Supplementary Material for:
Better late than never: Information retrieval from black holes
by Samuel L. Braunstein, Stefano Pirandola and Karol Życzkowski

I. STRUCTURE AND INTERPRETATION OF DECOUPLING THEOREMS

Decoupling theorems effectively describe the performance of random quantum error correction codes (QECC), whereby the quantum state to be protected $|\psi\rangle_{\rm input}$ is embedded into a larger 'code' Hilbert space by $|\psi\rangle \rightarrow |\psi\rangle \otimes |\phi_0\rangle$ followed by its encoding by a Haar-random unitary U acting on the code space. Thus

$$\text{QECC} : |\psi\rangle_{\rm input} \rightarrow |\Psi\rangle_{\rm code} = (U|\psi\rangle)_{\rm code}, \tag{7}$$

where $|\Psi\rangle_{\rm code}$ is the larger dimension code state and in the last expression we have suppressed the ancillary subsystem in standard state $|\phi_0\rangle$.

A subtlety we should mention is that the proof of the quantum decoupling theorem relies on the input state, which we wish to later locate, being entangled with a reference subsystem (ref). Thus, for example, Eq. (7) becomes

$$\mathbb{1} \otimes \text{QECC} : \sum_i |i\rangle_{\rm ref} \otimes |i\rangle_{\rm input} \rightarrow \sum_i |i\rangle_{\rm ref} \otimes (U|i\rangle)_{\rm code} \tag{8}$$

where we have suppressed normalization for convenience. This is a powerful step because it effectively allows us to utilize entanglement monogamy to precisely pin down where our encoded subsystem may be located.

With access to a sufficiently large piece of the code subspace, decoupling theorems tell us how well we can, in principle, reconstruct the original state (with its full entanglement to the reference in tact). In particular, for a k-qubit input state encoded into an n-qubit code state, there exists a threshold of $\frac{1}{2}(n+k)$ qubits above which access to more than this number of qubits of the code state allows near ideal reconstruction of the original state. More precisely, access to any $\frac{1}{2}(n+k)+c$ qubits from the code state allows reconstruction [1] of the original state with a mean fidelity of reconstruction (averaged over random encodings) bounded below by $1-2^{-c}$. Since the 'excess' unaccessed qubits of the code are not needed for the reconstruction, they have effectively decoupled from the original state and so the reconstruction protocol is unaffected by any errors that occur on these excess qubits. Thus the decoupling theorem quantifies the performance of random quantum error correcting codes.

Importantly, the proofs of (quantum) decoupling theorems are non-constructive. They only demonstrate the existence of a reconstructing unitary with the claimed performance, but do not say how it may be made.

II. DECOUPLING IN A CLASSICAL SETTING

Here we paraphrase the discussion in Ref. 1 for a simple version of a decoupling result in a classical setting. Consider a k-bit plaintext message randomly encoded into an n-bit ciphertext string. The codebook for this code will consist of 2^k n-bit random codewords and their associated k-bit messages. Obviously, anyone with access to the codebook and any specific encoded message will be able to exactly decode it. However, knowing the codebook allows one to do almost as well with just a few more than k bits of the encoded message (indeed *any* k plus a few bits) of the encoded message.

As noted in Ref. 1, given access to only $k+c$ bits of (and their location in) the encoded message, one can eliminate many of the potential entries in the codebook, thus narrowing down the possible message. To estimate the probability for this procedure to identify any particular message from $k+c$ bits we may treat matches as uniformly random (for our randomly generated codebook). The probability of a random match between $k+c$ bits from a specific encoded message and the identically located $k+c$ bits from any specific codeword in the codebook will then be $2^{-(k+c)}$. Given therefore that there are 2^k possible messages to distinguish between, the probability of failure to identify the correct message will be $2^k 2^{-(k+c)} = 2^{-c}$. Finally, the probability with which access to any $k+c$ bits of the encoded message (and the codebook) allows one to have successfully reconstructed the original plaintext message (what one might call the fidelity of reconstruction) is just $1-2^{-c}$.

We might note some significant differences between this classical decoupling and quantum decoupling results. First, for the case where $k \ll n$, classical decoupling allows many reconstructions of the original message from completely distinct subsets of bits from the encoded message — the classical information can be cloned in this manner. By contrast, for the analogous quantum encoding [as in Eq. (8)] access to $\frac{1}{2}(n+k)+c$ qubits of the encoded state are needed to achieve a reconstruction fidelity of $1-2^{-c}$. So quantum cloning is strictly prohibited. Nonetheless, any $\frac{1}{2}(n+k)+c$ qubits are adequate for this purpose. Second, in the classical setting the reconstruction protocol is trivial given access to the codebook, whereas the analogous reconstruction protocol in the quantum case is only shown to exist (the proof is non-constructive) but may require full knowledge of the encoding random unitary U, which would be an extreme burden in any even moderately high-dimensional scenario.

III. PENALTY FOR DISENTANGLEMENT ACROSS A HYPOTHETICAL BOUNDARY

Here we investigate the energy penalty one must pay for creating a disentangled (i.e., separable) state across a hypothetical boundary. We are not here going to consider the effects of real boundary conditions on the state of a quantum system, merely the effect of a constraint on the state space so as to exclude entangled states across a non-physical (fictive) boundary. Indeed, the equivalence principle has been argued [2] to imply that freely falling observers see nothing physical as they pass the event horizon.

Consider M coupled Harmonic oscillators with Hamiltonian

$$H = \frac{1}{2}\sum_{i=1}^{M} p_i^2 + \frac{1}{2}\sum_{i,j=1}^{M} K_{ij}\, x_i x_j, \tag{9}$$

where $[x_i, p_j] = i\,\delta_{ij}$ and K is a real symmetric (non-negative definite) matrix. The ground state wavefunction as a function of $\vec{x} \equiv (x_1, \ldots, x_M)^T$ is

$$\Psi(\vec{x}) = \frac{(\det\sqrt{K})^{1/4}}{\pi^{M/4}} \exp(-\vec{x}\cdot\sqrt{K}\cdot\vec{x}), \tag{10}$$

with ground state energy $\frac{1}{2}\mathrm{tr}\,(\sqrt{K})$.

