Prepared for submission to JHEP

The ghost in the radiation: Robust encodings of the black hole interior

Isaac Kim,[a,c] **Eugene Tang,**[b] **and John Preskill**[b]

[a] *Stanford Institute for Theoretical Physics, Stanford University, Stanford CA 94305, USA*
[b] *Institute for Quantum Information and Matter and Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena CA 91125, USA*
[c] *School of Physics, The University of Sydney, Sydney, Australia*

E-mail: isaac.kim@sydney.edu.au, eugene.tang@caltech.edu, preskill@caltech.edu

Abstract: We reconsider the black hole firewall puzzle, emphasizing that quantum error-correction, computational complexity, and pseudorandomness are crucial concepts for understanding the black hole interior. We assume that the Hawking radiation emitted by an old black hole is pseudorandom, meaning that it cannot be distinguished from a perfectly thermal state by any efficient quantum computation acting on the radiation alone. We then infer the existence of a subspace of the radiation system which we interpret as an encoding of the black hole interior. This encoded interior is entangled with the late outgoing Hawking quanta emitted by the old black hole, and is inaccessible to computationally bounded observers who are outside the black hole. Specifically, efficient operations acting on the radiation, those with quantum computational complexity polynomial in the entropy of the remaining black hole, commute with a complete set of logical operators acting on the encoded interior, up to corrections which are exponentially small in the entropy. Thus, under our pseudorandomness assumption, the black hole interior is well protected from exterior observers as long as the remaining black hole is macroscopic. On the other hand, if the radiation is not pseudorandom, an exterior observer may be able to create a firewall by applying a polynomial-time quantum computation to the radiation.

Contents

1 Introduction

The discovery that black holes emit Hawking radiation raised deep puzzles about the quantum physics of black holes [1]. What happens to quantum information that falls into a black hole, if that black hole subsequently evaporates completely and disappears? Is the information lost forever, or does it escape in the radiation emitted by the black hole,

albeit in a highly scrambled form that is difficult to decode? And if the information does escape, how? The struggle to definitively answer these questions has been a major theme of quantum gravity research during the 45 years since Hawking's pivotal discovery.

The AdS/CFT holographic correspondence provides powerful evidence indicating that quantum information really does escape from an evaporating black hole [2]. This correspondence, for which there is now substantial evidence, asserts that the process in which a black hole forms and then completely evaporates in an asymptotically anti-de Sitter bulk spacetime admits a dual description in terms of a conformally-invariant quantum field theory living on the boundary of the spacetime. In this dual description, the system evolves unitarily and therefore the process is microscopically reversible — on the boundary there is no gravity, no black hole, no place for information to hide. Since this observation applies to evaporating black holes that are small compared to the AdS curvature scale, it seems plausible that a similar conclusion should apply to more general spacetimes which are not asymptotically AdS, even though we currently lack a firm grasp of how quantum gravity works in that more general setting.

However, so far the holographic correspondence has not provided a satisfying picture of the *mechanism* that allows the information to escape from behind the black hole's event horizon. It is not even clear how the boundary theory encodes the experience of observers who cross the event horizon and visit the black hole interior.

That describing the inside of a black hole raises subtle issues was emphasized in 2012 by the authors known as AMPS [3]. Following AMPS, consider a black hole H that is maximally entangled with another system E which is outside the black hole, and suppose that B is a thermally occupied Hawking radiation mode which is close to the horizon and moving radially outward. Since the black hole is maximally entangled with E, the highly mixed state of B must be purified by a subsystem of E. But on the other hand, we expect that a freely falling observer who enters the black hole will not encounter any unexpected excitations at the moment of crossing the horizon; since field modes are highly entangled in the vacuum state, this means that B should be purified by a mode $\tilde{B}$ located inside the black hole. Now we have a problem, because it is not possible for the mixed state of B to be purified by both E and $\tilde{B}$. Something has to give! Were we to break the entanglement between B and $\tilde{B}$ for the sake of preserving the entanglement between B and E, the infalling observer would encounter a seething firewall at the horizon. This conclusion is hard to swallow, since for a macroscopic black hole we would expect semiclassical theory to be trustworthy at the event horizon, and the black hole solution to the classical Einstein equation has a smooth horizon, not a firewall.

To find a way out of this quandary, it is helpful to contemplate the thermofield double (TFD) state of two boundary conformal field theories, which we'll refer to as the left and right boundary theories. The TFD is an entangled pure state of the left and right boundaries, with the property that the marginal state of the right boundary (with the left boundary traced out) is a thermal state with temperature T, and likewise the marginal state of the left boundary (with the right boundary traced out) is thermal with the same temperature. The corresponding bulk geometry is a two-sided black hole. Both the left black hole and the right black hole are in equilibrium with a radiation bath at tempera-

ture T, and both have smooth event horizons. Furthermore, the two black holes have a shared interior — they are connected in the bulk by a non-traversable wormhole behind the horizon [4]. Here, the right black hole (let's call it H) is purified by another system (the left black hole E), and emits Hawking radiation, yet it has a smooth horizon. How can we reconcile this finding with the AMPS argument?

For the case of the two-sided black hole, there is an instructive answer [5]. The Hawking mode B outside the right black hole can be purified by both $\tilde{B}$ behind the horizon and by a subsystem of E, because E itself lies behind the horizon and $\tilde{B}$ is a subsystem of E! It is very tempting to suggest that a similar resolution of the AMPS puzzle applies to the case of a one-sided black hole H, which is entangled with a system E outside its horizon. That is, we may regard the black hole interior and the exterior system entangled with the black hole as two complementary descriptions of one and the same system. Indeed, we might imagine allowing E to undergo gravitational collapse, thereby obtaining a pair of entangled black holes, which, if we accept a conjecture formulated in [5], would be connected through the bulk by a non-traversable wormhole. The boundary dual of this bulk state, up to a one-sided transformation acting on one of the two boundaries, is a TFD, to which our previous discussion of the entanglement structure of the two-sided black hole ought to apply.

The idea that, for the case of a black hole H purified by the exterior system E, we may regard the black hole interior as related to E by a complicated encoding map, has been advocated, discussed, and criticized in much previous work [5–10]. We will revisit this issue in this paper, arguing that a proper resolution of the AMPS puzzle should invoke concepts that have received relatively short shrift in earlier discussions of the firewall problem, namely *quantum error correction, computational complexity, and pseudorandomness.*

The scenario described above, in which the black hole H has become maximally entangled with the exterior system E, might arise because the black hole actually formed long ago, and since then has radiated away more than half of its initial entropy. In that case E would be the Hawking radiation so far emitted during the black hole's lifetime, most of which is by now far away from the black hole. One could object that our proposal, that the black hole interior is related to E by a complicated encoding map, is too wildly non-local to be credible [8, 9, 11]. Why can't an exterior agent who interacts with the Hawking radiation send instantaneous signals to the black hole interior in flagrant violation of causality? And why can't such an agent access the encoded system $\tilde{B}$, breaking the entanglement between $\tilde{B}$ and B and hence creating excitations which can be detected by an observer who falls through the horizon?

Our answer is that such non-local operations are in principle possible, but are not accessible to observers whose computational abilities are bounded (a notion we make precise in Section 6); the operations required to disturb the interior mode are far too complex to be realizable in practice for any realistic observer. Thus, in spite of the extreme non-locality of the encoding map, violations of the semiclassical causal structure of the black hole spacetime are beyond the reach of any realistic exterior observer. This statement is most conveniently expressed using the language of quantum error correction and computational complexity. We will use $|S|$ to denote the size of a physical system S; by size we mean the number of qubits, so that $2^{|S|}$ is the dimension of the Hilbert space of S. We regard H as

the Hilbert space of black hole microstates, and E as the Hilbert space of the previously emitted radiation. For an old black hole H which is nearly maximally entangled with E, we show that a quantum error-correcting code can be constructed, in a subspace of EH, which describes the black hole interior. The logical operators of this code, which preserve the code subspace, are operators acting on the interior. We will argue that a code exists with the following property: Any operation on the radiation E that can be performed as a quantum computation whose size is polynomial in $|H|$ will commute with a set of logical operators of the code, up to corrections which are exponentially small in $|H|$. For this encoding, then, an observer outside a black hole can signal the interior only by performing an operation of super-polynomial complexity. Because the encoded interior is for all practical purposes invisible to the agent who roams the radiation system E, we call the code's logical operators *ghost operators.*[1]

To reach this conclusion, we make a nontrivial but reasonable assumption — that the radiation system E is *pseudorandom.* Note that if the state of EH is pure, and $|H| \ll |E|$, then the density operator ρ_E of E is not full rank, so that ρ_E is obviously distinguishable from the maximally mixed state σ_E. When we say that ρ_E is pseudorandom, we mean that ρ_E and σ_E are not *computationally* distinguishable. That is, suppose we receive a copy of ρ_E (or even polynomially many copies) and we are asked to determine whether the state is maximally mixed or not using a quantum circuit whose size is polynomial in $|H|$. If ρ_E is pseudorandom, then our probability of answering correctly exceeds $1/2$ by an amount which is exponentially small in $|H|$. Such pseuodrandom quantum states exist, and furthermore it has recently been shown [12] that they can be prepared by efficient quantum circuits, if one accepts a standard (and widely believed) assumption of post-quantum cryptography: That there exist one-way functions which are hard to invert using a quantum computer. Since black holes are notoriously powerful scramblers of quantum information [13], we think the assumption that ρ_E is pseudorandom is plausible, though undeniably speculative. Our main technical result shows that if ρ_E is pseudorandom, then a code with ghost logical operators must exist.

That the existence of quantum-secure one-way functions implies the hardness of *decoding* Hawking radiation had been pointed out earlier in [14] and [15]. But our statement goes further — it indicates that causality is well respected from the viewpoint of computationally bounded observers (as long as $|H|$ is large). The semi-classical causal structure of the evaporating black hole spacetime can be disrupted by an observer with sufficient computational power, but not by an observer whose actions can be faithfully modeled by a quantum circuit with size polynomial in $|H|$. On the other hand, interior observers, who in principle have access to H as well E, could plausibly perform nontrivial operations on the interior which are beyond the reach of the computationally bounded observer who acts on E alone.

Our main result can be regarded as a contribution to the theory of quantum error correction in a nonstandard setting. In the context of fault-tolerant quantum computation,

[1] The word "ghost" is sometimes used to describe unphysical degrees of freedom. That is not what we mean here. The ghost operators act on a system (the interior of a black hole) which is physical but *inaccessible* to observers outside the black hole who have reasonable computational power.

where the goal is to protect a quantum computer from noise due to uncontrolled interactions with the computer's environment, we usually consider noise which is weak and only weakly correlated. For example, we might model the noise using a Hamiltonian describing the interactions of the computer and environment, where each term in the Hamiltonian is small and acts on only a few of the computer's qubits. In our setting the "computer" is the system EH, and the "noise" results from the interactions of the computationally bounded observer with system E, while H is regarded as noiseless. In contrast to conventional quantum error correction, we allow the noise to be strong, highly correlated, and adversarially chosen, yet the logical system $\tilde{B}$ encoded in EH is well protected against this noise. To obtain this result, though, it is essential that the noise acts only on E and not on H, a departure from the usual model of fault tolerance in which all qubits are assumed to be noisy.

We note that the encoding of the black hole interior in EH is state dependent; that is, the way the system $\tilde{B}$ is embedded in EH depends on the initial state that underwent gravitational collapse to form a black hole. This state dependence of the encoding has sparked much discussion and consternation [8, 16]. What seems troubling is that operators which depend on the state to which they are applied are not linear operators acting on Hilbert space, and therefore can not be regarded as observables as described in the conventional quantum theory of measurement. Our view is that the tension arising from the state dependence of the encoded operator algebra signals that we do not yet have a fully satisfactory way to describe measurements performed inside black holes. We will not rectify this shortcoming in this paper.

Our argument about the robustness of the ghost logical operators makes no direct use of AdS/CFT technology. This may be viewed as either a strength or a weakness. The strength is that our results may be applicable to black holes in spacetimes which are asymptotically flat or de Sitter, and stand independently of any assumptions of holography. The weakness is that we have not presented evidence based on holographic duality which supports our conjecture.

There has been great recent progress toward resolving the discrepancy between Hawking's semiclassical analysis [1] and the Page curve [17–19] of an evaporating black hole, including formulas for the entropy of the radiation supported by explicit computations [20–22]. These results strengthen the evidence that black hole evaporation is a unitary process, and also point toward a resolution of the firewall problem in which the interior of a partially evaporated black hole is encoded in the Hawking radiation. This beautiful prior work, however, does not directly address how the profoundly nonlocal encoding of the interior in the radiation is compatible with the semiclassical causal structure of the black hole geometry. It is for that purpose that we hope our observations concerning the pseudorandomness of the radiation and the construction of ghost logical operators acting on the interior will prove to be relevant. Our main conclusion is that the encoded interior can be inaccessible to observers outside the black hole who have reasonable computational power. Establishing closer contact between our work and these recent computations is an important open problem.

The rest of this paper is structured as follows. In Section 2 we provide a non-technical summary of the paper. In Section 3 and 4, we review the notion of pseudorandomness in

both the classical and quantum setting; in Section 5, we argue that the Hawking radiation is a pseudorandom quantum state, and we explain in detail our computational model of the black hole.

In the remaining sections, we derive consequences of the pseudorandomness assumption, and explore their potential relevance to the black hole firewall problem. In Section 6, we show that it is computationally hard for an observer interacting with the early radiation E to distill the interior mode $\tilde{B}$ and carry it into the black hole. In Section 7, we show that the encoded system $\tilde{B}$ is protected against errors inflicted on E by any agent who performs a quantum operation with poly($|H|$) computational complexity and sufficiently small Kraus rank. In Section 8, we describe the construction of ghost logical operators acting on the black hole interior; these operators commute with all low-complexity operations applied to E by an agent O, provided that O's quantum memory is not too large. If the observable properties of the black hole interior are described by such ghost operators, we infer that the interior cannot be affected or detected by computationally bounded agents who interact with the Hawking radiation. The theory of ghost operators, which can be constructed for any approximate quantum error-correcting code, may also be of independent interest. In Section 9, we show that, if the state of the partially evaporated black hole has been efficiently generated, then an agent with access to both E and H can manipulate the encoded interior efficiently, and efficiently distill the encoded system $\tilde{B}$ to a small quantum memory. Section 10 contains our conclusions. Some technicalities are treated in the Appendices, and in Appendix D we discuss via an example how the construction of ghost logical operators may fail if the Hawking radiation is not pseudorandom.

2 Probing the radiation

In this section we'll provide a somewhat more explicit explanation of our main result, still skipping over technical details which will be laid out in later sections. The situation we consider is depicted in Figure 1. There, the unitary transformation $U_{\rm bh}$ describes the formation and subsequent partial evaporation of a black hole formed from infalling matter in a pure state $|\phi_{\rm matter}\rangle$, where E denotes the "early" Hawking radiation which has been emitted so far, H denotes the remaining black hole which has not yet evaporated, and B denotes Hawking quanta of the "late" radiation which has just been emitted from the black hole. We may assume for convenience that B is a single qubit — our conclusions would be the same if we considered B to be any system of constant dimension, independent of the size of E and H. The system P denotes an ancillary system called the "probe", which might represent, for example, ambient dust around the black hole. We will discuss the role of the probe in greater detail shortly, but for simplicity we may ignore its presence right now.

In the case of an "old" black hole H, which has already radiated away over half of its initial entropy and has become nearly maximally entangled with E, we have $|H| < |E|$. Because the lifetime of an evaporating black hole scales like the 3/2 power of its initial system size, we may regard the unitary transformation $U_{\rm bh}$ to be "efficient," meaning that it can be accurately described by a quantum circuit whose size increases only polynomially

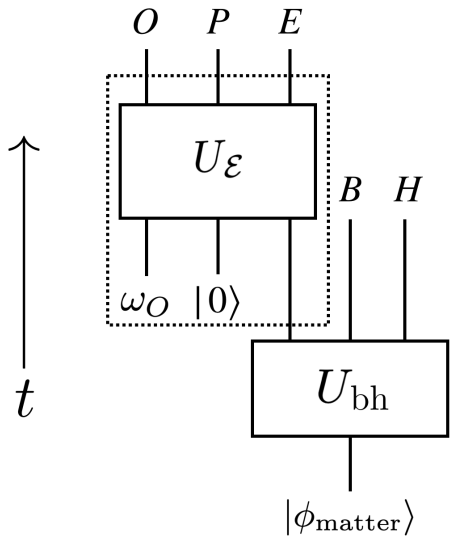

Figure 1. A black hole forms due to the gravitational collapse of an infalling state of matter. The black hole then evaporates for a while, emitting the "early" radiation E and the "late" radiation B; the formation and partial evaporation of the black hole are described by the unitary transformation $U_{\rm bh}$. An observer O interacts with the early radiation and a probe system P, where the unitary transformation $U_{\mathcal{E}}$ (enclosed by the dotted line) has quantum complexity which scales polynomially with the size $|H|$ of the remaining black hole (essentially its entropy $S_{\rm bh}$). If the radiation is pseudorandom, then O is unable to distinguish E from a perfectly thermal state.

with $|EHB|$. Our key assumption is that the efficient unitary $U_{\rm bh}$ creates a pseudorandom state of EB (see Section 5). The notion of a pseudorandom quantum state will be further discussed in Section 4.

In the context of the AMPS puzzle, the recently emitted system B should be purified by a system $\tilde{B}$ behind the horizon. We will explore the idea that this system $\tilde{B}$ is actually encoded in EH, the union of the black hole system H and the early radiation system E. Let us denote the state prepared by the unitary map $U_{\rm bh}$ as $|\Psi\rangle_{EHB}$, and consider its expansion

$$|\Psi\rangle_{EHB} = \sum_{ijk} \Psi_{ijk}\ |i\rangle_E \otimes |j\rangle_H \otimes |k\rangle_B. \tag{2.1}$$

There is a corresponding map $V_\Psi : \tilde{B} \to EH$ defined by

$$V_\Psi = \sqrt{d_B} \sum_{ijk} \Psi_{ijk} |i\rangle_E \otimes |j\rangle_H \otimes \langle k|_{\tilde{B}}, \tag{2.2}$$

where d_B is the dimension of B. If B is maximally mixed in the state $|\Psi\rangle$, then V_Ψ is an isometric map embedding $\tilde{B}$ in EH. We interpret V_Ψ as the encoding map of a quantum error-correcting code, which maps the interior system $\tilde{B}$ to the subspace of EH with which B is maximally entangled.

If $\tilde{T}_{\tilde{B}}$ is any operator acting on $\tilde{B}$, there is a corresponding "logical" operator T_{EH} acting on EH defined by

$$V_\Psi \tilde{T}_{\tilde{B}} = T_{EH} V_\Psi. \tag{2.3}$$

This logical operator is not uniquely defined, because equation (2.3) only specifies its action on the code space, the image of V_Ψ. We may say that T_{EH} is the "mirror operator" of $\tilde{T}_{\tilde{B}}$ determined by $|\Psi\rangle$, whose defining property is that T_{EH} and $\tilde{T}_{\tilde{B}}$ produce the same output when acting on the state $|\Psi\rangle$.

We wish to investigate whether an agent who interacts with only the radiation system E can manipulate the encoded system $\tilde{B}$. For that purpose we introduce an additional system O to represent an observer outside the black hole who interacts with E. This interaction is modeled by a unitary transformation $U_{\mathcal{E}}$ acting on OE, possibly followed by a simple measurement performed on O; for example one might measure all the qubits of O in a standard basis). The unitary transformation, but not the following measurement, is shown in Figure 1. After the interaction, but before O is measured, the joint state of $OEBH$ has evolved to

$$|\Psi'\rangle_{OEBH} = \left((U_{\mathcal{E}})_{OE} \otimes I_{BH}\right) \left(|\omega\rangle_O \otimes |\Psi\rangle_{EBH}\right), \tag{2.4}$$

where $|\omega\rangle_O$ is the initial state of O before O and E interact.

Our notation $U_{\mathcal{E}}$ for the unitary transformation is motivated by a widely used convention in the theory of quantum channels, in which $\mathcal{E}$ denotes a quantum noisy channel (a trace-preserving completely positive map), with the letter $\mathcal{E}$ indicating an "error" acting on the input to the channel. A quantum channel always admits a dilation (also called a purification), a unitary transformation which acts on the input system and an "environment," after which the environment is discarded. In our context, the noisy channel $\mathcal{E}$ acting on E arises from the action of the observer, and we may regard the observer's system O as the environment in the dilation of $U_{\mathcal{E}}$. In the following discussion, we will often omit the subscript OE on $(U_{\mathcal{E}})_{OE}$, leaving it implicit that $U_{\mathcal{E}}$ acts on the radiation system E and observer O.

Now we can appeal to a standard result in the theory of quantum error correction. In $|\Psi\rangle_{EHB}$ we regard B as a "reference system" which purifies the maximally mixed state of the encoded system $\tilde{B}$. Is there a recovery operator which can be applied to EH to correct the error induced by this noisy channel? In fact a recovery operator that corrects the error *exactly* exists if and only if the marginal state ρ'_{OB} of OB factorizes,

$$\rho'_{OB} = \rho'_O \otimes \rho'_B, \tag{2.5}$$

in which case we say the reference system B "decouples" from the environment O. Heuristically, the error can be corrected if and only if no information about the state of $\tilde{B}$ leaks to the environment O. There is also an approximate version of this statement [23, 24]. Roughly speaking (we will be more precise in Section 6), recovery with fidelity close to one is possible if and only if O and B are nearly uncorrelated after O and E interact.

Now consider the implications of our assumption that the Hawking radiation is pseuodorandom. As stated in Section 1, the marginal state ρ_{EB} is pseudorandom if ρ_{EB} cannot be distinguished from a maximally mixed state by any circuit with size polynomial in $|H|$, apart from an error exponentially small in $|H|$. We will show in Section 6 that, assuming $|O| \ll |H|$, if ρ_{EB} is pseudorandom and $U_{\mathcal{E}}$ is any polynomial-size unitary transformation, then O and B approximately decouple up to an error exponentially small in $|H|$. Therefore, apart from an exponentially small error, a computationally bounded observer O is unable to inflict an uncorrectable error on the encoded system $\tilde{B}$.

We can make a stronger assertion: It is possible to choose the logical operators acting on the encoded system to be robust *ghost operators*, which (acting on the code space) nearly commute with any operation applied by the computationally bounded observer O. Returning now for simplicity to the setting of exact correctability, we claim that if the error induced by $U_{\mathcal{E}}$ is correctable, then for any operator $\tilde{T}_{\tilde{B}}$ acting on system $\tilde{B}$, it is possible to choose the corresponding logical operator T_{EH} satisfying equation (2.3) such that

$$T_{EH} U_{\mathcal{E}} \left(I_O \otimes V_\Psi \right) = U_{\mathcal{E}} T_{EH} \left(I_O \otimes V_\Psi \right). \tag{2.6}$$

In this sense, the correctable errors have no effect on the ghost logical algebra. This claim is a special case of a more general statement about operator algebra quantum error correction (OAQEC) [25, 26]. Since we do not expect a black hole to provide an exact error-correcting code, we will need to analyze the case of approximate quantum error correction. Unfortunately, it does not seem straightforward to generalize the results of [25, 26] to the approximate setting. Instead, we present a self-contained construction of exact ghost logical operators in Section 8.1, without making direct use of known results from the theory of OAQEC, and then generalize the construction to the approximate setting in Section 8.2.

We will apply the approximate version of this result to the situation where $U_{\mathcal{E}}$ induces an approximately correctable error, thus inferring that the logical operators acting on the encoded system $\tilde{B}$ may be chosen so that they nearly commute with the actions of the computationally bounded observer O. We propose that these robust ghost operators are the logical operators acting on the black hole interior, and conclude that the interior is very well protected against the actions of any realistic observer who resides outside the black hole.

The statements about the indistinguishability of ρ_{EB} from a maximally mixed state, the decoupling of O from B, the correctability of $U_{\mathcal{E}}$, and the commuting action of $\tilde{T}_{EH}$ and $U_{\mathcal{E}}$ on the code space, are all approximate relations with exponentially small corrections. Therefore, we need to be mindful of these corrections in constructing our arguments. Fortunately, many relevant features of approximate quantum error-correction have been previously studied, and we make use of results from [23, 24] in particular.

