizon, the object is described by a and $a_{\rm far}$, the indices labeling field and string theoretic excitations over the semiclassical background spacetime. After hitting the stretched horizon, the information about the object will move to the index $\bar{a}$, labeling excitations of the stretched horizon. The information about the fallen object will then stay there, at least, for the thermalization (or scrambling) time of the stretched horizon, of order $Ml_{\rm P}^2 \ln(Ml_{\rm P})$. This allows for avoiding the inconsistency of quantum cloning in black hole physics [75]. Finally, the information in $\bar{a}$ will further move to k, which can (later) be extracted by an observer in the asymptotic region via the Hawking emission or mining process, as described in the previous two subsections.

We note that the statement that an object is in the semiclassical regime (i.e. represented by indices a and $a_{\rm far}$) does *not* necessarily mean that it is well described by semiclassical *field* theory. Specifically, it is possible that stringy effects become important before the object hits the stretched horizon. As an example, consider dropping an elementary particle of mass m ($\ll 1/l_*$) from $r = r_0$ with zero initial velocity. (Here, by elementary we mean that there is no composite structure at lengthscale larger than l_*.) The local energy and local radial momentum of the object will then

vary, as it falls, as:

$$E_{\rm loc} = m\sqrt{\frac{1-\frac{2Ml_{\rm P}^2}{r_0}}{1-\frac{2Ml_{\rm P}^2}{r}}}, \qquad p_{\rm loc} = -m\sqrt{\frac{\frac{2Ml_{\rm P}^2}{r}-\frac{2Ml_{\rm P}^2}{r_0}}{1-\frac{2Ml_{\rm P}^2}{r}}}. \tag{6.43}$$

The values of $E_{\rm loc} \approx -p_{\rm loc}$ get larger as r gets smaller, and for $m \gg 1/Ml_{\rm P}^2$ (which we assume here) become of order $1/l_*$ before the object hits the stretched horizon, i.e. at

$$r - 2Ml_{\rm P}^2 \simeq 2Ml_{\rm P}^2(ml_*)^2\left(1-\frac{2Ml_{\rm P}^2}{r_0}\right). \tag{6.44}$$

The Schwarzschild time it takes for the object to reach this point is only about $-4Ml_{\rm P}^2\ln(ml_*)$, much smaller than the time needed to reach the stretched horizon, $4Ml_{\rm P}^2\ln(Ml_{\rm P}^2/l_*)$. After the object reaches this point, i.e. when $E_{\rm loc} \approx -p_{\rm loc} \gtrsim 1/l_*$, stringy effects might become important; specifically, its Lorentz contraction saturates and transverse size grows with $E_{\rm loc}$ [84]. Note that this dependence of the description on the boost of a particle does not necessarily mean violation of Lorentz invariance—physics can still be fully Lorentz invariant.[18]

A schematic picture for the fate of an infalling object described above is given in Fig. 6.3. In a distant reference frame, the semiclassical description of the object is applicable only until it hits the stretched horizon, after which it is represented as excitations of the stretched horizon. On the other hand, according to general relativity (or the equivalence principle), the falling object does not experience anything other than smooth empty spacetime when it crosses the horizon, except for effects associated with curvature, which are very small for a black hole of mass $M \gg 1/l_{\rm P}$. If this picture is correct, then we expect there is a way to reorganize the dynamics of the stretched horizon such that the general relativistic smooth interior of the black hole becomes manifest. In the complementarity picture, this is achieved by performing an appropriate reference frame change. We now move on to discuss this issue.

6.3 Black Hole—An Infalling Description

In order to describe the fate of an infalling object using low energy language after it crosses the Schwarzschild horizon, we need to perform a change of the reference frame from a distant one, which we have been considering so far, to an infalling one which falls into the black hole with the object. In general, studying this issue is complicated by the fact that the general and precise

[18] It is illuminating to consider how these stringy effects appear in a two-particle scattering process in Minkowski space. For $\sqrt{s} \lesssim 1/l_*$, where s is the Mandelstam variable, there is a reference frame in which energies/momenta of *both* particles are smaller than $1/l_*$, guaranteeing that these effects are not important in the process. For $\sqrt{s} > 1/l_*$, on the other hand, at least one particle has an energy/momentum larger than $1/l_*$ in *any* reference frame, suggesting that stringy effects become important in scattering with such high $\sqrt{s}$.

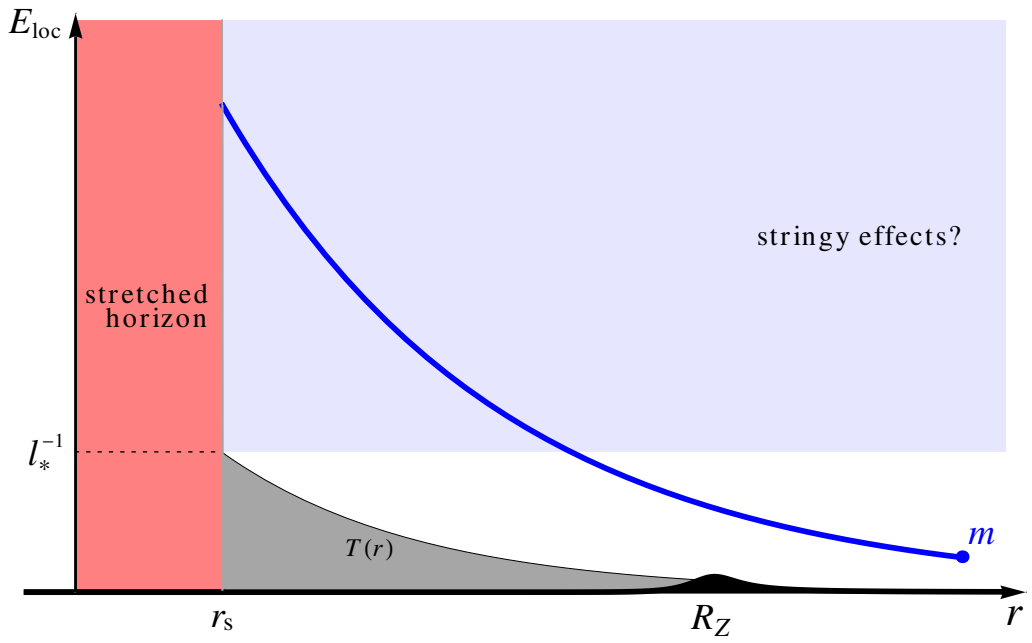


Figure 6.3: A schematic depiction of the fate of an elementary particle of mass m ($1/Ml_{\rm P}^2 \ll m \ll 1/l_*$) dropped into a black hole, viewed in a distant reference frame. As the particle falls, its local energy blueshifts and exceeds the string/cutoff scale $1/l_*$ before it hits the stretched horizon. After this point, stringy effects could become important, although the semiclassical description of the object may still be applicable. The object hits the stretched horizon at a Schwarzschild time of about $4Ml_{\rm P}^2 \ln(Ml_{\rm P}^2/l_*)$ after the drop. After this time, the semiclassical description of the object is no longer applicable, and the information about the object will be encoded in the index $\bar{a}$, representing excitations of the stretched horizon. (This information will further move to the vacuum index k later, so that it can be extracted by an observer in the asymptotic region via the Hawking emission or mining process.)

formulation of complementarity is not yet known, but we may still explore the expected physical picture based on some general considerations.

