Soft mode and interior operator in the Hayden-Preskill thought experiment

Beni Yoshida

Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada

We study the smoothness of the black hole horizon in the Hayden-Preskill thought experiment by using two particular toy models based on variants of Haar random unitary. The first toy model corresponds to the case where the coarse-grained entropy of a black hole is larger than its entanglement entropy. We find that, while the outgoing mode and the remaining black hole are entangled, the Hayden-Preskill recovery cannot be performed. The second toy model corresponds to the case where the system consists of low energy soft modes and high energy heavy modes. We find that the Hayden-Preskill recovery protocol can be carried out via soft modes whereas heavy modes give rise to classical correlations between the outgoing mode and the remaining black hole. We also point out that the procedure of constructing the interior partners of the outgoing soft mode operators can be interpreted as the Hayden-Preskill recovery, and as such, the known recovery protocol enables us to explicitly write down the interior operators. Hence, while the infalling mode needs to be described jointly by the remaining black hole and the early radiation in our toy model, adding a few extra qubits from the early radiation is sufficient to reconstruct the interior operators.

I. INTRODUCTION

Almost forty years since its formulation, the black hole information problem and its variants still shed new lights on deep conceptual puzzles in quantum gravity, and also provides useful insights to study strongly interacting quantum many-body systems [1]. While the ultimate solution of the problem could be obtained only by experimental observations, progresses can be made by utilizing thought experiments based on simple toy models [2,3]. In the last decade, two particular thought experiments on black hole dynamics have fascinated and puzzled theorists; the Almheiri-Marolf-Polchinski-Sully (AMPS) thought experiment [4] and the Hayden-Preskill thought experiment [5]. The AMPS thought experiment suggests that the smooth horizon in an old black hole, which is a consequence of the equivalence principle, may be inconsistent with monogamy of entanglement.[1] The Hayden-Preskill thought experiment poses questions concerning the absoluteness of the event horizon by suggesting that an object which has fallen into a black hole may be recovered. While there have been refined arguments and counterarguments on these conclusions in more realistic physical settings, essential features of the original works can be

reduced to very simple calculations based on Haar random unitary operators.

The main idea of this paper centers around a tension between the smooth horizon and the Hayden-Preskill thought experiment. The no-firewall postulate asserts the presence of entanglement between the infalling and outgoing Hawking pair. The recoverability in the Hayden-Preskill thought experiment requires the outgoing Hawking radiation to be entangled with a joint system of the early radiation and the reference qubits of the infalling quantum state. However, due to monogamy of entanglement, the no-firewall postulate and the recoverability look mutually incompatible.

Here we seek for a resolution to this tension by arguing that the "outgoing Hawking radiation" in the Hayden-Preskill and AMPS experiments are actually different degrees of freedom (d.o.f.). While it would be desirable to demonstrate such a separation of the Hilbert space of the outgoing mode via direct calculations on actual models of quantum gravity, our goal is more modest. In this paper, we will study a certain refinement of Haar random unitary dynamics. The unitary operator U preserves the total global $U(1)$ charge and acts as Haar random unitary operator in each subspace with fixed charge in a block diagonal manner. Although charges in black holes have led to intriguing puzzles in quantum gravity [7], it is not our primary goal to study the effect of global charges on the AMPS and Hayden-Preskill thought experiments. Our primary focus is on unitary dynamics which preserves energy. We utilize the block diagonal structure of $U(1)$-symmetric Haar random unitary to capture ergodic

---

[1]A related argument has appeared in [6].

dynamics which essentially acts as Haar random unitary on each small energy window. This is partly motivated from recent works where $U(1)$-symmetric local random unitary circuits successfully capture key properties of energy conserving systems such as an interplay of diffusive transport phenomena and the ballistic operator growth [8,9].

We will show that $U(1)$-symmetric modes are responsible for the Hayden-Preskill recovery whereas the nonsymmetric modes are responsible for correlations between the infalling and the outgoing Hawking pair. With an actual physical system with energy conservation in mind, we interpret the symmetric and non-symmetric modes as low energy (soft) and high energy (heavy) modes, respectively. Namely we claim that the Hayden-Preskill thought experiment can be carried out by using soft modes which are distinct from the Hawking radiation. Such low energy modes may be the pseudo-Goldstone mode which corresponds to the 't Hooft's gravitational mode [10]. Or perhaps they may correspond to soft gravitons due to spontaneous breaking of supertranslation symmetries [11]. In our toy model, however, the correlation between the infalling and outgoing Hawking modes is found to be purely *classical* as opposed to quantum correlations in the Hawking pair which would be seen by an infalling observer. Namely, the outgoing soft mode is found entangled with a joint of the remaining black hole and the early radiation, not with the remaining black hole itself.

We also discuss the construction of the interior partner operators of the outgoing Hawking mode. While the partners of the outgoing soft mode operators cannot be found in the remaining black hole, adding a few qubits from the early radiation to the remaining black hole is enough to construct the interior operators. The key observation is that the reconstruction of the interior operators can be seen as the Hayden-Preskill thought experiment, and hence the method from [12] can be used to explicitly write down the interior operators. This observation enables us to show that the black hole interior modes are robust against perturbations on the early radiation due to scrambling dynamics.

From the perspective of the AMPS puzzle, our result appears to suggest that interior soft operators need to be reconstructed in the early radiation instead of the remaining black hole. While this resonates with previous approaches bundled under "$A = R_B$" or "ER = EPR" [13–17], these run into various paradoxes [18–20] (See [13,21] for summaries) as the construction may lead to apparent nonlocal encoding between the interior and the early radiation. Relatedly, construction of interior operators is state-dependent, which may suffer from a number of potential inconsistencies with quantum mechanics. For further discussion, please see a selection (but by all means not a complete set) of recent work [14,17–19,22–28]. In an accompanying paper, we will debug these problems and present constructions of interior operators which are local (i.e., without involving the early radiation) and state-independent (i.e., no dependence on

the initial state of the black hole) by incorporating the effect of backreaction by the infalling observer explicitly [29].

The paper is organized as follows. In Sec. II, we warm up by studying the case where the black hole is entangled only through a subspace and its evolution is given by Haar random unitary. A corresponding physical situation is that the entanglement entropy S_E of a black hole is smaller than its coarse-grained entropy S_{BH}. The eternal AdS black hole corresponds to $S_E = S_{BH}$ whereas a one-sided pure state black hole corresponds to $S_E = 0$. We will see that taking $S_E < S_{BH}$ generates quantum entanglement between the outgoing mode and the remaining black hole, but the Hayden-Preskill recovery is no longer possible, highlighting their complementary nature. In Sec. III, we analyze the case where the black hole is entangled through a $U(1)$-symmetric subspace and its evolution is given by $U(1)$-symmetric Haar random unitary. Physically this corresponds to a black hole which is entangled with its partner through the subspace consisting of typical energy states at given temperature. In Sec. IV, we present concrete recovery protocols by following [12]. In Sec. V, we describe the procedure to construct the interior operators of the outgoing soft mode. In Sec. VI, we conclude with discussions.

Before delving into detailed discussions, we establish a few notations used throughout this paper. See Fig. 1. We will denote the Hilbert spaces for the input quantum state as A, the original black hole as B, the remaining black hole as C, and the late Hawking radiation as D. It is convenient to introduce the reference Hilbert space for the input quantum state. See [5,30] for detailed discussions on the use of the reference system. The reference Hilbert space is denoted by $\bar{A}$. The entangled partner of B is denoted by $\bar{B}$. The unitary dynamics U of a black hole acts on $AB \simeq CD$. The Hilbert space dimension of a subsystem R is denoted by d_R while the number of qubits on R is denoted by n_R. Entropies are computed as binary entropies.

The Hayden-Preskill thought experiment with Haar random unitary with various global symmetries was studied independently by Nakata, Wakakuwa and Koashi. They pointed out that, for $U(1)$ symmetry with generic input
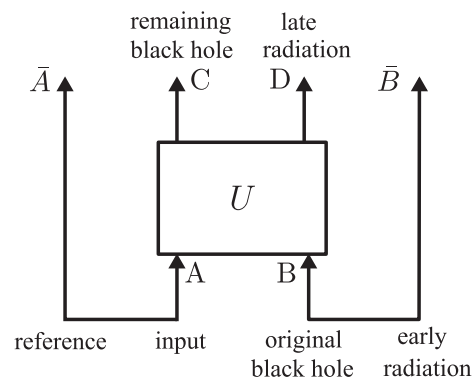

FIG. 1. The Hilbert space structure.

states, the recovery requires collecting extensive number of qubits. A similar conclusion is obtained in the Appendix for nonsymmetric input states. Upon completion of this work, we became aware of an independent work [31] which addresses the black hole evaporation process with distinction between hard and soft modes.

A. Summary of diagrammatic techniques

In this paper, we will extensively use diagrammatic tensor notations in order to express wave functions and operators as well as physical processes. Here we provide a brief tour of key properties for readers who are not familiar with these techniques.

Wave functions and operators are represented by

$$|\psi\rangle = \boxed{T} \qquad \langle\psi| = \boxed{T^*} \qquad O = \boxed{O} \tag{1}$$

which can be also explicitly written as

$$|\psi\rangle = \sum_j T_j |j\rangle \quad \langle\psi| = \sum_j T_j^* \langle j| \quad O = \sum_{ij} O_{ij} |i\rangle\langle j|. \tag{2}$$

By using these tensors as building blocks, one can express various physics in a graphical manner. For instance, an expectation value can be represented by

$$\langle\psi|O_n\cdots O_1|\psi\rangle = \begin{array}{c} \boxed{T^*} \\ \boxed{O_n} \\ \vdots \\ \boxed{O_1} \\ \boxed{T} \end{array}. \tag{3}$$

In order to associate a physical process to an equation like $\langle\psi|O_n\cdots O_1|\psi\rangle$, one needs to read it from the right to the left, i.e., the initial state $|\psi\rangle$ is acted by $O_1, O_2, \cdots$ sequentially and then is projected onto $|\psi\rangle$. In the diagrammatic notation, one needs to read the figure from the bottom to the top, i.e., the time flows upward in the diagram.