For the general argument sketched above we have assumed that the observer system satisfies $|O| \ll |H|$. But it is also instructive to consider a different scenario, in which the observer has access to an auxiliary probe system P. We now imagine that the probe P, which might have a size comparable to or larger than E, is prepared in a simple initial state and then interacts efficiently with E. After this interaction between E and P, the observer (still satisfying $|O| \ll |H|$), interacts with EP, performing an efficient quantum

computation that may be chosen adversarially. In this case, too, we can show under the same pseudorandomness assumption as before that the reference system B decouples from O, and that robust ghost logical operators can be constructed. For example, the probe might cause all of the qubits of E to dephase in a preferred basis, but the entanglement between $\tilde{B}$ and B would still be protected. The modification from the previously considered case is that now $\tilde{B}$ will be encoded in EHP rather than EH, and we conclude that the encoded black hole interior remains inaccessible to any computationally bounded observer O who examines the radiation and the probe, as long as the size of the observer's memory satisfies $|O| \ll |H|$.

Our conclusion that $\tilde{B}$ is difficult to decode or manipulate follows from the pseudorandomness of the Hawking radiation if the observer is computationally bounded and has access only to the radiation system E outside the black hole. But we might imagine that an observer who jumps into the black hole has access to the black hole degrees of freedom H as well as E. We show in Section 9 that an observer who has access to EH can efficiently manipulate and decode $\tilde{B}$, assuming only that the state $|\Psi\rangle_{EBH}$ was created by an efficient unitary process. In this sense, an interior observer can interact with the interior degrees of freedom, as one might expect. A similar remark applies to the fully evaporated black hole. If the final state after complete evaporation is a highly scrambled pure state of EB, where $|B| \ll |E|$, then the maximally mixed state of B is purified by a code subspace of E. If B has constant size, then the code state can be efficiently distilled and deposited in a small quantum memory, assuming only that the map from the infalling matter to the outgoing Hawking radiation is an efficient unitary process.

If an efficient measurement of EB *can* detect the correlation between E and B, then we may expect that an observer acting on E is able to interact efficiently with the black hole interior. In Appendix D, we show that, if a product observable $M_E \otimes N_B$ has an expectation value in the state $|\Psi\rangle_{EHB}$ that differs significantly from its expectation value in a maximally mixed state of EB, then there cannot be a complete set of ghost logical operators on EH commuting with M_E. It follows that, if M_E can be realized efficiently, low-complexity operations acting on the Hawking radiation *can* send a signal to the interior.

3 Classical pseudorandomness

Our argument that the black hole interior is inaccessible to computationally bounded exterior observers hinges on the hypothesis that the Hawking radiation emitted by an old black hole is pseudorandom. In this section we'll provide background about the concept of pseudorandomness, which some readers might find helpful.

As discussed in Section 1, we are interested in a black hole that is still macroscopic but has already been evaporating for longer than its Page time [27]. The state of the previously emitted radiation system EB is purified by the black hole system H, and by this time EB is much larger than H; therefore the microscopic state ρ_{EB} of EB has far lower rank than a thermal state. It must then be possible, at least in principle, to distinguish ρ_{EB} from a thermal state. But how, operationally, would an observer outside the black hole who interacts with the radiation be able to tell the difference?

To start with, it will be instructive to consider a simple classical model that captures some of the features of this setup — after we understand how the classical model works we'll be better prepared to analyze an analogous quantum model. Let's suppose that the emitted Hawking radiation is a classical bit string x of length n, which our observer is permitted to read. But this bit string is not chosen deterministically; rather, when the observer reads the radiation he actually samples from a probability distribution governing n-bit strings. We'll say that the state of the black hole is "thermal" if this distribution is the uniformly random distribution $p_I(x)$, where

$$p_I(x) = \frac{1}{2^n}, \qquad \forall x \in \{0,1\}^n. \tag{3.1}$$

But suppose the state of the black hole is described by a distribution that is in principle almost perfectly distinguishable from the uniform distribution. Can this state "fool" the observer, leading him to believe the distribution is uniform even though that is far from the case? See Figure 2.

To be more concrete, let's suppose the observer is assured that he is sampling from a distribution which is either the uniform distribution $p_I(x)$, or a different distribution $p_S(x)$ which is uniform on the subset of n-bit strings S:

$$p_S(x) = \begin{cases} 2^{-\alpha n}, & x \in S \\ 0, & x \notin S \end{cases}, \tag{3.2}$$

where $|S| = 2^{\alpha n}$ (with $0 < \alpha < 1$) is the number of strings contained in S. The observer samples once from the distribution, receiving x, and then executes a classical circuit C with x as an input, finally producing either the output 1 if he guesses that the distribution is p_S, or the output 0 if he guesses that the distribution is p_I.[2]

For n large, it is clear that C can be chosen so that the observer guesses correctly with a high success probability. Suppose, for example, that he outputs 1 if $x \in S$ and he outputs 0 if $x \notin S$. It the distribution is actually p_S, this guess is correct with probability 1. If the distribution is actually p_I, then the guess is correct unless x happens to lie in S "by accident," which occurs with probability $2^{-(1-\alpha)n}$. Therefore, for fixed α and large n, the probability of an incorrect guess is exponentially small in n.

However, depending on the structure of the set S, the circuit C that distinguishes p_S and p_I might need to be quite complex, making this strategy impractical if the observer has limited computational power. Suppose, for example, that the observer is unable to perform a computation with more than $\Lambda(n)$ gates, where $\Lambda(n)$ grows subexponentially with n — that is, $\Lambda(n) \leq \exp(f(n))$ where $f(n)$ scales sublinearly with n. Then we can show that the set S can be chosen such that this computationally bounded observer has only an exponentially small chance of distinguishing p_S and p_I; that is, his probability of guessing the distribution correctly is no better than $1/2 + 2^{-cn}$, where c is a positive constant. In that case, we say that the distribution p_S is *pseudorandom*.

[2]The conclusion we reach below would not change much if he were permitted to sample from the distribution a number of times polynomial in n.

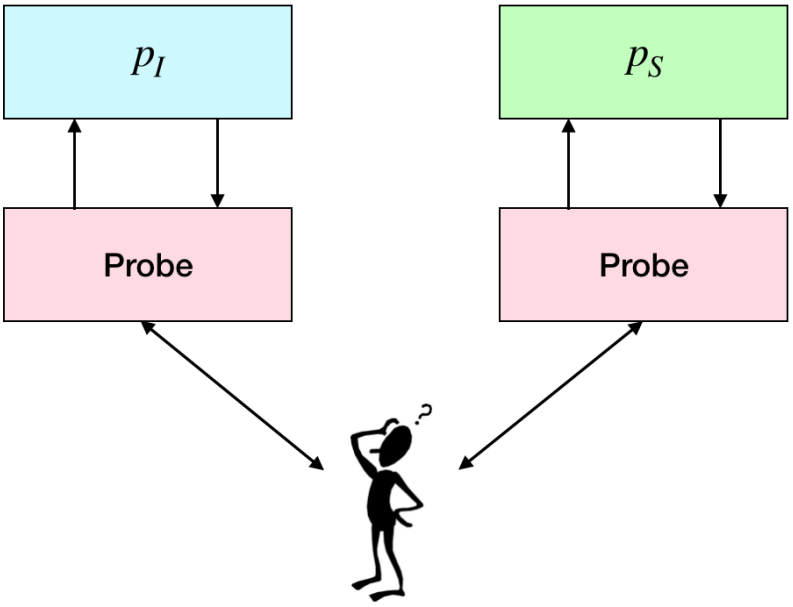

Figure 2. An observer samples from a distribution and attempts to decide whether the distribution is uniformly random or not.

To show that such a pseudorandom distribution p_S exists, we argue in two steps. In the first step, we consider some fixed circuit C, and denote by L_C the set of n-bit input strings for which C outputs 1 (we say that C "accepts" the strings in L_C). If the input x is chosen by sampling from $p_I(x)$, then C accepts x with probability

$$P_C(I) = \frac{|L_C|}{2^n}, \tag{3.3}$$

while if x is chosen by sampling from $p_S(x)$, then C accepts x with probability

$$P_C(S) = \frac{|S \cap L_C|}{2^{\alpha n}}.$$

Now suppose that S is chosen randomly from among all subsets of n-bit strings with cardinality $|S| = 2^{\alpha n}$. We can envision the possible strings as 2^n balls, of which the balls accepted by C are colored white, and the balls rejected by C are colored black, while S is a random sample containing $|S|$ of these balls. Suppose that the white balls constitute a fraction f of all the balls. Then, for n large, we expect that S also contains a fraction of white balls which is close to f. This intuition can be made precise using Hoeffding's inequality, from which we derive

$$\Pr\left(|P_C(S) - P_C(I)| \geq \epsilon\right) \leq e^{-2|S|\epsilon^2}, \tag{3.4}$$

where the probability is evaluated for the uniform distribution over all subsets with $|S|$ elements. Now we can choose ϵ to have the exponentially small value $\epsilon = |S|^{-1/4}$ (for example) to see that, if S is sampled uniformly with $|S|$ fixed, the probability that C accepts a sample from p_S is exponentially close to the probability that C accepts a sample

from p_I. We conclude that, not only is it possible to choose the subset S such that the fixed circuit C can barely distinguish p_S from p_I, but furthermore most choices for S with $|S| = 2^{\alpha n}$ have this property.

We have now completed the first step in our two-step argument. But so far we have only shown that S can be chosen such that p_S and p_I are hard to distinguish for one fixed circuit C. We wish to make a much stronger claim, that there is a choice for S such that p_S and p_I are nearly indistinguishable by *any* circuit with a number of gates subexponential in n.

To prove this stronger claim we proceed with the second step in the argument. For a collection of circuits $\{C_1, \ldots, C_N\}$, what is the probability that $|P_{C_i}(S) - P_{C_i}(I)| \geq \epsilon$, for at least one i? An upper bound on this probability follows from the union bound, which asserts that

$$P(A_1 \cup \cdots \cup A_N) \leq \sum_{i=1}^{N} P(A_i), \tag{3.5}$$

where $\{A_1, A_2, \ldots, A_N\}$ is any set of events. Using equation (3.4), we conclude that the probability that at least one of the N circuits distinguishes p_S from p_I with probability at least ϵ is no larger than $Ne^{-2|S|\epsilon^2}$.

How many possible circuits are there which act on the n-bit input x and contain m computation steps? In each step of the computation, we either input one of the bits of x or we execute a gate which is chosen from a set of G possible gates, where G is a constant. Our claim will hold if each gate in G has a constant number of input and output bits, so for simplicity let's assume that each gate has at most two input bits and generates a single output bit (like a NAND gate for example). Each two-bit gate acts on a pair of bits which are outputs from previous gates; this pair can be chosen in fewer than m^2 ways. Therefore, the total number $N(m)$ of size-m circuits can be bounded as

$$N(m) \leq \left((n+G)m^2\right)^m, \tag{3.6}$$

which implies

$$\log N(m) \leq m\left(2\log m + \log(n+G)\right). \tag{3.7}$$

Even if we choose an exponentially large circuit size $m = 2^{\gamma n}$ and an exponentially small error $\epsilon = 2^{-\delta n}$, we find that $N(m)e^{-2|S|\epsilon^2}$ is doubly exponentially small in n for $|S| = 2^{\alpha n}$ and $\alpha > \gamma + 2\delta$. Hence, if S is randomly chosen, it's extremely likely that the distributions p_S and p_I are indistinguishable by circuits of size $2^{\gamma n}$, up to an exponentially small error.

To summarize, we've shown that the set S can be chosen so that the probability distribution p_S has these properties:

1. Its entropy per bit α is a positive constant less than 1.

2. It is statistically distinguishable from p_I with an exponentially small failure probability.

3. If $\gamma < \alpha$, any circuit using at most $2^{\gamma n}$ gates almost always fails to distinguish p_S and p_I.

If n is macroscopic, the task of distinguishing p_S from p_I can be absurdly difficult, even if the entropy density α is quite small. Suppose, for example, that $n = 10^{23}$ is comparable to Avogadro's number, and $\alpha = 10^{-12}$. Choosing $\gamma = \delta = 10^{-13}$ we conclude that a circuit with $m = 2^{10^{10}}$ gates can distinguish p_S from p_I with a success probability no larger than $\epsilon = 2^{-10^{10}}$. Even if we could perform one gate per unit of Planck time and Planck volume, an unimaginably large spacetime region would be required to execute so large a circuit.

The existence of pseudorandom distributions was first suggested by Yao [28], and the construction we have described was discussed by Goldreich and Krawcyzk [29]. In our analysis we assumed that the observer executes a deterministic circuit, but it turns out that giving the observer access to a random number generator does not make his task any easier [29].

Up until now, we assumed that the observer performs a computation whose output is a single bit. But what if he obtains a k-bit output instead? Can we choose the set S so that for all circuits of bounded size the probability distribution governing the k output bits is very similar for input strings drawn from p_S and p_I? Our previous reasoning does not have to be modified much to handle this case. Now for the fixed circuit C, we denote by $L_C[y]$ the set of n-bit input strings for which C outputs the k-bit string y, and we denote by $P_C(I)[y]$, $P_C(S)[y]$ the probability that C outputs y when receiving as input a sample from p_I, p_S respectively. Now we envision the n-bit input strings as balls which can be colored in 2^k possible ways, corresponding to the 2^k possible values of the output y. Applying the previous argument to each color, we find that when S is chosen at random from among all subsets with cardinality $|S|$,

$$\text{Prob}\left[|P_C(S)[y] - P_C(I)[y]| \geq \epsilon\right] \leq e^{-2|S|\epsilon^2}, \tag{3.8}$$

for each output y. From the union bound, the probability that $|P_C(S)[y] - P_C(I)[y]|$ exceeds ϵ for at least one value of y is bounded above by $2^k e^{-2|S|\epsilon^2}$, which also provides an upper bound on the probability that the total variation distance between $P_C(S)$ and $P_C(I)$ exceeds $\epsilon' = 2^k\epsilon$. Therefore the total variation distance will be no larger than ϵ' with high probability as long as $2^{\alpha n}2^{-2k}\epsilon'^2$ is large, which means ϵ' can be exponentially small in n provided $2k < \alpha n$. We conclude that the pseudorandom input distribution and the uniformly random input distribution will yield exponentially close output distributions as long as the observer's output register is small compared to the entropy of S.

The preceding argument shows that, indeed, there are probability distributions which are computationally indistinguishable from the uniformly random distribution. But can we make such a distribution *efficiently*? It turns out that our argument for the computational hardness of distinguishing a pseudorandom distribution from a uniformly random distribution can be used to show that such distributions are typically hard to produce with polynomial-sized circuits.[3] However, there are distributions that can be created using polynomial-sized circuits which, under reasonable complexity-theoretic assumptions, are

[3]We thank Adam Bouland for emphasizing this point.

difficult to differentiate from the uniformly random distribution, for any polynomial-sized circuit. Such distributions can be generated by pseudorandom generators [30]. We will give a more complete description of such constructions for the quantum case in Section 4.

Efficient sampling from a (classical) pseudorandom distribution is analogous to the formation and partial evaporation of a black hole. Pseudorandom number generators consult a random "key" which is hidden from the adversary, and then compute a function which depends on the key. This function is chosen so that an output drawn from the resulting family of outputs indexed by the key is computationally indistinguishable from the output of a truly random function. In the case of the partially evaporated black hole, the key becomes a black hole microstate, and the key-dependent function evaluation becomes the chaotic unitary evolution of the evaporating black hole. An adversary samples the Hawking radiation, and attempts to determine whether the sample is drawn from a thermal distribution or not.

To properly discuss the evaporating black hole, we will need to consider the quantum version of pseudorandomness, to which we turn in the next two sections. But our simplified classical model of "Hawking radiation" is instructive. It teaches us that the (classical) adversary can interact with the (classical) radiation for a subexponential time (or even for the exponential time $2^{\gamma n}$ if γ is sufficiently small), without ever suspecting that the radiation is far from uniformly random. On the other hand, that conclusion may no longer apply if the adversary collects k bits of information where k is sufficiently large ($k > \alpha n/2$). Both of these features will pertain to the quantum version of our story.

4 Quantum pseudorandomness

Now consider the quantum version of the task described in the Section 3. Our observer receives a quantum state ρ, and is challenged to guess whether ρ is maximally mixed or not. For that purpose, he performs a quantum computation with ρ as input, and he outputs a single bit: 0 if he guesses ρ is maximally mixed and 1 otherwise.

Following our analysis of the classical case, let's first suppose that the observer executes a particular fixed quantum circuit. That means the observer measures a particular Hermitian observable A with unit operator norm. Suppose we try to fool the observer by providing as input a pure state $\rho = |\psi\rangle\langle\psi|$. How well can the observable A distinguish this pure state from the maximally mixed state?

Suppose that $|\psi\rangle$ is chosen uniformly at random from among all n-qubit pure states. Then Levy's lemma [31] says that

$$\Pr\left(\left|\langle\psi|A|\psi\rangle - \frac{\mathrm{Tr}(A)}{2^n}\right| \geq \epsilon\right) \leq e^{-c\,2^n\epsilon^2} \tag{4.1}$$

for some constant c, where the probability is evaluated with respect to the invariant Haar measure on the n-qubit Hilbert space. This means that, for n large, the pure state $|\psi\rangle$ can be chosen so that $|\psi\rangle$ and the maximally mixed state are exponentially difficult to distinguish using the observable A. Furthermore, most pure states have this property.

Even a pure quantum state can pretend to be maximally mixed, and the observer will not know the difference![4]

As in the classical case we can strengthen this claim: The state $|\psi\rangle$ can be chosen so that $|\psi\rangle$ is hard to distinguish from the maximally mixed state not just for one fixed quantum circuit, but for *any* quantum circuit of reasonable size. To carry out this step of the argument, we'll need an upper bound on the number of quantum circuits of specified size; here we confront the subtlety that quantum circuits, unlike classical ones, form a continuum, but this wrinkle poses no serious obstacle to completing the argument. If we settle for specifying the unitary transformation realized by a circuit with m gates to constant accuracy, it suffices to specify each gate to $O(\log m)$ bits of precision. Therefore, as in the classical case, the complete circuit can be specified by $O(m \log m)$ bits. It follows that, if m is subexponential in n, then the number $N(m)$ of circuits with size m is the exponential of a function which is subexponential in n. In contrast, the right-hand side of equation (4.1) is the exponential of an exponential function of n. Using the union bound, we conclude that if the pure state $|\psi\rangle$ is chosen uniformly at random, it will, with high probability, be hard to distinguish $|\psi\rangle$ from the maximally mixed state using *any* circuit of size subexponential in n.

On the other hand, if we were not concerned about the complexity of the observer's task then it would be easy to distinguish $|\psi\rangle$ from the maximally mixed state. The observer could perform a projective measurement with the two outcomes $\{E_0 = I - |\psi\rangle\langle\psi|,\ E_1 = |\psi\rangle\langle\psi|\}$, guessing that the input state is $|\psi\rangle$ if he obtains the outcome E_1, and guessing that the input state is maximally mixed if he obtains the outcome E_0. This strategy always succeeds if the input is $|\psi\rangle$, and fails with the exponentially small probability 2^{-n} if the input is maximally mixed. The trouble is that, for a typical pure state $|\psi\rangle$, this measurement is far too complex to carry out in practice.

A typical pure quantum state is somewhat analogous to the distribution p_S we described in Section 3. In both cases, it is hard for an observer who is limited to performing polynomial-size computations to tell that the state is not uniformly random, even though an observer with unlimited computational power can tell the difference. Furthermore, both examples are subject to the same criticism — it is computationally hard to sample uniformly from Haar measure (that is, to prepare a "typical" pure state), just as it is computationally hard in the classical setting to sample from the the distribution p_S. In the quantum setting, as for the classical setting, we may ask a more nuanced question: Can quantum states be prepared *efficiently* which are hard to distinguish from maximally mixed states? This more nuanced question is the relevant one as we contemplate the properties of the radiation emitted by a partially evaporated black hole, because the formation and subsequent complete evaporation of a black hole can occur in a time that scales like $S_{\rm bh}^{3/2}$, where $S_{\rm bh}$ is the initial black hole entropy. Hence, the preparation of the Hawking radiation can be simulated accurately by an efficient quantum circuit.

The answer is yes (under a reasonable assumption), as was shown recent by Ji, Liu,

[4] If two identical copies of $|\psi\rangle$ are available, then it is easy to distinguish the pure state $|\psi\rangle$ from the maximally mixed state by conducting a swap test. Here we assumed that only a single copy is available.

and Song [12]; pseudorandom quantum states *can* be prepared efficiently. The assumption we need is the the existence of a family of *quantum-secure pseudorandom functions* $\{\mathrm{PRF}_k\}_{k\in K}$. This means that each PRF_k can be efficiently computed, but it is difficult to distinguish a randomly sampled member of $\{\mathrm{PRF}_k\}$ from a truly random function with any efficient quantum algorithm. The set K is called the *key space* of the function family. The existence of such pseudorandom functions follows from the existence of quantum-secure one-way functions, an assumption which is standard in cryptography.

The key idea is that we can construct a pseudorandom quantum state as a superposition of computational basis states, where all basis states appear with equal weight except for a phase, and the phases appear to be random to a computationally bounded observer. Specifically, we consider a family of states $\{|\phi_k\rangle\}_{k\in K}$

$$|\phi_k\rangle = \frac{1}{\sqrt{N}} \sum_{x\in X} \omega_N^{\mathrm{PRF}_k(x)} |x\rangle, \tag{4.2}$$

where $N = 2^n$, $\omega_N = e^{2\pi i/N}$, $X = \{0, 1, 2, \ldots, N-1\}$, and $\{\mathrm{PRF}_k : X \to X\}_{k\in K}$ is a family of quantum-secure pseudorandom functions. We can show that a uniform mixture of the states $\{|\phi_k\rangle\}$ is computationally indistinguishable from the maximally mixed state.[5]

We may argue as follows. First we consider the family of *all* functions $f_{k'} : X \to X\}_{k'\in K'}$ indexed by key space K', and the corresponding family of pure states

$$|f_{k'}\rangle = \frac{1}{\sqrt{N}} \sum_{x\in X} \omega_N^{f_{k'}(x)} |x\rangle. \tag{4.3}$$

The first thing to note is that $\{|f_{k'}\rangle\}$ is information-theoretically indistinguishable from Haar-random; we state this fact for the reader's convenience in Lemma 4.1.

Lemma 4.1 ([12], Lemma 1)**.** *Let $\{|f_{k'}\rangle\}$ be the family of states defined in equation (4.3). Then, for m polynomial in n, the state ensemble $\{|f_{k'}\rangle^{\otimes m}\}$ is statistically indistinguishable from the ensemble $\{|\psi\rangle^{\otimes m}\}$ where $|\psi\rangle$ is Haar-random, up to a negligible error.*

Furthermore, the ensemble $\{|\phi_k\rangle\}$ cannot be efficiently distinguished from the ensemble $\{|f_{k'}\rangle\}$. If it could be, then we could leverage this fact to efficiently distinguish $\{\mathrm{PRF}_k\}$ from a family of random functions [12], contradicting our assumption that $\{\mathrm{PRF}_k\}$ is a quantum-secure pseudorandom function family. It now follows that the ensemble $\{|\phi_k\rangle\}$ cannot be efficiently distinguished from a maximally mixed state.

So far we have shown that a uniform mixture of the states $\{|\phi_k\rangle\}$ is pseudorandom; it remains to show that this mixture can be prepared efficiently. We start with a product of qubits, each in the state $|0\rangle$, and apply a Hadamard gate to each qubit to obtain the state

$$\frac{1}{\sqrt{N|K|}} \sum_{x\in X} \sum_{k\in K} |x\rangle|k\rangle. \tag{4.4}$$

[5] In fact, we can simplify the construction. It was shown in [32] that the same family of states is still pseudorandom if we replace the root of unity ω_N by -1.

Next, we apply the quantum Fourier transform (which has complexity polynomial in n) to another n-qubit register that is initialized in the state $|00\ldots01\rangle$, obtaining

$$\frac{1}{N\sqrt{|K|}} \sum_{x,y\in X} \sum_{k\in K} |x\rangle|k\rangle\omega_N^{y}|y\rangle. \tag{4.5}$$

Now, we compute $\mathrm{PRF}_k(x)$ and subtract modulo N from the y register. This computation can be done efficiently because by assumption the PRF_k is an efficiently computable function; the resulting state is

$$\frac{1}{N\sqrt{|K|}} \sum_{x,y\in X} \sum_{k\in K} |x\rangle|k\rangle\omega_N^{y}|y-\mathrm{PRF}_k(x)\rangle. \tag{4.6}$$

After shifting the summation index y, we have, up to a global phase,

$$\frac{1}{N\sqrt{|K|}} \sum_{x,y\in X} \sum_{k\in K} \omega_N^{\mathrm{PRF}_k(x)}|x\rangle|k\rangle\omega_N^{y}|y\rangle = \frac{1}{\sqrt{|K|}} \sum_{k} |\phi_k\rangle|k\rangle|\mathrm{QFT}\rangle, \tag{4.7}$$

where $|\mathrm{QFT}\rangle = N^{-1/2}\sum_{y\in X}\omega_N^{y}|y\rangle$. After the key $|k\rangle$ is discarded, the marginal state over the first register is the uniform mixture of $\{|\phi_k\rangle\}$. Thus, we have prepared this mixture efficiently.