Let us introduce a hypothetical boundary at index $b < M$. We assign all oscillators with indices $i \leq b$ as 'inside' this fictive boundary and all other oscillators as 'outside'. It is natural to partition the coupling matrix K into blocks as

$$K = \begin{pmatrix} K_{\rm in} & Q \\ Q^T & K_{\rm out} \end{pmatrix}. \tag{11}$$

where $K_{\rm in}$ is a $b \times b$ symmetric matrix and $K_{\rm out}$ is an $(M-b)\times(M-b)$ symmetric matrix. The Hamiltonian of Eq. (9) may be rewritten as

$$\begin{aligned} H \;=\; & \frac{1}{2}\big(\vec{p}_{\rm in}^{\,2} + \vec{x}_{\rm in}\cdot K_{\rm in}\cdot\vec{x}_{\rm in} + \vec{p}_{\rm out}^{\,2} + \vec{x}_{\rm out}\cdot K_{\rm out}\cdot\vec{x}_{\rm out}\big) \\ & + \vec{x}_{\rm in}\cdot Q\cdot\vec{x}_{\rm out}, \end{aligned} \tag{12}$$

where $\vec{x} = \vec{x}_{\rm in}\oplus\vec{x}_{\rm out}$ decomposes $\vec{x}$ into a b-dimensional vector $\vec{x}_{\rm in}$ and an $(M-b)$-dimensional vector $\vec{x}_{\rm out}$. This effectively decomposes the full M-oscillator Hilbert space $\mathcal{H}_{\rm total}$ into a tensor product $\mathcal{H}_{\rm total} = \mathcal{H}_{\rm in}\otimes\mathcal{H}_{\rm out}$.

Theorem: The general separable state across $\mathcal{H}_{\rm in}\otimes\mathcal{H}_{\rm out}$ with lowest energy for Hamiltonian (12) has energy above the ground state of

$$E_{\rm penalty} \equiv \frac{1}{2}\big[\mathrm{tr}\,(\sqrt{K_{\rm in}}) + \mathrm{tr}\,(\sqrt{K_{\rm out}}) - \mathrm{tr}\,(\sqrt{K})\big]. \tag{13}$$

We call this the minimal 'energy penalty' for ensuring the separability of a state across a hypothetical boundary.

Proof: A general separable state is just the convex sum over states of the form $\rho_{\rm in}\otimes\rho_{\rm out}$, where without loss of generality we may treat $\rho_{\rm in}$ and $\rho_{\rm out}$ as pure states. A lower bound to the energy expectation of a general separable state is therefore given by the lower bound for the energy expectation over a single such tensor product of pure states.

Consider now a general product of pure states. We may always write its wavefunction as a displaced product

$$\begin{aligned} &\Psi_{\rm prod}(\vec{x}_{\rm in}, \vec{x}_{\rm out}) \equiv \\ &D_{\rm in}(\vec{x}_{\rm in}^{\,0} + i\vec{p}_{\rm in}^{\,0})\Psi_{\rm in}(\vec{x}_{\rm in}) D_{\rm out}(\vec{x}_{\rm out}^{\,0} + i\vec{p}_{\rm out}^{\,0})\Psi_{\rm out}(\vec{x}_{\rm out}), \end{aligned} \tag{14}$$

where $\Psi_{\rm zero}\equiv\Psi_{\rm in}\Psi_{\rm out}$ is taken to have zero mean positions and momenta. The expectation $\langle H\rangle_{\rm prod}$ of Hamiltonian (12) with respect to the general state of Eq. (14) may now be rewritten as an expectation over this 'zero mean' state $\Psi_{\rm zero}$ as

$$\begin{aligned} &\frac{1}{2}\Big[\big\langle \vec{p}_{\rm in}^{\,2} + \vec{x}_{\rm in}\cdot K_{\rm in}\cdot\vec{x}_{\rm in} + \vec{p}_{\rm out}^{\,2} + \vec{x}_{\rm out}\cdot K_{\rm out}\cdot\vec{x}_{\rm out}\big\rangle_{\rm zero} \\ &+ \vec{p}_{\rm in}^{\,0\,2} + \vec{x}_{\rm in}^{\,0}\cdot K_{\rm in}\cdot\vec{x}_{\rm in}^{\,0} + \vec{p}_{\rm out}^{\,0\,2} + \vec{x}_{\rm out}^{\,0}\cdot K_{\rm out}\cdot\vec{x}_{\rm out}^{\,0} \\ &+ 2\,\vec{x}_{\rm in}^{\,0}\cdot Q\cdot\vec{x}_{\rm out}^{\,0}\Big]. \end{aligned} \tag{15}$$

Since K is non-negative definite, for any vector $\vec{x}^{\,0} = \vec{x}_{\rm in}^{\,0}\oplus\vec{x}_{\rm out}^{\,0}$ we have $\vec{x}^{\,0}\cdot K\cdot\vec{x}^{\,0}\geq 0$. Thus

$$\begin{aligned} &\langle H\rangle_{\rm prod} \geq \\ &\frac{1}{2}\big\langle \vec{p}_{\rm in}^{\,2} + \vec{x}_{\rm in}\cdot K_{\rm in}\cdot\vec{x}_{\rm in} + \vec{p}_{\rm out}^{\,2} + \vec{x}_{\rm out}\cdot K_{\rm out}\cdot\vec{x}_{\rm out}\big\rangle_{\rm zero}. \end{aligned} \tag{16}$$

Note that the right-hand-side is just the expectation of the sum of a pair of independent oscillators with individual ground state energies $\frac{1}{2}\mathrm{tr}\,(\sqrt{K_{\rm in}})$ and $\frac{1}{2}\mathrm{tr}\,(\sqrt{K_{\rm out}})$ respectively. Thus,

$$\langle H\rangle_{\rm prod} \geq \frac{1}{2}\big[\mathrm{tr}\,(\sqrt{K_{\rm in}}) + \mathrm{tr}\,(\sqrt{K_{\rm out}})\big]. \tag{17}$$

Further, since these independent product ground states have zero means, this lower bound is achieved. ■