The aim of this section is to argue that the existence of interior spacetime, as suggested by general relativity, does not contradict the unitarity of the Hawking emission and black hole mining processes, as described in the previous section in a distant reference frame. We do this by first arguing that there exists a reference frame—an infalling reference frame—in which the spacetime around a point on the Schwarzschild horizon appears as a large nearly flat region, with the curvature lengthscale of order $Ml_{\rm P}^2$. This is a reference frame whose origin falls freely from rest from a point sufficiently far from the black hole. We discuss how the description based on this reference frame is consistent with that in the distant reference frame, despite the fact that they apparently look very different, for example in spacetime locations of the vacuum degrees of freedom.

We then discuss how the system is described in more general reference frames, in particular a reference frame whose origin falls from rest from a point close to the Schwarzschild horizon. We

will also discuss (non-)relations of black hole mining by a near-horizon static detector and the—seemingly similar—Unruh effect in Minkowski space. The discussion in this section illuminates how general coordinate transformations may work at the level of full quantum gravity, beyond the approximation of quantum field theory in curved spacetime.

6.3.1 Emergence of interior spacetime—free fall from a distance

What does a reference frame really mean? According to the general complementarity picture described in Section 6.1, it corresponds to a foliation of a portion of spacetime which a single (hypothetical) observer can access. As discussed there, the procedure to erect such a reference frame should not depend on the background geometry in order for the framework to be applicable generally, and there is currently no precise, established formulation to do that (although there are some partially successful attempts; see, e.g., Ref. [8]). Here we focus only on classes of reference frames describing the same system with a fixed black hole background. This limitation allows us to bypass many of the issues arising when we consider the most general application of the complementarity picture.

In this subsection, we consider a class of reference frames which we call infalling reference frames. We argue that a reference frame in this class makes it manifest that the spacetime near the origin of the reference frame appears as a large approximately flat region when it crosses the Schwarzschild horizon, up to corrections from curvature of lengthscale $Ml_{\rm P}^2$. We discuss how the interior spacetime of the black hole can emerge through the complementarity transformation representing a change of reference frame from the distant to infalling ones. Consistency of the infalling picture described here with the distant frame description in Section 6.2 will be discussed in more detail in the next subsection.

We consider a reference frame associated with a freely falling (local Lorentz) frame, with its spatial origin p_0 following the worldline representing a hypothetical observer [33, 8]. In particular, we let the origin of the reference frame, p_0, follow the trajectory of a timelike geodesic, representing the observer who is released from rest at $r = r_0$, with r_0 sufficiently far from the Schwarzschild horizon, $r_0 - 2Ml_{\rm P}^2 \gtrsim Ml_{\rm P}^2$. According to the complementarity hypothesis, the system described in this reference frame does not have a (hot) stretched horizon at the location of the Schwarzschild horizon when p_0 crosses it. (The stretched horizon must have existed around the Schwarzschild horizon when p_0 was far away, $r_{p_0} - 2Ml_{\rm P}^2 \gtrsim O(Ml_{\rm P}^2)$, because the description in those earlier times must be approximately that of a distant reference frame, i.e. that discussed in the previous section.) In particular, the region around p_0 must appear approximately flat, i.e. up to small effects from curvature of order $1/M^2 l_{\rm P}^4$, until p_0 approaches the singularity.

In this infalling description, we expect that a "horizon" signaling the breakdown of the semi-classical description lies in the directions associated with "past-directed and inward" light rays

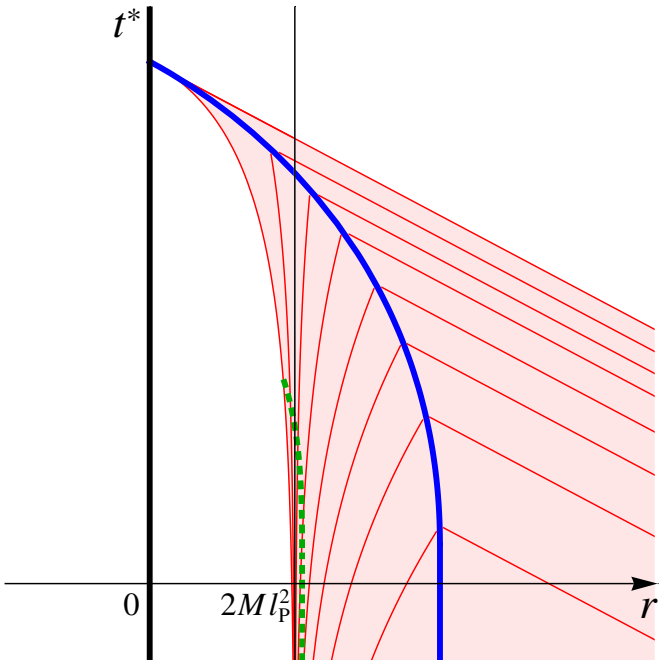


Figure 6.4: A sketch of an infalling reference frame in an Eddington-Finkelstein diagram: the horizontal and vertical axes are r and $t^* = t + r^* - r$, respectively, where r^* is the tortoise coordinate. The thick (blue) line denotes the spacetime trajectory of the origin, p_0, of the reference frame, while the thin (red) lines represent past-directed light rays emitted from p_0. The shaded area is the causal patch associated with the reference frame, and the dotted (green) line represents the stretched "horizon" as viewed from this reference frame.

(the directions with increasing r and decreasing t after p_0 crosses $r = 2Ml_{\rm P}^2$) as viewed from p_0; see Fig. 6.4.[19] As in the stretched horizon in a distant reference frame, this "horizon" emerges because of the "squeezing" of equal-time hypersurfaces; in particular, an observer following the trajectory of p_0 may probe only a tiny region near the Schwarzschild horizon for signals arising from this surface. (Note that $-r$ plays a role of time inside the Schwarzschild horizon.) Considering angular directions, this "horizon" has an area of order $M^2 l_{\rm P}^4$, and can be regarded as being located at distances of order $Ml_{\rm P}^2$ away from p_0 (with an appropriately defined distance measure on generic equal-time hypersurfaces in the infalling reference frame; see Section 6.3.2).

In analogy with the case of a distant frame description, we denote basis states for the general microstates in an infalling reference frame (before p_0 reaches the singularity) as

$$|\Psi_{\bar{\alpha}\,\alpha\,\alpha_{\rm far};\kappa}(M)\rangle, \tag{6.45}$$

where $\bar{\alpha}$ labels the excitations of the "horizon," and α, and $\alpha_{\rm far}$ are the indices labeling the

[19]This "horizon," as viewed from an infalling reference frame, should not be confused with the stretched, or Schwarzschild, horizon as viewed from a distant reference frame.

semiclassical excitations near and far from the black hole, conveniently defined; κ is the vacuum index in an infalling reference frame, representing degrees of freedom that cannot be resolved by semiclassical operators.[20] The complementarity transformation provides a map from the basis states in a distant description, Eq. (6.4), to those in an infalling description, Eq. (6.45), and vice versa. The general form of this transformation can be quite complicated, depending, e.g., on equal-time hypersurfaces taken in the two descriptions (which are in turn related with the general procedure of erecting reference frames by standard coordinate transformations within each causal patch). Here we consider how various indices are related under the transformation, focusing on the near black hole region.