The definition of a pseudorandom quantum state in reference [12] is really overkill for our purposes. Those authors are concerned with cryptographic applications, and therefore consider a definition (as stated in Lemma 4.1) where, for each value k of the key, m identical copies of $|\phi_k\rangle$ are available where m is polynomial in n. We will not encounter such scenarios in this paper. Therefore, we may instead adopt a simplified definition of pseudorandomness which is more suitable for the application to black hole physics. In Definition 6.1 below, the size $|H|$ of the remaining black hole parametrizes how difficult it is to distinguish radiation emitted from the partially evaporated black hole from the maximally mixed state. In this sense, the remaining black hole H serves as the key space of the pseudorandom radiation state. Even if $|H|$ is less than half of the initial black hole entropy, this task remains difficult so long as the remaining black hole H is macroscopic. Our hypothesis that the Hawking radiation is pseudorandom provides a way to formalize the idea that the Hawking radiation is effectively thermal even when the state of E has relatively low rank because $|H| \ll |E|$.

5 Is Hawking radiation pseudorandom?

We have now seen, in both the classical and quantum settings, that pseudorandom states exist. Though in principle these states are almost perfectly distinguishable from maximally mixed states, in practice no observer with reasonable computational power can tell the difference. Moreover, under standard cryptographic assumptions, there exist constructions of such states which can be efficiently prepared. But up to this point we have not addressed whether pseudorandom quantum states can be efficiently prepared in plausible physical processes like the evaporation of a black hole.

In the case of a black hole which forms from gravitational collapse and then *completely* evaporates, the resulting state of the emitted Hawking radiation, though highly scrambled, would *not* be pseudorandom. We take it for granted that the time evolution of the quantum state can be accurately approximated by a quantum circuit, which has size polynomial in the initial black hole entropy S_{bh} because the evaporation process takes a time $O\left(S_{\text{bh}}^{3/2}\right)$. We may consider a toy model of this process, in which the initial state $|\phi_{\text{matter}}\rangle$ of the collapsing matter is a product state of n qubits $|\phi_{\text{matter}}\rangle = |0\rangle^{\otimes n}$, and the final state after complete evaporation is $|\Psi_{\text{fin}}\rangle = U|\phi_{\text{matter}}\rangle$, where U is a unitary transformation constructed as a polynomial-size circuit. In this case, an observer could just execute this circuit in reverse, hence applying $U^\dagger$ to $|\Psi_{\text{fin}}\rangle$, and then measure the qubits in the standard basis, thus easily distinguishing $|\Psi\rangle$ from the maximally mixed state. We see that, if ρ is a state that can be prepared by a polynomial-size quantum circuit, yet is hard to distinguish from maximally mixed by polynomial-size circuits, then ρ cannot be pure.

Instead, we consider a *partially* evaporated black hole as in Figure 3. We imagine that the n-qubit state ρ_{EB} is prepared by applying a polynomial-size unitary circuit U_{bh} to the initial state $|0\rangle^{\otimes n}|0\rangle^{\otimes k}$ of EBH, where H is a k-qubit system, and then discarding H. In our toy model, EB is the Hawking radiation that has been emitted so far, and H is the remaining black hole. (Recall that B is a small portion of the emitted Hawking radiation whose properties we will investigate later; for the purpose of the present discussion we are only interested in the state of EB, the full radiation system.) If our observer had access to H as well as EB, he could easily tell that the state is not maximally mixed, but what if H is inaccessible?

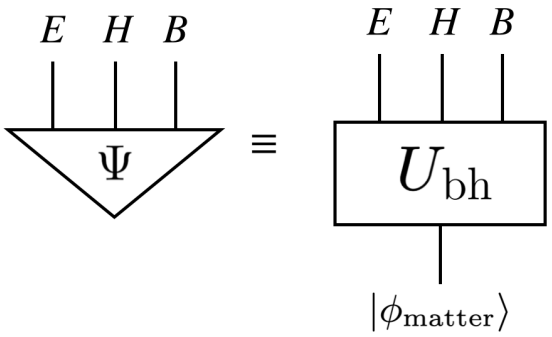

Figure 3. Our toy model of a partially evaporated black hole, where EB is the Hawking radiation emitted so far, and H is the remaining black hole. The initial state $|\phi_{\text{matter}}\rangle$ of the gravitationally collapsing matter is modeled as a product state. We conjecture that the unitary black hole dynamics prepares a pseudorandom state of EB.

We are particularly interested in the case where $1 \ll k = |H| < n = |EB|$, so that ρ_{EB} fails to have full rank, and must therefore be information-theoretically distinguishable from the maximally mixed state; this situation resembles the classical model discussed in Section 3, where the entropy of the distribution p_S is substantial but not maximal. Could the state ρ_{EB} of the Hawking radiation, which is prepared by unitary evolution of EBH for a time which is polynomial in $|EB|$, be pseudorandom?

This is a question about quantum gravity, and we don't know the answer for sure, but we can make a reasonable guess. We have already seen in Section 4 that quantum circuits exist that efficiently prepare pseudorandom quantum states. Since black holes are believed to be particularly potent scramblers of quantum information, it is natural to conjecture that the internal dynamics of a black hole can produce pseudorandom states as well. Indeed, we may expect similar behavior for the radiation emitted by other strongly chaotic quantum systems aside from black holes. Only for the case of black holes, though, where we face the daunting firewall puzzle, will our constructions of robust logical operators acting on EH seem to have a natural interpretation.

To better understand why the state of EB might be hard to distinguish from a maximally mixed state, we may suppose, for example, that ρ_H is maximally mixed so that the pure state of EBH has the form

$$|\Psi\rangle_{EBH} = \frac{1}{2^{|H|/2}} \sum_i |\psi_i\rangle_{EB} \otimes |i\rangle_H, \tag{5.1}$$

where the states $\{|\psi_i\rangle_{EB}\}$ are orthonormal. The marginal state of EB is then

$$\rho_{EB} = \frac{1}{2^{|H|}} \sum_i |\psi_i\rangle\langle\psi_i|. \tag{5.2}$$

Suppose the observer receives a state which is either ρ_{EB} or the maximally mixed state $\sigma_{EB} = I_{EB}/2^{|EB|}$. A natural test is as follows: The observer augments EB with the maximally mixed state of H (which is easy to prepare), and then measures the projection onto $|\Psi\rangle$. This can be done efficiently by applying $U_{\rm bh}^{-1}$ and then measuring in the standard basis. If the input state is σ_{EB}, the projection onto $|\Psi\rangle$ succeeds with probability $2^{-(|EB|+|H|)}$, while if the input state is ρ_{EB} the success probability is

$$\langle\Psi| \frac{\sum_i |\psi_i\rangle\langle\psi_i|}{2^{|H|}} \otimes \frac{I_H}{2^{|H|}} |\Psi\rangle = \frac{1}{2^{2|H|}}. \tag{5.3}$$

Thus this test distinguishes ρ_{EB} and σ_{EB}, but only with a probability that is exponentially small in H.

To conduct a better test we would somehow need to exploit the structure of the ensemble $\{|\psi_i\rangle_{EB}\}$. But if as we expect black holes are especially effective information scramblers, it is reasonable to suppose that the ensemble lacks any special properties that can be exploited by an observer who is limited to performing a polynomial-time quantum computation. If so, the Hawking radiation is pseudorandom, and the test we have described may be nearly optimal.

In the example above we have assumed that the radiation has infinite temperature. The actual behavior of a black hole evaporating in asymptotically flat spactime is more complicated — the temperature is actually finite, and in fact becomes hotter and hotter as the evaporation proceeds. Conceptually, though, the situation is similar to the idealized case of an black hole evaporating at infinite temperature. At early times, when $|H| \gg |EB|$, we expect the radiation emitted at a specified time to be information-theoretically indistinguishable from precisely thermal radiation at the same temperature. At late times,

when $|H| \ll |EB|$, the global state of the radiation is distinguishable in principle from a thermal state (with temperature varying according to the time of emission), but we assume that telling the difference is computationally hard because the radiation is highly scrambled.

We also note that the constructions in [12] reinforce earlier observations concerning the computational hardness of *decoding* the Hawking radiation [14, 15]. These authors considered the quantum state $|\Psi\rangle_{EBH}$ of an old black hole, and analyzed the task of extracting from the early radiation E the subsystem which is entangled with the recently emitted Hawking mode B. This task would be easy for an observer who has access to both E and H, but one can argue that there are efficiently preparable states of EBH for which this decoding task cannot be achieved by an observer who performs a polynomial-size quantum computation on E alone. Here, too, the hardness of decoding cannot be proven from first principles, but it follows from plausible complexity assumptions which are standard in "post-quantum" cryptography [14, 15]. Again, the existence of states that are hard to decode does not guarantee that a black hole creates such states, but we take it on faith that if efficient preparation of such states is possible, then a black hole will be up to the job.

To summarize, on the basis of these (admittedly speculative) considerations, we propose that for the quantum state $|\Psi\rangle_{EBH}$ of an old black hole, the state ρ_{EB} of the Hawking radiation is pseudorandom. If $|H| < |EB|$, then the rank of ρ_{EB} is not maximal, so that ρ_{EB} is distinguishable from a thermal state. In fact, an observer with access to H as well as EB could efficiently check that ρ_{EB} is not thermal. Furthermore, an observer without access to H could check that ρ_{EB} is not thermal by performing a quantum computation of exponential size on EB alone. But an observer outside the black hole, who performs a polynomial-size quantum computation on EB without access to H, will be able to distinguish ρ_{EB} from a thermal state with a success probability that is at best exponentially small in $|H|$. Our analysis of the robustness of the encoded black hole interior in the remainder of this paper will rest on this assumption.

This discussion highlights the importance of distinguishing the von Neumann entropy of the Hawking radiation from its thermodynamic entropy. After the Page time, the Von Neumann entropy of EB becomes far smaller than the von Neumann entropy of a perfectly thermal state, so one could in principle verify that the Hawking radiation is not perfectly thermal by measuring its von Neumann entropy. The existence of pseudorandom quantum states then implies that measuring the von Neumann entropy with a small error requires an operation of superpolynomial complexity [33]. One could imagine trying to measure the entropy of the radiation by, for example, withdrawing its thermal energy to operate a heat engine. If the radiation is pseudorandom, though, the radiation would be indistinguishable from thermal radiation in any efficient process, despite its low von Neumann entropy.

Recalling the construction of the pseudorandom state recounted in Section 4, we note [34] that the quantum Fourier transform can be executed with circuit depth $O(\log n)$, and that under plausible cryptographic assumptions the function PRF_k can be computed in depth polylog n [12]. Thus a pseudorandom state can be prepared in polylog n time. Plausibly, the state preparation can be achieved in a time comparable to the $O(\log n)$

scrambling time of a black hole, as one might naively expect.

6 Pseudorandomness and decoupling

In this section, we formalize our hypothesis that Hawking radiation is pseudorandom, and explore its implications regarding the firewall paradox [3]. Our analysis can be viewed as a refinement of the Harlow-Hayden argument [14, 15].

As formulated in [3] and summarized in Section 1, the firewall paradox highlights a conflict between the unitarity of black hole evaporation and the monogamy of entanglement. A possible resolution is that the interior mode $\tilde{B}$ that purifies a recently emitted Hawking mode B may actually be encoded in the radiation. On the face of it, this resolution flagrantly violates locality, and one wonders whether this violation of locality can be detected by an agent who first interacts with the radiation and then falls through the event horizon to visit the interior. We will argue that, provided the Hawking radiation is pseudorandom and the size of the observer is small compared to the black hole, the nonlocality is undetectable in practice because it would take an exponentially long time for the observer to distill the encoded interior mode before falling into the black hole.

We will first present a sketch of the argument in a simplified setting where the radiation interacts with a single observer who is significantly smaller than the remaining black hole. Later on we will extend the argument to the case where the observer has access to a large probe outside the horizon, whose size may be comparable to or even larger than the remaining black hole.

Recall our conventions: Let O denote the observer, H the remaining black hole, B the late outgoing mode, E the early radiation, and P the external probe. We will also refer to the joint system EB as the exterior radiation. All subsystems can be decomposed in terms of qubits, and our statements about computational complexity concern the number of steps in a computation executed by a universal quantum computer. Below and throughout the remainder of the paper, given an operator A, we will use $\|A\|_1 = \mathrm{Tr}(\sqrt{A^\dagger A})$ to denote the trace norm, $\|A\|_F = \sqrt{\mathrm{Tr}(A^\dagger A)}$ to denote the Frobenius norm, and $\|A\|$ to denote the operator norm.

First, we define what it means for the external radiation of the black hole to be pseudorandom.

Definition 6.1. *Let $|\Psi\rangle_{EBH}$ be the state of the black hole and the radiation. Let $\sigma_{EB} = I_{EB}/d_{EB}$ be the maximally mixed state of EB, and let $\rho_{EB} = \mathrm{Tr}_H(|\Psi\rangle\langle\Psi|)$. We say that the state $|\Psi\rangle_{EBH}$ is pseudorandom on the radiation EB, if there exists some $\alpha > 0$ such that*

$$|\Pr(\mathcal{M}(\rho_{EB}) = 1) - \Pr(\mathcal{M}(\sigma_{EB}) = 1)| \leq 2^{-\alpha|H|}, \tag{6.1}$$

for any two-outcome measurement $\mathcal{M}$ with quantum complexity polynomial in $|H|$, the size of the remaining black hole.

This definition captures the notion that no feasible measurement can tell the difference between ρ_{EB} and the maximally mixed state. A few remarks will help to clarify the

definition. (1) When we say a measurement of ρ_{EB} has polynomial quantum complexity, we mean it can be performed by executing a quantum circuit of polynomial size acting on EB, followed by a qubit measurement in the standard computational basis. Use of ancilla systems is also permitted in the measurement process, provided the ancilla is initialized in a product state. (2) Of particular interest is the value of the constant α that makes the bound in equation (6.1) tight for asymptotically large black holes. But because black holes are such effective information scramblers, we would expect a comparable value of α to apply also for black holes of moderate size. There is no obvious small parameter in the problem that would lead us to expect α to be small compared to 1. (3) This definition is appropriate for the case where the Hawking radiation has infinite temperature. As we remarked in Section 5, we expect the realistic case of finite-temperature radiation to be conceptually similar, and for similar conclusions to apply in that case. But we will stick with the infinite-temperature case for the rest of the paper to simplify our analysis.[6]

Let us now deduce a consequence of Definition 6.1. We introduce an observer subsystem O initialized in a state ω_O, and an ancilla subsystem P initialized in the product state $|0\rangle_P$. The main result of this section is the following: Suppose that $|\Psi\rangle_{EBH}$ is pseudorandom, and let ρ_{OPE} be any state of OPE obtained by applying a unitary of polynomial complexity to $\omega_O \otimes |0\rangle_P \otimes |\Psi\rangle_{EBH}$. Then the correlation between the observer and the early radiation is exponentially small in $|H|$ for any such state; *i.e.*,

$$\|\rho_{OB} - \rho_O \otimes \rho_B\|_1 \leq 6 \cdot 2^{-(\alpha|H|-|O|)}, \tag{6.2}$$

where we have now assumed that B is a single qubit. We will call (6.2) the *decoupling bound*, because it states that the observer O nearly decouples from the exterior radiation mode B, and therefore gains negligible information about the interior mode $\tilde{B}$ which is entangled with B. In Section 8 we leverage (6.2) to show that the interior mode $\tilde{B}$ can be regarded as an encoded subsystem of EH which is protected against all "low-complexity" errors, where "low-complexity" is shorthand for polynomial complexity.

Prior work [14, 15] has suggested that the decoupling bound holds when the size of the remaining black hole is an $O(1)$ fraction of the initial black hole entropy $S_{\rm bh}$. However, our conclusion goes further. Even if the majority of the initial black hole has evaporated, so that $|H| \ll |EB|$, the observer O and the late radiation B remain decoupled as long as the remaining black hole H is macroscopic and the observer's system O obeys $|O| \ll |H|$.

To derive the decoupling bound, we apply the pseudorandomness assumption to the setup described in Figure 1. The unitary $U_{\mathcal{E}}$ is applied to the radiation, probe, and the observer. Because the evaporation time of the black hole is polynomial in its size, and $U_{\mathcal{E}}$ is applied before the evaporation is complete, we may assume that $U_{\mathcal{E}}$ is applied in a polynomial time and therefore has polynomial complexity. We also assume that the initial

[6] In the finite temperature case, the entanglement between B and the rest of the system is no longer maximal. This causes an extra complication when we use the quantum error-correction technology in Section 7, because the encoding map V from B to EH defined by the state Ψ_{EBH} need not be exponentially close to an isometry. Instead we may replace V by the approximate isometry $V\rho_B^{-1/2}$, which slightly modifies the error bounds derived in Section 7 and 8. Similar techniques have been used in [7, 8, 10].

state of the observer ω_O is of low complexity, although this assumption is not crucial; we may take the state ω_O to be arbitrary, at the cost of a slightly weaker decoupling bound.

In order to bound the correlation between B and O, we consider a complete set of operators acting on OB. A convenient choice is the set of Pauli operators P_i acting on OB. By a Pauli operator acting on n qubits we mean a tensor product of n 2×2 Pauli matrices; there are 4^n such operators $\{P_i, i = 0, 1, 2, \ldots, 4^n - 1\}$ (where $P_0 = I$) whose phases can be chosen so that each P_i for $i \neq 0$ has eigenvalues ± 1, and the $\{P_i\}$ are orthogonal in the Froebenius norm: $\mathrm{Tr}\,(P_i P_j) = 2^n \delta_{ij}$. Here n is the number of qubits in OB. Because measurement of P_i is a low-complexity two-outcome measurement, it follows from the assumption that Ψ_{EBH} is pseudorandom that

$$|\mathrm{Tr}((\rho_{OB} - \sigma_{OB})P_i)| \leq 2^{-\alpha|H|} \tag{6.3}$$

for any Pauli operator P_i, where σ_{OB} is the state that results when the state ρ_{EB} measured by the observer is replaced by the maximally mixed state; see Figure 4.

To understand why equation (6.3) follows from pseudorandomness, note that we are modeling a measurement of EB by the observer O as a low-complexity unitary interaction between EB and O, followed by a simple measurement of the O register. Strictly speaking, then, we should allow the Pauli operator P_i to act only on O, not on OB. In effect, we are assuming that the observer's quantum memory contains $|OB|$ qubits rather than $|O|$ qubits, so that measuring a Pauli operator acting on OB is permitted. In our formulation of the pseudorandomness assumption, there is no restriction on the size of the observer's memory, only on the complexity of his operation. Therefore, assuming that the state of EB is pseudorandom, the observer's measurement will not distinguish ρ_{OB} from σ_{OB} even if the observer is permitted to measure B as well as O.

Using the completeness and orthogonality of the Pauli operators, we can bound the Frobenius distance between the two states as

$$\|\rho_{OB} - \sigma_{OB}\|_F^2 = \mathrm{Tr}\left((\rho_{OB} - \sigma_{OB})^2\right) \tag{6.4}$$

$$= 2^{-(|OB|)} \sum_i |\mathrm{Tr}((\rho_{OB} - \sigma_{OB})P_i)|^2 . \tag{6.5}$$

Because there are $4^{|OB|}$ Pauli operators, the right hand side is bounded by $2^{-2\alpha|H|}2^{|OB|}$. The trace distance is bounded by the Frobenius norm as

$$\|\rho\|_1 \leq \sqrt{\mathrm{rank}(\rho)}\, \|\rho\|_F, \tag{6.6}$$

for any operator ρ. Therefore we have

$$\|\rho_{OB} - \sigma_{OB}\|_1 \leq 2^{|OB|}2^{-\alpha|H|}, \tag{6.7}$$

because the rank of ρ_{OB} can be no larger than $2^{|OB|}$. From (6.7), one finds that

$$\begin{aligned} \|\rho_{OB} - \rho_O \otimes \rho_B\|_1 &\leq \|\rho_{OB} - \sigma_{OB}\|_1 + \|\sigma_{OB} - \rho_O \otimes \rho_B\|_1 \\ &\leq 2^{|OB|-\alpha|H|} + \|\sigma_{OB} - \rho_O \otimes \rho_B\|_1 \\ &= 2^{|OB|-\alpha|H|} + \|\sigma_O \otimes \sigma_B - \rho_O \otimes \rho_B\|_1 \\ &\leq 2^{|OB|-\alpha|H|} + \|\sigma_O - \rho_O\|_1 + \|\sigma_B - \rho_B\|_1 \\ &\leq 3 \times 2^{(|OB|-\alpha|H|)}. \end{aligned} \tag{6.8}$$

The first line is the triangle inequality. From the first line to the second line, we used (6.7). From the second line to the third line we used the fact that σ_{OB} is a product state over O and B. From the third line to the fourth line, we used the fact that

$$\begin{aligned} \|\sigma_O \otimes \sigma_B - \rho_O \otimes \rho_B\|_1 &\leq \|(\sigma_O - \rho_O) \otimes \sigma_B\|_1 + \|\rho_O \otimes (\sigma_B - \rho_B)\|_1 \\ &\leq \|\sigma_O - \rho_O\|_1 + \|\sigma_B - \rho_B\|_1, \end{aligned} \tag{6.9}$$

where the first line of (6.9) follows from the triangle inequality, and the second from the property that tracing out a subsystem cannot increase the trace distance. To reach the last line of (6.8), we again used the property that tracing out a subsystem cannot increase the trace distance. Finally, in the case where B is a single qubit, so that $|OB| = |O| + 1$, (6.8) becomes the decoupling bound (6.2). More generally, decoupling is satisfied whenever $|OB| \ll \alpha|H|$.

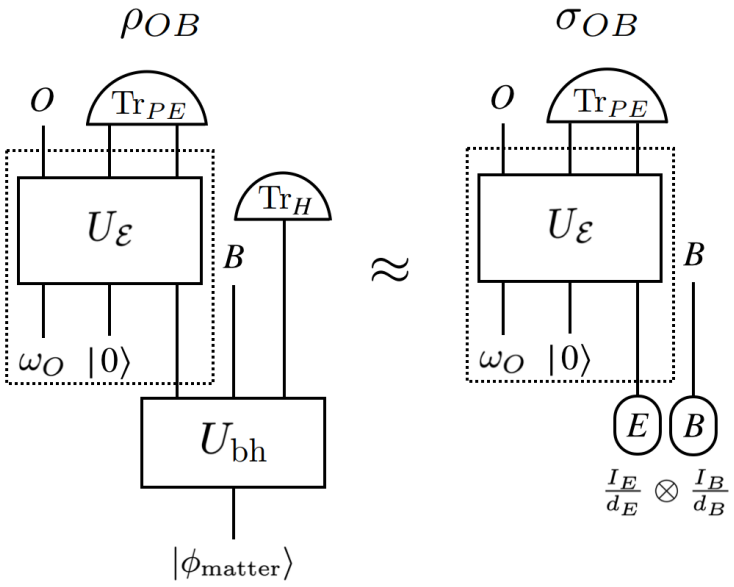

Figure 4. Graphical depiction of the decoupling bound, which follows from the pseudorandomness of the Hawking radiation emitted by an old black hole. On the left, a black hole forms from collapse and partially evaporates; the emitted radiation is EB and the remaining black hole is H. Then an observer O and probe P interact with the radiation subsystem E for a time that scales polynomially with the initial black hole entropy S_{bh}. On the right, the unitary transformation describing the interaction of OPE is the same as on the left, but the state of the Hawking radiation is replaced by a maximally mixed state of EB. The decoupling bound asserts that the final state of OB is the same in both cases, up to an error that is exponentially small in $|H|$, the size of the remaining black hole, provided that $|O| \ll |H|$.

Because two states close in trace distance cannot be distinguished well by any measurement, the decoupling bound implies that the state ρ_{OB} cannot be distinguished from the state σ_{OB} assuming that $|O| \ll |H|$. We thus conclude that any subsystem small compared to the remaining black hole H, even after interacting with the early radiation E, cannot be correlated with B; see Figure 4. In particular, an observer outside the black hole who

interacts with E for a polynomially bounded time remains decoupled from B, assuming that the Hawking radiation is pseudorandom.

This conclusion about the hardness of decoding follows from the pseudorandomness assumption for any computationally bounded observer who can access only system E. However, the decoding becomes easy if the observer has access to both E and H, as long as the state $|\Psi\rangle_{EBH}$ has polynomial complexity. For this case, we will describe an explicit decoding protocol in Section 9.