A key (yet sometimes confusing) feature of tensor diagrams is that the same tensor can represent different physical processes depending on which tensor indices are used as inputs and outputs. Let us look at a few important examples. An identity operator, $I = \sum_j |j\rangle\langle j|$, can be expressed as a straight line (i.e., a trivial tensor) since its inputs and outputs are the same:

$$I = \boxed{I} = \begin{array}{c}\uparrow\text{out}\\ \\ \uparrow\text{in}\end{array} \tag{4}$$

This diagram has one input leg (index) and one output leg. One may bend the line and construct the following diagram:

$$|\text{EPR}\rangle \propto \sum_j |j\rangle \otimes |j\rangle = \begin{array}{c}\uparrow\qquad\uparrow\\ \rule{1.5em}{0pt}\end{array} \tag{5}$$

which is the same trivial line, but with two output legs instead of one in and one out. This diagram represents an unnormalized EPR pair defined on $\mathcal{H}^{\otimes 2}$.

Another important example involves a transpose of an operator:

$$O = \boxed{O} = \boxed{O^T} \tag{6}$$

where

$$O = \sum_{i,j} O_{ij} |i\rangle\langle j| \qquad O^T = \sum_{i,j} O_{ji} |i\rangle\langle j|. \tag{7}$$

Here the transpose O^T exchanges the input and output of O. The original diagram with O represents a physical process where an arbitrary input wave function $|\psi\rangle$ is acted by O and $O|\psi\rangle$ appears as an output. The second diagram with O^T describes a physical process which involves three Hilbert spaces of the same size $\mathcal{H}^{\otimes 3} = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_3$:

$$\begin{array}{c}\text{EPR}\\ \boxed{O^T}\\ \text{EPR}\\ \mathcal{H}_1 \quad \mathcal{H}_2 \quad \mathcal{H}_3\end{array} \tag{8}$$

Here $\mathcal{H}_1$ supports an arbitrary input wave function whereas $\mathcal{H}_2 \otimes \mathcal{H}_3$ starts with the EPR pair. Then, the transpose O^T acts on $\mathcal{H}_2$, and then the system is projected onto the EPR pair on $\mathcal{H}_1 \otimes \mathcal{H}_2$. The outcome on $\mathcal{H}_3$ is $O|\psi\rangle$. One may represent this explicitly as the following equation:

$$(\langle\text{EPR}|_{12} \otimes I_3)(I_1 \otimes O_2^T \otimes I_3)|\psi\rangle_1 \otimes |\text{EPR}\rangle_{23} \propto O|\psi\rangle_3. \tag{9}$$

As the above examples suggest, one may interpret an operator as a quantum state and vice versa. To assign

physical interpretations to the diagram, we simply read it from the bottom to the top.

II. HAYDEN-PRESKILL WITH CODE SUBSPACES

In this section, we revisit the Hayden-Preskill thought experiment and the AMPS problem for quantum black holes which are entangled through "code subspaces."

Following the previous works, we will model a black hole as an n-qubit quantum system with $n = S_{BH}$ where S_{BH} is the coarse-grained entropy which is proportional to the area of the black hole. We will also assume that the dynamics of a black hole is given by a Haar random unitary operator. Let S_E be the entanglement entropy between the black hole and all the other d.o.f. including the early radiation. Previous discussions on the Hayden-Preskill thought experiment and the AMPS thought experiment concern maximally entangled black holes with $S_{BH} = S_E$. Here we will investigate a black hole with $S_{BH} > S_E$. Black holes before the Page time satisfy this condition. Similar situations have been previously considered by Verlinde and Verlinde [32]. In particular, they pointed out that taking $S_{BH} > S_E$ enables us to construct partner operators of the outgoing mode in the d.o.f. of the remaining black hole. Furthermore, they proposed a scenario to resolve the AMPS puzzle by using this effect.

The aim of this section is to study the consequence of taking $S_{BH} > S_E$ in the Hayden-Preskill thought experiment. We find that, if a black hole with $S_{BH} \gg S_E$ evolves under Haar random unitary operator, the Hayden-Preskill recovery cannot be performed unless one collects $O(n)$ qubits from the outgoing Hawking radiation. In fact, we find that the Hayden-Preskill recoverability and the smoothness of the horizon are mutually incompatible phenomena within the applicability of toy descriptions based on Haar random unitary operator.

A. Haar integral

The quantum state we are interested in is the following with $d_{\bar{A}} \le d_A$ and $d_{\bar{B}} \le d_B$:

$$= \frac{1}{\sqrt{d_{\bar{A}} d_{\bar{B}}}} U_{k\ell mo} |k\rangle_{\bar{A}} |\ell\rangle_{\bar{B}} |m\rangle_C |o\rangle_D \qquad (10)$$

where summations are implicit with $k = 1, \ldots d_{\bar{A}}$, $\ell = 1, \ldots, d_{\bar{B}}$, $m = 1, \ldots, d_C$, and $o = 1, \ldots, d_D$. Triangles represent normalized isometries. For instance, the input state in $\bar{A}A$ is given by

$$\frac{1}{\sqrt{d_{\bar{A}}}} \sum_{k=1}^{d_{\bar{A}}} |k\rangle_{\bar{A}} |k\rangle_A. \qquad (11)$$

Here $|k\rangle_A$ spans only a subspace of A. The choice of $|k\rangle_A$ is not important as the system evolves by Haar random unitary.

Each subsystem, $\bar{A}AB\bar{B}CD$, admits the following physical interpretation in the Hayden-Preskill thought experiment. A and $\bar{A}$ correspond to Hilbert spaces for the input quantum state and its reference system respectively. An input quantum information is drawn from $d_{\bar{A}}$-dimensional subspace $\bar{A}$ and is encoded into d_A-dimensional subspace A before thrown into a black hole. B corresponds to the coarse-grained Hilbert space of a black hole whereas $\bar{B}$ corresponds to the subspace where the black hole is entangled. The initial black hole satisfies $S_{BH} > S_E$ since

$$S_{BH} = \log d_B \qquad S_E = \log d_{\bar{B}}. \qquad (12)$$

Finally, C and D corresponds to the remaining black hole and the outgoing Hawking radiation respectively.

The Haar average formula with two Us and two $U^\dagger$s is

$$\int dU \, U_{i_1 j_1} U_{i_2 j_2} U^*_{i'_1 j'_1} U^*_{i'_2 j'_2}$$

$$= \frac{1}{d^2 - 1} \left(\delta_{i_1 i'_1} \delta_{i_2 i'_2} \delta_{j_1 j'_1} \delta_{j_2 j'_2} + \delta_{i_1 i'_2} \delta_{i_2 i'_1} \delta_{j_1 j'_2} \delta_{j_2 j'_1} \right) \qquad (13)$$

$$- \frac{1}{d(d^2 - 1)} \left(\delta_{i_1 i'_1} \delta_{i_2 i'_2} \delta_{j_1 j'_2} \delta_{j_2 j'_1} + \delta_{i_1 i'_2} \delta_{i_2 i'_1} \delta_{j_1 j'_1} \delta_{j_2 j'_2} \right) \qquad (14)$$

where $d = 2^n$. Approximating $d^2 - 1 \approx d^2$, this lets us compute the Haar average of $\text{Tr}\{\rho_C^2\}$ and $\text{Tr}\{\rho_{\bar{B}D}^2\}$:

$$\int dU \, \text{Tr}\{\rho_C^2\} = \frac{1}{d_{\bar{A}}^2 d_{\bar{B}}^2} \int dU \, U_{k\ell mo} U^*_{k\ell m'o} U_{k'\ell'm'o'} U^*_{k'\ell'mo'} \qquad (15)$$

$$\int dU \, \text{Tr}\{\rho_{\bar{B}D}^2\} = \frac{1}{d_{\bar{A}}^2 d_{\bar{B}}^2} \int dU \, U_{k\ell mo} U^*_{k\ell'mo'} U_{k'\ell'm'o'} U^*_{k'\ell m'o}, \qquad (16)$$

and after simple calculations of delta functions, we find

$$\int dU \, \text{Tr}\{\rho_C^2\} = \frac{1}{d_C} + \frac{1}{d_D d_{\bar{A}} d_{\bar{B}}} - \frac{d_C}{d^2} - \frac{1}{d d_C d_{\bar{A}} d_{\bar{B}}}$$

$$\approx \frac{1}{d_C} + \frac{1}{d_D d_{\bar{A}} d_{\bar{B}}} \tag{17}$$

$$\int dU \, \text{Tr}\{\rho_{\bar{B}D}^2\} = \frac{1}{d_D d_{\bar{B}}} + \frac{1}{d_C d_{\bar{A}}} - \frac{1}{d d_C d_{\bar{A}}} - \frac{1}{d d_D d_{\bar{A}}}$$

$$\approx \frac{1}{d_D d_{\bar{B}}} + \frac{1}{d_C d_{\bar{A}}} \tag{18}$$

where terms with $1/d$ factors are ignored.