In order to see how this separability penalty appears in a field theoretic setting consider a free scalar field with Hamiltonian

$$H = \frac{1}{2}\int d^3x\,\big[\pi^2 + (\vec{\nabla}\varphi)^2\big], \tag{18}$$

where $\pi = \partial_t\varphi$ is the conjugate momentum for the quantum field φ and these satisfy the equal-time canonical commutation relations

$$\big[\varphi(t,\vec{x}), \pi(t,\vec{x}')\big] = i\,\delta(\vec{x}-\vec{x}'). \tag{19}$$

Following Srednicki [3], we introduce a lattice of discrete points with equal spacing a in the radial direction. Furthermore, the field is placed in a spherical box of radius $A = (N+1)a$ and the field is taken to vanish at the (real) boundary at A. The field and its conjugate momentum can be decomposed into partial waves $\varphi_{j,lm}$ and $\pi_{j,lm}$ satisfying the equal time commutation relation

$$[\varphi_{j,lm}, \pi_{j',l'm'}] = i\,\delta_{jj'}\delta_{ll'}\delta_{mm'}, \tag{20}$$

where ja gives the discrete radial coordinate and $\{l, m\}$ label the partial waves' angular momentum. The discretized Hamiltonian then becomes $H = \sum_{l,m} H_{lm}$, with [3]

$$\begin{aligned} H_{lm} \;&=\; \frac{1}{2a}\sum_{j=1}^{N}\Bigg[\pi_{j,lm}^2 + (j+\frac{1}{2})^2\Big(\frac{\varphi_{j,lm}}{j} - \frac{\varphi_{j+1,lm}}{j+1}\Big)^2 \\ &\qquad + \frac{l(l+1)}{j^2}\varphi_{j,lm}^2\Bigg]. \end{aligned} \tag{21}$$

Numerical calculations of $E_{\rm penalty}$ from Eq. (13) for this discretized Hamiltonian yield

$$E_{\rm penalty} \simeq 0.05\,\frac{(N+1)^2}{a} = 0.05\,\frac{A^2}{a^3}, \tag{22}$$

where the hypothetical boundary index is chosen as $b = N/2$ (across a range of even N from 50 to 100). This penalty diverges as the cube of the ultraviolet regulator $1/a$. Thus we expect pure quantum states where entanglement has essentially vanished across a hypothetical boundary to have very large energies.

IV. UNIFORM ENTANGLEMENT

Because one of the key claims in the paper is about loss of trans-event horizon entanglement, we shall repeat the key calculation here for a black hole with trans-event horizon entanglement, but where, for simplicity, that entanglement is taken to be uniform. This allows us to repeat the analysis solely using results already available in the literature.

Consider black hole evaporation with uniform trans-event horizon entanglement as

$$\frac{1}{\sqrt{E}}\sum_{j=1}^{E}|j\rangle_{\rm int}\otimes|j\rangle_{\rm ext} \to \frac{1}{\sqrt{E}}\sum_{j=1}^{E}(U|j\rangle)_{\rm BR}\otimes|j\rangle_{\rm ext}. \tag{23}$$

Here $\log_2 E$ is the entropy of entanglement between the external (ext) neighborhood and the interior of the black hole. Except for the interpretation of the source of entanglement, this model has been recently analyzed by Ref. 1. We may therefore quote their key result in our terms: For any positive c, once $\frac{1}{2}S_{\rm BH} + \frac{1}{2}\log_2 E + c$ qubits have radiated away [this is just the $\frac{1}{2}(n+k)+c$ qubits required as discussed in the first section of this Supplementary Material], the trans-event horizon entanglement between the external neighborhood and the interior subsystems will have virtually vanished, with it appearing instead (with a fidelity of at least $1-2^{-c}$) as entanglement between the external neighborhood and the outgoing radiation. Here (as in our manuscript) c is a free parameter, but will be dwarfed by any of the entropies involved.

Repeating the argument from our manuscript, this loss must be delayed until the black hole has evaporated to roughly the Planck scale. (Indeed, section III of this Supplementary Material provides energy estimates for the departure from vacuum across the event horizon when trans-event horizon entanglement is lost.) Such a delay implies that roughly $S_{\rm BH}$ qubits must have already been radiated before such loss occurs, in which case

$$\log_2 E \approx S_{\rm BH}. \tag{24}$$

In section V below, we shall see that when the uniform entanglement of the above analysis is replaced with general trans-event horizon entanglement, the measure of entanglement $\log_2 E$ is replaced by the Rényi entropy $H^{(1/2)}(\rho_{\rm ext})$. This replacement is unchanged in the presence of in-fallen matter (also section VI).

V. FORMALISM FOR GENERAL ENTANGLEMENT

Note that all Rényi entropies are bounded above by the logarithm of the Hilbert space dimension, so $0 \le H^{(q)}(\rho_{\rm ext}) \le n \equiv S_{\rm BH}$ for the state we study. Of particular interest to us here will be two Rényi entropies for $q = \frac{1}{2}$, 2, so

$$\begin{aligned} H^{(1/2)}(\rho_{\rm ext}) \;&=\; \log_2\left[({\rm tr}\,\sqrt{\rho_{\rm ext}}\,)^2\right] \\ H^{(2)}(\rho_{\rm ext}) \;&=\; -\log_2\,({\rm tr}\;\rho_{\rm ext}^2\,). \end{aligned} \tag{25}$$

(In the limit of $q \to 1$ the Rényi entropy reduces to the more familiar von Neumann entropy.)

Our key result is based on our generalization (theorem below) of the decoupling theorem of Ref. 4. Consider now the tripartite state

$$\rho_{XYZ} = \rho_{XY_1Y_2Z}, \tag{26}$$

where the joint subsystems $Y = Y_1Y_2$ will be decomposed as either the radiation subsystem and interior black hole subsystem RB or vice-versa BR. This allows us to define

$$\sigma^U_{XY_2Z} \equiv {\rm tr}_{Y_1}\big(U_Y\;\rho_{XYZ}\;U_Y^\dagger\big). \tag{27}$$

In keeping with the naming convention of Ref. 4, we call the result below the Mother-in-law decoupling theorem.