7 Black hole as a quantum error-correcting code

In this section, we recast the findings in Section 6 in the language of quantum error correction. The quantum error correction point of view will prove to be useful in understanding more subtle thought experiments studied in Section 8. We will see that an old black hole, together with its previously emitted Hawking radiation, is a quantum error-correcting code with exotic properties that have not been noted in previous discussions of holographic quantum error-correcting codes [35, 36]. These properties hold if the Hawking radiation is pseudorandom. That a black hole can be viewed as a quantum error-correcting code is not new [17, 36–38]. What's new is that a black hole can protect quantum information against seemingly pernicious errors; we refer to these as "low-complexity errors," meaning errors inflicted by a malicious agent who performs a quantum computation on the Hawking radiation with complexity scaling polynomially in the size of the remaining black hole.

To explain this claim, it is useful to view the state of the black hole and the radiation as an encoding map from the interior mode $\tilde{B}$ into EH. That is, $|\Psi\rangle_{EHB}$ defines an isometric embedding of $\tilde{B}$ into EH. Recall that E denotes the early radiation, H denotes the remaining black hole, and B denotes a late outgoing mode. For simplicity, we assume that B is a single qubit, but the following results remain essentially unchanged for B of any constant size (small compared to H). The encoded system $\tilde{B}$ describes the mode in the black hole interior that is entangled with B.

We can define an (approximate) isometric embedding $V_\Psi : \mathcal{H}_{\tilde{B}} \to \mathcal{H}_{EH}$ of a single qubit $\tilde{B}$ into the subspace EH by

$$V_\Psi |i\rangle_{\tilde{B}} = 2(I_{EH} \otimes \langle\omega|_{B\tilde{B}})(|\Psi\rangle_{EHB} \otimes |i\rangle_{\tilde{B}}), \tag{7.1}$$

where $|\omega\rangle_{B\tilde{B}} = 2^{-1/2}(|00\rangle_{B\tilde{B}} + |11\rangle_{B\tilde{B}})$ denotes an EPR pair on $B\tilde{B}$; see Figure 5. While V_Ψ itself is not precisely an isometric embedding, it is exponentially close to one under the assumption that $|\Psi\rangle_{EBH}$ is pseudorandom on the exterior system EB, as specified in Definition 6.1. In Appendix A, we show that there exists an isometric embedding V such that

$$\|V - V_\Psi\| \leq 2 \cdot 2^{-\alpha|H|}, \tag{7.2}$$

where $\|\cdot\|$ denotes the operator norm. The isometry V then defines a code subspace that encodes $\tilde{B}$. For macroscopic observers (*i.e.*, $|O| \gg 1$), the error in (7.2) is negligible compared to the error in the decoupling bound (6.2). Although the norm in equation (7.2)

is the operator norm rather than the trace norm, that distinction need not concern us if $|B|$ is sufficiently small compared to $|H|$. Therefore we can ignore any differences between V and V_Ψ and use them interchangeably.

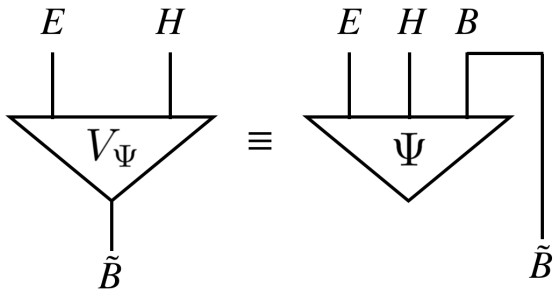

Figure 5. The definition of the encoding of $\tilde{B}$ into EH, with Ψ defined as in Figure 3.

We will now show that the isometry $V_\Psi : \mathcal{H}_{\tilde{B}} \to \mathcal{H}_{EH}$ defined above embeds $\tilde{B}$ into EH as a code subspace for which any low-complexity noise model acting on E is (approximately) correctable. By low-complexity error, we mean that the unitary process $U_\mathcal{E}$ in Figure 6 has complexity at most polynomial in $|H|$. Here the external observer O plays the role of the "environment" for the noise process acting on E and the probe P.

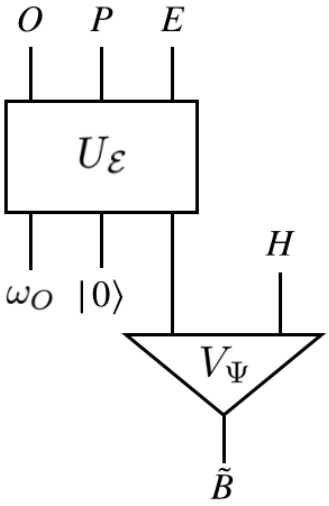

Figure 6. A black hole can be viewed as a quantum error-correcting code. By tracing out the observer O, we obtain a "noise model" $\mathcal{E}$ on the early radiation and the probe.

The error model depicted in Figure 6 is rather exotic compared to error models that are typically considered in discussions of quantum gravity and fault-tolerant quantum computing. For example, one widely studied error model is the "erasure model," wherein each qubit may be removed with some probability, and we know which qubits are removed. The performance of quantum codes against erasure errors arises, in particular, in studies of the holographic AdS/CFT dictionary [35, 36]; if a logical bulk operator can be "reconstucted"

on a portion of the boundary, that means that erasure of the complementary portion of the boundary is correctable for that logical operator. By the no-cloning theorem, no code can tolerate erasure of more than 50% of the qubits in the code block. In contrast, in our setup, erasure is correctable even if most of the qubits are removed. The catch is that the erased qubits must lie in E; removal of qubits in H is not allowed.

In studies of fault-tolerant quantum computing, the noise afflicting the physical qubits is usually assumed to be weak and weakly correlated. In a Hamiltonian formulation of the noise model, this means that each qubit in the computer is weakly coupled to a shared environment [39]. In contrast, for the noise model described by $U_{\mathcal{E}}$, the noise may act strongly on all the qubits in E; the only restriction is that the noise has quantum complexity scaling polynomially with $|H|$. Furthermore, how the noise acts depends on the initial state ω_O of the observer O, which may be chosen adversarially. Again, what makes successful error correction possible is that the subsystem H is assumed to be noiseless, an assumption that would be unrealistic for typical quantum computing hardware.

Codes that can protect against this malicious typle of noise are central to our proposed resolution of the firewall paradox. An old black hole provides such a code if its previously emitted radiation is pseudorandom. The code corrects errors successfully if the noise acting on E has low complexity and the remaining black hole H is noiseless, provided that the observer O is small compared to H.

7.1 Correcting low-complexity errors

For simplicity, we will first consider a scenario without the probe P shown in Figure 6. We will see that the error applied to the radiation system E is (approximately) correctable. In Section 7.2 we will explain how our conclusion changes when the probe P is included.

A central result in the theory of quantum error correction is the *information-disturbance relation*, which states that a code can protect quantum information from noise if and only if the "environment" of the noise channel $\mathcal{E}$ learns nothing about the logical information. More precisely, there is a physical process $\mathcal{R}$, the recovery process, which reverses the error:

$$\mathcal{R} \circ \mathcal{E} \approx \mathcal{I}, \tag{7.3}$$

where $\mathcal{I}$ is the identity operation, if and only if the "reference system" that purifies the quantum error-correcting code decouples from the environment . In Figure 6 (neglecting the probe P), the environment of the noise channel $\mathcal{E}$ acting on E is the observer O, and B is the reference system that purifies the encoded interior mode $\tilde{B}$. Therefore the necessary and sufficient condition for (approximate) correctability is the (approximate) decoupling of B and O,

$$\rho_{OB} \approx \rho_O \otimes \rho_B \tag{7.4}$$

where ρ_{OB} is the reduced density operator for OB. Here the approximation errors of equation (7.3) and equation (7.4) are related to each other by a constant factor. Therefore, using the decoupling bound (6.2), we can conclude that there exists a recovery process $\mathcal{R}$ that reverses $\mathcal{E}$ up to an error exponentially small in $|H|$, as long as $|O| \ll |H|$ and assuming that the Hawking radiation is pseudorandom.

Various formal statements that imply the existence of $\mathcal{R}$ in equation (7.3) are known; for the reader's convenience, we reproduce some of these results below. First, let us properly define what it means for a channel to be approximately correctable with respect to some code subspace — a more comprehensive discussion can be found in [23]. Let $S(\mathcal{H})$ denote the set of states on a Hilbert space $\mathcal{H}$. Suppose that we are given channels $\mathcal{E}, \mathcal{N} : S(\mathcal{H}) \to S(\mathcal{H})$. Fixing a state $\rho \in S(\mathcal{H})$, we define the *entanglement fidelity* between $\mathcal{E}$ and $\mathcal{N}$ with respect to ρ to be

$$F_\rho(\mathcal{E}, \mathcal{N}) = f\left[(\mathcal{E} \otimes \mathcal{I})(|\psi\rangle\langle\psi|), (\mathcal{N} \otimes \mathcal{I})(|\psi\rangle\langle\psi|)\right], \tag{7.5}$$

where $|\psi\rangle$ is a purification of ρ, and where

$$f(\rho, \tau) = \mathrm{Tr}\left(\sqrt{\sqrt{\tau}\rho\sqrt{\tau}}\right) \tag{7.6}$$

is the usual fidelity between states ρ and τ. To quantify the closeness of two channels, we use the worst-case entanglement fidelity to define the *Bures distance*, given by

$$\mathfrak{B}(\mathcal{E}, \mathcal{N}) = \max_\rho \sqrt{1 - F_\rho(\mathcal{E}, \mathcal{N})}; \tag{7.7}$$

we sometimes define a more restricted notion of the Bures distance, where we maximize over states in some specified subspace. In discussions of error correction, we say that a noise channel $\mathcal{E}$ is ϵ-correctable with respect to a code subspace $\mathcal{C} \subseteq \mathcal{H}$ if there exists a recovery channel $\mathcal{R}$ such that

$$\mathfrak{B}(\mathcal{R} \circ \mathcal{E}, \mathcal{I}) \leq \epsilon, \tag{7.8}$$

where the maximization in the Bures metric is over all code states ρ with support on $\mathcal{C}$.

The Bures metric is bounded above and below by the trace norm as

$$2\mathfrak{B}^2(\mathcal{E}, \mathcal{N}) \leq \max_\rho \|(\mathcal{E} \otimes \mathcal{I})(|\psi\rangle\langle\psi|) - (\mathcal{N} \otimes \mathcal{I})(|\psi\rangle\langle\psi|)\|_1 \leq 2\sqrt{2}\mathcal{B}(\mathcal{E}, \mathcal{N}). \tag{7.9}$$

The norm in the middle is essentially the diamond-norm distance between the channels $\mathcal{E}$ and $\mathcal{N}$ [40], except that for the purpose of characterizing error correction the maximization is over code states only. Applying this inequality and tracing out the purifying system, the ϵ-correctability of a channel $\mathcal{E}$ implies that we have

$$\max_\rho \|(\mathcal{R} \circ \mathcal{E})(\rho) - \rho\|_1 \leq 2\sqrt{2}\epsilon, \tag{7.10}$$

where again the maximization is over code states.

As mentioned previously, an important result characterizing approximate correctability is the information-disturbance trade-off, which we now state quantitatively. Let $\mathcal{E} : S(\mathcal{H}_A) \to S(\mathcal{H}_A)$ be a noise channel acting on a system A, and let $V : \mathcal{H}_A \to \mathcal{H}_F \otimes \mathcal{H}_A$ be an isometry which purifies $\mathcal{E}$; i.e.,

$$\mathcal{E}(\rho) = \mathrm{Tr}_F(V\rho V^\dagger). \tag{7.11}$$

Hence F is the environment of the channel; we have resisted the temptation to denote the environment by E to avoid confusion with our convention that E denotes a subsystem of the Hawking radiation. Then the *complementary channel* $\widehat{\mathcal{E}} : S(\mathcal{H}_A) \to S(\mathcal{H}_F)$ is defined by

$$\widehat{\mathcal{E}}(\rho) = \mathrm{Tr}_A(V\rho V^\dagger). \tag{7.12}$$

A special case of interest is the identity channel $\mathcal{I}$. Taking the environment to be 1-dimensional, the complementary channel to the identity channel is simply the (partial) trace

$$\widehat{\mathcal{I}}(\rho) = \mathrm{Tr}_A(\rho). \tag{7.13}$$

Then the information-disturbance trade-off states the following:

Theorem 7.1 ([23], Theorem 1)**.** *Let $\mathcal{C} \subseteq \mathcal{H}_A$ be a code subspace. Let $\mathcal{E} : S(\mathcal{H}_A) \to S(\mathcal{H}_A)$ be an error channel. Then*

$$\inf_{\mathcal{R}} \mathfrak{B}\left(\mathcal{R} \circ \mathcal{E}, \mathcal{I}\right) = \inf_{\mathcal{R}'} \mathfrak{B}\left(\widehat{\mathcal{E}}, \mathcal{R}' \circ \mathrm{Tr}\right), \tag{7.14}$$

where the infimums are taken over all channels $\mathcal{R} : S(\mathcal{H}_A) \to S(\mathcal{H}_A)$, and $\mathcal{R}' : \mathbb{R} \to S(\mathcal{H}_F)$.

Note that a channel $\mathcal{R}' : \mathbb{R} \to S(\mathcal{H}_F)$ is just state preparation on the channel environment $\mathcal{H}_F$, i.e., every such channel $\mathcal{R}'$ is uniquely identified with a state $\sigma_F \in S(\mathcal{H}_F)$ such that

$$(\mathcal{R}' \circ \mathrm{Tr})(\rho) = \mathrm{Tr}(\rho)\, \sigma_F, \tag{7.15}$$

so we can equivalently write

$$\inf_{\mathcal{R}'} \mathfrak{B}\left(\widehat{\mathcal{E}}, \mathcal{R}' \circ \mathrm{Tr}\right) = \inf_{\sigma_F} \mathfrak{B}\left(\widehat{\mathcal{E}}, \sigma_F \otimes \mathrm{Tr}\right). \tag{7.16}$$

Now let's see what equation (7.14) tells us in the context of the black hole error-correcting code defined by the (approximate) isometry V_Ψ. Let $\tilde{\rho}_{\tilde{B}}$ be a logical state and let $\tilde{\rho}_{\tilde{B}B}$ be a purification. The isometry V_Ψ then embeds $\tilde{\rho}_{\tilde{B}B}$ as a (purified) code state ρ_{EHB}:

$$\rho_{EHB} = V_\Psi \tilde{\rho}_{\tilde{B}B} V_\Psi^\dagger. \tag{7.17}$$

Let $\mathcal{E} : S(\mathcal{H}_E) \to S(\mathcal{H}_E)$ be an arbitrary channel acting on E such that some purification $U_\mathcal{E}$ of $\mathcal{E}$ has low-complexity (see the set-up described in Figure 1). Let

$$\sigma_{OEHB} = U_\mathcal{E}(\omega_O \otimes \rho_{EHB})U_\mathcal{E}^\dagger \tag{7.18}$$

denote the overall post-evolution state. To apply Theorem 7.1, let us consider the error channel $\mathcal{E} \otimes \mathcal{I}_H$. Then the environment of the channel $\mathcal{E} \otimes \mathcal{I}_H$ is the observer subsystem O,

and the complementary channel $\widehat{\mathcal{E} \otimes \mathcal{I}_H}$ maps $S(\mathcal{H}_{EH})$ to $S(\mathcal{H}_O)$. From (7.18), the state obtained from ρ_{EHB} after the application of $\widehat{\mathcal{E} \otimes \mathcal{I}_H}$ is precisely given by

$$\left(\widehat{\mathcal{E} \otimes \mathcal{I}_H} \otimes \mathcal{I}_B\right)(\rho_{EHB}) = \mathrm{Tr}_{EH}(\sigma_{OEHB}) = \sigma_{OB}. \tag{7.19}$$

Since σ_{OEHB} was a state obtained through acting on the black hole code state ρ_{EHB} with a low-complexity unitary, it follows by the pseudorandom hypothesis that the decoupling bound (6.2) holds. Therefore we have

$$\|\sigma_{OB} - \sigma_O \otimes \sigma_B\|_1 \leq 6 \cdot 2^{-(\alpha|H|-|O|)}. \tag{7.20}$$

Finally, since $U_{\mathcal{E}}$ is supported away from B, we have $\sigma_B = \rho_B$, and so

$$\left\|\left(\widehat{\mathcal{E} \otimes \mathcal{I}_H} \otimes \mathcal{I}_B\right)(\rho_{EHB}) - \sigma_O \otimes \rho_B\right\|_1 \leq 6 \cdot 2^{-(\alpha|H|-|O|)}. \tag{7.21}$$

This holds for all code states, so (7.21), together with the first inequality in (7.9), implies that we have

$$\inf_{\sigma_O} \mathfrak{B}\left(\widehat{\mathcal{E} \otimes \mathcal{I}_H}, \sigma_O \circ \mathrm{Tr}_{EH}\right) \leq \sqrt{3} \cdot 2^{-(\alpha|H|-|O|)/2}. \tag{7.22}$$

Therefore, the channel $\mathcal{E}$ is approximately correctable by Theorem 7.1. We state this as a Lemma.

Lemma 7.2. *Let V_Ψ be the approximate isometric embedding defined by the state Ψ_{EHB}. Let $\mathcal{E}$ be an error channel on E with purification $U_{\mathcal{E}}$. Suppose that the decoupling bound (6.2) holds. Then $\mathcal{E}$ is ϵ-correctable for V_Ψ, where*

$$\epsilon = \sqrt{3} \cdot 2^{-(\alpha|H|-|O|)/2}, \tag{7.23}$$

if B is a single qubit. For general $|B|$, we have

$$\epsilon = \sqrt{\frac{3}{2}} \cdot 2^{-(\alpha|H|-|OB|)/2}. \tag{7.24}$$

Note that the recovery operator $\mathcal{R}$ acts on EH rather than E. The same will be true for the ghost logical operators we construct in Section 8.

7.2 Including the probe

We would now like to consider a modified scenario in which both the observer O and a probe P interact with the Hawking radiation system E, as indicated in Figure 1. We cannot simply absorb P into O, because we will continue to insist that O is small compared to H, while we wish to allow P to be comparable to H in size, or even larger. In this modified scenario, the unitary purification $U_{\mathcal{E}}$ of the noise model acts on OPE rather than OE. This change does not alter the conclusion that O and B decouple if $U_{\mathcal{E}}$ has low complexity. Therefore, just as before, there is a recovery map that reverses the effect of the noise on the encoded state. What changes is that now the recovery map acts on PEH rather than EH.

We emphasize that if the probe P is sufficiently large, then P need not decouple from B, even if $U_{\mathcal{E}}$ has low complexity. To understand why not, suppose P has the same size as the system E and that the channel $\mathcal{E}$ swaps P and E. Before this swap, B is entangled with the code space embedded in EH; therefore after the swap (a low-complexity operation), B is entangled with PH. More realistically, we might imagine that P is a cloud of dust surrounding the black hole, and that $|P| \gg |E|$. After the dust interacts with the Hawking radiation, the encoding of $\tilde{B}$ will be modified, so that B is entangled with a code subspace of PEH rather than a subspace of EH [9].

However, any subsystem of OP which is small compared to H will decouple from B, as long as $U_{\mathcal{E}}$ has low complexity, and assuming that the Hawking radiation is pseudorandom. The only way to distill the encoded state into a small subsystem is to perform a high complexity operation. Hence, if only low-complexity operations are allowed, we need not worry about a scenario in which the encoded version of $\tilde{B}$ outside the horizon is decoded into a small system, and then falls into the black hole to meet its twin in the interior. This is essentially the observation of Harlow and Hayden [14], later extended by Aaronson [15]. Our analysis goes further by clarifying that the encoded state is hard to distill even when the remaining black hole H is much smaller than E, as long as H is macroscopic and assuming that the Hawking radiation is pseudorandom.

One might wonder whether the encoded mode can be easily extracted if the probe system P is prepared in a carefully chosen state [41]. Our conclusion is that any such initial state of P would need to have exponential complexity, an unlikely property for the dust surrounding an evaporating black hole. One might also ask what happens if all the qubits in the early radiation system E are measured in the standard basis by the observer. Surely this *would* disrupt the encoded interior of the black hole. But in our model the number of radiation qubits that can be measured is limited by the size $|O|$ of the observer's memory, and the interior will stay well protected as long as $|O|$ is much smaller than $|H|$.

It is also instructive to view the system O in a different way. Up to now we have regarded O as a potentially malicious agent who attempts to damage the encoded interior of the black hole by acting on its exterior. More prosaically, we can think of O as an abstract purifying space which is introduced for convenience so that we can describe the noise channel $\mathcal{E}$ using its purification, the unitary transformation $U_{\mathcal{E}}$. From that point of view, limiting the size $|O|$ of the "observer" O is just a convenient way of restricting the form of the quantum channel $\mathcal{E}$. Specifically, the rank of the marginal density operator ρ_O after $U_{\mathcal{E}}$ is applied is called the Kraus rank (or simply the rank) of the channel $\mathcal{E}$. This rank can be no larger than the dimension of system O, namely $2^{|O|}$, which we have assumed to be small compared to the dimension $2^{|H|}$ of the Hilbert space of black hole microstates. Thus our conclusion can be restated: If the Hawking radiation is pseudorandom and H is macroscopic, then the quantum error-correcting code protects the encoded version of $\tilde{B}$ against any noise channel acting on PE that has both low complexity *and* low rank.

An advantage of this viewpoint is that one might otherwise be misled into interpreting $|O|$ as the physical size of an actual observer. More accurately, it can be regarded as the effective size of the quantum memory of a physical object. This distinction is significant. For an object of specified mass, the largest possible quantum memory is achieved by a

black hole of that mass, but the memory size of a quantum computer typically falls far short of that optimal value, because most of its mass is locked into the rest mass of atomic nuclei and unavailable for information processing purposes. Furthermore, the mass per unit volume of a typical quantum computer is far smaller than a black hole's. Therefore it is reasonable to expect that the effective Hilbert space dimension of system O (and hence the Kraus rank of the channel $\mathcal{E}$) is far smaller than the Hilbert space dimension of a black hole with the same circumference as system O.

Up until now we have mostly focused on the hardness of decoding the black hole interior mode by acting on the Hawking radiation outside the black hole, concluding that distilling the encoded system to a small quantum memory is computationally hard if the remaining black hole is macroscopic. In section 8 we will turn to a more subtle question: Can a low-complexity operation acting on the Hawking radiation system E create an excitation near the black hole horizon that could be detected by an infalling observer who falls into the black hole? Here too we will argue that the answer is no. This is a nontrivial extension beyond what we have found so far — on the face of it, perturbing a quantum state is a far easier task than depositing the state in a compact quantum memory.

Bousso emphasized that if the interior mode $\tilde{B}$ is encoded in EH, and if effective quantum field theory on curved spacetime is a good approximation in regions of low curvature, then the vacuum near the black hole horizon would need to be "frozen" [9]. That is, neither a small agent O acting on E nor a large probe P interacting with E could disrupt the entanglement of $\tilde{B}$ with B and hence create an excitation localized near the horizon. We agree with this conclusion, provided that $|H| \gg 1$ and that the interactions of OP with E have quantum complexity scaling polynomially with $|H|$. Interactions with the large probe may alter how the black hole interior is encoded in the radiation and probe, but they do not disrupt the frozen vacuum.

Once $|H|$ is O(1), large corrections to effective field theory may be expected. Furthermore, the semiclassical structure of spacetime may no longer be applicable in the regime where operations of superpolynomial complexity are allowed; these high-complexity operations could tear spacetime apart. In particular, our expectation that an agent acting on E should be unable to influence the black hole interior might be flagrantly violated if the agent can perform high-complexity operations. We should grow accustomed to the notion that for effective field theory to be an accurate approximation, we require not only geometry with low curvature and states with low energy, but also operations with low complexity and low Kraus rank.

To investigate whether the semiclassical causal structure is robust with respect to low-complexity operations we will need to develop some additional formalism, specifically the theory of *ghost logical operators*; in the context of an old black hole, these may be viewed as operators which act on the black hole interior. We would like to understand, given that the interior is encoded in the Hawking radiation outside the black hole, why low-complexity operations acting on the Hawking radiation produce no detectable excitations inside the black hole. We turn to that task next.

8 Theory of ghost logical operators

So far, we have argued that the late radiation system B remains decoupled from any sufficiently small subsystem of the early radiation E and the probe P, when the observer O performs a low-complexity operation on EP. Therefore an infalling observer with reasonable computational power is prevented from extracting the encoded interior mode before jumping into the black hole. But what if the observer settles for the seemingly easier task of disrupting the interior rather than decoding it? In this section we will show that an algebra of *ghost logical operators* can be constructed acting on the interior mode, with the property that low-complexity operations performed outside the black hole nearly commute with the ghost algebra. Hence, if these ghost operators are regarded as operations that can be performed by an observer inside the black hole, we may conclude that the interior is well protected against the actions of malicious agents outside the black hole.