B. Recoverability

The Hayden-Preskill thought experiment concerns recoverability of the input quantum state on $\bar{A}$ by having an access to both the early radiation $\bar{B}$ and the outgoing mode D. As such, the recoverability of a quantum state can be studied by asking whether one can distill EPR pairs between $\bar{A}$ and $D\bar{B}$. One can analyze the recoverability by studying quantum correlations between $\bar{A}$ and $\bar{B}D$ in $|\Psi\rangle$ in Eq. (10). Hence, as a measure of recoverability, we will use the Rényi-2 mutual information, defined by $I^{(2)}(\bar{A}, \bar{B}D) = S_{\bar{A}}^{(2)} + S_{\bar{B}D}^{(2)} - S_{\bar{A}\bar{B}D}^{(2)}$ by following [30].[2]

We can compute the Rényi-2 mutual information between $\bar{A}$ and $\bar{B}D$:

$$\int dU \, 2^{I^{(2)}(\bar{A}, \bar{B}D)} \approx d_{\bar{A}} \frac{d_D d_{\bar{A}} d_{\bar{B}} + d_C}{d_D d_{\bar{B}} + d_C d_{\bar{A}}}. \tag{19}$$

Let us look at recoverability in three regimes.

(a) (small d_D) For $d\frac{1}{d_{\bar{A}} d_{\bar{B}}} \gg d_D^2$, we find

$$d_D d_{\bar{A}} d_{\bar{B}} \ll d_C, \qquad d_D d_{\bar{B}} \ll d_C d_{\bar{A}} \tag{20}$$

leading to

$$2^{I^{(2)}(\bar{A}, \bar{B}D)} \approx 1, \tag{21}$$

implying that $\bar{A}$ and $\bar{B}D$ are not correlated.

(b) (intermediate d_D) For $d\frac{d_{\bar{A}}}{d_{\bar{B}}} \gg d_D^2 \gg d\frac{1}{d_{\bar{A}} d_{\bar{B}}}$, we find

$$d_D d_{\bar{A}} d_{\bar{B}} \gg d_C, \qquad d_D d_{\bar{B}} \ll d_C d_{\bar{A}} \tag{22}$$

leading to

$$2^{I^{(2)}(\bar{A}, \bar{B}D)} \approx \frac{d_D^2 d_{\bar{A}} d_{\bar{B}}}{d}. \tag{23}$$

[2]Strictly speaking, the Rényi-2 mutual information is not an entanglement monotone in general. See [33] for details.

The mutual information increases by two as the number of qubits in D increases by one. In order for this regime to be present, we need $d_{\bar{A}}^2 \gg 1$.

(c) (large d_D) For $d_D^2 \gg d\frac{d_{\bar{A}}}{d_{\bar{B}}}$, we find

$$d_D d_{\bar{A}} d_{\bar{B}} \gg d_C, \qquad d_D d_{\bar{B}} \gg d_C d_{\bar{A}} \tag{24}$$

$$2^{I^{(2)}(\bar{A}, \bar{B}D)} \approx d_{\bar{A}}^2, \tag{25}$$

implying that the correlation is nearly maximal.

C. Smoothness

Next let us turn our attention to the AMPS problem. Here we are interested in whether partners of outgoing mode operators on D can be reconstructed on the remaining black hole C or not. This is possible if and only if C and D retain strong quantum correlations. Hence, as a measure of smoothness of the horizon, we will use $I^{(2)}(C, D)$.

We compute the Haar average of the mutual information between C and D:

$$\int dU \, 2^{I^{(2)}(C,D)} \approx \frac{d_C d_D^2}{d_D d_{\bar{A}} d_{\bar{B}} + d_C}. \tag{26}$$

We also derive the mutual information between $\bar{A}\bar{B}$ and D in order to illustrate the monogamy of entanglement. The two mutual information are related as follows

$$2^{I^{(2)}(\bar{A}\bar{B}, D)} = \frac{2^{S_D^{(2)}}}{2^{I^{(2)}(C,D)}} \approx \frac{d_D^2}{2^{I^{(2)}(C,D)}} \tag{27}$$

where we approximated ρ_D by a maximally mixed state.

(a) (small d_D) For $d\frac{1}{d_{\bar{A}} d_{\bar{B}}} \gg d_D^2$, we find

$$2^{I^{(2)}(C,D)} \approx d_D^2, \qquad 2^{I^{(2)}(\bar{A}\bar{B}, D)} \approx 1 \tag{28}$$

implying that $I^{(2)}(\bar{A}\bar{B}, D) \approx 0$ and does not increase as d_D increases.

(b) (intermediate and large d_D) For $d_D^2 \gg d\frac{1}{d_{\bar{A}} d_{\bar{B}}}$, we find

$$2^{I^{(2)}(C,D)} \approx \frac{d}{d_{\bar{A}} d_{\bar{B}}}, \qquad 2^{I^{(2)}(\bar{A}\bar{B}, D)} \approx \frac{d_D^2 d_{\bar{A}} d_{\bar{B}}}{d} \tag{29}$$

implying that $I^{(2)}(\bar{A}\bar{B}, D)$ becomes large as the number of qubits in D increases while $I^{(2)}(C, D)$ remains unchanged.

D. Physical interpretation

Recall that the initial black hole satisfies $S_{\text{BH}} = \log d_B$ and $S_E = \log d_{\bar{B}}$ in our toy model. For the recovery in the Hayden-Preskill thought experiment, we need

$$d_D \geq \sqrt{d_A d_{\bar{A}}} \sqrt{\frac{d_B}{d_{\bar{B}}}}. \qquad (30)$$

When a black hole is maximally entangled with $d_B = d_{\bar{B}}$, the above lower bound reduces to

$$d_D \geq \sqrt{d_A d_{\bar{A}}} \qquad (31)$$

reproducing the result in [12].

Here we are interested in cases with $d_B \gg d_{\bar{B}}$ where the coarse-grained entropy is larger than the entanglement entropy. Namely, if the number of qubits in B and $\bar{B}$ satisfy $n_B - n_{\bar{B}} \sim O(n)$, then Bob needs to collect an extensive number of qubits. As such, if $d_B \gg d_{\bar{B}}$, simple recovery is not possible (unless one collects an extensive number of qubits). If we take $d_{\bar{B}} = 1$, we have $d_D \geq d_A \sqrt{d_B} \sim O(\sqrt{d})$ implying that Bob needs to collect more than a half of the total qubits for reconstruction as pointed out by Page [2].

The failure of recovery for $d_B \gg d_{\bar{B}}$ can be understood from the calculation of $I^{(2)}(C, D)$. When d_D is small, most of qubits in D are entangled with the remaining black hole C and do not reveal any information about $\bar{A}\bar{B}$. Once $2^{I^{(2)}(C,D)}$ reaches a stationary value of $\frac{d}{(d_{\bar{A}} d_{\bar{B}})^2}$, entanglement between D and $\bar{A}\bar{B}$ starts to develop. From the perspective of the firewall puzzle, the increase of $I^{(2)}(C, D)$ appears to suggest that partner operators of the outgoing mode D can be found on the remaining black hole C. Verlinde and Verlinde employed this mechanism as a possible resolution of the firewall puzzle and made an intriguing relation to theory of quantum error-correction [32]. These observations illustrate that Hayden-Preskill and the smoothness of the horizon are mutually complementary phenomena when U is a Haar random unitary which thoroughly mixes the Hilbert space AB. Namely, the presence of partner operators in C requires that D is correlated with C whereas the recoverability of an input quantum state in the Hayden-Preskill thought experiment requires that D is entangled with $\bar{A}\bar{B}$. In fact, the tradeoff between the smoothness and recoverability is strikingly sharp; once D becomes large enough to start releasing information about $\bar{A}$ via increase of $I^{(2)}(\bar{A}, \bar{B}D)$, the growth of $I^{(2)}(C, D)$ stops.

Finally, we make a comment on the complexity of performing recovery protocols in the Hayden-Preskill thought experiment. For $d_B = d_{\bar{B}}$, a simple recovery protocol is known to exist [12]. When applying this method to the case with large d_D, it is crucial to identify d.o.f. in D which is not entangled with C. Under chaotic dynamics of a black hole, it is plausible to expect that such d.o.f. become non-local inside D and require complex operations. Hence, performing the Hayden-Preskill recovery may be unphysical when $d_B \gg d_{\bar{B}}$. In fact, it may be more correct to say that D corresponds to simple d.o.f. which can be accessed easily from the outside whereas C corresponds to complex ones.

III. HAYDEN-PRESKILL WITH $U(1)$ SYMMETRY

In this section, we study the Hayden-Preskill thought experiment in the presence of conserved quantities. For simplicity of discussions, we will consider systems with $U(1)$ global symmetry where the total spin in the z-direction is preserved. We will find that the Hayden-Preskill recovery is possible only if an input quantum state is embedded into a subspace with fixed charges.

A. $U(1)$-symmetric system

Our motivations to study $U(1)$-symmetric systems are two-fold. The first obvious motivation is that we want to address the Hayden-Preskill recoverability and the AMPS problem in the presence of symmetries. The second, less obvious, motivation is to study the same set of questions for quantum systems which conserve total energy.

Let us illustrate the second point. In the discussions from the previous section, we treated qubits on B as coarse-grained d.o.f. of a black hole with $S_{\text{BH}} = \log d_{\bar{B}}$. Instead, one might want to interpret B as physical qubits on the boundary quantum system and the subspace $\bar{B}$ as a typical energy subspace of a black hole at finite temperature. Namely, if we consider the entangled AdS black hole, this amounts to assuming that $S_E = S_{\text{BH}} = \log d_{\bar{B}}$ while $\log d_B$ qubits are placed at UV. In this interpretation, our calculation would suggest that the recovery in the Hayden-Preskill thought experiment is not possible for a maximally entangled black hole at finite temperature. However, this conclusion is weird as a physical process akin to the Hayden-Preskill recovery has been recently found [34]. It has been also argued that scrambling in a sense of decay of out-of-time order correlator is sufficient to perform recovery protocols even at finite temperature [12]. Hence, something must be wrong in this interpretation.