Generalized decoupling theorem:

$$
\begin{aligned}
&\left(\int_{U\in U(Y)} dU\, \left\|\sigma^U_{XY_2Z} - \sigma^U_X \otimes \sigma^U_{Y_2Z}\right\|_1\right)^2 \\
\le\;& \mathrm{tr}\,\rho_X^{2\nu}\ \mathrm{tr}\,\rho_Z^{2\mu}\Big\{\Big[\mathrm{tr}\,\rho^2_{XZ}(\rho_X^{-2\nu}\otimes\rho_Z^{-2\mu}) \\
&\qquad - 2\,\mathrm{tr}\,\rho_{XZ}(\rho_X^{1-2\nu}\otimes\rho_Z^{1-2\mu}) \\
&\qquad + \mathrm{tr}\,\rho_X^{2-2\nu}\ \mathrm{tr}\,\rho_Z^{2-2\mu}\Big] \\
&\quad + \frac{Y_2}{Y_1}\Big[\mathrm{tr}\,\rho^2_{XYZ}(\rho_X^{-2\nu}\otimes\rho_Z^{-2\mu}) \\
&\qquad + \mathrm{tr}\,\rho_X^{2-2\nu}\ \mathrm{tr}\,\rho^2_{YZ}\ \rho_Z^{-2\mu}\Big]\Big\} \qquad (28) \\
\le\;& \frac{Y_2}{Y_1}\,\mathrm{tr}\,\rho_X^{2\nu}\ \mathrm{tr}\,\rho_Z^{2\mu}\Big[\mathrm{tr}\,\rho^2_{XYZ}(\rho_X^{-2\nu}\otimes\rho_Z^{-2\mu}) \\
&\qquad + \mathrm{tr}\,\rho_X^{2-2\nu}\ \mathrm{tr}\,\rho^2_{YZ}\ \rho_Z^{-2\mu}\Big] \qquad (29) \\
\le\;& 2\,\frac{Y_2}{Y_1}\,2^{H_X+H_Z}, \qquad (30)
\end{aligned}
$$

where $H_A \equiv H^{(1/2)}(\rho_A)$, $0 \le 2\nu, 2\mu \le 1$, and the trace norm is defined by $\|X\|_1 \equiv \mathrm{tr}\,|X|$. Recall from our manuscript, that we reuse subsystem labels for Hilbert space dimensionalities, thus here Y_2/Y_1 denotes the ratio of their Hilbert space dimensions. Note that here, to go from Eq. (28) to Eq. (29), we would require $\rho_{XZ} = \rho_X \otimes \rho_Z$; and to go from Eq. (29) to Eq. (30), we would require ρ_{XYZ} is pure and we take $2\nu = 2\mu = \frac{1}{2}$.

Proof: Using the Cauchy-Schwarz inequality we may write

$$
\begin{aligned}
&\left\|\sigma^U_{XY_2Z} - \sigma^U_X \otimes \sigma^U_{Y_2Z}\right\|_1 \qquad (31)\\
\le\;& \left\|\rho_X^{\nu}\otimes \mathbb{1}_{Y_2}\otimes\rho_Z^{\mu}\right\|_2 \\
&\times\left\|\rho_X^{-\nu}\otimes\rho_Z^{-\mu}(\sigma^U_{XY_2Z} - \sigma^U_X\otimes\sigma^U_{Y_2Z})\right\|_2,
\end{aligned}
$$

where without loss of generality we may assume that ρ_X^{ν} and ρ_Z^{μ} are invertible; then using the methods already outlined in Ref. 4 the results are easily obtained. ■
We note that the statement of the result reduces to the conventional decoupling theorem for the choice $\nu = 0$ and subsystem Z is one-dimensional.

Of particular interest here is the case where $2\nu = \frac{1}{2}$ and $\rho_{\mathrm{ext},Y}$ is pure, which gives

$$
\int_{U\in U(Y)} dU\, \left\|\sigma^U_{\mathrm{ext},Y_2} - \sigma^U_{\mathrm{ext}}\otimes\sigma^U_{Y_2}\right\|_1 \le \left(2\,\frac{Y_2}{Y_1}\,2^{H_{\mathrm{ext}}}\right)^{\frac{1}{2}}, \tag{32}
$$

with $H_{\mathrm{ext}} \equiv H^{(1/2)}(\rho_{\mathrm{ext}})$.

Now $1 - F(\rho,\sigma) \le \frac{1}{2}\|\rho-\sigma\|_1$, where the fidelity is defined by $F(\rho,\sigma) \equiv \|\sqrt{\rho}\sqrt{\sigma}\|_1$. As a consequence, the fidelity with which the initial trans-event horizon entanglement is encoded within the combined ext, Y_1 subsystem is bounded below by $1-\sqrt{2^{H_{\mathrm{ext}}}Y_2/Y_1}$. Now allowing this in turn to be bounded from below by $1-2^{-c}$ and choosing $Y_1 = R$ and $Y_2 = B$ and recalling that $BR = 2^{S_{\mathrm{BH}}}$ gives the result quoted in our manuscript [Eq. (3) there].

Interestingly, the opposite choice $Y_1 = B$ and $Y_2 = R$ tells us that, for any positive c, for fewer than $\frac{1}{2}[S_{\mathrm{BH}} - H^{(1/2)}(\rho_{\mathrm{ext}})] - c$ qubits radiated away, the initial trans-event horizon entanglement remains encoded between the external neighborhood and the interior subsystems with fidelity of at least $1-2^{-c}$. This effectively gives the number of qubits that must be radiated before trans-event horizon entanglement *begins* to be reduced from its initial value. Of particular interest is the case when $H^{(1/2)}(\rho_{\mathrm{ext}}) \approx S_{\mathrm{BH}}$ for which we would conclude that the trans-event horizon entanglement begins to be depleted by radiation almost immediately.

VI. ENTANGLEMENT LOSS IN THE PRESENCE OF IN-FALLEN MATTER

The unitary evaporation of an entangled black hole in the presence of in-fallen matter was argued in our manuscript to be described by

$$
\begin{aligned}
&\frac{1}{\sqrt{K}}\sum_{i=1}^{K}|i\rangle_{\mathrm{ref}}\otimes\sum_j\sqrt{p_j}\,(|i\rangle\otimes|j\rangle\oplus 0)_{\mathrm{int}}\otimes|j\rangle_{\mathrm{ext}} \qquad (33)\\
\to\;& \frac{1}{\sqrt{K}}\sum_{i=1}^{K}|i\rangle_{\mathrm{ref}}\otimes\sum_j\sqrt{p_j}\,[U(|i\rangle\otimes|j\rangle\oplus 0)]_{BR}\otimes|j\rangle_{\mathrm{ext}},
\end{aligned}
$$

where $\log_2 K \equiv S_{\mathrm{matter}}$ is the number of qubits of quantum information in the in-fallen matter.