Following arguments from [11], consider an operator T which acts on the interior mode. Because the corresponding encoded operator acting on the Hawking radiation is highly scrambled, the commutator of this encoded operator with a generic simple operator acting on the radiation has no reason to be small. It seems, then, that an external observer should be able to perturb the interior mode easily [9, 11]. Can this conclusion be evaded by constructing the encoded operators suitably? For two-side black holes in AdS/CFT, Papadodimas and Raju argued that "mirror operators" with the desired properties can be constructed [7, 10], but no satisfactory construction is known for evaporating black holes.

Within our simple toy model of evaporating black hole, we can construct analogues of the mirror operators. Assume that the decoupling bound (6.2) holds. Then, as we will see, for every operator $\tilde{T}_B$ acting on some outgoing mode B, there exists a "mirror operator" T_{EH} acting on EH which satisfies the following conditions:

$$\begin{aligned} \tilde{T}_B|\Psi\rangle &\approx T_{EH}|\Psi\rangle, \\ [T_{EH}, E_a]|\Psi\rangle &\approx 0, \end{aligned} \tag{8.1}$$

where $\{E_a\}$ is a set of operators that a computationally bounded external observer can apply on the radiation, and $|\Psi\rangle$ is the state of the radiation and the black hole. The equations (8.1) hold up to an error exponentially small in $|H|$. The first line implies that one can (in principle if not in practice) certify entanglement between an outgoing radiation mode and an abstract subsystem specified by the operators $\{T_{EH}\}$.[7] Therefore, these operators satisfy the right measurement statistics expected for sensibly defined interior operators. The second line implies that these operators approximately commute with all the operators that the external observer can apply. The fact that T_{EH} commutes with $\{E_a\}$ holds as an operator equation on all the states in the code subspace. Therefore, the subsystem specified by the mirror operators $\{T_{EH}\}$ is fully entangled with the late outgoing radiation modes while also being effectively "space-like separated" from the external observer. That is, the external observer can disrupt the semiclassical causal structure of the black hole only by applying operations of superpolynomial complexity to the radiation.

[7] For example, one could perform Bell tests using the Pauli operators acting on B and its mirror.

In our construction, it is important to properly characterize the set $\{E_a\}$ of operators that the exterior observer can apply to the radiation. If we view the observer, the black hole, and the exterior radiation as a closed system, we ought to model the entire evolution as a unitary process. In order to enforce the unitarity of this process, the operator applied by the observer to the radiation should depend on the initial state of the observer, as in Figure 7.

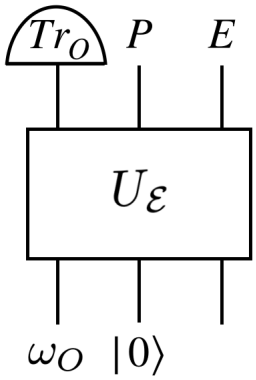

Figure 7. The operator applied by an exterior observer to the Hawking radiation depends on the observer's initial state ω_O, the probe's initial state $|0\rangle$, and the joint unitary transformation $U_{\mathcal{E}}$.

In this scenario, the set of operations that the observer can apply to the radiation is not completely arbitrary. Specifically, any such operation must be of the following form:

$$\rho_{PE} \mapsto \mathrm{Tr}_O\big(U_{\mathcal{E}}(\omega_O \otimes \rho_{PE})U_{\mathcal{E}}^\dagger\big), \tag{8.2}$$

wherein the only freedom available to the observer is the choice of the initial state ω_O. Because the observer is part of a system that is governed by the laws of physics, the observer's actions are determined entirely by that initial state, not by the global unitary process. One may view equation (8.2) as a quantum channel that acts on PE with a Kraus representation and corresponding dilation given by

$$\begin{aligned} \rho_{PE} &\mapsto \sum_a E_a \rho_{PE} E_a^\dagger \\ &= \mathrm{Tr}_O \left(\sum_{a,b} (|a\rangle_O \otimes E_a)\, \rho_{PE}\, (\langle b|_O \otimes E_b^\dagger) \right), \end{aligned} \tag{8.3}$$

where $\sum_a E_a^\dagger E_a = I$, and $\{|a\rangle\}$ is an orthonormal basis for O. Therefore, $E_a \rho_{PE} E_a^\dagger$ can be thought as a (subnormalized) post-selected state in which the state of the observer after interacting with the radiation is $|a\rangle_O$. Up to normalization, the operator that the observer applied on the radiation would be E_a in that case. While we do not know the exact details about $\{E_a\}$, within our model we have the following non-trivial constraints:

1. The cardinality of the set $\{E_a\}$ is bounded above by d_O, where $d_O = 2^{|O|}$ is the dimension of the observer's Hilbert space.

2. The global unitary evolution $U_{\mathcal{E}}$ has a complexity polynomial in the black hole entropy $\sim |H|$.

The construction of the mirror operators rests on the observation that V_Ψ defines the embedding map of a quantum error-correcting code that can protect quantum information against "environmental noise" caused by the observer O; see Figure 6. The error model induced by the observer is different from conventional error models that are typically considered in discussions of fault-tolerant quantum computing. For one, the error is applied only on the radiation E and probe P, not the remaining black hole H. Secondly, $U_{\mathcal{E}}$ can apply any operation to the radiation with complexity polynomial in $|H|$. In contrast, more conventional noise models such as the depolarizing channel or the amplitude damping channel typically result from a brief interaction between the environment and the system of interest.

We have already seen in Section 7 that the encoding map V_Ψ protects quantum information against this exotic error model; this conclusion follows from the decoupling condition, which in turn is a consequence of the pseudorandomness of Hawking radiation as discussed in Section 6. Our next task is to relate this robustness against low-complexity noise to the claim in equation (8.1). The formalization and proof of this statement is the main technical contribution of this section.

Before diving into details in the following subsections, let us summarize the conclusion. Consider an error model in which one applies either a channel $\mathcal{E}(\cdot) = \sum_a E_a(\cdot)E_a^\dagger$ or the identity channel, each occurring with nonzero probability. If a quantum error-correcting code V_Ψ can correct such errors, then there is a complete set of logical operators that commutes with all the errors $\{E_a\}$ when acting on the code space; see Figure 8. That is, for any operator $\tilde{T}$ acting on the abstract logical space, there exists a corresponding logical operator T acting identically on the code subspace such that T satisfies the following intertwining condition for all E_a:

$$TE_aV_\Psi \approx E_aTV_\Psi \approx E_aV_\Psi\tilde{T}. \tag{8.4}$$

These logical operators are special because the commutation relation holds as an operator equation acting on all the states in the code subspace. By mapping the isometry V_Ψ back to the state $|\Psi\rangle$, we arrive at equation (8.1) and Figure 8. Note that this is a stronger statement than saying that the commutator of T and E_a has a vanishing expectation value in the code subspace, *i.e.*,

$$V_\Psi^\dagger TE_aV_\Psi \approx V_\Psi^\dagger E_aTV_\Psi. \tag{8.5}$$

In Section 8.1, we will prove (8.4) in the exactly correctable setting. We will then generalize the construction to the approximate case in Section 8.2.

Equation (8.4) also arises in the theory of Operator Algebra Quantum Error-Correction (OAQEC) [25, 26]. However, in that context, one normally considers a logical operator T which annihilates the orthogonal complement of the code space. A novelty of our discussion is that we will allow T to have support extending beyond the code space. In that case, it is delicate to ensure that the action of T on states outside the code space is consistent with (8.4). More importantly, OAQEC was formulated in [25, 26] for the case of exact quantum

error-correction. Our discussion in Section 8.1 is self-contained, and generalizes readily to the approximate setting, as we show in Section 8.2.

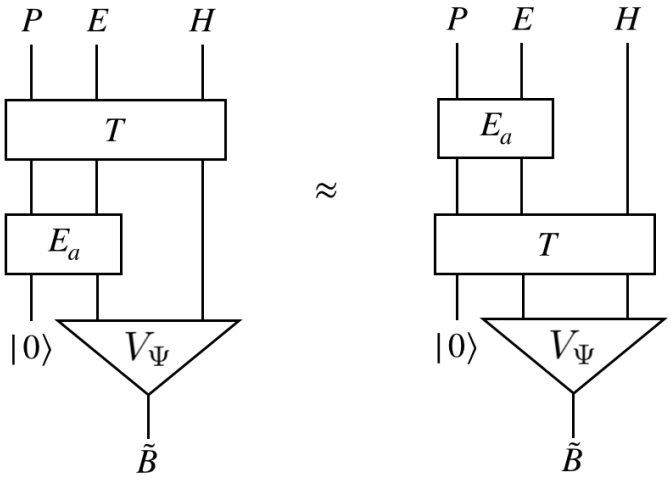

Figure 8. Acting on any code state, the ghost logical operator T (approximately) commutes with any "error" in the set $\{E_a\}$.

8.1 Exact ghost operators

Let $\tilde{\mathcal{H}}$ be an abstract logical Hilbert space, and consider an encoding $V : \tilde{\mathcal{H}} \to \mathcal{C} \subseteq \mathcal{H}$, where $\mathcal{C}$ denotes the code subspace embedded within the larger physical Hilbert space $\mathcal{H}$. Given a Hilbert space $\mathcal{H}$, we will let $S(\mathcal{H})$ denote the state space of $\mathcal{H}$, i.e., the set of all density operators supported on $\mathcal{H}$. Let $\mathcal{E}$ be a correctable error channel for $\mathcal{C}$, which we can write in a Kraus representation as

$$\mathcal{E}(\rho) = \sum_{a=1}^{|K|} E_a \rho E_a^\dagger, \tag{8.6}$$

where we denote the set of Kraus operators as $K = \{E_a\}$. A given channel will of course have many different Kraus representations; the choice of representation will not matter in the exact case, since the set of exactly correctable errors is closed under linear combinations, but we will have to be careful in the analysis of the approximate case in section 8.2. In this section, we will fix an arbitrary Kraus representation K for $\mathcal{E}$.

As a convention, we will denote quantities in $\tilde{\mathcal{H}}$ with tildes, and quantities in $\mathcal{H}$ without. Let

$$\tilde{T} = \sum_{k=1}^{r} \lambda_k \tilde{P}_k \tag{8.7}$$

be a normal operator on $\tilde{\mathcal{H}}$, with (distinct) eigenvalues $\{\lambda_k\}$, where each $\mathcal{P}_k$ is the spectral projector onto the corresponding eigenspace. For ease of notation, given any projector P, we will denote the corresponding range subspace as $[P]$, *i.e.*, $[P] = \mathrm{Im}(P)$.

Consider the encoded subspace $F_k = \mathrm{Im}(V\tilde{P}_k)$ of each eigenspace, and define

$$[P_k] = \mathrm{span}\left\{ E_a|\phi\rangle \;\middle|\; E_a \in K,\ |\phi\rangle \in F_k \right\}. \tag{8.8}$$

Note that $[P_k]$ is the subspace generated by the set of all correctable errors, i.e., the span of K, acting on the encoded eigenspace F_k. These subspaces are well-defined since linear combinations of correctable errors remain correctable, and the Knill-Laflamme conditions [42] imply that subspaces corresponding to distinct eigenvalues will be orthogonal. We can then define a normal operator $T : \mathcal{H} \to \mathcal{H}$ by

$$T = \sum_{k=1} \lambda_k P_k, \tag{8.9}$$

where each P_k is the corresponding projector onto $[P_k]$.

Definition 8.1. *Given any normal operator $\tilde{T} : \tilde{\mathcal{H}} \to \tilde{\mathcal{H}}$, we will call the operator $T : \mathcal{H} \to \mathcal{H}$ obtained through the above construction the* pseudo-ghost operator *corresponding to $\tilde{T}$.*

For any $|\chi_j\rangle \in F_j$ and any error operator $E_a \in K$, the action of the pseudo-ghost operator T is such that

$$TE_a|\chi_j\rangle = \sum_{k=1}^{r} \lambda_k P_k E_a|\chi_j\rangle = \lambda_j E_a|\chi_j\rangle = E_a\hat{T}|\chi_j\rangle. \tag{8.10}$$

Here $\hat{T}$ can be any logical operator for $\tilde{T}$, which therefore satisfies $\hat{T}|\chi_j\rangle = \lambda_j|\chi_j\rangle$.

These pseudo-ghost operators satisfy $TE_a = E_a\hat{T}$ acting on the code space, and so do the ghost operators that we wish to construct. However, note that a pseudo-ghost operator T will not necessarily act as a logical operator for $\tilde{T}$ since we might not have $F_k \subseteq [P_k]$ if the identity is not among the Kraus operators. The operator T will not act correctly on the code subspace unless each of the encoded eigenspaces for $\tilde{T}$ are contained within the corresponding eigenspace for T. Our definition of a ghost logical operator should stipulate that T is logical, as well as requiring $[T, E_a] = 0$ acting on the code space.

Definition 8.2. *Let $T : \mathcal{H} \to \mathcal{H}$ be a logical operator for $\tilde{T}$. We say that T is a* ghost logical operator *for $\tilde{T}$ if*

$$TE_a|\psi\rangle = E_aT|\psi\rangle \tag{8.11}$$

for all $E_a \in K$ and $|\psi\rangle \in \mathcal{C}$. Given a pseudo-ghost operator T, we say that T is extensible *if it admits an extension onto $\mathcal{H}$ such that it becomes a logical operator for $\tilde{T}$.*

Clearly the extension of any extensible pseudo-ghost operator will define a corresponding ghost logical operator. With the above definitions, it is simple to give a concise criterion for when pseudo-ghost operators extend to ghost logical operators in the exact setting.

Lemma 8.3. *Let T be a pseudo-ghost operator. Then T is extensible if and only if*

$$\langle\chi_j|E|\chi_i\rangle = 0, \qquad (i \neq j) \tag{8.12}$$

for all $E \in K$, $|\chi_i\rangle \in F_i$, $|\chi_j\rangle \in F_j$.

Proof. To see necessity, suppose that T is extensible, and let T' denote its logical extension. Because T' is a logical operator for $\tilde{T}$, it must satisfy

$$T'|\chi_i\rangle = \lambda_i|\chi_i\rangle. \tag{8.13}$$

Let $E \in K$ be arbitrary. Left multiplying by $\langle\chi_j|E^\dagger$, we get

$$\lambda_i\langle\chi_j|E^\dagger|\chi_i\rangle = \langle\chi_j|E^\dagger T'|\chi_i\rangle = \sum_{k=1}^{r}\lambda_k\langle\chi_j|E^\dagger P_k|\chi_i\rangle = \lambda_j\langle\chi_j|E^\dagger|\chi_i\rangle. \tag{8.14}$$

Here we have used $E|\chi_j\rangle \in [P_k]$, and noted that T and T' have the same action on $[P_k]$; we also used $P_kE|\chi_j\rangle = \delta_{kj}\lambda_jE|\chi_j\rangle$. If $\lambda_i \neq \lambda_j$, then we must have $\langle\chi_j|E^\dagger|\chi_i\rangle = 0$. Taking the complex conjugate, we obtain equation (8.12).

Conversely, suppose that for all $i \neq j$ and all correctable errors we have $\langle\chi_i|E|\chi_j\rangle = 0$. We must extend the action of T to each encoded eigenvector $|\chi_i\rangle \in \mathcal{C}$. The relations $\langle\chi_i|E|\chi_j\rangle = 0$ imply that $|\chi_i\rangle$ is orthogonal to the subspaces $[P_j]$ for $j \neq i$. There are two possible cases, either $|\chi_i\rangle \in [P_i]$, for which $T|\chi_i\rangle = \lambda_i|\chi_i\rangle$ is already well-defined and we are done, or else there exists a component of $|\chi_i\rangle$ lying in the subspace orthogonal to $\bigoplus_{k=1}^{r}[P_k]$.

Let $|\chi_i^\perp\rangle$ denote the normalized component of $|\chi_i\rangle$ orthogonal to $[P_i]$. Then we extend the subspace $[P_i]$ to $[P_i']$ by defining the projector

$$P_i' = P_i + |\chi_i^\perp\rangle\langle\chi_i^\perp|. \tag{8.15}$$

Note that the new subspace $[P_i']$ contains within it $[P_i]$ and remains orthogonal to $[P_j]$ for $j \neq i$. Moreover, we have $|\chi_i\rangle \in [P_i']$. We can now define an extension of T with the projector P_i' in place of P_i. Then the extension T' satisfies

$$T'|\chi_i\rangle = \lambda_i|\chi_i\rangle. \tag{8.16}$$

We may repeat this procedure with an orthogonal basis $\{|\chi_k\rangle\}$ for $\mathcal{C}$ until we are left with an extension which acts as a logical operator for $\tilde{T}$. □

We will be primarily interested in the case where there exists a full set of ghost logical operators. We say that there exists a *complete set* of ghost logical operators if for every normal operator $\tilde{T} : \tilde{\mathcal{H}} \to \tilde{\mathcal{H}}$, there exists a corresponding ghost logical operator T. In what follows, given a channel $\mathcal{E}$, we will let $\mathcal{E}_\mathcal{I}$ denote the channel

$$\mathcal{E}_\mathcal{I} = \mathcal{I}/2 + \mathcal{E}/2, \tag{8.17}$$

where $\mathcal{I}$ is the identity channel. That is, in the channel $\mathcal{E}_\mathcal{I}$, with probability $1/2$ $\mathcal{E}$ is applied, and with probability $1/2$ nothing happens.

Theorem 8.4. *Let $\mathcal{E}$ be a correctable channel with Kraus operators K. Then a complete set of ghost logical operators for $\mathcal{E}$ exists if and only if $K \cup \{I\}$ is a correctable set, i.e., if and only if $\mathcal{E}_\mathcal{I}$ is a correctable channel.*

Proof. Suppose that $K \cup \{I\}$ is a correctable set. Then the Knill-Laflamme conditions for $K \cup \{I\}$ imply that the hypotheses of Lemma 8.3 are satisfied so that every pseudo-ghost operator is extensible to a ghost logical operator. It follows that there exists a complete set of ghost logical operators.

Conversely, suppose that there exists a complete set of ghost logical operators. Let $|\psi\rangle, |\phi\rangle \in \mathcal{C}$ be two mutually orthogonal code states, and let $|\tilde{\psi}\rangle = V^\dagger|\psi\rangle$ and $|\tilde{\phi}\rangle = V^\dagger|\phi\rangle$ be the corresponding pre-images in $\tilde{\mathcal{H}}$. Define the operators

$$\tilde{T}_1 = |\tilde{\phi}\rangle\langle\tilde{\phi}| - |\tilde{\psi}\rangle\langle\tilde{\psi}|, \tag{8.18}$$

and

$$\tilde{T}_2 = |\tilde{\phi} + \tilde{\psi}\rangle\langle\tilde{\phi} + \tilde{\psi}| - |\tilde{\phi} - \tilde{\psi}\rangle\langle\tilde{\phi} - \tilde{\psi}|, \tag{8.19}$$

where $|\tilde{\phi} \pm \tilde{\phi}\rangle = 2^{-1/2}(|\tilde{\phi}\rangle \pm |\tilde{\psi}\rangle)$. By assumption, there exist ghost logical operators T_1 and T_2 corresponding to $\tilde{T}_1$ and $\tilde{T}_2$. Now let $E_a, E_b \in K \cup \{I\}$. Then we have

$$\begin{aligned} \langle\psi|E_a^\dagger E_b|\phi\rangle &= \langle\psi|E_a^\dagger E_b T_1|\phi\rangle && (8.20)\\ &= \langle\psi|T_1 E_a^\dagger E_b|\phi\rangle && (8.21)\\ &= -\langle\psi|E_a^\dagger E_b|\phi\rangle, && (8.22) \end{aligned}$$

where the first line follows due to the fact that $|\phi\rangle$ is an eigenvector for T_1 with eigenvalue 1, the second line follows from the defining equations (8.11) for the ghost operators, together with the fact that T_1 is self-adjoint, and the last line follows from the fact that $|\psi\rangle$ is an eigenvector for T_1 with eigenvalue -1. This implies that $\langle\psi|E_a^\dagger E_b|\phi\rangle = 0$.

Repeating the same argument for T_2, we have

$$\begin{aligned} \langle\phi - \psi|E_a^\dagger E_b|\phi + \psi\rangle &= \langle\phi - \psi|E_a^\dagger E_b T_2|\phi + \psi\rangle && (8.23)\\ &= \langle\phi - \psi|T_2 E_a^\dagger E_b|\phi + \psi\rangle && (8.24)\\ &= -\langle\phi - \psi|E_a^\dagger E_b|\phi + \psi\rangle, && (8.25) \end{aligned}$$

which implies that

$$0 = \langle\phi|E_a^\dagger E_b|\phi\rangle - \langle\psi|E_a^\dagger E_b|\psi\rangle. \tag{8.26}$$

Since ϕ and ψ were arbitrary, this holds for any pair of orthogonal states.

Let $\{|j\rangle\}$ be an orthonormal basis for $\mathcal{C}$ and define $\lambda_{ab} = \langle\psi|E_a^\dagger E_b|\psi\rangle$ for an arbitrary state $|\psi\rangle \in \mathcal{C}$. Then it follows that we have

$$\langle i|E_a^\dagger E_b|j\rangle = \lambda_{ab}\delta_{ij}, \tag{8.27}$$

so that the Knill-Laflamme conditions for $K \cup \{I\}$ are satisfied. Therefore, $K \cup \{I\}$ is a correctable set of errors. □

8.2 Approximate ghost operators

In this section, we discuss how the ghost logical operators can be constructed for *approximate* quantum error-correcting codes. We need to consider this case because we inferred in Section 7 that the errors due to low-complexity operations on the radiation system E are correctable approximately (with a residual error exponentially small in $|H|$) rather than exactly. Although the uncorrected error is exponentially small, the Hilbert space is exponentially large, so we need to do a careful analysis to check that the ghost logical operators commute with the errors apart from exponentially small effects.

It turns out the strategy that we pursued in the exact setting also works in the approximate setting. To get started, we will construct *approximate ghost projectors* $\{\mathcal{P}_i\}$ that play the same role as the $\{P_i\}$ in the previous section.

Definition 8.5. *Let $\{|\tilde{i}\rangle\}$ be an orthonormal basis for $\tilde{\mathcal{H}}$, and suppose V is an encoding isometry. We define (approximate)* ghost projectors *with respect to this basis, denoted $\mathcal{P}_i$, to be the orthogonal projectors onto the* positive *eigenspace of*

$$\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp}), \tag{8.28}$$

where $|i\rangle = V|\tilde{i}\rangle$ for $|\tilde{i}\rangle \in \tilde{\mathcal{H}}$, and where

$$\rho_{i,\perp} = \frac{1}{\dim \tilde{\mathcal{H}} - 1} \sum_{j \neq i} |j\rangle\langle j|. \tag{8.29}$$

The motivation behind this definition follows from the fact that $\mathcal{P}_i$ is an operator that can optimally distinguish $\mathcal{E}(|i\rangle\langle i|)$ from $\mathcal{E}(\rho_{i,\perp})$, according to the Holevo-Helstrom theorem [40]. Because the effect of the channel $\mathcal{E}$ can be reversed up to a small error, it nearly preserves the orthogonality of $|i\rangle\langle i|$ and $\rho_{i,\perp}$; therefore, $\mathcal{P}_i$ can distinguish the two states almost perfectly. This suggests that $\mathcal{P}_i$, up to a small error, projects $\mathcal{E}(|i\rangle\langle i|)$ to a state close to $\mathcal{E}(|i\rangle\langle i|)$ and nearly annihilates $\mathcal{E}(\rho_{i,\perp})$. In the following two lemmas, we prove these claims rigorously. In Lemma 8.6, we show that $\mathcal{P}_i E_a|i\rangle \approx E_a|i\rangle$, and in Lemma 8.7 we show that $\mathcal{P}_i E_a|j\rangle \approx 0$, for $i \neq j$, where E_a is any Kraus operator of the channel $\mathcal{E}$.

If $\mathcal{E}$ is an ϵ-correctable channel then we have

$$\max_{\rho} \|(\mathcal{R} \circ \mathcal{E})(\rho) - \rho\|_1 \leq 2\sqrt{2}\epsilon := \tilde{\epsilon}, \tag{8.30}$$

as given by equation (7.10). Let us define $\tilde{\epsilon} = 2\sqrt{2}\epsilon$ to minimize factors of $2\sqrt{2}$. Then we can obtain the following bound:

Lemma 8.6. *Let $\mathcal{E}$ be an ϵ-correctable channel and let $\mathcal{P}_i$ be the corresponding ghost projector with respect to some basis. Then we have*

$$\||E_a|i\rangle - \mathcal{P}_i E_a|i\rangle\|_2^2 \leq 2\sqrt{2}\epsilon := \tilde{\epsilon}, \tag{8.31}$$

where $\||\phi\rangle\|_2 := \sqrt{\langle\phi|\phi\rangle}$.