Imagine that equal-time hypersurfaces in the two—distant and infalling—reference frames agree at some time $t = t_0$ in the spacetime region near but outside the surface where the stretched horizon exists if viewed from the distant reference frame. (Note that the stretched horizon has physical substance only in a distant reference frame.) We are interested in how basis states in the two descriptions transform between each other in the timescale of the fall of the infalling reference frame. The time here can be taken as the proper time at p_0 in each reference frame [33, 8], which is approximately the Schwarzschild time for the distant reference frame. In this case, the relevant timescale is $t - t_0 \lesssim O(Ml_{\rm P}^2 \ln(Ml_{\rm P}))$ in the distant reference frame, while $t - t_0 \lesssim O(Ml_{\rm P}^2)$ in the infalling reference frame.

As discussed in Section 6.2.6, in the distant reference frame, an object dropped from some r_0 with $r_0 - 2Ml_{\rm P}^2 \approx O(Ml_{\rm P}^2)$ is first represented by a and then by $\bar{a}$ after it hits the stretched horizon. On the other hand, in the infalling frame, the object is represented by the index α throughout, first as a semiclassical excitation outside the Schwarzschild horizon and then as a semiclassical excitation inside the Schwarzschild horizon, implying that the object does not find anything special at the horizon. Here, we have assumed that p_0 follows (approximately) the trajectory of the falling object. This suggests that a portion of the α index representing excitations in the interior of the black hole is transformed into the $\bar{a}$ index in the distant description (and vice versa) under the complementarity transformation; i.e., the interior of the black hole accessible from the infalling reference frame is encoded in the excitations of the stretched horizon in the distant reference frame. Note that the amount of information needed to reconstruct the interior (in the semiclassical sense) is much smaller than the Bekenstein-Hawking entropy [11, 46]—the logarithm of the dimension of the relevant Hilbert space is of order $(\mathcal{A}/l_{\rm P}^2)^q$ with $q < 1$.

In the exterior spacetime region, the portion of the α index representing excitations there, as well as the $\alpha_{\rm far}$ index, are mapped to the corresponding a and $a_{\rm far}$ indices, and vice versa (after matching the equal-time hypersurface in the two descriptions through appropriate time evolutions). Because equal-time hypersurfaces foliate the causal patch, excitations in the far exterior

[20] After p_0 hits the singularity, the system as viewed from the infalling reference frame can only be represented by "singularity states": intrinsically quantum gravitational states that do not allow for a spacetime interpretation [33].

region naturally have trans-Planckian energies in the infalling description. However, as discussed in Section 6.2.6, this does not mean that the semiclassical description is invalid—objects may still be described as excitations in the semiclassical spacetime, although stringy effects may become important. Indeed, we expect that the semiclassical description is applicable in the far exterior region even in the infalling reference frame, because of the absence of the "squeezing" effect described above which leads to the breakdown of the semiclassical picture.

We emphasize that the construction of the interior spacetime described here does not suffer from the paradoxes discussed in Refs. [4, 73, 74]. By labeling states in terms of excitations, we are in a sense representing the interior spacetime already in the distant description. (The interpretation, however, is different. In the distant description, the relevant excitations must be regarded as those of the stretched horizon.) In fact, we do not find any inconsistency in postulating that the dynamics of an infalling object is described by the corresponding Hamiltonian in the semiclassical theory in a sufficiently small region around p_0, to the extent that microscopic details of interactions with κ degrees of freedom are neglected. Namely, we do not find any inconsistency in postulating that physics at the classical level is well described by general relativity.

Finally, we discuss where the fine-grained vacuum degrees of freedom represented by κ must be viewed as being located in the infalling description. Because of the lack of an obvious static limit, it is not straightforward to answer to this question. Nevertheless, it seems natural to expect, in analogy with the case of a distant description, that most of the degrees of freedom are located close to the "horizon" (in terms of a natural distance measure in which the distance between the "horizon" and p_0 is of order $Ml_{\rm P}^2$). In fact, we expect that the number of κ degrees of freedom existing around p_0 within a distance scale sufficiently smaller than $Ml_{\rm P}^2$ is of $O(1)$ or smaller, since the time and length scales of the system characterizing local deviations from Minkowski space (as viewed from the infalling reference frame) are both of order $Ml_{\rm P}^2$. As in the case of the distant description, we expect that the κ degrees of freedom do not extend significantly to the far exterior region, since the existence of the black hole does not affect the spacetime there much.[21]

6.3.2 Consistency between the distant and infalling descriptions

In analyzing a black hole system in a distant reference frame, we argued that the microscopic information about the black hole, represented by the k index, is distributed according to the gravitational thermal entropy calculated using semiclassical field theory. In particular, on the

[21]Note that the descriptions in the two reference frames are already different at the semiclassical level. For example, the backreaction of a detector click in a distant reference frame is described as an absorption of a particle in the thermal bath, while in an infalling reference frame it is described as an emission of a particle, with the difference arising from different definitions of energy in the two reference frames [85]. The reference frame dependence discussed here is much more drastic, however—the spacetime locations of physical degrees of freedom are different in the two reference frames.

Schwarzschild (or stretched) horizon, this information has a Planckian density: one qubit per area of order $l_{\rm P}^2$ on the horizon (or per volume of order $l_{\rm P}^3$ if we take into account the "thickness" of the stretched horizon, $\sim l_{\rm P}$). On the other hand, we have just argued that in an infalling reference frame, the spacetime distribution of the microscopic information (now represented by the κ index) is different. In particular, the spatial density of the information around the Schwarzschild horizon, when the origin of the reference frame passes through it, is very small: one qubit per volume of order $(Ml_{\rm P}^2)^3$. How can we reconcile these two seemingly very different perspectives?

In this subsection, we consider this problem and argue that despite the fact that the spacetime distribution of the microscopic information depends on the reference frame one chooses to describe the system, the answers to any operationally well-defined question one obtains in different reference frames are consistent with each other. As an example most relevant to our discussion, we consider a physical detector hovering at a constant Schwarzschild radius $r = r_{\rm d}$ ($> 2Ml_{\rm P}^2$). In a distant description, the spatial density of the microscopic information, represented by k, is large at the location of the detector when $r_{\rm d} - 2Ml_{\rm P}^2 \ll Ml_{\rm P}^2$. Such a detector (or a system of detectors) can thus be used for black hole mining: accelerated extraction of energy and information from the black hole. In an infalling reference frame, however, the density of the microscopic information, represented by κ, is very small at the detector location, at least when the origin of the reference frame, p_0, passes nearby. This implies that the rate of extracting information from spacetime cannot be much faster than $1/Ml_{\rm P}^2$ around p_0 in the infalling description, reflecting the fact that the spacetime appears approximately flat there. How are these two descriptions consistent?