It is now straightforward to apply the generalized decoupling theorem above to show that, for an arbitrary positive number c, when the number of qubits radiated reaches

$$
\log_2 R = S_{\mathrm{BH}} - \frac{1}{2}\chi^{(1/2)} + c, \tag{34}
$$

then the trans-event horizon entanglement has effectively vanished and instead has been transferred to entanglement between the external neighborhood and the outgoing radiation, with a fidelity of at least $1-2^{-c}$. Recall from our manuscript that the number of unentangled qubits initially within the black hole is roughly quantified by

$$
\chi^{(q)} \equiv S_{\mathrm{BH}} - S_{\mathrm{matter}} - H^{(q)}(\rho_{\mathrm{ext}}) \ge 0, \tag{35}
$$

with q of order unity.

Repeating the argument from our manuscript, unless this occurs when $\log_2 R \approx S_{\mathrm{BH}}$ then a noticeable violation of the equivalence principle will occur. This implies that

$$
\chi^{(1/2)} \lll S_{\mathrm{BH}}, \tag{36}
$$

or equivalently, that

$$S_{\rm BH} \approx H^{(1/2)}(\rho_{\rm ext}) + S_{\rm matter} \approx H^{(1/2)}(\rho_{\rm ext}), \qquad (37)$$

since from 't Hooft's bound, the entropic content of matter is only a vanishingly small fraction of the thermodynamic entropy of the black hole, i.e., $S_{\rm matter} \lll S_{\rm BH}$.

VII. WHERE'S THE PLANCK SCALE?

In the first part of the manuscript, we show that a black hole's thermodynamic entropy must be very well approximated by its entropy of entanglement across the event horizon. The proof relied on preservation of the equivalence principle prior to the black hole having evaporated to the Planck scale. However, what defines the beginning of the Planck scale for black holes?

A universal feature of black holes is their thermodynamic entropy or (essentially up to a constant prefactor) their surface area. We shall therefore use the entropy (in bits) as a measure of size of a black hole. As we wish to avoid making claims about the physics of Planck scale black holes, we shall suppose there is some size, above which Planck scale effects are negligible (in particular, above which the predictions of the equivalence principle are left in tact). Stated conversely, we shall suppose that entry into the Planck scale regime, where effects on the equivalence principle begin to become non-negligible, occurs at some generic size (or equivalently entropy). In particular, we take this entry into the Planck scale for black holes of entropy smaller than

$$S_{\rm BH}^{\rm Planckian} \lesssim 2^p. \qquad (38)$$

It will turn out that virtually any choice for p makes no difference to our analysis since the entry-point entropy so defined will be dwarfed by those of that of typical large black holes (e.g., a stellar mass black hole has thermodynamic entropy of $10^{10^{77}}$).

To see how the argument runs, we must determine the thermodynamic entropy of a black hole evaporating according to Eq. (33). We shall suppose the von Neumann entropy computed from Eq. (33) is a good estimate for the thermodynamic entropy. Evolution corresponds to the radiation subsystem R becoming an ever larger portion of the initial black hole Hilbert space and the remaining black hole interior subsystem B becoming an ever shrinking portion, subject to the constraint that

$$2^{S_{\rm BH}} = BR, \qquad (39)$$

where one should recall that we reuse subsystem labels as their corresponding Hilbert space dimensionalities. During evaporation, a simple upper bound to the von Neumann entropy of the black hole interior $S(B)$ is given by the logarithm of its dimensionality. Hence

$$S(B) \leq S_{\rm BH} - \log_2 R. \qquad (40)$$

For a lower bound we can use the negative logarithm of the so-called purity

$$S(B) \geq -\log_2 \langle\!\langle {\rm tr}\,(\rho_B^U)^2 \rangle\!\rangle_U, \qquad (41)$$

where ρ_B^U is the reduced substate on the black hole interior of Eq. (33), and $\langle\!\langle \cdots \rangle\!\rangle_U$ denotes averaging over the random unitary U with the Haar measure. Using standard methods [4], the purity can be easily estimated in the latter stages of evaporation to be

$$\langle\!\langle {\rm tr}\,(\rho_B^U)^2 \rangle\!\rangle_U \simeq \frac{B(R^2-1)}{(BR)^2-1} \simeq \frac{R}{2^S_{\rm BH}}, \qquad (42)$$

and hence $S(B) \gtrsim S_{\rm BH} - \log_2 R$. These bounds imply that during the latter stages of evaporation via Eq. (33), a black hole's von Neumann entropy will be

$$S(B) \simeq S_{\rm BH} - \log_2 R. \qquad (43)$$

Combining Eqs. (38) and (43) we see that for a large black hole to have evaporated to just above the Planck scale (prior to any need to invoke Planck-scale physics), it must have emitted virtually all its initial entropy as Hawking radiation. In other words, as one approaches the Planck scale, one has to very high precision that

$$\log_2 R \approx S_{\rm BH}. \qquad (44)$$

VIII. INFORMATION RETRIEVAL FROM PURE-STATE BLACK HOLES

As noted in our manuscript, the description of an evaporating black hole via

$$|i\rangle_{\rm int} \to (U|i\rangle)_{\rm RB}. \qquad (45)$$

is not new. This was originally formulated [5] assuming that all the in-falling matter (and the black-hole itself) was in a pure state $|i\rangle$. In other words, it is assumed that initially there is no trans-event horizon entanglement. It should be noted, that prior to Ref. 6 this evolution was not connected to or claimed to be supported by any microscopic mechanism.

The original analysis suggested that a 'discernible information' (corresponding to the deficit of the entropy of a subsystem from its maximal value) would yield a suitable metric for information content in the radiation [5]. In order to find the "typical" behavior of an evaporating black hole it calculated the mean discernible information averaged over random unitaries [5].

Starting with a pure-state interior, the mean discernible information of the radiation remains almost zero until half the qubits of the initial black hole had been radiated, after which it rises at the rate of roughly two bits for every qubit radiated [5]. This behavior suggests that first entanglement is created, followed by dense coding [7] of *classical* information about the initial state.