The error in the aforementioned argument can be traced back to the approximation of the black hole dynamics by Haar random unitary. Since Haar random unitary does not conserve energy, it brings quantum states on the input
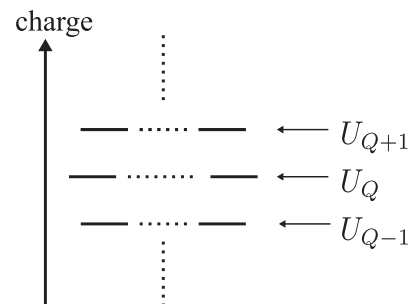

FIG. 2. $U(1)$-symmetric Haar random unitary as a toy model of energy conserving dynamics.

Hilbert space AB to outside of the window of typical energy states, and hence recovering quantum states becomes harder. This observation motivates us to consider the AMPS problem and the Hayden-Preskill thought experiment by considering some version of Haar random unitary dynamics which captures physics of energy conserving systems.

One can mimic such a situation by considering $U(1)$-symmetric Haar random unitary where the total $U(1)$ charges are conserved. To be concrete, a $U(1)$-symmetric system can be modeled as a set of n qubits where basis states can be expressed as an n-binary string and its total charge is defined as the number of 1s:

$$|i_1, \ldots, i_n\rangle \qquad m = \sum_{k=1}^{n} i_k \qquad i_k = 0, 1. \qquad (32)$$

We will consider the $U(1)$-symmetric Haar random unitary:
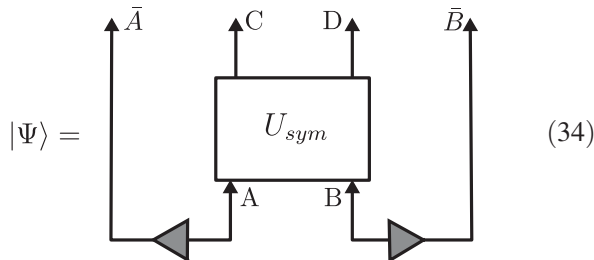
$$U = \bigoplus_{m=0}^{n} U_m \qquad \mathcal{H} = \bigoplus_{m=0}^{n} \mathcal{H}_m \qquad (33)$$

where U_m is independently Haar random acting on each fixed-charge subspace $\mathcal{H}_m$ (see Fig. 2).

With energy conservation in mind, our goal was to consider a toy model of scrambling dynamics which mixes eigenstates with roughly equal energies. By viewing each fixed-charge subspace $\mathcal{H}_m$ as eigenstates with roughly equal energies, a $U(1)$-symmetric Haar random unitary can capture dynamics which throughly mixes quantum states from the small typical energy window. In a realistic quantum system, the Hilbert space structure does not decompose into a diagonal form in an exact manner. Here we hope to capture some salient feature of energy conserving systems in this simplified toy model.

B. Haar integral

Let us denote the local charge on R by m_R (i.e., the total number of 1's in R). We will consider the cases where A and B have fixed charge m_A and m_B respectively. The quantum state of our interest is

$$|\Psi\rangle = \qquad (34)$$

where filled triangles represent normalized isometries onto fixed-charge subspaces. The input quantum states on AB can be spanned in $\mathcal{H}_{\text{in}} = \mathcal{H}_A^{(m_A)} \otimes \mathcal{H}_B^{(m_B)}$ where $\mathcal{H}_R^{(m_R)}$ represents a subspace of $\mathcal{H}_R$ with fixed charge m_R.

Define $d_{\bar{A}} \equiv \dim(\mathcal{H}_A^{(m_A)})$ and $d_{\bar{B}} \equiv \dim(\mathcal{H}_B^{(m_B)})$. The input state in $\bar{A}A$ is given by

$$\frac{1}{\sqrt{d_{\bar{A}}}} \sum_{k=1}^{d_{\bar{A}}} |k\rangle_{\bar{A}} |k_{\text{sym}}\rangle_A \qquad (35)$$

where $|k_{\text{sym}}\rangle$ are states with fixed charges. For instance, one may consider $|100\rangle$, $|010\rangle$, $|001\rangle$ for $n_A = 3$ and $m_A = 1$.

After a $U(1)$-symmetric unitary evolution, the total charge $m = m_A + m_B$ is conserved. The output Hilbert space is (assuming $n_C \geq m \geq n_D$)

$$\mathcal{H}_{\text{out}} = \bigoplus_{Q=0}^{n_D} \mathcal{H}_C^{(m-Q)} \otimes \mathcal{H}_D^{(Q)}. \qquad (36)$$

It is convenient to define

$$d_C^{(m-Q)} = \dim(\mathcal{H}_C^{(m-Q)}) \qquad d_D^{(Q)} = \dim(\mathcal{H}_D^{(Q)})$$

$$d = \dim(\mathcal{H}_{\text{out}}) = \sum_{Q=0}^{n_D} d_C^{(m-Q)} d_D^{(Q)}. \qquad (37)$$

We have $d_C^{(m-Q)} = \binom{n_C}{m-Q}$ and $d_D^{(Q)} = \binom{n_D}{Q}$.

Much of the analysis resembles the one in the previous section. The only complication is the treatment of delta functions when the Hilbert space does not have a direct product structure. The quantum state $|\Psi\rangle$ of the Hayden-Preskill thought experiment can be expressed as follows

$$|\Psi\rangle = \frac{1}{\sqrt{d_{\bar{A}} d_{\bar{B}}}} U_{k\ell(s,t)} |k\rangle_{\bar{A}} |\ell\rangle_{\bar{B}} |(s,t)\rangle_{CD} \qquad (38)$$

where (s, t) indicates that summations over (s, t) should be taken according to Eq. (36). We find

$$\int dU \, \text{Tr}\{\rho_C^2\}$$

$$= \frac{1}{d_{\bar{A}}^2 d_{\bar{B}}^2} \int dU \, U_{k\ell(s,t)} U_{k\ell(s',t)}^* U_{k'\ell'(s',t')} U_{k'\ell'(s,t')}^* \qquad (39)$$

and a similar equation for $\text{Tr}\{\rho_{\bar{B}D}^2\}$. In using the Haar formula, we need to apply delta functions to (s, t). For instance, the first term of the Haar integral in $\text{Tr}\{\rho_C^2\}$ is

$$\frac{1}{d_{\bar{A}}^2 d_{\bar{B}}^2} \cdot (\text{number of } k, \ell, k', \ell') \cdot (\text{number of } (s, t) \quad \text{and}$$

$$(s', t') \quad \text{with} \quad s = s') \qquad (40)$$

$$= \frac{1}{d_{\bar{A}}^2 d_{\bar{B}}^2} \cdot d_{\bar{A}}^2 d_{\bar{B}}^2 \cdot \sum_{Q=0}^{n_D} d_C^{(m-Q)} (d_D^{(Q)})^2 = \sum_{Q=0}^{n_D} d_C^{(m-Q)} (d_D^{(Q)})^2. \qquad (41)$$

It is convenient to define

$$W_C \equiv \sum_{Q=0}^{n_D} (d_C^{(m-Q)})^2 \cdot d_D^{(Q)} \qquad W_D \equiv \sum_{Q=0}^{n_D} d_C^{(m-Q)} \cdot (d_D^{(Q)})^2.$$

(42)

We find

$$\int dU \, \mathrm{Tr}\{\rho_C^2\} \approx \frac{W_D}{d^2} + \frac{W_C}{d^2 d_{\bar{A}} d_{\bar{B}}}$$

$$\int dU \, \mathrm{Tr}\{\rho_{\bar{B}D}^2\} \approx \frac{W_C}{d^2 d_{\bar{B}}} + \frac{W_D}{d^2 d_{\bar{A}}}$$

(43)

after ignoring terms suppressed by $1/d$.

C. Recoverability

Let us find the criteria for recovery. If n_D is large enough such that

$$\frac{W_D}{d_{\bar{A}}} \gg \frac{W_C}{d_{\bar{B}}},$$

(44)

we will have

$$\mathrm{Tr}\{\rho_{\bar{A}}^2\} = \frac{1}{d_{\bar{A}}} \quad \mathrm{Tr}\{\rho_C^2\} \approx \frac{W_D}{d^2} \quad \mathrm{Tr}\{\rho_{\bar{B}D}^2\} \approx \frac{W_D}{d^2 d_{\bar{A}}}$$

(45)

implying $I^{(2)}(\bar{A}, \bar{B}D) \approx 2\log_2 d_{\bar{A}}$. Hence the recovery will be possible. As such, we need to find the condition on n_D such that Eq. (44) holds.

Let us write W_C and W_D as follows:

$$W_C = \sum_{Q=0}^{n_D} W_C^{(Q)} \qquad W_D = \sum_{Q=0}^{n_D} W_D^{(Q)}$$

(46)

where

$$W_C^{(Q)} \equiv \binom{n_C}{m-Q}^2 \binom{n_D}{Q} \qquad W_D^{(Q)} \equiv \binom{n_C}{m-Q} \binom{n_D}{Q}^2.$$

(47)

We look for the condition for the following inequality:

$$\frac{W_D^{(Q)}}{d_{\bar{A}}} \gg \frac{W_C^{(Q)}}{d_{\bar{B}}}.$$

(48)

By writing it down explicitly, we have

$$\frac{\binom{n_D}{Q}}{d_{\bar{A}}} \gg \frac{\binom{n_C}{m-Q}}{\binom{n_B}{m_B}}.$$

(49)

When $Q \geq m_A$ and $n_D \geq n_A$, we have $m - Q \leq m_B$ and $n_C \leq n_B$. So, the right-hand side (RHS) is smaller than unity. Hence it suffices to take

$$\binom{n_D}{Q} \gg d_{\bar{A}}$$

(50)

in order to satisfy Eq. (48).