In the distant description, the rate of extracting microscopic information about the black hole is at most of order one qubit per Schwarzschild time $1/T_{\rm H} = 8\pi Ml_{\rm P}^2$ *per channel*, regardless of the location of the detector [82]—the acceleration of information extraction occurs not because of a higher speed of information extraction in each channel but because of an increased number of channels available by immersing the detector deep into the zone. This implies that each single detector, which we define to act on a single channel, "clicks" once (i.e. extracts of $O(1)$ qubits) per a Schwarzschild time of order $8\pi Ml_{\rm P}^2$.

Now, consider describing such a detector in an infalling reference frame whose origin p_0 is released at $r = 2Ml_{\rm P}^2 + O(Ml_{\rm P}^2)$ from rest, at an angular location close to the detector. To understand the relevant kinematics, we adopt the near-horizon Rindler approximation: for $r > 2Ml_{\rm P}^2$

$$\rho \approx 2\sqrt{2Ml_{\rm P}^2(r - 2Ml_{\rm P}^2)}, \qquad \omega \approx \frac{t}{4Ml_{\rm P}^2}, \tag{6.46}$$

in terms of which the metric is given by

$$ds^2 \approx -\rho^2 d\omega^2 + d\rho^2 + r(\rho)^2 d\Omega. \tag{6.47}$$

As is well-known, this metric can be written in the Minkowski form

$$ds^2 \approx -dT^2 + dZ^2 + r(T,Z)^2 d\Omega, \tag{6.48}$$

by introducing the coordinates

$$T = \rho \sinh \omega, \qquad Z = \rho \cosh \omega, \tag{6.49}$$

which can be extended into the $r < 2Ml_{\rm P}^2$ region. Our setup corresponds to the situation in which the detector follows a trajectory of a constant ρ:

$$\rho = \rho_{\rm d} \ll Ml_{\rm P}^2, \tag{6.50}$$

while the origin of the reference frame p_0—or the (fictitious) observer—is at a constant Z:

$$Z = Z_{\rm o} \approx O(Ml_{\rm P}^2). \tag{6.51}$$

Note that while we *approximate* the geometry by flat space, given by Eq. (6.47) or (6.48), the actual system has small nonzero curvature with lengthscale of order $Ml_{\rm P}^2$.

As discussed above, the detector extracts an $O(1)$ amount of information in each time interval of

$$\Delta\omega \approx O\left(\frac{1}{4Ml_{\rm P}^2 T_H}\right) \approx O(1), \tag{6.52}$$

while the "observer," p_0, and the detector meet (or pass by each other) at

$$\begin{pmatrix} \omega \\ \rho \end{pmatrix} = \begin{pmatrix} \operatorname{arccosh} \frac{Z_{\rm o}}{\rho_{\rm d}} \\ \rho_{\rm d} \end{pmatrix} \equiv \begin{pmatrix} \omega_* \\ \rho_* \end{pmatrix}. \tag{6.53}$$

This implies that in the Minkowski coordinates—i.e. as viewed from the infalling observer p_0—the detector clicks only once in each time/space interval of

$$\Delta T \approx \Delta\omega \frac{\partial T}{\partial \omega}\bigg|_{\omega=\omega_*,\rho=\rho_*} \approx Z_{\rm o} \approx O(Ml_{\rm P}^2), \tag{6.54}$$

$$\Delta Z \approx \Delta\omega \frac{\partial Z}{\partial \omega}\bigg|_{\omega=\omega_*,\rho=\rho_*} \approx Z_{\rm o} \approx O(Ml_{\rm P}^2), \tag{6.55}$$

around p_0. This is precisely what we expect from the equivalence principle: the spacetime appears approximately flat when viewed from an infalling observer, up to curvature effects with lengthscale of $Ml_{\rm P}^2$. While the detector clicks of order $\ln(Ml_{\rm P})$ times within the causal patch of the infalling reference frame, all these clicks occur at distances of order $Ml_{\rm P}^2$ away from p_0, where we expect a higher density of κ degrees of freedom. The two descriptions—distant and infalling—are therefore

consistent, despite the fact that the spacetime distributions of the microscopic information about the black hole—represented by k and κ, respectively—are different in the two reference frames.

While we have so far discussed the case in which a physical detector is located close to the Schwarzschild horizon, the conclusion is the same in the case of spontaneous Hawking emission. In this case, since Hawking particles appear as semiclassical excitations only at $r - 2Ml_{\rm P}^2 \gtrsim Ml_{\rm P}^2$ with local energies of order $1/Ml_{\rm P}^2$, the consistency of the two descriptions is in a sense obvious. Alternatively, one can regard this case as the $\rho_{\rm d} \approx Ml_{\rm P}^2$ limit of the previous analysis. While the Rindler approximation is strictly valid only for ρ sufficiently smaller than $Ml_{\rm P}^2$, qualitative results are still valid for $\rho_{\rm d} \approx Ml_{\rm P}^2$; in particular, the estimates in Eqs. (6.54, 6.55) are valid at an order of magnitude level.

6.3.3 Other reference frames—free fall from a nearby point

In this subsection, we consider how the black hole is described in a class of reference frames whose origin follows a timelike geodesic released from rest at $r = r_0$, where r_0 is *close to* the Schwarzschild horizon, $r_0 - 2Ml_{\rm P}^2 \ll Ml_{\rm P}^2$.[22] We argue that the description in these reference frames does not look similar to either the distant or infalling description discussed before, and yet it is consistent with both of them.[23]

To understand how the black hole appears in such a reference frame, let us consider a setup similar to that in Section 6.3.2—a physical detector hovering at a constant Schwarzschild radius $r = r_{\rm d}$—and see how this system is described in the reference frame. As in Section 6.3.2, we may adopt the Rindler approximation, in which Eq. (6.51) is now replaced by

$$Z = Z_{\rm o} \ll Ml_{\rm P}^2. \tag{6.56}$$

This implies that as viewed from this reference frame, the detector clicks once in each time/space interval of

$$\Delta T \approx \Delta Z \approx Z_{\rm o} \ll Ml_{\rm P}^2. \tag{6.57}$$

Here, we have assumed that $\rho_{\rm d} < Z_{\rm o}$. Since each detector click extracts an $O(1)$ amount of information from spacetime, which we expect not to occur in Minkowski space, this implies that the spacetime cannot be viewed as approximately Minkowski space over a region beyond lengthscale $Z_{\rm o}$. In particular, in contrast with the case in an infalling reference frame (with $Z_{\rm o} \gtrsim O(Ml_{\rm P}^2)$),

[22] In a full geometry in which the black hole is formed by collapsing matter, the trajectory of the origin, p_0, of such a reference frame corresponds to a fine-tuned one in which p_0 stays near outside of the Schwarzschild horizon for long time due to large outward velocities at early times. (Here, we have focused only on the relevant branch in the full quantum state; see, e.g., footnote 4.)

[23] Note that we use the term "infalling reference frame" exclusively for reference frames discussed in Sections 6.3.1 and 6.3.2, i.e. the ones in which p_0 starts from rest at r_0 with $r_0 - 2Ml_{\rm P}^2 \gtrsim O(Ml_{\rm P}^2)$.

the spacetime region around p_0 in this reference frame does not appear nearly flat over lengthscale of $Ml_{\rm P}^2$ when p_0 crosses the Schwarzschild horizon.