In order to get a much clearer picture of quantum information flow in Eq. (45) we can rely on the decoupling theorem. In particular, entangling the state of the in-fallen matter with some distant reference (ref) subsystem, allows one to track the flow of quantum information [1, 8]. In this way Eq. (45) becomes

$$\frac{1}{\sqrt{K}}\sum_{i=1}^{K}|i\rangle_{\rm ref}\otimes|i\rangle_{\rm int}\to\frac{1}{\sqrt{K}}\sum_{i=1}^{K}|i\rangle_{\rm ref}\otimes(U|i\rangle)_{\rm BR}, \quad (46)$$

recall that $\log_2 K\equiv S_{\rm matter}$ is the number of qubits describing the quantum state of the matter used to form the otherwise pure-state black hole. Using the decoupling theorem [4] we may show that, for any positive number c, prior to $\frac{1}{2}(S_{\rm BH}-S_{\rm matter})-c$ qubits having been radiated, the quantum information about the in-fallen matter is encoded within the black hole interior with fidelity at least $1-2^{-c}$; whereas after a further $S_{\rm matter}+2c'$ qubits have been radiated, for arbitrary positive c', the information about the in-fallen matter is encoded within the radiation with fidelity at least $1-2^{-c'}$ (see also Ref. 1 for this latter result). The quantum information about the in-fallen matter naively appears to leave in a narrow 'pulse' at the radiation emission rate; this pulse occurs just as half of the black hole's qubits have radiated away.

Now consider what happens if additional matter is dumped into the black hole after its creation. Following Ref. 1, we model this process via cascaded random unitaries on the black hole interior — one unitary before each radiated qubit. (Naturally, any such analysis relies on a very short global thermalization time for the black hole. An assumption which was not needed for any of the results quoted in our manuscript itself.) Within the pure-state black hole of Eq. (46), it was argued [1] that *after* half of the initial qubits had radiated away, any information about matter subsequently falling into the black hole would be "reflected" immediately at roughly the radiation emission rate [1]. By contrast, in the early stages of evaporation information about matter subsequently thrown in would only begin to emerge after half of the initial qubits of the black hole had radiated away [1]. These very different *behaviors* in the first and second halves of its life suggest that such a black hole acts almost as two different species: as storage during the first half of its radiated qubits and as a reflector during the second half.

A subtle flaw to this argument of Ref. 1 is due to the omission of the fact that a black hole's entropy is non-extensive, e.g., scaling as the square of the black hole's mass M^2 for the Schwarzshild family of black holes: for every k qubits dumped into such a black hole, the entropy typically increases by $O(kM)\gg k$. Likewise, the number of unentangled qubits within the (initially pure-state) black hole will increase by $O(kM)$. Therefore, within the cascaded unitary pure-state black hole, the reflection described in Ref. 1 would not begin immediately, but only after a large delay in time of $O(kM^2)$. Notwithstanding the delay, the pure-state black hole behaves effectively as two distinct species as described above.

Because this behavior seems so bizarre it is worth going back over the key assumptions that went into it: i) that the behavior of a specific unitary in Eq. (45) is well described by Haar averages over all random unities; ii) that the number of qubits comprising the initial black hole Hilbert space is $n\simeq S_{\rm BH}$. (These two assumptions are discussed in some detail in our manuscript.) Finally, iii) that the black hole is *initially* in a pure state up to a negligible amount of entanglement that may come from the matter content. In fact, it is this last assumption which is weakest and at odds with the well known quantum physics of condensed matter systems and rigorous results from axiomatic field theory as discussed in our manuscript.

IX. INFORMATION RETRIEVAL FROM AN ENTANGLED BLACK HOLE

Here we give explicit statements of results from decoupling summarized in our manuscript.

Applying the decoupling theorem [4] to the entangled-state black hole of Eq. (33) allows us to show that, for any positive number c, for all but the final $S_{\rm matter}+\frac{1}{2}\chi^{(2)}+c$ qubits radiated, the information about the in-fallen matter is encoded in the combined space of the external neighborhood and black hole interior with fidelity at least $1-2^{-c}$. Similarly, for any positive c', for all but the initial $S_{\rm matter}+\frac{1}{2}\chi^{(2)}+c'$ qubits radiated, this information is encoded in the combined radiation and external neighborhood subsystems with fidelity at least $1-2^{-c'}$. In addition, at all times this information is encoded with unit fidelity within the joint radiation and interior subsystems.

In other words, between the initial and final roughly $S_{\rm matter}+\frac{1}{2}\chi^{(2)}$ qubits radiated, the information about the in-fallen matter is effectively *deleted* from each individual subsystem [8, 9], instead being encoded in any two of the three of subsystems (consisting of the out-going radiation, the external neighborhood, and the black hole interior). During this time, the information about the in-fallen matter is to an excellent approximation encoded within the perfect correlations of a *quantum one-time pad* [8, 10] of these three subsystems.

Furthermore, using our generalized decoupling theorem we may show that, for any positive c'', that prior to the first $\frac{1}{2}\chi^{(1/2)}-c''$ qubits radiated, the information about the in-fallen matter is still encoded solely within the black hole interior, with a fidelity of at least $1-2^{-c''}$. Similarly, for any positive c''', within the final $\frac{1}{2}\chi^{(1/2)}-c'''$ qubits radiated, the information about the in-fallen matter is encoded within the out-going radiation, with a fidelity of at least $1-2^{-c'''}$. Combining these with the above results we see that both the encoding and decoding of the tripartite quantum one-time pad occur during the radiation of roughly $S_{\rm matter} + \frac{1}{2}(\chi^{(2)} - \chi^{(1/2)}) \simeq S_{\rm matter}$ qubits, i.e., the black hole's quantum one-time pad encoding (and decoding) occurs at roughly the radiation emission rate.

How does this entangled-state description of black hole evaporation respond to matter subsequently swallowed after its formation? Instead of the two distinct behaviors of storage and reflection found in the pure-state black hole, here, any additional qubits thrown in will immediately begin to be encoded into the tripartite one-time pad. The decoding into the radiation subsystem of the information about *all* the in-fallen matter will only occur at the very end of the evaporation. (The non-extensive increase in black hole entropy is taken up as entanglement with the external neighborhood so no further delays occur.) Thus, instead of behaving almost as two distinct species, a highly entangled-state black hole has one principle behavior — forming a tripartite quantum one-time pad between the black hole interior, the external neighborhood and the radiation from the black hole, with release of that information only at the end of the evaporation.