While one cannot easily satisfy Eq. (48) by taking large n_D for very small Q, contributions from such cases are negligibly small. In fact, for large n_C, we have

$$W_C^{(Q)} \simeq W_C^{(0)} \binom{n_D}{Q} \epsilon^{2Q} \qquad W_D^{(Q)} \simeq W_D^{(0)} \binom{n_D}{Q}^2 \epsilon^Q$$

(51)

with $\epsilon = \frac{p}{1-p}$. Since both $W_C^{(Q)}$ and $W_D^{(Q)}$ are (approximately) proportional to a binomial distribution and its square respectively, contributions to W_C and W_D are dominated by $Q \sim O(n_D)$. For such Q, it suffices to take $n_D \gg n_A$ in order for Eq. (48) to hold. Hence, we conclude that $n_D \gg n_A$ is sufficient for the recovery to be possible. We will present concrete recovery protocols in Sec. IV.

D. Smoothness

Next, let us compute the mutual information $I(C, D)$. For simplicity of discussion, we focus on the case where $d_A = 1$, i.e., with no input state. In this case, ρ_{CD} is a maximally mixed state with charge m:

$$\rho_{CD} = \frac{1}{\binom{n}{m}} \sum_{k=1}^{\binom{n}{m}} |k_{\mathrm{sym}}\rangle\langle k_{\mathrm{sym}}|.$$

(52)

This density matrix can be decomposed into a black diagonal form:

$$\rho_{CD} = \sum_{Q=0}^{n_D} \mathrm{Pr}(Q)\rho_{CD}^{(Q)} \qquad \rho_{CD}^{(Q)} = \sigma_C^{(m-Q)} \otimes \sigma_D^{(Q)}$$

(53)

where $\sigma_C^{(m-Q)}$ and $\sigma_D^{(Q)}$ represent maximally mixed states of charge $m - Q$ and Q respectively. The probability weight is given by

$$\mathrm{Pr}(Q) = \frac{\binom{n_D}{Q}\binom{n_C}{m-Q}}{\binom{n}{m}}.$$

(54)

For large n_C, $\mathrm{Pr}(Q)$ can be approximated by a binomial distribution with $p = m/n$. The mutual information is given by

$$I(A, B) = -\sum_{Q=0}^{n_D} \mathrm{Pr}(Q) \log \mathrm{Pr}(Q) \simeq \frac{1}{2} \log n_D.$$

(55)

Since the variance of the binomial distribution is $\sim\sqrt{n_D}$, we can approximate it as a distribution over $\sqrt{n_D}$-level states which can be encoded in $\sim\frac{1}{2}\log n_D$ bits. The above argument suggests that these "charge-bits" are strongly correlated with those on C. However, as the block diagonal form suggests, the correlation between C and D is purely classical.

E. Physical interpretation

Let us interpret the $U(1)$-symmetric Haar random unitary as an energy conserving dynamics. Then, the Hilbert space B corresponds to physical qubits of a quantum system while $\bar{B}$ corresponds to the coarse-grained Hilbert space which is also the typical energy subspace at finite temperature. The black hole is maximally entangled with the early radiation in this interpretation; $S_{\mathrm{BH}} = S_E = \log d_{\bar{B}}$. The Hawking radiation D contains symmetric and nonsymmetric modes which can be interpreted as soft and heavy modes respectively. When the input is symmetric with fixed charge, we found that the recovery is possible by collecting $O(1)$ qubits. However, as shown in Appendix, when the input is nonsymmetric with variance in charge values, we found that the recovery requires n_D to be extensive. With energy conserving systems in mind, this implies that input quantum states should be encoded in soft modes for recovery. We also see that the correlation between C and D is purely classical and results from charge conservation. With energy conserving systems in mind, it corresponds to correlations of heavy modes under energy conservation. Therefore, we interpret these heavy modes as Hawking quanta whereas soft modes are some entity responsible for the Hayden-Preskill recovery and scrambling dynamics.

From the perspective of the AMPS problem, calculations in this section motivate us to consider two different kinds of operators on D. The off-diagonal operators are the ones which change the local charge in D. The partners of those off-diagonal operators can be identified in C as operators which decrease and/or increase the total charge. This is due to classical (diagonal) correlations between C and D. On the other hand, the diagonal operators are the ones which leave the total charge unchanged up to phases. Such operators can be explicitly written as follows:

$$W = \sum_n e^{i\theta_n}|n\rangle\langle n|. \tag{56}$$

Unlike off-diagonal operators, the interior partners of diagonal operators W cannot be found in C. Hence, the nonlocality problem remains for partners of symmetric (soft) modes.

IV. RECOVERY VIA SOFT MODE

The goal of this section is to show that the Hayden-Preskill recovery can be performed via symmetric modes

only. While we will study Haar random dynamics, we believe that similar conclusions hold for any "scrambling" systems in a sense of [12] where scrambling is defined with respect to out-of-time order correlation functions.

We have discussed the recoverability in the Hayden-Preskill thought experiment without presenting explicit recovery protocols. The original work by Hayden and Preskill was essentially an existence proof of recovery protocols when the dynamics is given by Haar random unitary. Recently the author and Kitaev have constructed simple recovery protocols which work for any scrambling systems whose out-of-time order correlation functions decay [12]. Similar recovery protocols can be applied to our $U(1)$-symmetric toy model. For simplicity of discussion, we will focus on a probabilistic recovery protocol. A deterministic protocol can be also constructed by following [12]. Since the analysis in this section is a simple extension of the original work, we keep the presentation brief.

Consider the following quantum state:
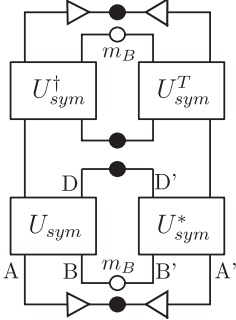
$$|\Phi\rangle \equiv \quad$$ $$\tag{57}$$

where Bob prepared a particular quantum state on $A'\bar{A}'$ which is identical to the one on $A\bar{A}$, and applied the complex conjugate U_{sym}^* on B' and A'. The initial quantum state on BB' is a maximally entangled symmetric state: $\frac{1}{\sqrt{d_{\bar{B}}}}\sum_{\ell=1}^{d_{\bar{B}}}|\ell_{\mathrm{sym}}\rangle_B \otimes |\ell_{\mathrm{sym}}^*\rangle_{B'}$. In the diagram, the unfilled dot with m_B represents a normalized projection onto the subspace with total charge m_B. Bob has an access to $B'D$ (or $DD'C'\bar{A}'$), and his goal is to distill EPR pairs on $\bar{A}\bar{A}'$.

Bob's strategy is to perform a projection onto EPR pairs on DD'. Denoting the projector by $\Pi_{\mathrm{EPR}}^{(DD')}$, the probability of measuring EPR pairs is

$$P_{\mathrm{EPR}} = \langle\Phi|\Pi_{\mathrm{EPR}}^{(DD')}|\Phi\rangle = \quad$$

$$\tag{58}$$

where the filled dots in the middle represent projectors onto EPR pairs. Recovery is successful if Bob can distill EPR pairs on $\bar{A}\bar{A}'$. Let us denote the fidelity of the distillation,

conditioned on the measurement of EPR pairs on DD', by F_{EPR}. The probability of measuring EPR pairs on both DD' and $\bar{A}\bar{A}'$ is

$$P_{\text{EPR}}F_{\text{EPR}} = \langle\Phi|\Pi_{\text{EPR}}^{(DD')}\Pi_{\text{EPR}}^{(\bar{A}\bar{A}')}|\Phi\rangle = \quad\text{(diagram)}.$$

(59)

Both P_{EPR} and $P_{\text{EPR}}F_{\text{EPR}}$ can be explicitly computed. From the above diagrams, we notice

$$P_{\text{EPR}} = \text{Tr}(\rho_{\bar{B}D}^2)\frac{d_{\bar{B}}}{d_D} \quad P_{\text{EPR}}F_{\text{EPR}} = \text{Tr}(\rho_C^2)\frac{d_{\bar{B}}}{d_D d_{\bar{A}}}. \quad (60)$$

For $U(1)$-symmetric Haar random unitary, we obtain

$$P_{\text{EPR}} \simeq P_{\text{EPR}}F_{\text{EPR}} \simeq \frac{W_D d_{\bar{B}}}{d_{\bar{A}} d_D} \quad (61)$$

where subleading terms are suppressed for $n_D \gg n_A$. Hence, upon postselection, we have $F_{\text{EPR}} \simeq 1$ implying nearly perfect recovery.

The aforementioned recovery protocol can be modified to use only the symmetric mode. Namely, by applying a projection onto maximally entangled symmetric states on DD', EPR pairs can be distilled on $\bar{A}\bar{A}'$. Let us denote the projector onto entangled states with charge Q by $\Pi_Q^{(DD')}$. This projector is related to the EPR projector by

$$\Pi_{\text{EPR}}^{(DD')} = \frac{1}{2^{n_D}}\sum_{Q=0}^{n_D}\binom{n_D}{Q}\Pi_Q^{(DD')}. \quad (62)$$

Let us denote the probability amplitude for measuring $\Pi_Q^{(DD')}$ by P_Q. For $U(1)$-symmetric Haar, we find

$$P_Q \simeq P_Q F_{\text{EPR}} \simeq \binom{n_C}{m-Q}\binom{n_D}{Q}\frac{d_{\bar{B}}}{d^2 d_{\bar{A}}} \quad (63)$$

for large n_D which satisfies Eq. (49). Hence, the Hayden-Preskill recovery can be carried out via symmetric modes.

V. CONSTRUCTION OF INTERIOR OPERATOR

Finally, we discuss the construction of partner operators that would describe the interior mode in our toy model. The partner of nonsymmetric (heavy) operators on D can be easily constructed on the remaining black hole C due to the
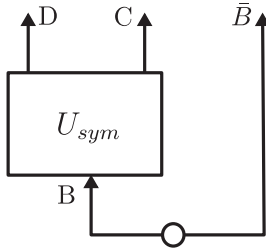
classical correlations between C and D. As such, we will focus on the partner of symmetric (soft) operators.