At a technical level, this difference arises from the fact that the relative boost of p_0 with respect to the distant reference frame when p_0 approaches the detector

$$\gamma = \frac{1}{\sqrt{1-v_{\rm rel}^2}} = \sqrt{\frac{1-\frac{2Ml_{\rm P}^2}{r_0}}{1-\frac{2Ml_{\rm P}^2}{r_d}}}, \tag{6.58}$$

is very different in the two reference frames. In an infalling reference frame γ is huge, $\approx O(Ml_{\rm P}^2/\rho_{\rm d})$, while in the reference frame considered here $\gamma \approx O(Z_{\rm o}/\rho_{\rm d})$, which is not as large as that in the infalling case. In the infalling reference frame of Sections 6.3.1 and 6.3.2, the huge boost of $\gamma \approx O(Ml_{\rm P}^2/\rho_{\rm d})$ "stretched" the interval between detector clicks to time/length scales of order $Ml_{\rm P}^2$. Here, this "stretching" makes only a small region around p_0, with lengthscale of order $Z_{\rm o}$ ($\ll Ml_{\rm P}^2$), look nearly flat at any given time.

We may interpret this result to mean that in the reference frame under consideration, the "horizon" (as viewed from this reference frame) is located at a distance of order $Z_{\rm o}$ away from p_0, so that detector clicks occur near or "on" this surface. (In the latter case, the detector click events must be viewed as occurring in the regime outside the applicability of the semiclassical description; in particular, they can only be described as complicated quantum gravitational processes occurring on the "horizon.") Since we expect that microscopic information about the black hole (analogous to k and κ in the distant and infalling reference frames, respectively) is located near and on the "horizon," there is no inconsistency that detector clicks extract microscopic information from the black hole.

One might be bothered by the fact that in this reference frame spacetime near the Schwarzschild horizon does not appear large, $\approx O(Ml_{\rm P}^2)$, nearly flat space, and consider that this implies the non-existence of a large black hole interior as suggested by general relativity. This is, however, not correct. The existence of *a* reference frame in which spacetime around the Schwarzschild horizon appears as a large nearly flat region—in particular, the existence of an infalling reference frame discussed in Sections 6.3.1 and 6.3.2—already ensures that an infalling physical object/observer does not experience anything special, e.g. firewalls, when it/he/she crosses the Schwarzschild horizon. The analysis given here simply says that the spacetime around the Schwarzschild horizon does *not always* appear as a large nearly flat region, even in a reference frame whose origin falls freely into the black hole. This extreme relativeness of descriptions is what we expect from complementarity.

6.3.4 (Non-)relations with the Unruh effect in Minkowski space

It is often thought that the system described above is similar to an accelerating detector existing in Minkowski space, based on a similarity of geometries between the two setups. If this were true

at the full quantum level, it would mean that the description in an *inertial* reference frame in Minkowski space must possess a "horizon," at which the semiclassical description of the system breaks down. Does this make sense?

Here we argue that physics of a detector held near the Schwarzschild horizon, given above in Section 6.3.3, is, in fact, different from that of an accelerating detector in Minkowski space. The intuition that the two must be similar comes from the (wrong) perception that the detector located near the Schwarzschild horizon feels a high blueshifted Hawking temperature, $\approx 1/\rho_{\rm d} \gg 1/Ml_{\rm P}^2$, which makes the detector click at a high rate, while the spacetime curvature there is very small, with lengthscale $\approx Ml_{\rm P}^2$, so that such a tiny curvature must not affect the system. This intuition, however, is flawed by mixing up two different pictures—the system as viewed at the location of the detector and as viewed in the asymptotic region.

Suppose we represent all quantities as defined in the asymptotic region. The temperature a detector feels is then of order $1/Ml_{\rm P}^2$ and the timescale for detector clicks is $T \approx O(Ml_{\rm P}^2)$ for *any* $r_{\rm d} > 2Ml_{\rm P}^2$. On the other hand, the energy density of the black hole region is of order $M/(Ml_{\rm P}^2)^3$, so that the curvature lengthscale L is estimated as

$$\frac{1}{L^2} \sim G_{\rm N} \frac{M}{(Ml_{\rm P}^2)^3} \sim \frac{1}{(Ml_{\rm P}^2)^2}. \tag{6.59}$$

This implies that

$$T \sim L \sim O(Ml_{\rm P}^2); \tag{6.60}$$

namely, curvature is expected to give *an* $O(1)$ *effect* on the dynamics of the detector response.

The same conclusion can also be reached when we represent all the quantities in the static frame at the detector location. In this case, the temperature the detector feels is of order $1/Ml_{\rm P}^2\chi$, where $\chi = \sqrt{1 - 2Ml_{\rm P}^2/r_{\rm d}}$ is the redshift factor, so that $T \approx O(Ml_{\rm P}^2\chi)$. On the other hand, the energy density of the black hole region is given by $\sim (M/\chi)/(Ml_{\rm P}^2)^3\chi$, so that the "blueshifted curvature length" L is given by

$$\frac{1}{L^2} \sim G_{\rm N} \frac{M/\chi}{(Ml_{\rm P}^2)^3\chi} \sim \frac{1}{(Ml_{\rm P}^2\chi)^2}. \tag{6.61}$$

This yields

$$T \sim L \sim O(Ml_{\rm P}^2\chi), \tag{6.62}$$

again implying that curvature provides an $O(1)$ effect on the dynamics.

It is, therefore, no surprise that the physics of a near-horizon detector in Section 6.3.3 differs significantly from that of an accelerating detector in Minkowski space experiencing the Unruh effect [78]. In fact, we consider, as we naturally expect, that an inertial frame description in Minkowski space does *not* have a horizon, implying that no information about spacetime is extracted by an accelerating detector, despite the fact that it clicks at a rate controlled by the

acceleration a, $T \approx O(1/a)$, in the detector's own frame. This is indeed consistent with the idea that any information must be accompanied by energy. In the black hole case, the detector mines the black hole, i.e. its click extracts energy from the black hole spacetime, while in the Minkowski case the energy needed to excite the detector comes entirely from the force responsible for the acceleration of the detector—the detector does not mine energy from Minkowski space. We conclude that blueshifted Hawking radiation and Unruh radiation in Minkowski space are very different as far as the information flow is concerned.

Does this imply a violation of the equivalence principle? The equivalence principle states that gravity is the same as acceleration, and the above statement might seem to contradict this principle. This is, however, not true. The principle demands the equivalence of the two only at a point in space in a given coordinate system, and the descriptions of the two systems discussed above—a black hole and Minkowski space—are indeed the same in an infinitesimally small (or lengthscale of order l_*) neighborhood of p_0. The principle does not require that the descriptions must be similar in regions away from p_0, and indeed they are very different: there is a "horizon" at a distance of order $Z_{\rm o}$ from p_0 in the black hole case while there is no such thing in the Minkowski case. And it is precisely in these regions that the detector clicks to extract (or non-extract) information from the black hole (Minkowski) spacetime. In quantum mechanics, a system is specified by a quantum state which generally encodes global information on the equal-time hypersurface. It is, therefore, natural that the equivalence principle, which makes a statement only about a point, does not enforce the equivalence between physics of blueshifted Hawking radiation and of the Unruh effect in Minkowski space at the fully quantum level.