Proof. Note that, by the monotonicity of the trace norm, we have

$$\|\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})\|_1 \geq \|(\mathcal{R}\circ\mathcal{E})(|i\rangle\langle i| - \rho_{i,\perp})\|_1. \tag{8.32}$$

We can use the fact that the recovery map $\mathcal{R}$ nearly succeeds in recovering the original state. By the triangle inequality,

$$\begin{aligned} 2 = \||i\rangle\langle i| - \rho_{i,\perp}\|_1 \leq{}& \|(\mathcal{R}\circ\mathcal{E})(|i\rangle\langle i| - \rho_{i,\perp})\|_1 \\ &+ \||i\rangle\langle i| - (\mathcal{R}\circ\mathcal{E})(|i\rangle\langle i|)\|_1 \\ &+ \|\rho_{i,\perp} - (\mathcal{R}\circ\mathcal{E})(\rho_{i,\perp})\|_1. \end{aligned} \tag{8.33}$$

Therefore,

$$\|\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})\|_1 \geq 2 - 2\tilde{\epsilon}. \tag{8.34}$$

Moreover, we have

$$\begin{aligned} \|\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})\|_1 &= \mathrm{Tr}(2\mathcal{P}_i\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})) \\ &\leq 2\mathrm{Tr}(\mathcal{P}_i\mathcal{E}(|i\rangle\langle i|)). \end{aligned} \tag{8.35}$$

The first line above follows by decomposing $\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})$ into its positive and negative parts. Because the operator is traceless, the trace of the positive part is equal to the trace of the negative part, up to a minus sign. Since the trace distance is equal to the sum of the absolute value of the positive and negative trace, and because these values are the same, we arrive at the first identity. The second line then follows from the fact that $\mathrm{Tr}(\mathcal{P}_i\mathcal{E}(\rho_{i,\perp})) \geq 0$.

Therefore, we get the following bound:

$$\begin{aligned} 1 - \tilde{\epsilon} &\leq \mathrm{Tr}(\mathcal{P}_i\mathcal{E}(|i\rangle\langle i|)) \\ &= \sum_a \langle i|E_a^\dagger \mathcal{P}_i E_a|i\rangle \\ &= \sum_a q_{ia}\langle\psi_{ia}|\mathcal{P}_i|\psi_{ia}\rangle \\ &= 1 - \sum_a q_{ia}(1 - \langle\psi_{ia}|\mathcal{P}_i|\psi_{ia}\rangle), \end{aligned} \tag{8.36}$$

where we define

$$|\psi_{ia}\rangle = \frac{E_a|i\rangle}{\sqrt{\langle i|E_a^\dagger E_a|i\rangle}}, \tag{8.37}$$

and $q_{ia} = \langle i|E_a^\dagger E_a|i\rangle$. Note that $\sum_a q_{ia} = 1$ since $\mathcal{E}$ is trace-preserving. Therefore, we get

$$1 - \langle\psi_{ia}|\mathcal{P}_i|\psi_{ia}\rangle \leq \frac{\tilde{\epsilon}}{q_{ia}} \tag{8.38}$$

by noting that the last line of equation (8.36) contains a sum of non-negative terms. Since the sum is $\leq \tilde{\epsilon}$, each individual term must be $\leq \tilde{\epsilon}$ as well. Substituting in the expressions for q_{ia} and $|\psi_{ia}\rangle$, this inequality becomes

$$\langle i|E_a^\dagger E_a|i\rangle - \langle i|E_a^\dagger \mathcal{P}_i E_a|i\rangle \leq \tilde{\epsilon}, \tag{8.39}$$

which is equivalent to equation (8.31). □

Lemma 8.7. *Under the same hypothesis as Lemma 8.6, if $i \neq j$, then*

$$\|\mathcal{P}_i E_a |j\rangle\|_2^2 \leq (\dim \mathcal{C})\, \tilde{\epsilon}. \tag{8.40}$$

Proof. Note that

$$\begin{aligned} 2 - 2\tilde{\epsilon} &\leq \|\mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})\|_1 \\ &= \mathrm{Tr}(2\mathcal{P}_i \mathcal{E}(|i\rangle\langle i| - \rho_{i,\perp})) \\ &\leq 2 - 2\mathrm{Tr}(\mathcal{P}_i \mathcal{E}(\rho_{i,\perp})), \end{aligned} \tag{8.41}$$

where we've used equation (8.34) in the first line, and equation (8.35) in the second. The last line follows from the fact that $\mathcal{P}_i \leq I$. It follows that

$$\frac{1}{\dim \mathcal{C} - 1} \sum_{j \neq i} \mathrm{Tr}(\mathcal{P}_i \mathcal{E}(|j\rangle\langle j|)) = \mathrm{Tr}(\mathcal{P}_i \mathcal{E}(\rho_{i,\perp})) \leq \tilde{\epsilon}, \tag{8.42}$$

and therefore, we have

$$\mathrm{Tr}(\mathcal{P}_i \mathcal{E}(|j\rangle\langle j|)) \leq (\dim \mathcal{C})\, \tilde{\epsilon}, \tag{8.43}$$

for all $j \neq i$. Expanding in terms of the Kraus operators of the channel $\mathcal{E}$, this becomes

$$\sum_a \mathrm{Tr}(\mathcal{P}_i E_a |j\rangle\langle j| E_a^\dagger \mathcal{P}_i) = \sum_a \|\mathcal{P}_i E_a |j\rangle\|_2^2 \leq (\dim \mathcal{C})\tilde{\epsilon}, \tag{8.44}$$

where the first equality holds because $\mathcal{P}_i$ is a projector. Equation (8.40) then follows. □

At this point, we can follow the construction we used for the case of exact ghost operators. Let $\mathcal{C}$ be a code subspace and let $\mathcal{E}$ be an error channel such that $\mathcal{E}_\mathcal{I}$ is ϵ-correctable. Then by Lemmas 8.6 and 8.7, we have

$$\|E_a|i\rangle - \mathcal{P}_i E_a |i\rangle\|_2^2 \leq 2\tilde{\epsilon}, \quad \text{and} \quad \|\mathcal{P}_i E_a |j\rangle\|_2^2 \leq 2(\dim \mathcal{C})\, \tilde{\epsilon}, \tag{8.45}$$

where each E_a is a Kraus operators for $\mathcal{E}$, or the identity. Note that the extra factor of 2 comes from the fact that the Kraus operators for $\mathcal{E}_\mathcal{I}$ are given by $\{E_a/\sqrt{2}\} \cup \{I/\sqrt{2}\}$, where each E_a is a Kraus operator for $\mathcal{E}$.

Given a normal operator $\tilde{T} : \tilde{\mathcal{H}} \to \tilde{\mathcal{H}}$ defined by

$$\tilde{T} = \sum_k \lambda_k |\tilde{k}\rangle\langle\tilde{k}|, \tag{8.46}$$

we define the operator

$$T = \sum_k \lambda_k \mathcal{P}_k, \tag{8.47}$$

where each $\mathcal{P}_k$ is a ghost projector with respect to the given eigenbasis for $\tilde{T}$. Then the operator T satisfies

$$\begin{aligned}
\|TE_a|j\rangle - \lambda_j E_a|j\rangle\|_2 &= \left\| \sum_k \lambda_k \mathcal{P}_k E_a|j\rangle - \lambda_j E_a|j\rangle \right\|_2 \\
&\leq \left\| \sum_{k\neq j} \lambda_k \mathcal{P}_k E_\ell|j\rangle \right\|_2 + \|\lambda_j \mathcal{P}_j E_\ell|j\rangle - \lambda_j E_a|j\rangle\|_2 \\
&\leq \sum_{k\neq j} |\lambda_k|\, \|_2 \mathcal{P}_k E_a|j\rangle\|_2 + |\lambda_j|\, \|\mathcal{P}_j E_a|j\rangle - E_a|j\rangle\|_2 \\
&\leq \sqrt{2(\dim \mathcal{C})\tilde{\epsilon}} \sum_{k\neq j} |\lambda_k| + |\lambda_j| \sqrt{2\tilde{\epsilon}} \\
&\leq \sqrt{2(\dim \mathcal{C})\tilde{\epsilon}} \|\tilde{T}\|_1,
\end{aligned} \tag{8.48}$$

where $\|\tilde{T}\|_1$ is the trace norm of $\tilde{T}$. Now, let $\hat{T} = V\tilde{T}V^\dagger$, where V is the code embedding. Then for a general code state $|\psi\rangle = \sum_j c_j|j\rangle$, we have

$$\begin{aligned}
\left\| TE_a|\psi\rangle - E_a\hat{T}|\psi\rangle \right\|_2 &\leq \sum_j |c_j| \cdot \|TE_a|j\rangle - \lambda_j E_a|j\rangle\|_2 \\
&\leq \sqrt{2(\dim \mathcal{C})\tilde{\epsilon}} \|\tilde{T}\|_1 \sum_j |c_j| \\
&\leq (\dim \mathcal{C}) \|\tilde{T}\|_1 \sqrt{2\tilde{\epsilon}}.
\end{aligned} \tag{8.49}$$

A slightly weaker, but more convenient bound in terms of the operator norm of $\tilde{T}$ can be given as

$$\left\| TE_a|\psi\rangle - E_a\hat{T}|\psi\rangle \right\|_2 \leq (\dim \mathcal{C})^2 \|\tilde{T}\| \sqrt{\left(4\sqrt{2}\right)\epsilon}, \tag{8.50}$$

which we can also express as a bound on the difference of two operators in the operator norm:

$$\left\| TE_aV - E_aV\tilde{T} \right\| \leq 2^{5/4} (\dim \mathcal{C})^2 \|\tilde{T}\| \sqrt{\epsilon}, \tag{8.51}$$

where we have used $\tilde{\epsilon} = 2\sqrt{2}\epsilon$. Note that the bound (8.51) holds for any Kraus representation $\{E_a\}$ of $\mathcal{E}$.

The bound (8.51) motivates the following definition:

Definition 8.8. *Let $(\mathcal{E}, K)$ be a noise channel $\mathcal{E}$ equipped with a given Kraus representation $K = \{E_a\}$. Let $\tilde{T} : \tilde{\mathcal{H}} \to \tilde{\mathcal{H}}$ be a normal operator. We say that T is a δ-approximate ghost operator for $\tilde{T}$ with respect to $(\mathcal{E}, K)$ if we have*

$$\|TE_aV - E_aV\tilde{T}\| \leq \|\tilde{T}\|\delta, \tag{8.52}$$

where $E_a \in K \cup \{I\}$ is either a Kraus operator for $\mathcal{E}$, or the identity. We say that a ghost operator T is universal *if equation (8.52) holds for every Kraus representation of $\mathcal{E}$.*

Now we are ready to prove the analog of Theorem 8.4 in the approximate setting. As before, we say that there exists a complete set of ϵ-approximate ghost operators if there exists an ϵ-approximate ghost logical operator for every normal operator on $\tilde{\mathcal{H}}$.

Theorem 8.9. *Let $\mathcal{C}$ be a code subspace and suppose that $\mathcal{E}_{\mathcal{I}}$ is ϵ-correctable for $\mathcal{C}$. Then there exists a complete set of δ-approximate universal ghost operators, where*

$$\delta = 2^{5/4}(\dim \mathcal{C})^2\sqrt{\epsilon}. \tag{8.53}$$

For the sake of completeness, we also prove a converse of this result (Theorem B.3) in Appendix B. These results collectively can be seen as a generalization of the standard theorems of operator algebra quantum error-correction [25, 26] to the approximate setting.

Proof. Suppose that $\mathcal{E}_{\mathcal{I}}$ is ϵ-correctable for $\mathcal{C}$. Then equation (8.51) shows that T defined by equation (8.47) is a δ-approximate ghost operator for any normal operator $\tilde{T}$, where $\delta = 2^{5/4}(\dim \mathcal{C})^2\sqrt{\epsilon}$. The construction of the ghost projector $\mathcal{P}_k$, and therefore also the construction of T, depends only on the channel $\mathcal{E}$ and not on any particular Kraus representation; it follows that T is universal. □

8.3 Firewall revisited

We have now seen that, by assuming that the state of the Hawking radiation system EB is pseudorandom, we may infer that low-complexity operations on E are approximately correctable; the code space $\tilde{B}$ that purifies the late radiation system B is protected against low-complexity operations on E. Correctability in turn implies that a complete set of ghost logical operators acting on EH, which nearly commute with all low-complexity operations on E, can be constructed.

Let us now reconsider the potential implications of the existence of ghost logical operators in the context of the black hole firewall problem. First, we assemble the results we have derived thus far to determine the value of δ for which the ghost logical operators are δ-approximate. Under the pseudorandomness assumption equation (6.1), we saw in Lemma 7.2 that low-complexity operations are ϵ-correctable for $\epsilon = \sqrt{3/2}\cdot 2^{-(\alpha|H|-|OB|)/2}$. Since the code space dimension is $\dim \mathcal{C} = 2^{|B|}$, equation (8.53) says that the ghost operators are δ-approximate for

$$\delta = 2^{5/4}2^{2|B|}\sqrt{\epsilon} = 2\cdot 3^{1/4}\ 2^{2|B|}\ 2^{-(\alpha|H|-|OB|)/4} = 2\cdot 3^{1/4}\ 2^{-(\alpha|H|-|O|-9|B|)/4}. \tag{8.54}$$

Thus, δ becomes exponentially small for asymptotically large $|H|$, $|O|$, and $|B|$, provided $|O|, |B| \ll |H|$. We could, for example, consider an encoded interior and an observer with size scaling linearly with $|H|$, and still have a complete set of ghost logical operators commuting with all low-complexity operations on E, up to exponentially small errors.

This conclusion followed only from the assumption that the state of EB is pseudorandom — we needed no other special properties of black holes to derive it. We might, in fact, expect the same pseudorandomness assumption to hold not just for black holes but also for other strongly chaotic quantum systems. But a black hole *is* special, because it has an event horizon, and it is because of the event horizon that we expect the late radiation

system B to be entangled with modes behind the horizon as well as with a subspace of EH; thus arises the black hole firewall problem. To ease the firewall problem, we propose using the ghost logical operators to describe (a portion of) the black hole interior. We would not make such a proposal for describing the "interior" of a burning lump of coal.

Pleasingly, under this proposal, it is hard for an agent who acts on the radiation to create a firewall, or to otherwise influence the black hole interior apart from exponentially small effects. To create an excitation behind the horizon, the agent outside the black hole must perform an operation of superpolynomial complexity.

We might want to allow the observer to perform a quantum computation on E which is chosen from a long list of possible unitary transformations. The observer's freedom to choose can be encoded in the observer's initial state ω_O, as depicted in Figure 7. If there are multiple observers $\{O_1, O_2, \ldots O_m\}$, all interacting with E, we can group them all together into a collective observer $O = O_1 O_2 \ldots O_m$. We may construct a complete set of ghost logical operators acting on the encoded black hole interior, consistently shared by all the observers, provided that $|O|, |B| \ll |H|$.

To be more concrete, suppose we want the black hole interior to be protected against any unitary transformation acting on E chosen from amongst a collection of N possible unitaries $\mathcal{U} = \{U_a\}_{a=1}^N$. We can model this situation by considering a conditional unitary transformation, controlled by an ancilla register in the observer's possession. To ensure that we can apply Theorem 8.9 we will add the identity transformation $U_0 = I_E$ to the list of possibilities, and envision that the observer applies

$$U_{\mathcal{U}} = \sum_{a=0}^{N} |a\rangle\langle a|_O \otimes (U_a)_E, \tag{8.55}$$

where each $|a\rangle_O$ is a computational basis state and $2^{|O|} = N+1$. Thus U_a is applied by fixing the initial state of the O register to be $|a\rangle_O$; see Figure 9.

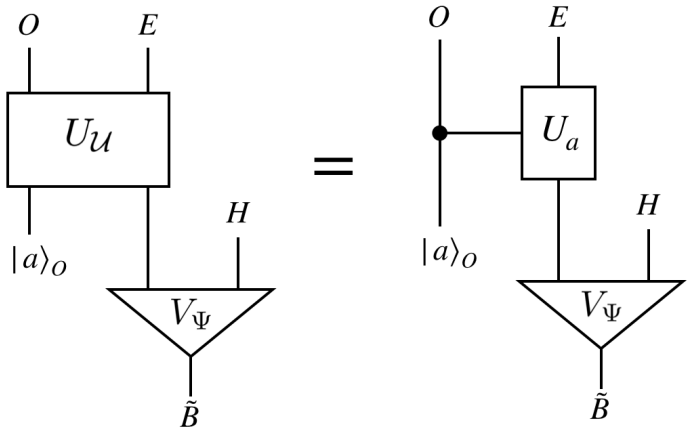

Figure 9. The action of the observer as a controlled unitary tranformation.

Our construction of a complete set of ghost logical operators applies — assuming the Hawking radiation is pseudorandom — if $U_{\mathcal{U}}$ has complexity polynomial in $|H|$. This will

be assured if the cardinality N of the list of unitaries is polynomial in $|H|$. The unitary

$$\Lambda_a(U_a) = |a\rangle\langle a|_O \otimes (U_a)_E \tag{8.56}$$

for which a non-trivial unitary acting on E is triggered only by the basis state $|a\rangle_O$, has polynomial quantum complexity if U_a does — we show in Lemma C.1 that, if we fix the complexity of U_a, then $\Lambda_a(U_a)$ can be implemented to precision ϵ with a circuit of $O(N^2 \log^4(1/\epsilon))$ two-gubit gates. Furthermore, the overall operator

$$U_{\mathcal{U}} = \prod_{a=0}^{N} \Lambda_a(U_a) \tag{8.57}$$

is a product of $N+1$ such unitaries, and thus has complexity at worst a factor of $N+1$ larger. Therefore, if $N = \text{poly}(|H|)$, then $U_{\mathcal{U}}$ can be executed to exponential precision with a circuit of size $\text{poly}(|H|)$.

The unitary transformation $U_{\mathcal{U}}$ is a dilation of the quantum channel

$$\mathcal{E}_{\mathcal{U}}(\rho) = \frac{1}{N+1} \sum_{a=0}^{N} U_a \rho U_a^\dagger \tag{8.58}$$

acting on E, with Kraus operators $\{U_a\}_{a=0}^N$. Because $U_{\mathcal{U}}$ has polynomial complexity, under the pesudorandomness assumption a complete set of δ-approximate ghost logical operators can be constructed, with δ given by equation (8.54). In other words, for each unitary U_a that the observer might apply, U_a commutes with all ghost logical operators up to an exponentially small error. Hence no matter which low-complexity unitary the observer applies, the encoded black hole interior is hardly affected at all.

This conclusion is summarized by the following theorem:

Theorem 8.10. *Suppose that the decoupling bound (6.2) holds. Let $V : \mathcal{H}_{\tilde{B}} \to \mathcal{H}_{EH}$ denote the black hole code embedding. Let $\mathcal{U} = \{U_a\}_{a=1}^N$ denote an arbitrary set of $N = \text{poly}(|H|)$ unitaries acting on the early radiation E, where each unitary has complexity $\text{poly}(|H|)$. Then there exists a complete set of logical operators $\mathcal{L} \subseteq \mathcal{B}(\mathcal{H}_{EH})$ for the black hole code such that for all $T \in \mathcal{L}$, and all $U_a \in \mathcal{U}$, we have*

$$\|[U_a, T]V\| \leq 2\delta' \|\tilde{T}\|, \tag{8.59}$$

and

$$\|TV - V\tilde{T}\| \leq \delta' \|\tilde{T}\|, \tag{8.60}$$

where $\tilde{T}$ is the operator on $\tilde{B}$ corresponding to T, and

$$\delta' = 8 \cdot 6^{1/4}\ 2^{-\alpha|H|/4} (N+1)^{3/4} \tag{8.61}$$

if $\tilde{B}$ is a single qubit ($|B| = 1$).

Proof. Let us model the observer O on the Hilbert space $\mathcal{H}_O = \mathbb{C}^{N+1}$, so that $2^{|O|} = N+1$. In Lemma C.1, we show that the conditional unitary $U_{\mathcal{U}}$ defined in equation (8.55) can be approximated to exponential accuracy with a circuit of size poly($|H|$) if each U_a has complexity poly($|H|$) and $N = \text{poly}(|H|)$; therefore, under the pseudorandomness assumption, $U_{\mathcal{U}}$ is ϵ-correctable with

$$\epsilon = \sqrt{\frac{3}{2}} \cdot 2^{-(\alpha|H|-|OB|)/2}, \tag{8.62}$$

and hence there exists a complete set of δ-approximate ghost logical operators for $U_{\mathcal{U}}$ with

$$\delta = 2 \cdot 3^{1/4}\ 2^{-(\alpha|H|-|O|-9|B|)/4}, \tag{8.63}$$

or

$$\delta = 8 \cdot 6^{1/4}\ 2^{-\alpha|H|/4} 2^{|O|/4} = 8 \cdot 6^{1/4}\ 2^{-\alpha|H|/4} (N+1)^{1/4} \tag{8.64}$$

if $|B| = 1$. The Kraus operators for the channel $\mathcal{E}_{\mathcal{U}}$ in equation (8.58) are $\{U_a/\sqrt{N+1}\}$; hence

$$\|TU_aV - U_aV\tilde{T}\| \leq \|\tilde{T}\|\delta\sqrt{N+1} = \|\tilde{T}\|\delta'. \tag{8.65}$$

This, together with Lemma B.2, gives the desired results equation (8.59) and equation (8.60). □

Note that, although δ' in equation (8.61) could be exponentially small even for superpolynomial N, we required $N = \text{poly}(|H|)$ because only in that case have we shown that the conditional unitary $U_{\mathcal{U}}$ has complexity poly($|H|$); we needed this property for the pseudorandomness assumption to imply that the observer is unable to distinguish the state of EB from a maximally mixed state.

We have inferred the existence of ghost logical operators which act on EH. It should also be possible to realize a non-trivial logical operator as a physical operator acting on E alone, but only if that operator is computationally complex to construct. For instance, suppose that $W : \mathcal{H}_E \to \mathcal{H}_E$ is a unitary logical operator that can be accurately approximated by a quantum circuit of polynomial size. Then there exists a ghost logical operator T that fails to commute with W acting on the code space. Since W has polynomial complexity, this contradicts Theorem 8.10, and we conclude that no such W can exist. This conclusion resonates with the observations of Bouland, Fefferman, and Vazirani, who argued that in the context of AdS/CFT duality, the dictionary relating the black hole exterior and interior should be computationally complex [43, 44].

On the other hand, if a quantum circuit is allowed to act on H as well as E, and if B has constant size, then any logical operator on the code space *can* be realized efficiently. We show this in Section 9.

8.4 State dependence

The (approximate) encoding isometry $V_\Psi : \mathcal{H}_{\tilde{B}} \to \mathcal{H}_{EH}$ is determined by the pure quantum state Ψ_{EHB} of the black hole H and its emitted Hawking radiation EB. This state, and

hence the encoding map, depends on the initial microstate of the infalling matter that collapsed to form the black hole. Therefore, the encoded interior of the black hole is said to be "state dependent" [7, 10].

If black hole evaporation is unitary, and the event horizon is smooth because the black hole interior is encoded in the radiation, then state dependence of the encoding seems to be unavoidable; if the quantum information encoded in the initial state is preserved in the final state of the fully evaporated black hole, then how the late radiation emitted after the Page time is entangled with the early radiation emitted before the Page time must depend on that initial state. This state dependence of the encoding is nonetheless troubling [8, 9, 16]. If the experiences of observers who fall through the event horizon are described by the logical operators of the code, and these logical operators are state dependent, then the observers inside the black hole seem to be capable of measuring nonlinear operators acting on Ψ_{EHB}, rather than linear operators as in the standard theory of quantum measurement. This ability to measure nonlinear properties of the state could lead to inconsistencies. We regard this as an unresolved issue, reflecting our incomplete understanding of how to describe measurements conducted behind black hole horizons.

But the state-dependent encoding of the black hole interior is not sufficient by itself to solve the black hole firewall problem.[8] If the Hawking radiation is thoroughly scrambled, then we expect that the interior mode that purifies B can be decoded by acting on E alone after the Page time [45], and therefore that the logical operators of the code may also be chosen to act on E alone. If T and S are two noncommuting logical operators, where S acts on E, then an observer (Bob) outside the black hole who applies S could in principle alter the outcome of a measurement of T performed by an observer (Alice) inside the black hole. Thus Bob can send an instantaneous message to Alice, in apparent violation of relativistic causality.