Can we reconcile the information retrieval behavior of the pure-state black hole with its entangled counterpart? Naively, if the pure-state black hole analysis were run on twice as many qubits, but stopped just after the information about the in-fallen matter had escaped as a narrow pulse then there would be broad agreement between the two types of black hole. This doubling of the number of qubits would make some crude sense if we supposed that the pure-state black hole was not making a split between interior and exterior at the event horizon, but somewhat further out at some arbitrary boundary where trans-boundary entanglement would not be participating in the evaporation. The dimensionality of the Hilbert space within this extended boundary would then be dominated by the product of the dimensionality of the original black hole interior, and the nearby external neighborhood entangled with them. This would be roughly twice the number of qubits within the black hole interior itself. Once the original number of qubits had evaporated away (now half the total for our extended boundary pure-state black hole) the black hole interior would be exhausted of Hilbert space and evaporation would cease. This suggests that despite the general incompatibility between the two types of black hole, a pure-state analysis, if thoughtfully set up, could capture important features of information retrieval from an entangled-state black hole.

X. HEURISTIC FLOW VIA CORRELATIONS

The rigorous results from our manuscript may be heuristically visualized by following how the correlations with the distant reference system behave. For a pure tripartite state XYZ, these correlations satisfy

$$C(X\!:\!Y) + C(X\!:\!Z) = S(X), \tag{47}$$

Here $S(X)$ is the von Neumann entropy for subsystem X and $C(X:Y) \equiv \frac{1}{2}[S(X) + S(Y) - S(X,Y)]$, one-half the quantum mutual information, is a measure of correlations between subsystems X and Y. Relation (47) is additive for a pure tripartite state, so the correlations with subsystem X smoothly move from subsystems Y to Z and vice-versa.

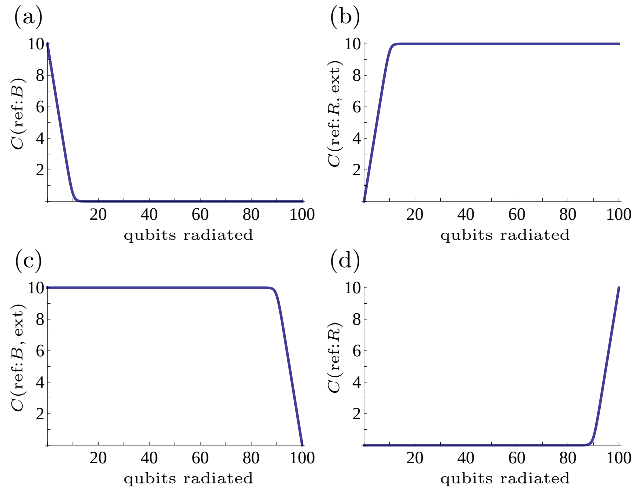

FIG. 1: Correlations to the reference subsystem as a function of the number of qubits radiated ($\log_2 R$). Correlations between the reference (ref) subsystem and: (a) black hole interior, B; (b) radiation, R, and external (ext) neighborhood; (c) black hole interior and external neighborhood; and (d) radiation alone. Note that, as expected from Eq. (47), the sum of C's in subplots (a) and (b) is a constant, as is that of subplots (c) and (d). In each subplot, the in-fallen matter consists of $S_{\rm matter} = 10$ qubits and the black hole initially consists of $\log_2 BR = 100$ qubits with $\chi^{(q)} = 0$. (Entropies are evaluated using base-two logarithms.)

For simplicity, here we restrict ourselves to the case where

$$\rho_{\rm ext} = \frac{1}{M}\sum_{j=1}^{M} |j\rangle_{\rm ext\ ext}\langle j|, \tag{48}$$

and where we assume no excess unentangled qubits, i.e., $\chi^{(q)} = 0$. Thus, the initial number of qubits within the black hole interior is given by $\log_2 N = \log_2(BR) = S_{\rm matter} + \log_2 M$, for $S_{\rm matter}$ qubits of in-fallen matter. We computed the above measure of correlations, Eq. (47), from von Neumann entropies approximated using the average purity (see next section); numerical calculations showed this as a good approximation for systems of even a few qubits. Fig. 1 shows a typical scenario: A black hole is assumed to be created from in-fallen matter comprising $S_{\rm matter}$ qubits of information and negligible excess unentangled qubits. Within the first $S_{\rm matter}$ qubits radiated, information about the in-fallen matter (a) vanishes from the black hole interior at roughly the radiation emission rate and (b) appears in the joint radiation and external neighborhood subsystem. From then until just before the final $S_{\rm matter}$ qubits are radiated, the in-fallen matter's information is encoded in a tripartite state, involving the radiation, external neighborhood and interior subsystems, subplots (b) and (c). In the final $S_{\rm matter}$ qubits radiated the information about the in-fallen matter is released from its correlations and appears in the radiation subsystem alone, subplot (d). This qualitative picture is in excellent agreement with the results from the decoupling theorem and its generalization.

Evaluation of purities

In order to approximate the computation of the correlation measure described above, we use a lower bound for a subsystem with density matrix ρ

$$\langle\langle S(\rho)\rangle\rangle \geq -\langle\langle \log_2 p(\rho)\rangle\rangle \geq -\log_2\langle\langle p(\rho)\rangle\rangle. \qquad (49)$$

Here $S(\rho) = -\mathrm{tr}\,\rho\log_2\rho$ is the von Neumann entropy of ρ, $p(\rho) = \mathrm{tr}\,\rho^2$ is its purity, and here $\langle\langle\cdots\rangle\rangle$ denotes averaging over random unitaries with the Haar measure. The former inequality above is a consequence of the fact that the Rényi entropy is a non-increasing function of its argument [11], and the latter follows from the concavity of the logarithm and Jensen's inequality. We may estimate the von Neumann entropies required then by the rather crude approximation $\langle\langle S(\rho)\rangle\rangle \approx -\log_2\langle\langle\, p(\rho)\rangle\rangle$, which turns out to be quite reasonable for spaces with even a few qubits.