6.3.5 Complementarity: general covariance in quantum gravity

We have argued that unitary information transfer described in Section 6.2, associated with Hawking emission and black hole mining, is consistent with the existence of the interior spacetime suggested by general relativity. We can summarize important lessons we have learned about quantum gravity through this study in the following three points:

- In *a fixed reference frame*, the microscopic information about spacetime, in this case about a black hole, may be viewed as being associated with specific spacetime locations. In particular, for a (quasi-)static description of a system, these degrees of freedom are distributed according to the gravitational thermal entropy calculated using semiclassical field theory. The distribution of these degrees of freedom—which we may call "constituents of spacetime"—controls how they can interact with the degrees of freedom in semiclassical theory, e.g. matter and radiation in semiclassical field theory.

- The spacetime distribution of the microscopic information, however, changes if we adopt a

different reference frame to describe the system. In this sense, the "constituents of spacetime" are *not* anchored to spacetime; they are associated with specific spacetime locations only after the reference frame is fixed. In particular, no reference frame independent statement can be made about where these degrees of freedom are located in spacetime. We may view this as a manifestation of the holographic principle [11, 12]—gauge invariant degrees of freedom in a quantum theory of gravity live in some "holographic space."

- Despite the strong reference frame dependence of the location of the microscopic degrees of freedom, the answers to any physical question are consistent with each other when asked in different reference frames. In particular, when we change the reference frame, the distribution of the microscopic degrees of freedom (as well as some of the semiclassical degrees of freedom) is rearranged such that this consistency is maintained.

These items are basic features of general coordinate transformations at the level of full quantum gravity, beyond the approximation of semiclassical theory in curved spacetime. In particular, they provide important clues about how complementarity as envisioned in Refs. [33, 8] may be realized at the microscopic level.

6.4 Summary—A Grand Picture

The relation between the quantum mechanical view of the world and the spacetime picture of general relativity has never been clear. The issue becomes particularly prominent in a system with a black hole. Quantum mechanics suggests that the black hole formation and evaporation processes are unitary—a black hole appears simply as an intermediate (gigantic) resonance between the initial collapsing matter and final Hawking radiation states. On the other hand, general relativity suggests that a classical observer falling into a large black hole does not feel anything special at the horizon. These two, seemingly unrelated, assertions are surprisingly hard to reconcile. With naive applications of standard quantum field theory on curved spacetime, one is led to the conclusion that unitarity of quantum mechanics is violated [55] or that an infalling observer finds something dramatic (firewalls) at the location of the horizon [4, 73, 74, 68].

We have argued that a potential resolution to this puzzle lies in how a semiclassical description of the system—quantum theory of matter and radiation on a fixed spacetime background—arises from the microscopic theory of quantum gravity. While a semiclassical description employs an *exact* spacetime background, the quantum uncertainty principle implies that there is no such thing—there is an intrinsic uncertainty for background spacetime for any finite energy and momentum. This implies, in particular, that at the microscopic level there are many different ways to arrive at the same background for the semiclassical theory, within the precision allowed by quantum mechanics. This is the origin of the Bekenstein-Hawking (and related, e.g. Gibbons-Hawking [86])

entropy. The semiclassical picture is obtained after coarse-graining these degrees of freedom representing the microscopic structure of spacetime, which we called the vacuum degrees of freedom. More specifically, any result in semiclassical theory is a statement about the maximally mixed ensemble of microscopic quantum states consistent with the specified background within the required uncertainty [5].

This picture elucidates why the purely semiclassical calculation of Ref. [55] finds a violation of unitarity. At the microscopic level, formation and evaporation of a black hole are processes in which information in the initial collapsing matter is converted into that in the vacuum degrees of freedom, which is later transferred back to semiclassical degrees of freedom, i.e. Hawking radiation. Since semiclassical theory is incapable of describing microscopic details of the vacuum degrees of freedom (because it describes them as already coarse-grained, Bekenstein-Hawking entropy), the *description* of the black hole formation and evaporation processes in semiclassical theory violates unitarity at all stages throughout these processes. This, of course, does not mean that the processes are non-unitary at the fundamental level.

In order to address the unitary evolution and explore its relation with the existence or non-existence of the interior spacetime, we therefore need to discuss the properties of the vacuum degrees of freedom. While the theory governing the detailed microscopic dynamics of these degrees of freedom is not yet fully known, we may include them in our description in the form of a new index—vacuum index—carried by the microscopic quantum states (which we denoted by k and κ) in addition to the indices representing excitations in semiclassical theory and of the stretched horizon. We have argued that these degrees of freedom show peculiar features, which play key roles in addressing the paradoxes discussed in Refs. [4, 73, 74]:

Extreme relativeness:

In a fixed reference frame, vacuum degrees of freedom may be viewed as distributed (nonlocally) over space. The spacetime distribution of these degrees of freedom, however, changes if we adopt a different reference frame—they are not anchored to spacetime, and rather live in some "holographic space." This dependence on the reference frame occurs in a way that the answers to any physical question are consistent with each other when asked in different reference frames. Together with the reference frame dependence of (some of the) semiclassical degrees of freedom, discussed in the earlier literature [3, 40], this comprises basic features of how general coordinate transformations work in the full theory of quantum gravity.

Spacetime-matter duality:

The vacuum degrees of freedom exhibit dual properties of spacetime and matter (even in a description in a single reference frame): while these degrees of freedom are interpreted as

representing the way the semiclassical spacetime is realized at the microscopic level, their interactions with semiclassical degrees of freedom make them look like thermal radiation. (At a technical level, the Hilbert space labeled by the vacuum index and that by semiclassical excitations do not factor.) In a sense, these degrees of freedom are neither spacetime nor matter/radiation, as can be seen from the fact that their spacetime distribution changes as we change the reference frame, and that their detailed dynamics cannot be treated in semiclassical theory (as was done in Refs. [4, 73, 74]). This situation reminds us of wave-particle duality, which played an important role in early days in the development of quantum mechanics—a quantum object exhibited dual properties of waves and particles, while the "true" (quantum) description did not fundamentally rely on either of these classical concepts.

These features make the existence of the black hole interior consistent with unitary evolution, in the sense of complementarity [3] as envisioned in Refs. [33, 8]. In particular, a large nearly flat spacetime region near the Schwarzschild horizon becomes manifest in a reference frame whose origin follows a free-fall trajectory starting from rest from a point sufficiently far from the black hole.

It is often assumed that two systems related by the equivalence principle, e.g. a static detector held near the Schwarzschild horizon and an accelerating detector in Minkowski space, must reveal similar physics. This is, however, not true. Since the equivalence principle can make a statement only about a point at a given moment in a given reference frame, while a system in quantum mechanics is specified by a state which generally encodes global information on the equal-time hypersurface, there is no reason that physics of the two systems must be similar beyond a point in space. In particular, a detector reacts very differently to blueshifted Hawking radiation and Unruh radiation in Minkowski space at the microscopic level—it extracts microscopic information about spacetime in the former case, while it does not in the latter.