While we agree that such acausal signaling is possible in principle, we insist that the computational complexity of the task should be considered. Under the assumption that the Hawking radiation is pseudorandom, we have found that, in order to signal Alice, Bob must apply an operation to E with complexity superpolynomial in $|H|$, if Alice's observables are the ghost logical operators we have constructed. Though possible, such an operation is infeasible in practice if the black hole H is macroscopic; therefore the semiclassical causal structure of the spacetime is respected.

9 Inside the black hole

Under our pseudorandomness assumption, an observer who acts on the early radiation system E can affect the encoded interior of a black hole only by applying an operation with superpolynomial complexity. However, an agent who has access to the black hole system H as well as E can manipulate the interior efficiently. Here we construct an efficient unitary circuit $\bar{U}_{EH}$, acting on EH, that perturbs the encoded interior. Our construction makes use of an efficient quantum circuit that realizes the unitary $U_{\rm bh}$ that describes the formation and partial evaporation of a black hole. This unitary creates a state in which

[8] We thank Raphael Bousso for raising this issue.

B is maximally entangled with a subspace of EH; if the circuit that implements $U_{\rm bh}$ is efficient, then $\bar{U}_{EH}$ can be implemented efficiently as well. We will also see that an agent with access to EH can efficiently decode the interior, distilling the code subspace of EH to a small quantum memory.

Suppose we are given a unitary operator U_{BEH} which realizes the map

$$U_{BEH}|0\rangle_B|0\dots0\rangle_{EH} = \frac{1}{\sqrt{2}}(|0\rangle_B|\psi_0\rangle_{EH} + |1\rangle_B|\psi_1\rangle_{EH}), \tag{9.1}$$

where B is a single qubit, and EH is n qubits. By applying the circuits that implement U_{BEH} and $U_{BEH}^\dagger$ on an ancillary register, together with some additional gates acting on the ancilla and EH, we will apply a unitary operator $\bar{U}_{EH}$ acting on EH with the property that

$$\begin{aligned}\bar{U}_{EH}|\psi_0\rangle_{EH} &= v_{00}|\psi_0\rangle_{EH} + v_{10}|\psi_1\rangle_{EH},\\ \bar{U}_{EH}|\psi_1\rangle_{EH} &= v_{01}|\psi_0\rangle_{EH} + v_{11}|\psi_1\rangle_{EH},\end{aligned} \tag{9.2}$$

where

$$v = \begin{pmatrix} v_{00} & v_{01} \\ v_{10} & v_{11} \end{pmatrix} \tag{9.3}$$

is some chosen 2×2 unitary matrix. That is, $\bar{U}_{EH}$ applies an arbitrary "logical" unitary transformation on the two-dimensional "code space" spanned by $\{|\psi_0\rangle_{EH}, |\psi_1\rangle_{EH}\}$.

The protocol is explained in two steps. First, we describe a probabilistic protocol which applies $\bar{U}_{EH}$ with success probability $\frac{1}{4}$. Next, using the probabilistic protocol, we build a deterministic protocol which applies $\bar{U}_{EH}$ with probability 1. The first protocol applies a unitary $U_{a_1a_2}$ and $U_{a_1a_2}^\dagger$ once each. Here, $U_{a_1a_2}$ is a unitary acting on an ancillary register $a = a_1a_2$ and can be realized by applying the circuit that implements U_{BEH} on register a_1 and a_2. The register B is replaced with a_1 and the register EH is replaced with a_2. The second protocol applies $U_{a_1a_2}$ and $U_{a_1a_2}^\dagger$ three times each. We also use some additional gates, which are also efficient.

For the probabilistic protocol, consider the following sequence of operations:

1. Initialize a in the $|0\dots0\rangle$ state.
2. Apply $U_{a_1a_2}$.
3. Apply a swap between a_2 and EH.
4. Apply the single-qubit operation v^T to a_1.
5. Apply $U_{a_1a_2}^\dagger$.
6. Measure the a register in the computational basis, and postselect on measuring the all-0 bit string.

Applying this protocol for $U_{BEH} = U_{\rm bh}$, and taking the initial state to be $|\phi_{\rm matter}\rangle = |00\dots0\rangle$, we obtain the circuit diagram in Figure 10.

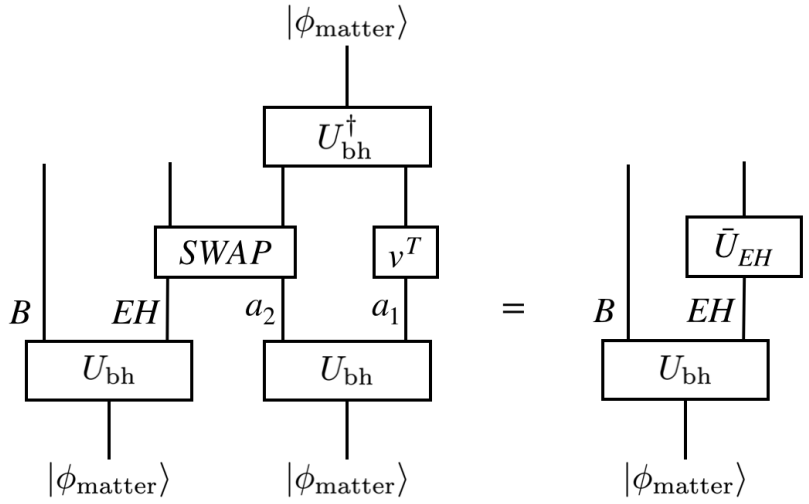

Figure 10. A probabilistic protocol which (with success probability 1/4) applies an arbitrary unitary operator v to $\tilde{B}$, the encoded interior partner of B. Here $\bar{U}_{EH}$ denotes v acting on the code subspace of EH.

Let's analyze what happens when this protocol is executed. Suppose the state of EH is an arbitrary pure quantum state $|\psi\rangle_{EH}$. After the second step, we have

$$\frac{1}{\sqrt{2}}\left(|0\rangle_{a_1}|\psi_0\rangle_{a_2}+|1\rangle_{a_1}|\psi_1\rangle_{a_2}\right)|\psi\rangle_{EH}. \tag{9.4}$$

Now expand $|\psi\rangle_{EH}$ in an orthonormal basis that includes both $|\psi_0\rangle$ and $|\psi_1\rangle$; the remaining 2^n-2 elements of the basis set are labeled $|\psi_i\rangle$ from $i=2$ to $i=2^n-1$, so that

$$|\psi\rangle=\sum_i \lambda_i|\psi_i\rangle. \tag{9.5}$$

After the third step, we obtain

$$\frac{1}{\sqrt{2}}(|0\rangle_{a_1}|\psi_0\rangle_{EH}+|1\rangle_{a_1}|\psi_1\rangle_{EH})|\psi\rangle_{a_2}, \tag{9.6}$$

which after the fourth step becomes

$$\begin{aligned}&\frac{1}{\sqrt{2}}((v_{00}|0\rangle_{a_1}+v_{01}|1\rangle_{a_1})|\psi_0\rangle_{EH}+(v_{10}|0\rangle_{a_1}+v_{11}|1\rangle_{a_1})|\psi_1\rangle_{EH})|\psi\rangle_{a_2}\\ =&\frac{1}{\sqrt{2}}(|0\rangle_{a_1}\left(v_{00}|\psi_0\rangle_{EH}+v_{10}|\psi_1\rangle_{EH}\right)+|1\rangle_{a_1}\left(v_{01}|\psi_0\rangle_{EH}+v_{11}|\psi_1\rangle_{EH}\right))|\psi\rangle_{a_2}.\end{aligned} \tag{9.7}$$

Now we want to study what happens after we carry out the fifth and the sixth step. Instead of explicitly applying $U_{a_1a_2}^{\dagger}$, it is more convenient to think about an orthogonal measurement in a basis that includes $U_{a_1a_2}|0\ldots0\rangle_a=\frac{1}{\sqrt{2}}(|0\rangle_{a_1}|\psi_0\rangle_{a_2}+|1\rangle_{a_1}|\psi_1\rangle_{a_2})$. After projecting onto this state, we obtain the (subnormalized) state

$$\begin{aligned}&\frac{1}{2}(\lambda_0\left(v_{00}|\psi_0\rangle_{EH}+v_{10}|\psi_1\rangle_{EH}\right)+\lambda_1\left(v_{01}|\psi_0\rangle_{EH}+v_{11}|\psi_1\rangle_{EH}\right))\\ =&\frac{1}{2}\left(\left(v_{00}\lambda_0+v_{01}\lambda_1\right)|\psi_0\rangle_{EH}+\left(v_{10}\lambda_0+v_{11}\lambda_1\right)|\psi_1\rangle_{EH}\right),\end{aligned} \tag{9.8}$$

which aside from the normalization factor of $1/2$ is equivalent to applying v to the code vector $\lambda_0|\psi_0\rangle_{EH} + \lambda_1|\psi_1\rangle_{EH}$. Hence, $\bar{U}_{EH}$ is applied with success probability $1/4$.

Now we explain how to upgrade this probabilistic operation to a unitary quantum circuit that applies $\bar{U}_{EH}$ deterministically. For this purpose, we use the *oblivious amplitude amplification* technique introduced by Berry *et al.*; see Lemma 3.6 of [46]. For the reader's convenience, we restate this result.

Lemma 9.1. *(Oblivious amplitude amplification) Let V' and V be unitary matrices on $\mu + n$ qubits and n qubits respectively, and let $\theta \in (0, \pi/2)$. Suppose that for any n-qubit state $|\psi\rangle$,*

$$V'|0^\mu\rangle|\psi\rangle = \sin(\theta)|0^\mu\rangle V|\psi\rangle + \cos(\theta)|\Phi^\perp\rangle, \tag{9.9}$$

where $(|0^\mu\rangle\langle 0^\mu| \otimes I)|\Phi^\perp\rangle = 0$. Let $R = 2|0^\mu\rangle\langle 0^\mu| \otimes I - I$ and $S = -V'RV'^\dagger R^\dagger$. Then,

$$S^\ell V'|0^\mu\rangle|\psi\rangle = \sin((2\ell+1)\theta)|0^\mu\rangle V|\psi\rangle + \cos((2\ell+1)\theta)|\Phi^\perp\rangle. \tag{9.10}$$

In our case, V' is the unitary process described in the first five steps, V is $\bar{U}_{EH}$, $|0^\mu\rangle$ is $|0\ldots0\rangle_a$, and $\sin(\theta) = \frac{1}{2}$. Therefore, $\theta = \frac{\pi}{6}$, and we can choose $\ell = 1$ to apply V deterministically. For this choice of ℓ, it suffices to apply V' twice and its inverse once to achieve V. For each V', we apply $U_{a_1a_2}$ and its inverse $U^\dagger_{a_1a_2}$ once each (as well as other simple unitary operations). In total, then, we can deterministically apply $\bar{U}_{EH}$ by using $U_{a_1a_2}$ three times and $U^\dagger_{a_1a_2}$ three times. In particular, the entire circuit is efficient if $U_{a_1a_2}$ is. Applying this protocol for $U_{BEH} = U_{\text{bh}}$, we obtain the circuit diagram in Figure 11.

More generally, suppose that the register B contains $|B| > 1$ qubits, so that the code subspace of EH has dimension $2^{|B|}$. A probabilistic protocol for applying an arbitrary unitary transformation to the code space can be constructed that closely follows the construction for a single qubit, but now with success probability $2^{-2|B|}$. In particular, using the probabilistic protocol and oblivious amplitude amplification we can approximate any two-qubit gate ($|B| = 2$) acting on the code space accurately and efficiently. From a universal set of such two-qubit gates, we can build a logical unitary circuit. Hence any low-complexity operation on the code space can be realized as a low-complexity quantum circuit acting on EH.

If we can perform logical gates on the code space, then we can also decode the logical state, distilling it to a small quantum memory in our possession. To be concrete, suppose the code space is two-dimensional. To decode, it suffices to prepare an ancilla qubit b in an arbitrary state, and then perform a SWAP operation on b and the encoded qubit. For this purpose we can use the quantum circuit identity shown in Figure 12, where SWAP is constructed from controlled-X, controlled-Z, and Hadamard gates. The Hadamard gates act on b, and the C-X and C-Z gates act with b as the control qubit and the code space as the target qubit.

Suppose we have a circuit acting on EH that applies X to the code. We can replace each gate in that circuit by a controlled gate, with b as the control qubit. The resulting circuit applies C-X with b as the control qubit, and if the circuit for X is efficient, so is the circuit for C-X. Likewise, we can turn an efficient circuit acting on EH that applies

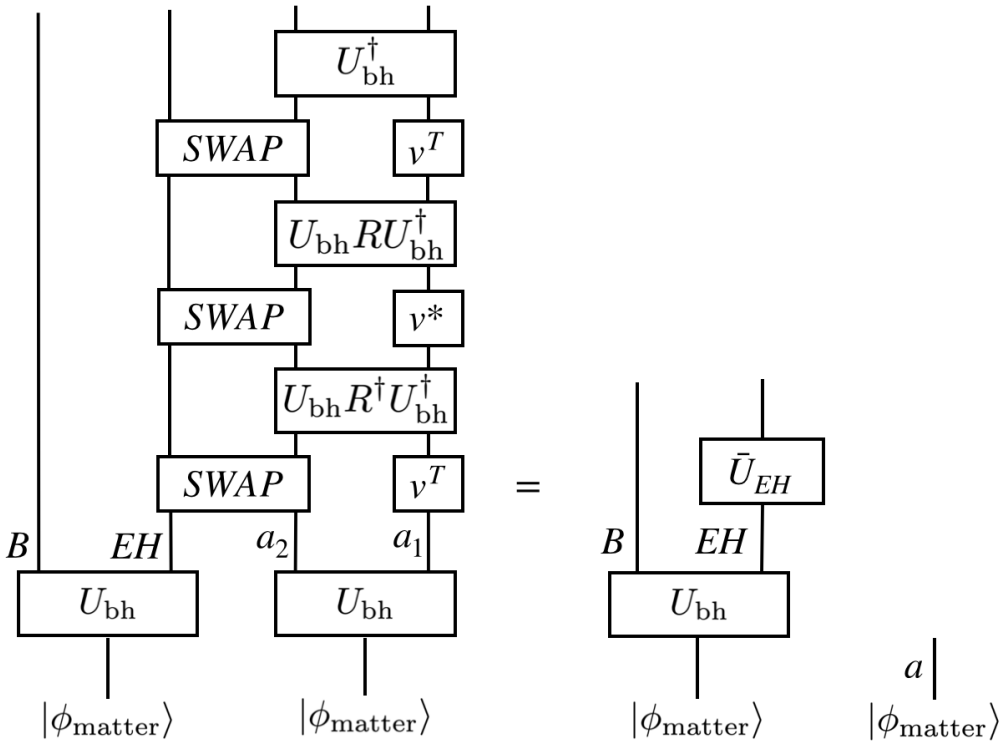

Figure 11. A deterministic circuit which applies an arbitrary unitary operator v to $\tilde{B}$, the encoded interior partner of B. Here $\bar{U}_{EH}$ denotes v acting on the code subspace of EH. Note that the final $U_{\rm bh}^\dagger$ and v^T acting on the ancilla can be removed without changing how the circuit acts on the code space.

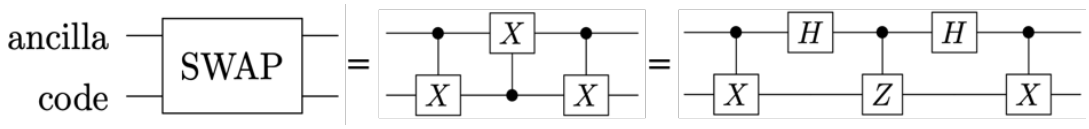

Figure 12. A two-qubit SWAP gate can be expressed in terms of Hadamard gates, controlled-X gates, and a controlled-Z gate. If there are efficient circuits for the X and Z gates acting on the code space, we may replace the gates in these circuits by gates controlled by an ancilla qubit, and use this identity to build a circuit that swaps the logical qubit in the code space with the ancilla qubit.

Z to the code into an efficient circuit for C-Z. Using the circuit identity, we obtain an efficient circuit acting on b and EH that swaps the encoded information into b. Using this realization of the SWAP gate, the entangled state of B with the encoded interior mode $\tilde{B}$ becomes an entangled state of B and b.

Note that this construction of logical gates, and of the decoding circuit, can also be applied to the fully evaporated black hole. After the evaporation is complete, H is gone, but any Hawking radiation qubit B is entangled with a highly scrambled subspace of E, a large system composed of all the other radiation quanta. Because the evolution of the initial infalling matter to the final outgoing Hawking radiation is described by an efficient

unitary transformation U, we have seen how $U_{a_1 a_2}$ and $U^\dagger_{a_1 a_2}$ can be used three times each to construct either X or Z acting on the encoded qubit. By replacing the gates in U by controlled gates, we can construct the SWAP operator, and hence distill the encoded qubit which is entangled with B into a small quantum memory efficiently.

10 Conclusion

From a purely quantum information perspective, the results in this paper apply to a tripartite pure state Ψ_{EHB}, where $|E| \gg |H| \gg |B|$. Our central assumption, from which all else follows, is that the marginal state ρ_{EB} is *pseudorandom* — *i.e.*, cannot be distinguished from a maximally mixed state with a bias better than $2^{-\alpha|H|}$ by any quantum computation with complexity polynomial in $|H|$. From this assumption, it follows that if a unitary transformation with complexity poly($|H|$) acts on E and an observer O, then B and O *decouple* in the resulting state Ψ'_{OEHB}, *i.e.*, $\rho'_{OB} \approx \rho'_O \otimes \rho'_B$ up to an error $O\left(2^{-\alpha|H|+|O|+|B|}\right)$. Here $\alpha = O(1)$ is a positive constant.

The state Ψ_{EHB} also defines an encoding map $V_\Psi : \mathcal{H}_{\tilde{B}} \to \mathcal{H}_{EH}$, whose image is a subspace of EH that is nearly maximally entangled with B. From the decoupling condition we can infer that the encoded system $\tilde{B}$ is hard to decode if $\alpha|H|-|O|-|B| \gg 1$; the observer can distill $\tilde{B}$ to a small subsystem only by performing an operation with complexity superpolynomial in $|H|$. Furthermore, if the observer O performs any quantum computation on E with complexity poly($|H|$), there is a recovery operator $\mathcal{R}$ acting on EH that corrects this "error" with fidelity $F = 1 - \epsilon$ where $\epsilon = O\left(2^{-\alpha|H|/2+|O|/2+|B|/2}\right)$. Here the size $|O|$ of the observer O may be interpreted as the number of qubits in O's quantum memory, or equivalently as the Kraus rank of the quantum channel applied to E by O.

The existence of such a recovery operator $\mathcal{R}$ has a further implication. We can construct a complete set of *ghost logical operators* for $\tilde{B}$ acting on EH; if O applies a quantum channel to E with complexity poly($|H|$), then these ghost operators commute with all the Kraus operators of the channel, up to an error $O\left(2^{-\alpha|H|/4+|O|/4+9|B|/4}\right)$. Thus the ghost operators fail to detect the action of any observer who performs an operation on E with complexity poly($|H|$).

For quantum informationists, these results may be viewed as a contribution to the theory of operator algebra quantum error-correcction in the approximate setting. What can be said about their potential physical consequences?

The existence of pseudorandom quantum states that can be prepared by quantum circuits with depth $O(\text{polylog}|H|)$ follows from standard assumptions used in post-quantum cryptography [12]. Because black holes are efficient scramblers of quantum information, it is plausible that a pseudorandom state can be efficiently prepared by an evaporating black hole, where the black hole microstates of H provide the concealed "key" of the state. A similar remark may apply to other strongly chaotic systems as well. In the setting of black holes, our conclusion about the hardness of decoding the Hawking radiation of an old black hole builds on the work of Harlow and Hayden [14] by highlighting the role of pseudorandomness, and by clarifying that the the condition $|H| \gg 1$ already ensures that decoding is hard - even if $|H|$ is much smaller than $|E|$.

We require in addition that $|H|$ is sufficiently large compared to the size $|O|$ of the observer's quantum memory, though we may allow the observer to wield a large probe system P which interacts with E, where $|P| \gg |H|, |O|$. In that case, the system $\tilde{B}$ becomes encoded in PEH rather than EH. However, the conclusion that $\tilde{B}$ cannot be efficiently distilled to a subsystem of size $|O|$ still applies for $|O| \ll \alpha|H|$, if B has constant size. Therefore, no agent with reasonable computational power can decode $\tilde{B}$ and carry it into the black hole without incurring a substantial backreaction on the black hole geometry.

To evade the black hole firewall problem, it has been proposed that (part of) the interior of an old black hole past its Page time is actually encoded in the radiation system E emitted long ago. This encoding is profoundly nonlocal and therefore potentially problematic — why can't an agent far outside the black hole who acts on E send instantaneous messages to observers who are inside, or even create a firewall at the event horizon? Our view is that computational complexity should be invoked to reconcile the nonlocal encoding of the interior with the semiclassical causal structure of the black hole geometry.

The finding that ghost logical operators can be constructed when the Hawking radiation is pseudorandom fits neatly with this viewpoint. We propose that the observables accessible to observers inside the black hole are described by these ghost logical operators, though admittedly we have no compelling general basis for this claim other than to address the firewall problem. If we accept the claim, it follows that an agent outside the black hole can create detectable excitations behind the horizon only by performing operations of superpolynomial complexity. This conclusion, though based on different arguments, meshes with the proposal by Bouland *et al.* [43, 44], that the dictionary relating the black hole interior to its exterior in the context of AdS/CFT duality must be computationally complex.

In our discussion, the encoding map relating the interior system $\tilde{B}$ to the early radiation E and remaining black hole H depends on the microstate of the initial collapsing body from which the black hole formed. It can also depend on how the observer interacts with the radiation [47]. Specifically, an observer who controls a large probe system P that comes into contact with E is empowered to alter the encoding substantially. But modifying the code does not help the observer to decode the radiation or to send a message to the interior — achieving either task by acting on E requires an operation with complexity superpolynomial in $|H|$.

Once an observer falls through the event horizon, the interior of the black hole should become accessible. From our point of view, this interior observer can interact not just with E but also with H, which makes the task of manipulating the interior far easier. Indeed, for a code space of constant dimension, arbitrary unitary transformations on the code space can be realized by quantum circuits acting on EH with complexity $\text{poly}(|EH|)$.

It is a familiar notion that, even in a theory of quantum gravity, local effective field theory on a curved background can provide an excellent approximation when the spacetime curvature is sufficiently small and the energy is sufficiently low. The story of ghost logical operators indicates that further constraints may need to be satisfied for physics to be approximately local: operations must have sufficiently low complexity and Kraus rank. Operations with high complexity and/or high rank can tear spacetime apart.

Our description of the robust encoded interior of an old black hole highlights the effectiveness of quantum error-correction against a nonstandard noise model. In the setting of fault-tolerant quantum computing, we normally seek an encoding that can protect against weakly correlated errors with a relatively low error rate. Here, though, the "noise" inflicted by our observer O on the early radiation system E is strong and chosen adversarially. As long as this noise process has computational complexity poly($|H|$) and sufficiently small Kraus rank, the encoded system $\tilde{B}$ can be restored with high fidelity, and the ghost logical operators are barely affected at all. What makes this protection possible is that, although E is treated very harshly, the "key space" H is assumed to be noiseless. Perhaps related ideas can be exploited to protect quantum information in other physically relevant settings.

Acknowledgments

We thank Adam Bouland, Raphael Bousso, Anne Broadbent, Juan Maldacena, and Geoff Penington for valuable discussions. IK's work was supported by the Simons Foundation It from Qubit Collaboration and by the Australian Research Council via the Centre of Excellence in Engineered Quantum Systems (EQUS) project number CE170100009. Part of this work was done during IK's visit to the Galileo Galilei Institute during the "Entanglement in Quantum Systems" workshop. ET and JP acknowledge funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant PHY-1733907), the Simons Foundation It from Qubit Collaboration, the DOE QuantISED program (DE-SC0018407), and the Air Force Office of Scientific Research (FA9550-19-1-0360). ET acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] S. W. Hawking, Comm. Math. Phys. **43**, 199 (1975).

[2] J. M. Maldacena, Adv. Theor. Math. Phys. **2**, 231 (1998), http://arxiv.org/abs/hep-th/9711200v3.

[3] A. Almheiri, D. Marolf, J. Polchinski, and J. Sully, JHEP **2013**, 62 (2013).

[4] J. Maldacena, JHEP **2003**, 021 (2003).