The main result of this section is the observation that construction of interior partner operators can be interpreted as the Hayden-Preskill recovery problem in disguise. This observation enables us to show that the black hole interior modes are robust against perturbations on the early radiation due to scrambling dynamics. Namely, even if almost all the qubits, except a few, in the early radiation are damaged, interior partner operators can be still constructed.

We assume $d_A = 1$ although our construction works well for cases with $d_A > 1$ too.

A. Interior from Hayden-Preskill

Interestingly, reconstruction of the interior operators can be performed by using a procedure similar to the Hayden-Preskill recovery. Let us begin by defining what we mean by interior partner operators. The quantum state of our interest is as follows:
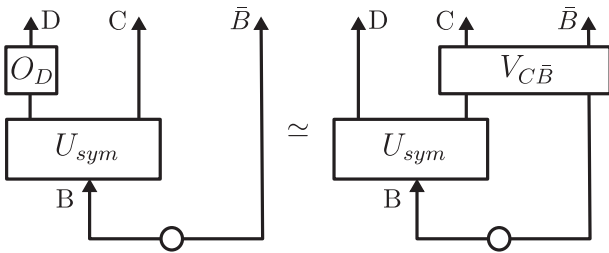
$$|\Psi\rangle = \quad\text{(diagram)} \quad (64)$$

where D is the outgoing mode and C is the remaining black hole. Our task is the following; given a symmetric (diagonal) operator O_D, find the partner operator $V_{C\bar{B}}$ such that

$$(O_D \otimes I_{C\bar{B}})|\Psi\rangle \simeq (I_D \otimes V_{C\bar{B}})|\Psi\rangle. \quad (65)$$

Graphically the above equation reads

$$\text{(diagram)} \simeq \text{(diagram)}.$$

(66)

Note that the existence of a partner operator is guaranteed. The question concerns how to write it down.

To address this problem, it is convenient to interpret the above quantum state $|\Psi\rangle$ as a map from D to $C\bar{B}$ where (symmetric) quantum states on D are encoded into $C\bar{B}$. To make this interpretation more concrete, let us deform the diagrams in Eq. (66) in the following manner:

$$(67)$$
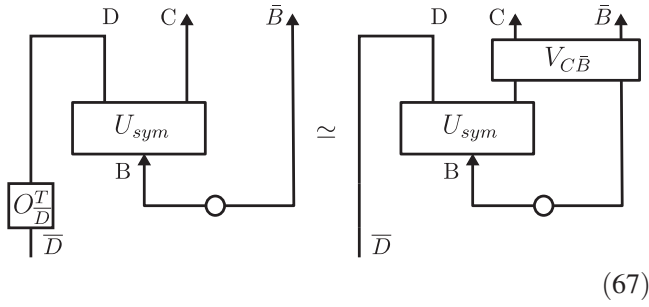
Here the original diagrams had three upward "output" legs whereas the deformed graphs use the D-index as an "input" by bending it downward. The new diagrammatic equation represents a map from $O_{\bar{D}}^T$ to a partner operator $V_{C\bar{B}}$. The map itself can be explicitly expressed as follows:

$$\bar{D} \to C\bar{B} : \qquad\qquad (68)$$
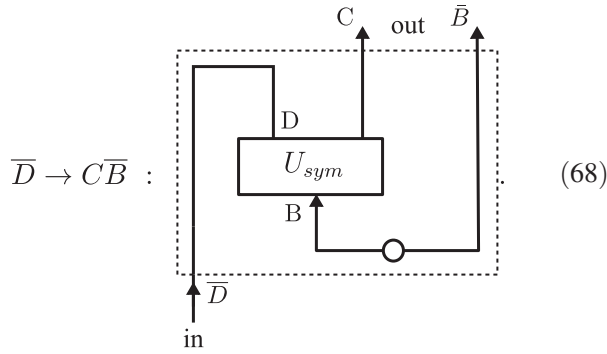
This is an isometric embedding (preserving inner products) from a smaller Hilbert space $\bar{D}$ onto a larger one $C\bar{B}$. It is worth recalling that $\bar{B}$ and C correspond to the early radiation and the remaining black hole respectively. We can associate a physical process to the deformed diagram by reading it from the bottom to the top. Namely, it begins with a system consisting of an arbitrary initial state on $\bar{D}$ and a maximally entangled symmetric state on $B\bar{B}$. Then the system evolves by U_{sym}, and is projected onto the EPR pair on $\bar{D}D$ to obtain an output wave function on $C\bar{B}$.

One can further simplify the above map (and make the relation to the Hayden-Preskill problem more explicit) by rotating the box of the unitary operator U_{sym} by 180 degrees. Carefully redrawing the diagram, we obtain the following(s):

$$(69)$$
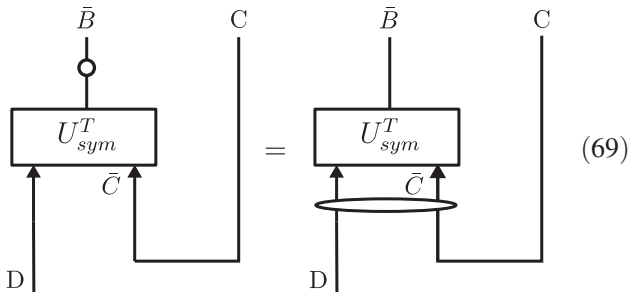
Here U_{sym}^T represents the transpose of U_{sym}, which results from flipping the diagram upside down. The circle on $\bar{B}$ represent a projection onto a symmetric subspace with fixed charge. An elongated circle also represents a projection onto a symmetric subspace with fixed charge. Note that the projector does not factor on $D \otimes \bar{C}$. Also note that the projector commutes with symmetric operator O_D and U_{sym}^T.

Our task is to reconstruct a partner of O_D^T on $\bar{B}C$. Of course, it is possible to find a partner operator on the early radiation $\bar{B}$ by simply time-evolving O_D^T by U_{sym}^T. However, the construction of a partner operator $V_{C\bar{B}}$ is not unique since $\bar{B}C$ is larger than D. The nonuniqueness of the interior partner operator is closely related to the fact that the above quantum state $|\Psi\rangle$ can be interpreted as a quantum error-correcting code where (symmetric) quantum states on D are encoded into $C\bar{B}$.

Here we want to find an alternative representation which involves as few qubits on $\bar{B}$ as possible. A key observation is that, in the above diagram, the outgoing mode D can be interpreted as an input Hilbert space for the Hayden-Preskill thought experiment where the unitary evolution is given by U_{sym}^T. To be explicit, let $\bar{B}_0$ be some small subsystem of the early radiation $\bar{B}$ which contains $n_{\bar{B}_0} \sim O(1)$ qubits such that $n_{\bar{B}_0} \gg n_D$. Let $\bar{B}_1$ be the complement of $\bar{B}_0$ in $\bar{B}$. The recoverability in the Hayden-Preskill thought experiment for symmetric modes implies that there exists a partner operator $V_{\bar{B}_0 C}$ supported on $\bar{B}_0 C$:

$$(70)$$
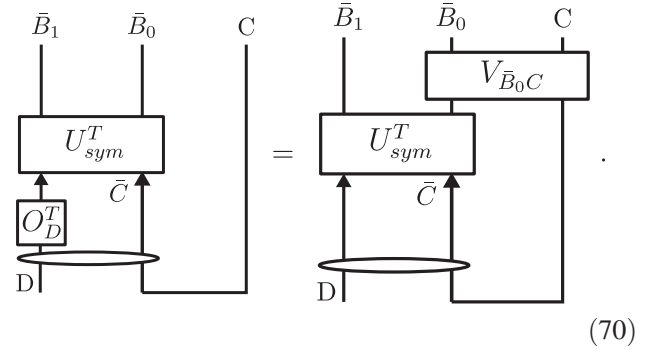
While we assumed $d_A = 1$, the above procedure works for $d_A > 1$ cases since $\bar{B}_0$ can be chosen arbitrarily.

We have observed that the reconstruction of the interior operators is essentially the Hayden-Preskill thought experiment. This suggests the following result:

(i) While the partner operator of the outgoing soft mode D cannot be found in the remaining black hole C, adding a few extra qubits $\bar{B}_0$ from the early radiation $\bar{B}$ to C is enough to construct the interior operator. Due to the scrambling nature of U_{sym}^T, one may choose any set of a few qubits $\bar{B}_0$ in order to construct the interior operator as long as out-of-time order correlation functions between D and $\bar{B}_0$ decay.

B. Fault-tolerance

From the outside quantum mechanical viewpoint, the AMPS problem (or the nonlocality problem) originates from the fact that $I(C, D)$ is small and the partner operator of D cannot be constructed in the remaining black hole C. The construction of interior operators via the Hayden-Preskill recovery protocols enables us to construct a partner operator on $C\bar{B}_0$ where $\bar{B}_0$ is a small subsystem with $|\bar{B}_0| \gtrsim |D|$. While we still need to include a few extra qubits from the early radiation $\bar{B}$, the construction is "almost" inside the remaining black hole C.

In fact, this observation sheds a new light on questions concerning the robustness of the black hole interior against perturbations on the early radiation. Some previous works attempted to resolve the AMPS puzzle by using the concept of quantum circuit complexity by drawing distinctions between simple and complex quantum operations [15,35]. Namely, it has been argued that action of simple quantum operations on the early radiation should not disturb the black hole interior operators [15]. Heuristic explanations for such fault-tolerant "encoding" of interior operators have been presented by using an analogy with quantum error-correcting code [32,36–40]. However concrete physical origin of robustness of the black hole interior mode was not discussed from the outside quantum mechanical perspective.