Although traditional methods [12] may be used to compute these purities, a much simpler approach is to use the approach from Ref. 4. In particular, for a typical purity of interest we use the following decomposition

$$\begin{aligned}\mathrm{tr}\,\sigma^{U\,2}_{R,\mathrm{ext}} &= \mathrm{tr}\big(\sigma^U_{R,\mathrm{ext}}\otimes\sigma^U_{R',\mathrm{ext}'}\,\mathcal{S}_{R,\mathrm{ext};R',\mathrm{ext}'}\big) \qquad (50)\\ &= \mathrm{tr}\big(\rho_{\mathrm{ref},BR,\mathrm{ext}}\otimes\rho_{\mathrm{ref}',B'R',\mathrm{ext}'}\\ &\quad\times U^\dagger_{BR}\otimes U^\dagger_{B'R'}\,\mathcal{S}_{R;R'}\,U_{BR}\otimes U_{B'R'}\,\mathcal{S}_{\mathrm{ext};\mathrm{ext}'}\big)\end{aligned}$$

where $\mathcal{S}_{A;A'}$ is the swap operator between subsystems A and A', similarly, $\mathcal{S}_{AB;A'B'} = \mathcal{S}_{A;A'}\mathcal{S}_{B;B'}$. Then the average over the Haar measure is accomplished by an application of Schur's lemma [4]

$$\begin{aligned}&\langle\langle U^\dagger_A\otimes U^\dagger_{A'}\,\mathcal{S}_{A_2;A_2'}\,U_A\otimes U_{A'}\rangle\rangle\\ &= \frac{A_2(A_1^2-1)}{A^2-1}\,\mathbb{1}_{A;A'} + \frac{A_1(A_2^2-1)}{A^2-1}\,\mathcal{S}_{A;A'}. \qquad (51)\end{aligned}$$

This approach allows us to straight-forwardly compute the required purities as

$$\begin{aligned}p(\mathrm{ref}) &= \frac{1}{K},\quad p(\mathrm{ext}) = \frac{1}{N},\quad p(\mathrm{ref,ext}) = \frac{1}{KN},\\ p(R) &= \frac{1}{(BR)^2-1}\Big(R(B^2-1) + \frac{B(R^2-1)}{KN}\Big), \qquad (52)\\ p(R,\mathrm{ext}) &= \frac{1}{(BR)^2-1}\Big(\frac{R(B^2-1)}{N} + \frac{B(R^2-1)}{K}\Big),\end{aligned}$$

with $p(B,\mathrm{ext})$ and $p(B,\mathrm{ext})$ given by the above expressions under the exchange $R\leftrightarrow B$, similarly the exchange $K\leftrightarrow N$ gives us expressions for $p(\mathrm{ref},R)$, etc.

XI. BLACK HOLES VERSUS LUMPS OF COAL

Bekenstein [13] tells of a thought experiment he attributes to Sidney Coleman: A cold piece of coal (initially in its ground state) is illuminated by a laser beam. The system is thus prepared in a pure state and radiates thermally after the laser is switched off. Eventually the lump of coal returns to its initial state, so presumably the radiation subsystem has merely encoded any information in the subtle correlations between the individual thermal photons. Can this differ from the overall behavior of a unitarily evaporating black hole?

Such a hot coal model of a black hole will correspond very closely to the pure-state model of a black hole. As such, information about the state of the laser beam that has heated up the coal will become accessible from the radiation field shortly after half of the total number of thermal photons (each carrying roughly one bit's worth of information) have radiated away. This behavior, however, will be very different from the entangled black hole analyzed in our manuscript. For such highly entangled black holes there is another component of the system to include in the dynamics: The entanglement across the boundary corresponding to the event horizon. This entanglement is not merely static as it would be across a fixed boundary, but must itself escape from the black hole in order for the boundary itself to shrink. As was uncovered in our manuscript, entangled black holes encode the information about the in-fallen matter into a quantum one-time pad. The information is in principle accessible from any two of three subsystems (the interior of the black hole, the modes just external to the black hole but entangled with it across the event horizon and the Hawking radiation itself) within a very short time after the black hole begins to radiate. Once that encoding into the quantum one-time pad has occurred, this information becomes inaccessible from any one of these subsystems alone (and in particular from the Hawking radiation).

Only when the quantum one-time pad becomes decoded will the full information become accessible within the Hawking radiation. For a highly entangled black hole, as shown in our manuscript, this occurs within the final and vanishingly small fraction of the black hole's lifetime. Before this time, the Hawking radiation is completely uncorrelated from the information about the in-fallen matter. This behavior is therefore very different from that of information return from a hot coal.

The authors gratefully acknowledge H.-J. Sommers's original calculation of Eq. (52) and several fruitful discussions with him, Netta Cohen and Manas Patra.

[1] P. Hayden and J. Preskill, JHEP **2007**(09), 120 (2007).
[2] L. Susskind, L. Thorlacius and J. Uglum, Phys. Rev. D **48**, 3743 (1993).
[3] M. Srednicki, Phys. Rev. Lett. **71**, 666 (1993).
[4] A. Abeyesinghe, I. Devetak, P. Hayden and A. Winter, Proc. R. Soc. A **465**, 2537 (2009).
[5] D. N. Page, Phys. Rev. Lett. **71**, 3743 (1993).
[6] S. L. Braunstein and M. K. Patra, Phys. Rev. Lett. **107**, 071302 (2011).
[7] C. H. Bennett and S. J. Wiesner, Phys. Rev. Lett. **69**, 2881 (1992).
[8] S. L. Braunstein and A. K. Pati, Phys. Rev. Lett. **98**, 080502 (2007).
[9] D. Kretschmann, D. Schlingemann, and R. F. Werner, J. Funct. Analysis, **255**, 1889 (2008).
[10] D. W. Leung, Quantum Inf. Comput. **2**, 14 (2001).
[11] I. Bengtsson and K. Życzkowski, Geometry of Quantum States: An Introduction to Quantum Entanglement. (Cambridge University Press, Cambridge, 2006).
[12] P. A. Mello, J. Phys. A **23**, 4061 (1990).
[13] J. D. Bekenstein, Contemp. Phys. **45**, 31 (2004).