While our study has focused on a system with a black hole, we do not see any reason why the basic picture we arrived at does not apply to more general cases. We find it enlightening that our results indicate specific properties for the microscopic degrees of freedom that play a crucial role in the emergence of spacetime at the fundamental level. Unraveling the detailed dynamics of these degrees of freedom would be a major step toward obtaining a complete theory of quantum gravity. As a first step, it seems interesting to study implications of our picture for the case that spacetime approaches anti-de Sitter space in the asymptotic region, in which we seem to know a little more [1]. It would also be interesting to explore implications of our picture for cosmology, e.g. along the lines of Refs. [33, 32, 44].

Bibliography

[1] J. M. Maldacena, Int. J. Theor. Phys. **38**, 1113 (1999) [Adv. Theor. Math. Phys. **2**, 231 (1998)] doi:10.1023/A:1026654312961 [hep-th/9711200].

[2] E. Witten, Adv. Theor. Math. Phys. **2**, 253 (1998) [hep-th/9802150].

[3] L. Susskind, L. Thorlacius and J. Uglum, Phys. Rev. D **48**, 3743 (1993) [hep-th/9306069].

[4] A. Almheiri, D. Marolf, J. Polchinski and J. Sully, JHEP **02**, 062 (2013) [arXiv:1207.3123 [hep-th]].

[5] Y. Nomura and S. J. Weinberg, JHEP **10**, 185 (2014) [arXiv:1406.1505 [hep-th]].

[6] Y. Nomura, F. Sanches and S. J. Weinberg, JHEP **1504**, 158 (2015) doi:10.1007/JHEP04(2015)158 [arXiv:1412.7538 [hep-th]].

[7] Y. Nomura, F. Sanches and S. J. Weinberg, Phys. Rev. Lett. **114**, 201301 (2015) doi:10.1103/PhysRevLett.114.201301 [arXiv:1412.7539 [hep-th]].

[8] Y. Nomura, J. Varela and S. J. Weinberg, Phys. Lett. B **733**, 126 (2014) [arXiv:1304.0448 [hep-th]].

[9] R. Bousso, "A Covariant entropy conjecture," JHEP **9907**, 004 (1999) doi:10.1088/1126-6708/1999/07/004 [hep-th/9905177].

[10] R. Bousso, JHEP **9906**, 028 (1999) doi:10.1088/1126-6708/1999/06/028 [hep-th/9906022].

[11] G. 't Hooft, Salamfest 1993:0284-296 [gr-qc/9310026].

[12] L. Susskind, J. Math. Phys. **36**, 6377 (1995) doi:10.1063/1.531249 [hep-th/9409089].

[13] R. Bousso and N. Engelhardt, "New Area Law in General Relativity," Phys. Rev. Lett. **115**, no. 8, 081301 (2015) doi:10.1103/PhysRevLett.115.081301 [arXiv:1504.07627 [hep-th]].

[14] R. Bousso and N. Engelhardt, "Proof of a New Area Law in General Relativity," Phys. Rev. D **92**, no. 4, 044031 (2015) doi:10.1103/PhysRevD.92.044031 [arXiv:1504.07660 [gr-qc]].

[15] S. A. Hayward, Phys. Rev. D **49**, 6467 (1994). doi:10.1103/PhysRevD.49.6467

[16] S. A. Hayward, Class. Quant. Grav. **15**, 3147 (1998) doi:10.1088/0264-9381/15/10/017 [gr-qc/9710089].

[17] A. Ashtekar and B. Krishnan, Phys. Rev. D **68**, 104030 (2003) doi:10.1103/PhysRevD.68.104030 [gr-qc/0308033].

[18] A. Ashtekar and G. J. Galloway, Adv. Theor. Math. Phys. **9**, no. 1, 1 (2005) doi:10.4310/ATMP.2005.v9.n1.a1 [gr-qc/0503109].

[19] S. W. Hawking, "Gravitational radiation from colliding black holes," Phys. Rev. Lett. **26**, 1344 (1971). doi:10.1103/PhysRevLett.26.1344

[20] J. D. Bekenstein, "Black holes and the second law," Lett. Nuovo Cim. **4**, 737 (1972). doi:10.1007/BF02757029

[21] J. D. Bekenstein, "Black holes and entropy," Phys. Rev. D **7**, 2333 (1973). doi:10.1103/PhysRevD.7.2333

[22] J. M. Bardeen, B. Carter and S. W. Hawking, "The Four laws of black hole mechanics," Commun. Math. Phys. **31**, 161 (1973). doi:10.1007/BF01645742

[23] J. D. Bekenstein, "Generalized second law of thermodynamics in black hole physics," Phys. Rev. D **9**, 3292 (1974). doi:10.1103/PhysRevD.9.3292

[24] S. W. Hawking, "Black hole explosions," Nature **248**, 30 (1974). doi:10.1038/248030a0

[25] S. W. Hawking, "Particle Creation by Black Holes," Commun. Math. Phys. **43**, 199 (1975) Erratum: [Commun. Math. Phys. **46**, 206 (1976)]. doi:10.1007/BF02345020

[26] S. Ryu and T. Takayanagi, Phys. Rev. Lett. **96**, 181602 (2006) doi:10.1103/PhysRevLett.96.181602 [hep-th/0603001].

[27] A. Lewkowycz and J. Maldacena, JHEP **1308**, 090 (2013) doi:10.1007/JHEP08(2013)090 [arXiv:1304.4926 [hep-th]].

[28] V. E. Hubeny, M. Rangamani and T. Takayanagi, JHEP **0707**, 062 (2007) doi:10.1088/1126-6708/2007/07/062 [arXiv:0705.0016 [hep-th]].

[29] F. Sanches and S. J. Weinberg, "A Holographic Entanglement Entropy Conjecture for General Spacetimes," arXiv:1603.05250 [hep-th].

[30] A. C. Wall, Class. Quant. Grav. **31**, no. 22, 225007 (2014) doi:10.1088/0264-9381/31/22/225007 [arXiv:1211.3494 [hep-th]].

[31] D. N. Page, Phys. Rev. Lett. **71**, 1291 (1993) [gr-qc/9305007].

[32] Y. Nomura, JHEP **1111**, 063 (2011) doi:10.1007/JHEP11(2011)063 [arXiv:1104.2324 [hep-th]].

[33] Y. Nomura, Found. Phys. **43**, 978 (2013) doi:10.1007/s10701-013-9729-1 [arXiv:1110.4630 [hep-th]].

[34] N. Engelhardt and A. C. Wall, JHEP **1403**, 068 (2014) doi:10.1007/JHEP03(2014)068 [arXiv:1312.3699 [hep-th]].

[35] T. Hartman and J. Maldacena, JHEP **1305**, 014 (2013) doi:10.1007/JHEP05(2013)014 [arXiv:1303.1080 [hep-th]].

[36] M. Headrick and T. Takayanagi, Phys. Rev. D **76**, 106013 (2007) doi:10.1103/PhysRevD.76.106013 [arXiv:0704.3719 [hep-th]].

[37] T. Faulkner, A. Lewkowycz and J. Maldacena, JHEP **1311**, 074 (2013) doi:10.1007/JHEP11(2013)074 [arXiv:1307.2892 [hep-th]].

[38] R. Bousso and N. Engelhardt, Phys. Rev. D **93**, no. 2, 024025 (2016) doi:10.1103/PhysRevD.93.024025 [arXiv:1510.02099 [hep-th]].