Our construction of interior operators suggests that robustness of the black hole interior arises from scrambling dynamics. The essential point is that the small subsystem $\bar{B}_0$ in $\bar{B}$ can be chosen in an arbitrary manner as long as the out-of-time order correlation functions between $\bar{B}_0$ and D decay [12,30]. In particular, let us consider a scenario where some large perturbations are added on the early radiation $\bar{B}$ which would damage all the qubits on $\bar{B}_1$ ($n - \log|\bar{B}_0|$ qubits). The construction of the interior operator is immune to such a drastic error since it does not involve any qubits from $\bar{B}_1$. In this sense, our construction is naturally fault-tolerant against perturbations on the early radiation $\bar{B}$. This unusual robustness of the encoding of interior operators results from the very fact that they can be constructed via the Hayden-Preskill recovery protocols.

The remaining question concerns how to write down the interior operators explicitly. On a formal level, if the time evolution U_{sym}^T is a scrambling unitary with the decay of out-of-time order correlation functions, the interior operators can be explicitly constructed by running the recovery protocol proposed in [12]. We are currently working on specific quantum systems and observe that the concept of operator growth perspective plays important roles in the construction. This will be presented elsewhere.

VI. DISCUSSIONS

In this paper, we addressed the tension between the smoothness of the horizon and the recoverability in the

Hayden-Preskill thought experiment by using a toy model with energy conservation. Within the validity of the toy model, our calculation suggests that the Hawking radiation corresponds to heavy modes whereas the Hayden-Preskill thought experiment must concern soft modes only. The correlation between the remaining black hole and the outgoing radiation is found to be classical. The classical correlation remains due to the energy conservation and due to the fact that the black hole in our toy model is entangled with the early radiation only through soft modes. Our toy model suggests that the off-diagonal correlation decoheres since the phases of heavy modes are scrambled by chaotic dynamics in soft modes. Finally, we observed that the procedure of reconstructing the soft part of the infalling mode can be interpreted as the Hayden-Preskill recovery protocol. As such, while the description of the infalling mode may require the early radiation, only a few extra qubits will be sufficient. In the reminder of the paper, we present discussions on relevant topics.

A. AMPS puzzle and scrambling

Our construction of interior operators via the Hayden-Preskill recovery phenomenon sheds a new light on the AMPS puzzle through the lens of quantum information scrambling. For simplicity of discussion, let us consider a maximally entangled black hole at infinite temperature which is represented by n copies of EPR pairs. To recap briefly, the AMPS puzzle concerns an apparent tension between descriptions by the infalling observer and the outside observer. From the perspective of the outside observer, the outgoing mode D is entangled with some d.o.f. in the early radiation $\bar{B}$. From the perspective of the infalling observer, the same outgoing mode D must be entangled with some interior mode, leading to violation of the monogamy of quantum entanglement.

The intriguing lesson from our construction of interior operators is that whoever possesses a tiny portion $\bar{B}_0$ of the early radiation $\bar{B}$ will be able to distill a qubit which is entangled with the outgoing mode D. Namely, if $\bar{B}_0$ is included to C as d.o.f. which the infalling observer can touch, she can distill an EPR pair between D and $\bar{B}_0 C$. On the other hand, if $\bar{B}_0$ is left untouched by the infalling observer, the outgoing mode D is entangled with $\bar{B} = \bar{B}_0 \bar{B}_1$ and the outside observer can distill an EPR pair between D and $\bar{B}$. See Fig. 3 for schematic illustration. In fact, these statements can be made quantitative. Decay of out-of-time order correlation functions implies that $I^{(2)}(D, \bar{B}_0 C)$ is nearly maximal [30]. Since $I^{(2)}(D, \bar{B}_0 C) + I^{(2)}(D, \bar{B}_1) = 2 S_D^{(2)}$, this suggests that $I^{(2)}(D, \bar{B}_1)$ is close to zero. Namely, the infalling observer may reconstruct a partner operator on $\bar{B}_0 C$ while the outside observer cannot reconstruct it on $\bar{B}_1$.

It is worth emphasizing that $\bar{B}_0$ can be any subsystem of $\bar{B}$ as long as $|\bar{B}_0| \gtrsim |D|$. Hence, from the perspective
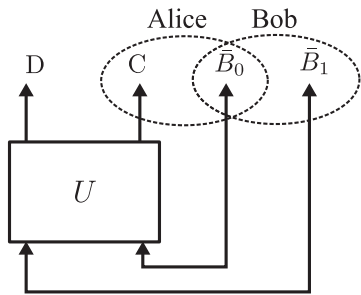
FIG. 3. The AMPS puzzle and scrambling. Alice, an infalling observer, can distill an EPR pair by accessing $\bar{B}_0 C$ while Bob, an outside observer, can distill an EPR pair by accessing $\bar{B} = \bar{B}_0 \bar{B}_1$. Note that $\bar{B}_0$ can be any subsystem of the early radiation $\bar{B}$ as long as $|\bar{B}_0| \gtrsim |D|$ due to scrambling property of U.

of the infalling observer, it is rather easy to disentangle the outgoing mode D from the early radiation $\bar{B}$ as she needs to touch only a few qubits in $\bar{B}$. In fact, it turns out that quite generic perturbations to the black hole by the infalling observer will disentangle the outgoing mode D from the early radiation $\bar{B}$ *without* ever accessing the early radiation $\bar{B}$. In an accompanying paper, based on this disentangling phenomena, we will propose a possible resolution of the AMPS problem in a manner which is free from the nonlocality problem and the state-dependence problem [29].

B. Soft modes $\simeq$ codewords

Our toy model crucially relies on the assumption that there is a clear separation between heavy and soft modes. The key insight behind this simplification is that a few thermodynamic quantities, such as energy and charge, determine the underlying classical geometry. This is essentially the statement of the no hair theorem of classical black holes. However, there are many black hole microstates which are consistent with the given classical geometry. The soft mode, discussed in the toy model, aims to capture all of these extremely low energy d.o.f. While it is unclear to us to what extent this toy model can capture the actual physics, it is concrete and simple enough to make theoretically verifiable predictions.

One interesting point is that a simple toy model with $U(1)$ symmetry can be interpreted as a energy conserving system and naturally gives rise to heavy and soft modes. In a more generic setting, we may imagine an approximate decomposition of the full Hilbert space into a block diagonal form:

$$\mathcal{H} \approx \bigoplus_{E,Q} \mathcal{H}_{E,Q} \tag{71}$$

where E represents the energy and Q represents the charge, angular momentum and other relevant macroscopic

quantities. Each subspace $\mathcal{H}_{E,Q}$ defines a Hilbert space for the classical geometry determined by a set of E, Q. Heavy operators correspond to those which moves between different subspaces whereas soft modes correspond to d.o.f. inside $\mathcal{H}_{E,Q}$. In realistic situations, such a decomposition into the block diagonal form will be an approximate one. Also there are ambiguities on which d.o.f. should be treated as soft modes. For instance, depending on the problems of interest and energy/time scales as well as dimensionality, matter on the bulk may be considered as either soft or hard mode. Our toy model aims to capture the idealistic limit where the decomposition becomes exact with sharp distinction between heavy and soft modes.

It is worth recalling that separation of soft and heavy modes plays important roles in a number of problems in quantum gravity. To add a more speculative comment, the quantum error-correcting property in the AdS/CFT correspondence is a manifestation of such separation of energy scales [41,42]. In this interpretation, the geometry (E, Q) determines the codeword subspace $\mathcal{H}_{E,Q}$ while the low energy modes correspond to different codeword states in a quantum error-correcting code determined by (E, Q). The "errors" in this quantum error-correcting code $\mathcal{H}_{E,Q}$ are heavy operators which moves the system to the outside of the codeword subspace $\mathcal{H}_{E,Q}$. In this sense, our toy model is an attempt to apply the idea of quantum error-correction to dynamical problems in quantum gravity. Hence, we believe that our approach of using the $U(1)$-symmetric toy model, despite being very simple, is applicable to a wide variety of interesting questions in quantum gravity.

At this moment, however, it is unclear to us how the black hole evaporates and eventually gets entangled only through soft modes. On one hand, if we assume that the underlying geometry changes adiabatically during evaporation, then it is reasonable to assume that the fluctuation of energy, or more generically heavy modes, is greatly suppressed. On the other hand, as is clear from the calculation of $I(C, D)$ in our toy model, the process of emitting the Hawking radiation does introduce fluctuations whose energy scale is much larger than soft modes by definition. Hence, in order for our toy model to be applicable, there needs to be some physical mechanism to suppress the energy variance in a dynamical manner.

C. Factorization of Hilbert space

Throughout the paper we used a toy model that represents a black hole as a system of qubits. This is a drastic simplification building on two nontrivial assumptions. First, it is assumed that the black hole Hilbert space is discrete. Second, it is assumed that the Hilbert space of the black hole is factorizable. The first assumption is relatively well justified as it stems from the finiteness of the black hole entropy. On the other hand, the second assumption is incorrect in a strict sense.

Whether the Hayden-Preskill recovery (as well as the AMPS argument) is applicable to systems with non-factorizable Hilbert spaces is an interesting problem. We envision that this question can be ultimately answered by using operator algebraic approaches, defining Hilbert spaces from operators instead of starting from a given Hilbert space. For such an extension, we would need to define entanglement and recoverability in operator alge-braic languages. While providing a full-fledged answer to this question is clearly beyond the scope of this paper, our analysis may be viewed as a first step toward this question. In this paper, we discussed the Hayden-Preskill recovery problem in the presence of symmetries. A symmetric subspace, even if it is embedded on a factorizable Hilbert space, is known to be nonfactorizable. We studied conditions under which the recovery is possible and saw that distinction between symmetric and nonsymmetric operators is crucial. This observation hints that the criteria on recoverability may be stated purely in terms of operators without direct use of entanglement and the structure of a given Hilbert space.