[5] J. Maldacena and L. Susskind, Fortschritte der Physik **61**, 781 (2013).

[6] L. Susskind (2014), arXiv:1402.5674.

[7] K. Papadodimas and S. Raju, JHEP **2013**, 212 (2013).

[8] D. Harlow, JHEP **2014**, 55 (2014).

[9] R. Bousso, Phys. Rev. Lett. **112**, 041102 (2014).

[10] K. Papadodimas and S. Raju, Phys. Rev. D **93**, 084049 (2016).

[11] A. Almheiri, D. Marolf, J. Polchinski, D. Stanford, and J. Sully, JHEP **2013**, 18 (2013).

[12] Z. Ji, Y.-K. Liu, and F. Song, in *Advances in Cryptology – CRYPTO 2018* (Springer International Publishing, Cham, 2018), pp. 126–152, ISBN 978-3-319-96878-0.

[13] Y. Sekino and L. Susskind, JHEP **2008**, 065 (2008).

[14] D. Harlow and P. Hayden, JHEP **2013**, 85 (2013).

[15] S. Aaronson (2016), arXiv:1607.05256.

[16] D. Marolf and J. Polchinski, JHEP **2016**, 8 (2016).

[17] P. Hayden and G. Penington, JHEP **2019**, 7 (2019), ISSN 1029-8479, URL https://doi.org/10.1007/JHEP12(2019)007.

[18] G. Penington (2019), arXiv:1905.08255.

[19] A. Almheiri, N. Engelhardt, D. Marolf, and H. Maxfield (2019), arXiv:1905.08762.

[20] A. Almheiri, R. Mahajan, J. Maldacena, and Y. Zhao (2019), arXiv:1908.10996.

[21] G. Penington, S. H. Shenker, D. Stanford, and Z. Yang (2019), arXiv:1911.11977.

[22] A. Almheiri, T. Hartman, J. Maldacena, E. Shaghoulian, and A. Tajdini (2019), arXiv:1911.12333.

[23] C. Bény and O. Oreshkov, Phys. Rev. Lett. **104**, 120501 (2010).

[24] S. T. Flammia, J. Haah, M. J. Kastoryano, and I. H. Kim, Quantum **1**, 4 (2017).

[25] C. Bény, A. Kempf, and D. W. Kribs, Phys. Rev. Lett. **98**, 100502 (2007).

[26] C. Bény, A. Kempf, and D. W. Kribs, Phys. Rev. A **76**, 042303 (2007).

[27] D. N. Page, Phys. Rev. Lett. **71**, 1291 (1993).

[28] A. C. Yao, in *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)* (1982), pp. 80–91.

[29] O. Goldreich and H. Krawczyk, in *Proceedings of the 9th Annual International Cryptology Conference on Advances in Cryptology* (Springer-Verlag, Berlin, Heidelberg, 1990), CRYPTO '89, pp. 113–127, ISBN 3-540-97317-6.

[30] S. Arora and B. Barak, *Computational Complexity: A Modern Approach* (Cambridge University Press, New York, NY, USA, 2009), 1st ed., ISBN 0521424267, 9780521424264.

[31] P. Hayden, D. W. Leung, and A. Winter, Comm. Math. Phys. **265**, 95 (2006).

[32] Z. Brakerski and O. Shmueli, in *TCC* (2019).

[33] A. Gheorghiu and M. J. Hoban, arXiv preprint arXiv:2002.12814 (2020).

[34] R. Cleve and J. Watrous, in *Proceedings 41st Annual Symposium on Foundations of Computer Science* (IEEE, 2000), pp. 526–536.

[35] A. Almheiri, X. Dong, and D. Harlow, JHEP **2015**, 163 (2015).

[36] F. Pastawski, B. Yoshida, D. Harlow, and J. Preskill, JHEP **2015**, 149 (2015).

[37] E. Verlinde and H. Verlinde, JHEP **2013**, 107 (2013).

[38] B. Yoshida, JHEP **2019**, 132 (2019).

[39] J. Preskill, Quantum Information and Computation **13**, 0181 (2013).

[40] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, New York, NY, USA, 2011), 10th ed., ISBN 1107002176, 9781107002173.

[41] J. Oppenheim and B. Unruh, JHEP **2014**, 120 (2014).

[42] E. Knill and R. Laflamme, Physical Review A **55**, 900 (1997).

[43] A. Bouland, B. Fefferman, and U. Vazirani, arXiv preprint arXiv:1910.14646 (2019).

[44] L. Susskind, arXiv preprint arXiv:2003.01807 (2020).

[45] P. Hayden and J. Preskill, JHEP **2007**, 120 (2007).

[46] D. W. Berry, A. M. Childs, R. Cleve, R. Kothari, and R. D. Somma, in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2014), STOC '14, p. 283–292.

[47] B. Yoshida (2019), arXiv:1910.11346.

[48] A. Y. Kitaev, A. Shen, M. N. Vyalyi, and M. N. Vyalyi, *Classical and quantum computation*, 47 (American Mathematical Soc., 2002).

[49] A. Barenco, C. H. Bennett, R. Cleve, D. P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. A. Smolin, and H. Weinfurter, Physical review A **52**, 3457 (1995).

A Approximate Embedding

Lemma A.1. *Let $|\Psi\rangle_{EBH}$ be a pseudo-random state (see Definition 6.1), where B is a single qubit. Then the operator V_Ψ defined by equation (7.1) is an approximate embedding, i.e., there exists an embedding V such that*

$$\|V - V_\Psi\| \leq 2 \cdot 2^{-\alpha|H|}. \tag{A.1}$$

Proof. Let $\rho_{EBH} = |\Psi\rangle\langle\Psi|_{EBH}$. Applying the decoupling inequality (6.2) without the presence of an observer (i.e., taking $|O| = 0$), we see that ρ_B is nearly maximally mixed, i.e.,

$$\|\rho_B - \frac{1}{2} I_B\|_1 \leq 2^{-\alpha|H|}. \tag{A.2}$$

Equivalently, this implies that

$$\|V_\Psi^\dagger V_\Psi - I_{\tilde{B}}\|_1 \leq 2 \cdot 2^{-\alpha|H|} := \epsilon. \tag{A.3}$$

Now, let $U\Sigma_\Psi W^\dagger = V_\Psi$ be the singular value decomposition for V_Ψ. Let us denote the singular values of V_Ψ as $\{\sigma_k\}$. Then (A.3) implies that we have $|\sigma_k^2 - 1| \leq \epsilon$. Since $|\sigma_k + 1| \geq 1$ (the singular values are nonnegative real numbers), we then also have

$$|\sigma_k - 1| \leq \epsilon \cdot |\sigma_k + 1|^{-1} \leq \epsilon. \tag{A.4}$$

Now, let Σ denote the matrix with the same shape as Σ_Ψ whose diagonal values are all equal to 1. Define $V = U\Sigma W^\dagger$, and note that V is an isometric embedding since $V^\dagger V = I_{\tilde{B}}$. Finally, we have

$$\begin{aligned} \|V - V_\Psi\| &= \|U(\Sigma - \Sigma_\Psi)W^\dagger\| && \text{(A.5)} \\ &\leq \|U\| \cdot \|W^\dagger\| \cdot \|\Sigma - \Sigma_\Psi\| && \text{(A.6)} \\ &\leq \epsilon, && \text{(A.7)} \end{aligned}$$

where the last inequality follows since all singular values of $\Sigma - \Sigma_\Psi$ are bounded above by ϵ by construction. □

B Complete Set of Ghost Operators Implies Correctability

In this Appendix, we prove a converse to Theorem 8.9, showing that if a quantum error-correcting code $\mathcal{C}$ has a complete set of δ-approximate ghost logical operators for a channel

$$\mathcal{E}(\rho) = \sum_{a=1}^{r} E_a \rho E_a^\dagger \tag{B.1}$$

with a set of r Kraus operators $K = \{E_a\}$, then the channel $\mathcal{E}_\mathcal{I}$ with Kraus operators $K \cup \{I\}$ is ϵ-correctable for $\mathcal{C}$, where $\epsilon = O(|K|\sqrt{(\dim \mathcal{C})\,\delta})$.

For this purpose, we will use the approximate version of the Knill-Laflamme error-correction conditions studied by Bény and Oreshkov [23]; these may be expressed in the form

$$P E_a^\dagger E_b P = \lambda_{ab} P + B_{ab}, \tag{B.2}$$

where P is the projector on the code space $\mathcal{C}$, λ_{ab} is a density matrix (a non-negative Hermitian operator with trace 1), and for each a and b, B_{ab} is an operator mapping $\mathcal{C}$ to $\mathcal{C}$. For $B_{ab} = 0$, these are the usual Knill-Laflamme conditions for exact correctability [42]. If B_{ab} is small, the Knill-Laflamme conditions are approximately satisfied, and a recovery operator $\mathcal{R}$ exists that corrects the channel $\mathcal{E}$ acting on the code space, up to a small error ϵ as in equation (7.8).

A relation between B_{ab} and ϵ was derived in [23]. We define maps Λ and $\mathcal{B}$ by

$$\Lambda(\rho) = \sum_{a,b=1}^{r} \lambda_{ab} \mathrm{Tr}(\rho) |a\rangle\langle b|, \quad \text{and} \quad \mathcal{B}(\rho) = \sum_{a,b=1}^{r} \mathrm{Tr}(\rho B_{ab}) |a\rangle\langle b|, \tag{B.3}$$

respectively. Consider the Bures distance $\mathfrak{B}(\Lambda + \mathcal{B}, \Lambda)$ defined as in equation (7.7), with the maximum taken over all code states ρ. Then the noise channel $\mathcal{E}$ is ϵ-correctable for the code $\mathcal{C}$ if and only if $\mathfrak{B}(\Lambda + \mathcal{B}, \Lambda) \leq \epsilon$ [23].

We may estimate this Bures distance as in equation (7.9), finding

$$2\mathfrak{B}^2(\Lambda + \mathcal{B}, \Lambda) \leq \max_\rho \left\| (\mathcal{B} \otimes \mathcal{I})(|\psi\rangle\langle\psi|) \right\|_1, \tag{B.4}$$

where $|\psi\rangle$ is a purification of the logical density operator ρ. Using equation (B.3), we obtain

$$\begin{aligned} \left\| (\mathcal{B} \otimes \mathcal{I})(|\psi\rangle\langle\psi|) \right\|_1 &= \left\| \sum_{a,b=1}^{r} \langle\psi|B_{ab}|\psi\rangle |a\rangle\langle b| \right\|_1 \\ &\leq r^2 \max_{a,b} |\langle\psi|B_{ab}|\psi\rangle| \\ &\leq r^2 \max_{a,b} \|B_{ab}\| \\ &\leq r^2 (\dim \mathcal{C}) \max_{a,b} \|B_{ab}\|_{\max}. \end{aligned} \tag{B.5}$$

Here the entry-wise max norm of $\|A\|_{\max}$ of a matrix A is defined as the largest (in absolute value) entry of the matrix in the computational basis; *i.e.*,

$$\|A\|_{\max} = \max_{i,j} |\langle i|A|j\rangle|, \tag{B.6}$$

and we used an inequality relating the operator and max norms,

$$\|B_{ab}\| \leq (\dim \mathcal{C})\|B_{ab}\|_{\max}. \tag{B.7}$$

We can now prove:

Lemma B.1. *The channel*

$$\mathcal{E}(\rho) = \sum_{a=1}^{r} E_a \rho E_a^\dagger \tag{B.8}$$

is ϵ-correctable with respect to the code $\mathcal{C}$, with

$$\epsilon = r\sqrt{\frac{1}{2}(\dim \mathcal{C})\,\delta}, \tag{B.9}$$

if there is a density operator λ_{ab} and an orthonormal basis $\{|i\rangle\}$ for the code space such that for all i and j

$$\left|\langle i|E_a^\dagger E_b|j\rangle - \delta_{ij}\lambda_{ab}\right| \leq \delta. \tag{B.10}$$

Proof. According to the Bény-Oreshkov criterion [23], the channel is ϵ-correctable if $\mathfrak{B}^2(\Lambda + \mathcal{B}, \Lambda) \leq \epsilon^2$, and from equations (B.4) and (B.5) we have

$$\mathfrak{B}^2(\Lambda + \mathcal{B}, \Lambda) \leq \frac{1}{2}r^2(\dim \mathcal{C}) \max_{a,b} \|B_{ab}\|_{\max} \leq \frac{1}{2}r^2(\dim \mathcal{C})\delta, \tag{B.11}$$

where we derived the last inequality from the definition of the $\|\cdot\|_{\max}$ norm and equation (B.10). This proves the Lemma. □

We will use the following Lemma in the proof of Theorem 8.10, as well as in the proof of Theorem B.3 below.

Lemma B.2. *Let $\mathcal{C}$ be a code subspace with code projector P. Let T be an δ-approximate ghost operator for the channel $\mathcal{E}$ and the set of Kraus operators K. Then*

$$\|[T, E]P\| \leq 2\delta\|\tilde{T}\| \tag{B.12}$$

for all $E \in K$.

Proof. Let V be the code embedding. By definition of the ghost operator, we have

$$\|TEV - EV\tilde{T}\| \leq \delta\|\tilde{T}\|, \tag{B.13}$$

for all $E \in K \cup \{I\}$. Taking $E = I$ gives

$$\|TV - V\tilde{T}\| \leq \delta\|\tilde{T}\|. \tag{B.14}$$

Then we have

$$\begin{aligned}\|[T,E]V\| = \|TEV - ETV\| &= \|TEV - EV\tilde{T} + EV\tilde{T} - ETV\| \\ &\leq \|TEV - EV\tilde{T}\| + \|EV\tilde{T} - ETV\| \\ &\leq 2\delta\|\tilde{T}\| + \|E\| \cdot \|V\tilde{T} - TV\| \\ &\leq 2\delta\|\tilde{T}\|,\end{aligned} \tag{B.15}$$

where in the last line we used equation (B.14) and the fact that $\|E\| \leq 1$ since $E^\dagger E \leq I$ implies $\|E^\dagger E\| = \|E\|^2 \leq 1$. We can now obtain equation (B.12) if we can replace V in equation (B.15) by P. This is justified because, for any operator A, we have

$$\|AP\| = \|AVV^\dagger\| \leq \|AV\| \cdot \|V^\dagger\| \leq \|AV\|, \tag{B.16}$$

where we have used $\|V^\dagger\| \leq 1$ in the last line since V is an isometric embedding. □

With these Lemmas in hand, we can proceed to prove:

Theorem B.3. *Suppose that there exists a complete set of δ-approximate ghost logical operators for the channel $\mathcal{E}$ and its set of Kraus operators $K = \{E_a\}$. Then $\mathcal{E}_\mathcal{I}$ is ϵ-correctable for the code $\mathcal{C}$, where*

$$\epsilon = (|K| + 1)\sqrt{2(\dim \mathcal{C})\,\delta}. \tag{B.17}$$

Proof. Suppose that there exists a complete set of δ-approximate ghost logical operators for $\mathcal{E}$ with respect to some Kraus decomposition $K = \{E_a\}_{a=1}^r$. We will also define $E_0 = I$.

Given any two orthogonal code states $|\psi\rangle, |\phi\rangle \in \mathcal{C}$, let us define the operators $\tilde{T}_1$ and $\tilde{T}_2$ as in the proof of Theorem 8.4. Note that $\|\tilde{T}_1\| = \|\tilde{T}_2\| = 1$. Let T_1 and T_2 be their respective δ-approximate ghost operators. Then, for $0 \leq a, b \leq r$, we get

$$\begin{aligned}&\left|2\langle\psi|E_a^\dagger E_b|\phi\rangle\right| = \left|\langle\psi|E_a^\dagger E_b T_1|\phi\rangle - \langle\psi|T_1 E_a^\dagger E_b|\phi\rangle\right| \\ &= \left|\langle\psi|E_a^\dagger E_b T_1|\phi\rangle - \langle\psi|E_a^\dagger T_1 E_b|\phi\rangle + \langle\psi|E_a^\dagger T_1 E_b|\phi\rangle - \langle\psi|T_1 E_a^\dagger E_b|\phi\rangle\right| \\ &\leq \left|\langle\psi|E_a^\dagger E_b T_1|\phi\rangle - \langle\psi|E_a^\dagger T_1 E_b|\phi\rangle\right| + \left|\langle\psi|E_a^\dagger T_1 E_b|\phi\rangle - \langle\psi|T_1 E_a^\dagger E_b|\phi\rangle\right| \\ &\leq \|\left(E_b T_1 - T_1 E_b\right)|\phi\rangle\|\||E_a|\psi\rangle\| + \||E_b|\phi\rangle\|\|\left(T_1 E_a - E_a T_1\right)|\psi\rangle\| \\ &\leq 4\delta,\end{aligned} \tag{B.18}$$

where in the second-to-last line we used the Schwarz inequality, and in the the last line we used Lemma B.2 and the fact that $\|E_a\| \leq 1$. Therefore we have

$$\left|\langle\psi|E_a^\dagger E_b|\phi\rangle\right| \leq 2\delta. \tag{B.19}$$

Repeating the same argument for $\tilde{T}_2$, we likewise get

$$\left|\langle\phi - \psi|E_a^\dagger E_b|\phi + \psi\rangle\right| \leq 2\delta. \tag{B.20}$$

Then we have

$$\begin{aligned}
&\left|\langle\phi|E_a^\dagger E_b|\phi\rangle - \langle\psi|E_a^\dagger E_b|\psi\rangle\right| \\
=\ &\left|\langle\phi|E_a^\dagger E_b|\phi\rangle - \langle\psi|E_a^\dagger E_b|\psi\rangle + \langle\phi|E_a^\dagger E_b|\psi\rangle - \langle\psi|E_a^\dagger E_b|\phi\rangle - \langle\phi|E_a^\dagger E_b|\psi\rangle + \langle\psi|E_a^\dagger E_b|\phi\rangle\right| \\
\leq\ &\left|\langle\phi|E_a^\dagger E_b|\phi\rangle - \langle\psi|E_a^\dagger E_b|\psi\rangle + \langle\phi|E_a^\dagger E_b|\psi\rangle - \langle\psi|E_a^\dagger E_b|\phi\rangle\right| + 2\left|\langle\phi|E_a^\dagger E_b|\psi\rangle\right| \\
\leq\ &2\left|\langle\phi-\psi|E_a^\dagger E_b|\phi+\psi\rangle\right| + 4\delta \\
\leq\ &8\delta. \qquad \text{(B.21)}
\end{aligned}$$

Now consider an orthonormal basis $\{|i\rangle, i = 0, 1, 2, \ldots, \dim\mathcal{C} - 1\}$, for the code space and define $\lambda_{ab} = \langle 0|E_a^\dagger E_b|0\rangle$. Noting that in the equations (B.19) and (B.21), $|\phi\rangle$ and $|\psi\rangle$ can be any two elements of the orthonormal basis, we see that

$$|\langle i|E_a^\dagger E_b|j\rangle| \leq 2\delta \qquad \text{(B.22)}$$

for $i \neq j$, while

$$|\langle i|E_a^\dagger E_b|i\rangle - \lambda_{ab}| \leq 8\delta. \qquad \text{(B.23)}$$

Thus we find that the approximate Knill-Laflamme conditions for $\mathcal{E}_\mathcal{I}$ are satisfied:

$$\frac{1}{2}\left|\langle i|E_a^\dagger E_b|j\rangle - \lambda_{ab}\delta_{ij}\right| \leq 4\delta. \qquad \text{(B.24)}$$

Note that the factor of $1/2$ comes from the normalization of the Kraus operators for $\mathcal{E}_\mathcal{I}$. From Lemma B.1, this implies that $\mathcal{E}_\mathcal{I}$ is ϵ-correctable for $\mathcal{C}$, where

$$\epsilon = (|K| + 1)\sqrt{2(\dim\mathcal{C})\,\delta}. \qquad \text{(B.25)}$$

□

C Complexity of Controlled Unitary

Lemma C.1. *Let U be a unitary of circuit complexity k with respect to some universal 2-qubit gate set $\mathcal{G}$. Given an ancillary system of n qubits, let $\Lambda_m(U)$ be the operator controlled on the state $|m\rangle$, where $0 \leq m < 2^n$, i.e.,*

$$\Lambda_m(U)(|\ell\rangle \otimes |x\rangle) = |\ell\rangle \otimes U^{\delta_{\ell m}}|x\rangle. \qquad \text{(C.1)}$$

Then given any $\epsilon > 0$, the operator $\Lambda_m(U)$ can be implemented with ϵ-precision with circuit complexity $O\left(4^n k \log^4(k/\epsilon)\right)$.

Proof. Let $U = U_k \cdots U_1$ be a decomposition of U into elements of $\mathcal{G}$. To implement $\Lambda_m(U)$ to ϵ-precision, it suffices to implement $\Lambda_m(U_i)$ to ϵ/k-precision for each $1 \leq i \leq k$. Since each U_i is a 2-qubit gate, it follows that $\Lambda_m(U_i)$ is supported on at most $n + 2$ qubits. By the Solovay-Kitaev theorem [48], each $\Lambda_m(U_i)$ can be implemented to ϵ/k-precision with $O(4^n \log^4(k/m))$ gates from $\mathcal{G}$. It follows that U itself can be implemented to ϵ-precision with $O(4^n k \log^4(k/\epsilon))$ gates.

□

The scaling with n can be considerably improved using circuit constructions from [49], but Lemma C.1 will suffice for our purposes.

D What if the radiation is not pseudorandom?

The central assumption of this paper is that the state of the Hawking radiation EB emitted by a partially evaporated black hole is pseudorandom. Here we ask what happens if this assumption is broken in a particular way.

Suppose B is a single qubit and the pure state of EBH is

$$|\Psi\rangle_{EBH} = \frac{1}{\sqrt{2}}(|0\rangle_B|\psi_0\rangle_{EH} + |1\rangle_B|\psi_1\rangle_{EH}). \tag{D.1}$$

Consider a Hermitian operator M_E acting on E such that $M_E \otimes Z_B$ can be efficiently measured, where Z_B is the Pauli-Z operator acting on B. Suppose that

$$\langle M_E \otimes Z_B\rangle_\Psi - \langle M_E\rangle_\Psi \langle Z_B\rangle_\Psi = c, \tag{D.2}$$

where the subscript Ψ indicates that the expectation value is evaluated in the global state $|\Psi\rangle_{EBH}$, or equivalently in the marginal state ρ_{EB}. Note that $c = 0$ if ρ_{EB} is maximally mixed. Therefore, by definition, if ρ_{EB} is pseudorandom, then c must be exponentially small in $|H|$. It follows that if c is a nonzero constant, independent of $|H|$, then ρ_{EB} is not pseudorandom (though the converse is not necessarily true).

We will now show that, if $c \neq 0$ there cannot be a complete set of logical operators that commute with M_E acting on the code space spanned by $\{|\psi_0\rangle_{EH}, |\psi_1\rangle_{EH}\}$. Note that because the marginal state ρ_B is maximally mixed, we have $\langle Z_B\rangle_\Psi = 0$, and therefore

$$2c = 2\langle M_E \otimes Z_B\rangle_\Psi = \langle\psi_0|M_E|\psi_0\rangle - \langle\psi_1|M_E|\psi_1\rangle. \tag{D.3}$$

Consider a Hermitian operator X_L on EH that acts on the code basis states $\{|\psi_0\rangle_{EH}, |\psi_1\rangle_{EH}\}$ like the Pauli-X operator:

$$X_L|\psi_0\rangle = |\psi_1\rangle, \quad X_L|\psi_1\rangle = |\psi_0\rangle, \tag{D.4}$$

and notice that

$$\langle\psi_1|[X_L, M_E]|\psi_0\rangle = \langle\psi_0|M_E|\psi_0\rangle - \langle\psi_1|M_E|\psi_1\rangle = 2c \neq 0. \tag{D.5}$$

This shows that the commutator $[X_L, M_E]$ is $O(1)$ acting on the code space. Thus no logical Pauli-X operator commutes with M_E acting on the code space, and in particular there can be no complete set of ghost logical operators commuting with M_E.

For this argument we chose the operator acting on B to be Z_B, but a similar argument works for any Hermitian operator acting on B. Suppose N_B is a Hermitian operator acting on B such that

$$\langle M_E \otimes N_B\rangle_\Psi - \langle M_E\rangle_\Psi \langle N_B\rangle_\Psi = c \neq 0. \tag{D.6}$$

Since N_B is Hermitian, we can diagonalize it in a certain basis, and we may assume without loss of generality that N_B is traceless. (If N_B is not traceless, we may replace N_B by $N'_B = N_B - \mathrm{Tr}\,(N_B)\,(I/2)$ without modifying equation (D.6).) In the basis in which it is diagonal, then, N_B is equal to Z_B up to a nonzero multiplicative constant.