[39] For reviews, see e.g. J. Preskill, in *Blackholes, Membranes, Wormholes and Superstrings*, ed. S. Kalara and D. V. Nanopoulos (World Scientific, Singapore, 1993) p. 22 [hep-th/9209058];

[40] L. Susskind and J. Lindesay, *An Introduction to Black Holes, Information and the String Theory Revolution: The Holographic Universe* (World Scientific, Singapore, 2005).

[41] C. R. Stephens, G. 't Hooft and B. F. Whiting, Class. Quant. Grav. **11**, 621 (1994) [arXiv:gr-qc/9310006].

[42] P. A. M. Dirac, *Lectures on Quantum Mechanics*, (Belfer Graduate School of Science, Yeshiva University, New York, 1964).

[43] B. S. DeWitt, Phys. Rev. **160**, 1113 (1967).

[44] Y. Nomura, Phys. Rev. D **86**, 083505 (2012) [arXiv:1205.5550 [hep-th]].

[45] Y. Nomura, J. Varela and S. J. Weinberg, Phys. Rev. D **87**, 084050 (2013) [arXiv:1210.6348 [hep-th]].

[46] Y. Nomura and S. J. Weinberg, arXiv:1310.7564 [hep-th].

[47] See, e.g., M. Van Raamsdonk, Gen. Rel. Grav. **42**, 2323 (2010) [Int. J. Mod. Phys. D **19**, 2429 (2010)] [arXiv:1005.3035 [hep-th]].

[48] See, e.g., R. M. Wald, *General Relativity* (The University of Chicago Press, Chicago, 1984).

[49] R. Bousso, Rev. Mod. Phys. **74**, 825 (2002) [hep-th/0203101].

[50] É. É. Flanagan, D. Marolf and R. M. Wald, Phys. Rev. D **62**, 084035 (2000) [hep-th/9908070].

[51] H. Ollivier, D. Poulin and W. H. Zurek, Phys. Rev. Lett. **93**, 220401 (2004) [arXiv:quant-ph/0307229]; R. Blume-Kohout and W. H. Zurek, Phys. Rev. A **73**, 062310 (2006) [arXiv:quant-ph/0505031].

[52] D. N. Page, Phys. Rev. Lett. **44**, 301 (1980).

[53] D. N. Page, Phys. Rev. Lett. **71**, 3743 (1993) [hep-th/9306083].

[54] Y. Nomura, J. Varela and S. J. Weinberg, arXiv:1207.6626 [hep-th].

[55] S. W. Hawking, Phys. Rev. D **14**, 2460 (1976).

[56] R. M. Wald, Phys. Rev. D **21**, 2742 (1980).

[57] Y. Aharonov, A. Casher and S. Nussinov, Phys. Lett. B **191**, 51 (1987); T. Banks, A. Dabholkar, M. R. Douglas and M. O'Loughlin, Phys. Rev. D **45**, 3607 (1992) [hep-th/9201061].

[58] S. B. Giddings, Phys. Rev. D **46**, 1347 (1992) [hep-th/9203059].

[59] F. Dyson, Institute for Advanced Study report (1976), unpublished.

[60] J. Preskill, in *Blackholes, Membranes, Wormholes and Superstrings*, ed. S. Kalara and D. V. Nanopoulos (World Scientific, Singapore, 1993) p. 22 [hep-th/9209058]; S. B. Giddings, in *Particles, Strings and Cosmology*, ed. J. Bagger *et al.* (World Scientific, Singapore, 1996) p. 415 [hep-th/9508151].

[61] S. D. Mathur, Class. Quant. Grav. **26**, 224001 (2009) [arXiv:0909.1038 [hep-th]].

[62] See, e.g., G. 't Hooft, Nucl. Phys. B **335**, 138 (1990), and references therein.

[63] O. Lunin and S. D. Mathur, Nucl. Phys. B **623**, 342 (2002) [hep-th/0109154].

[64] S. B. Giddings, Class. Quant. Grav. **28**, 025002 (2011) [arXiv:0911.3395 [hep-th]]; Phys. Rev. D **85**, 124063 (2012) [arXiv:1201.1037 [hep-th]].

[65] R. Brustein, arXiv:1209.2686 [hep-th].

[66] See, e.g., D. N. Page, hep-th/9305040, and references therein.

[67] W. K. Wootters and W. H. Zurek, Nature **299**, 802 (1982).

[68] See also S. L. Braunstein, arXiv:0907.1190v1 [quant-ph].

[69] Y. Nomura and J. Varela, arXiv:1211.7033 [hep-th].

[70] For reviews, see e.g. A. H. Guth, Phys. Rept. **333**, 555 (2000) [arXiv:astro-ph/0002156]; A. Vilenkin, J. Phys. A **40**, 6777 (2007) [arXiv:hep-th/0609193]; S. Winitzki, Lect. Notes Phys. **738**, 157 (2008) [arXiv:gr-qc/0612164]; A. Linde, Lect. Notes Phys. **738**, 1 (2008) [arXiv:0705.0164 [hep-th]].

[71] A. A. Starobinskiĭ and S. N. Churilov, Zh. Eksp. Teor. Fiz. **65**, 3 (1973).

[72] D. N. Page, Phys. Rev. D **13**, 198 (1976).

[73] A. Almheiri, D. Marolf, J. Polchinski, D. Stanford and J. Sully, JHEP **09**, 018 (2013) [arXiv:1304.6483 [hep-th]].

[74] D. Marolf and J. Polchinski, Phys. Rev. Lett. **111**, 171301 (2013) [arXiv:1307.4706 [hep-th]].

[75] P. Hayden and J. Preskill, JHEP **09**, 120 (2007) [arXiv:0708.4025 [hep-th]]; Y. Sekino and L. Susskind, JHEP **10**, 065 (2008) [arXiv:0808.2096 [hep-th]].

[76] Y. Takahashi and H. Umezawa. Collective Phenomena **2**, 55 (1975).

[77] J. B. Hartle and S. W. Hawking, Phys. Rev. D **13**, 2188 (1976).

[78] W. G. Unruh, Phys. Rev. D **14**, 870 (1976).

[79] J. M. Bardeen, Phys. Rev. Lett. **46**, 382 (1981); R. Balbinot, Class. Quant. Grav. **1**, 573 (1984).

[80] G. Dvali, Fortsch. Phys. **58**, 528 (2010) [arXiv:0706.2050 [hep-th]].

[81] W. G. Unruh and R. M. Wald, Phys. Rev. D **25**, 942 (1982).

[82] A. R. Brown, Phys. Rev. Lett. **111**, 211301 (2013) [arXiv:1207.3342 [gr-qc]].

[83] W. H. Zurek, Phys. Rev. Lett. **49**, 1683 (1982); D. N. Page, Phys. Rev. Lett. **50**, 1013 (1983).

[84] L. Susskind, Phys. Rev. D **49**, 6606 (1994) [hep-th/9308139].

[85] W. G. Unruh and R. M. Wald, Phys. Rev. D **29**, 1047 (1984).

[86] G. W. Gibbons and S. W. Hawking, Phys. Rev. D **15**, 2738 (1977).