APPENDIX: U(1)-SYMMETRIC HAYDEN-PRESKILL; NONSYMMETRIC INPUTS

Here we consider the Hayden-Preskill thought experi-ment with $U(1)$-symmetric time-evolution when an input quantum state is nonsymmetric. To be specific, we will consider the input states which are superpositions of the following states with different charges:

$$|\tilde{j}\rangle = |1\cdots10\cdots\rangle \qquad j = 0,\cdots,Q \qquad \text{(A1)}$$

where the first j entries are 1's. So we have $n_A = Q$ and $n_B = n - Q$. The quantum state of our interest is

$$|\Psi\rangle = \qquad \text{(A2)}$$

where the red triangles on $A\bar{A}$ corresponds to $\frac{1}{\sqrt{Q+1}}\sum_{j=0}^{Q}|j\rangle_{\bar{A}}\otimes|\tilde{j}\rangle_A$. Let us denote the Hilbert space with total charge $m_B + a$ by $\mathcal{H}_a$ and the Haar random unitary acting on it by U_a.

By directly computing $\text{Tr}\{\rho_C^2\}$ and $\text{Tr}\{\rho_{\bar{B}D}^2\}$, one can show that the reconstruction does not work for some states. To understand which states can be reconstructed, however, we need an additional argument. We will find that the diagonal information about the total charge j can be reconstructed from both C and $\bar{B}D$. This implies $I^{(2)}(A, \bar{B}D), I^{(2)}(A, C) \gtrsim \log(Q + 1)$. Noting that $I^{(2)}(A, \bar{B}D) + I^{(2)}(A, C) = 2\log(Q + 1)$, we conclude $I^{(2)}(A, \bar{B}D)$, $I^{(2)}(A, C) \approx \log(Q + 1)$. This suggests that the off-diagonal phase information (anything except the total charge) cannot be reconstructed from either C or $\bar{B}D$.

Below we describe how to distinguish different values of j. Our argument works for $\epsilon = \frac{p}{1-p} < 1$, but we believe that a similar conclusion applies to cases with $\epsilon \simeq 1$. Consider the following quantum state

$$|\Psi^{(j)}\rangle = \qquad \text{(A3)}$$

For sufficiently large n_D, the Haar average of $\text{Tr}\{(\rho_C^{(j)})^2\}$ is given by

$$\int dU \text{Tr}\{(\rho_C^{(j)})^2\} \approx F_j \qquad F_j = \frac{\sum_{q=0}^{n_D}\binom{n_C}{m_B+j-q}\binom{n_D}{q}^2}{\binom{n}{m_B+j}^2} \qquad \text{(A4)}$$

for sufficiently large n_D. Noting that $\epsilon F_j \approx F_{j+1}$, the Rényi-2 entropy in C differs by $\log \epsilon$ when the total charge differ by one. Hence, measuring the Rényi-2 entropy in C (or $D\bar{B}$) is sufficient to learn the value of j. We would like to note that the above calculation concerns the Haar average of $\text{Tr}\{(\rho_C^{(j)})^2\}$. In order for the value of $\text{Tr}\{(\rho_C^{(j)})^2\}$ to reliably distinguish values of j, the statistical variance of $\text{Tr}\{(\rho_C^{(j)})^2\}$ must be small. For Haar random ensemble, the variance is suppressed by the total Hilbert space dimension.[3]

---

[3]Readers might think that measurement of total charge on C could reveal the value of j. Indeed, if the value of j differs by one, the expectation value of the total charge on C differs by $\frac{n_C}{n}$ since the charge will be uniformly distributed after the time-evolution. However, since the variance of the charge is large, distinguishing different values of j may require a large number of charge measurements.

[1] S. Hawking, Particle creation by black holes, Commun. Math. Phys. **43**, 199 (1975).

[2] D. N. Page, Average Entropy of a Subsystem, Phys. Rev. Lett. **71**, 1291 (1993).

[3] D. N. Page, Information in Black Hole Radiation, Phys. Rev. Lett. **71**, 3743 (1993).

[4] A. Almheiri, D. Marolf, J. Polchinski, and J. Sully, Black holes: Complementarity or firewalls?, J. High Energy Phys. 02 (2013) 062.

[5] P. Hayden and J. Preskill, Black holes as mirrors: Quantum information in random subsystems, J. High Energy Phys. 09 (2007) 120.

[6] S. L. Braunstein, S. Pirandola, and K. Życzkowski, Better Late than Never: Information Retrieval from Black Holes, Phys. Rev. Lett. **110**, 101301 (2013).

[7] N. Arkani-Hamed, LubošMotl, A. Nicolis, and C. Vafa, The string landscape, black holes and gravity as the weakest force, J. High Energy Phys. 06 (2007) 060.

[8] V. Khemani, A. Vishwanath, and D. A. Huse, Operator Spreading and the Emergence of Dissipative Hydrodynamics Under Unitary Evolution with Conservation Laws, Phys. Rev. X **8**, 031057 (2018).

[9] T. Rakovszky, F. Pollmann, and C. W. von Keyserlingk, Diffusive Hydrodynamics of Out-of-Time-Ordered Correlators with Charge Conservation, Phys. Rev. X **8**, 031058 (2018).

[10] G. 't Hooft, Strings from gravity, Phys. Scr. **T15**, 143 (1987).

[11] A. Strominger, On bms invariance of gravitational scattering, J. High Energy Phys. 07 (2014) 152.

[12] B. Yoshida and A. Kitaev, Efficient decoding for the Hayden-Preskill protocol, arXiv:1710.03363.

[13] A. Almheiri, D. Marolf, J. Polchinski, D. Stanford, and J. Sully, An apologia for firewalls, J. High Energy Phys. 09 (2013) 018.

[14] K. Papadodimas and S. Raju, An infalling observer in AdS/CFT, J. High Energy Phys. 10 (2013) 212.

[15] J. Maldacena and L. Susskind, Cool horizons for entangled black holes, Fortschr. Phys. **61**, 781 (2013).

[16] L. Susskind, New concepts for old black holes, arXiv:1311.3335.

[17] R. Bousso, Complementarity is not enough, Phys. Rev. D **87**, 124023 (2013).

[18] D. Marolf and J. Polchinski, Gauge-Gravity Duality and the Black Hole Interior, Phys. Rev. Lett. **111**, 171301 (2013).

[19] R. Bousso, Firewalls from double purity, Phys. Rev. D **88**, 084035 (2013).

[20] S. B. Giddings, Nonviolent nonlocality, Phys. Rev. D **88**, 064023 (2013).

[21] D. Harlow, Jerusalem lectures on black holes and quantum information, Rev. Mod. Phys. **88**, 015002 (2016).

[22] R. Bousso, Violations of the Equivalence Principle by a Nonlocally Reconstructed Vacuum at the Black Hole Horizon, Phys. Rev. Lett. **112**, 041102 (2014).

[23] D. Marolf and J. Polchinski, Violations of the born rule in cool state-dependent horizons, J. High Energy Phys. 01 (2016) 008.

[24] D. Harlow, Aspects of the Papadodimas-Raju proposal for the black hole interior, J. High Energy Phys. 11 (2014) 055.

[25] K. Papadodimas and S. Raju, Black Hole Interior in the Holographic Correspondence and the Information Paradox, Phys. Rev. Lett. **112**, 051301 (2014).

[26] K. Papadodimas and S. Raju, State-dependent bulk-boundary maps and black hole complementarity, Phys. Rev. D **89**, 086010 (2014).

[27] S. Raju, Smooth causal patches for AdS black holes, Phys. Rev. D **95**, 126002 (2017).

[28] D. L. Jafferis, Bulk reconstruction and the Hartle-Hawking wavefunction, arXiv:1703.01519.

[29] B. Yoshida, Firewalls vs. scrambling, arXiv:1902.09763.

[30] P. Hosur, X.-L. Qi, D. A. Roberts, and B. Yoshida, Chaos in quantum channels, J. High Energy Phys. 02 (2016) 004.

[31] Y. Nomura, Reanalyzing an evaporating black hole, Phys. Rev. D **99**, 086004 (2019).

[32] E. Verlinde and H. Verlinde, Black hole entanglement and quantum error correction, J. High Energy Phys. 10 (2013) 107.

[33] B. Yoshida and N. Y. Yao, Disentangling Scrambling and Decoherence via Quantum Teleportation, Phys. Rev. X **9**, 011006 (2019).

[34] P. Gao, D. L. Jafferis, and A. C. Wall, Traversable wormholes via a double trace deformation, J. High Energy Phys. 12 (2017) 151.

[35] D. Harlow and P. Hayden, Quantum computation vs. firewalls, J. High Energy Phys. 06 (2013) 085.

[36] A. Almheiri, X. Dong, and D. Harlow, Bulk locality and quantum error correction in AdS/CFT, J. High Energy Phys. 04 (2015) 163.

[37] A. Almheiri, Holographic quantum error correction and the projected black hole interior, arXiv:1810.02055.

[38] E. Verlinde and H. Verlinde, Behind the horizon in AdS/CFT, arXiv:1311.1137.

[39] E. Verlinde and H. Verlinde, Passing through the firewall, arXiv:1306.0515.

[40] A. Goel, H. T. Lam, G. J. Turiaci, and H. Verlinde, Expanding the black hole interior: Partially entangled thermal states in SYK, J. High Energy Phys. 02 (2019) 156.

[41] A. Almheiri, X. Dong, and D. Harlow, Bulk locality and quantum error correction in AdS/CFT, J. High Energy Phys. 04 (2015) 163.

[42] F. Pastawski, B. Yoshida, D. Harlow, and J. Preskill, Holographic quantum error-correcting codes: Toy models for the bulk/boundary correspondence, J. High Energy Phys. 06 (2015) 149.