Published for SISSA by Springer

Received: December 7, 2023
Accepted: May 27, 2024
Published: July 3, 2024

Page curves and replica wormholes from random dynamics

Jan de Boer[a], **Jildou Hollander**[a] **and Andrew Rolph**[a,b]

[a]*Institute for Theoretical Physics, University of Amsterdam, PO Box 94485, 1090 GL Amsterdam, The Netherlands*
[b]*Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels, Belgium*

E-mail: j.deboer@uva.nl, j.s.hollander@uva.nl, andrew.d.rolph@gmail.com

Abstract: We show how to capture both the non-unitary Page curve and replica wormhole-like contributions that restore unitarity in a toy quantum system with random dynamics. The motivation is to find the simplest dynamical model that captures this aspect of gravitational physics. In our model, we evolve with an ensemble of Hamiltonians with GUE statistics within microcanonical windows. The entropy of the averaged state gives the non-unitary curve, the averaged entropy gives the unitary curve, and the difference comes from matrix index contractions in the Haar averaging that connect the density matrices in a replica wormhole-like manner.

Keywords: AdS-CFT Correspondence, Black Holes, Random Systems

ArXiv ePrint: 2311.07655

https://doi.org/10.1007/JHEP07(2024)023

JHEP07(2024)023

Contents

1 Introduction

The black hole information problem is a tension between unitarity in quantum gravity and the insensitivity of Hawking radiation to initial conditions [1, 2]. Pure state black holes evaporate away to mixed-state thermal radiation. The von Neumann entropy of black hole radiation is sensitive to this tension because Hawking radiation is thermal and evaporation implies that the entropy increases monotonically, while unitarity requires it to return to zero [3–5]. A resolution to the problem, as far as calculating a unitary Page curve, recognises that the exact state and the semiclassical state prepared by the dominant gravitational saddle are not the same, and in particular, their entropies can be quite different. This was first seen in holographic systems through the quantum extremal surface formula [6–8], which can calculate the entropy of the exact state in the boundary CFT [9–13], then generalised to non-holographic systems with the island formula which can be derived by including novel connected gravitational saddles in the replica trick called replica wormholes [14, 15].

In a way, AdS/CFT itself is a resolution to the black hole information paradox. For small AdS black holes, the information paradox is the prediction that a pure state black hole will evolve to a mixed state thermal gas, and AdS/CFT is a resolution in the sense that the

JHEP07(2024)023

dual description of evaporation is manifestly unitary, so the von Neumann entropy is simply zero at all times.[1] This resolution is unsatisfying, but why? This is because it is not clear how unitarity is restored in the bulk. In principle, to know everything about AdS black hole evaporation, one can simply take a high energy CFT state, let it evolve, and then probe it. In practice, the state is enormously complex and it is impossible to semiclassically determine microscopic details, for example, how the exact dynamics of the late time radiation purifies the early radiation. The same objection holds for the island formula: it gives us unitary Page curves, but it is not obvious how the Euclidean gravitational path integral from which it is derived tells us anything about exactly how the late time radiation purifies the early radiation.

To make progress towards this fuller resolution, let's ask: what, for an evaporating AdS black hole, are the necessary and sufficient conditions on the holographic dual to get a unitary Page curve?[2] Requiring exact unitarity in a quantum system with Hilbert space dimension $|\mathcal{H}|$ imposes $2|\mathcal{H}|^2$ equality constraints on the matrix elements of the Hamiltonian, whereas a unitary Page curve is a single curve, so exact unitary is presumably overkill. Does only a subsector of the theory need to be unitary and, if so, which? Can we replace unitary time evolution with a time-dependent completely positive and trace-preserving (CPTP) map and still get a unitary Page curve? Can exact properties of the Hamiltonian be replaced with statistical properties and, if so, which? For this last point, there is a hint that this is true, because (1) replica wormholes give unitary Page curves, (2) replica wormholes are gravitational saddles, and (3) AdS gravity, without stringy corrections, is conjectured to be best described by a statistical ensemble (along the lines developed in [16–21], see also [22]). Answering these questions will tell us which parts of the non-gravitational dual are responsible for the late time purification of Hawking radiation, and which are irrelevant.

In this paper, to deepen our understanding of the underlying mechanics of unitary black hole evaporation, we show how to capture both the non-unitary Page curve and replica wormhole-like contributions that restore unitarity in a toy quantum system with random dynamics. Semiclassical gravity gives us two curves for the radiation entanglement entropy: the Hawking curve and, after including replica wormholes, the unitary Page curve. Our claim is that chaotic dynamics is, with a few other basic assumptions, sufficient to capture both. If semiclassical gravity captures only the statistical properties of the exact UV theory, then those statistical properties alone should be sufficient. As far as these two curves are concerned, neither complex gravitational systems nor holography are needed.

Our model has a factorised Hilbert space of the system $\mathcal{H}_{\mathsf{A}}$ and its environment $\mathcal{H}_{\mathsf{B}}$. One can imagine the system representing a gravitational system with a black hole, or a lab with a piece of coal. Within the full spectrum, we consider a high-energy microcanonical window, which we take to spanned by N product states, one of which has the energy in the system, so that the environment is in its vacuum state,

$$\mathrm{Tr}_{\mathsf{A}}(|\psi_1\rangle\langle\psi_1|) = |0\rangle\langle 0|_{\mathsf{B}}\,. \tag{1.1}$$

[1] The paradox is cleanest in this set-up, but we can also couple the boundary CFT to a radiation bath to get the standard non-trivial Page curve.

[2] The Page curve in this context is the renormalised von Neumann entropy of the whole boundary CFT for a small isolated AdS black hole, or of the bath if the holographic CFT is coupled to one.

JHEP07(2024)023

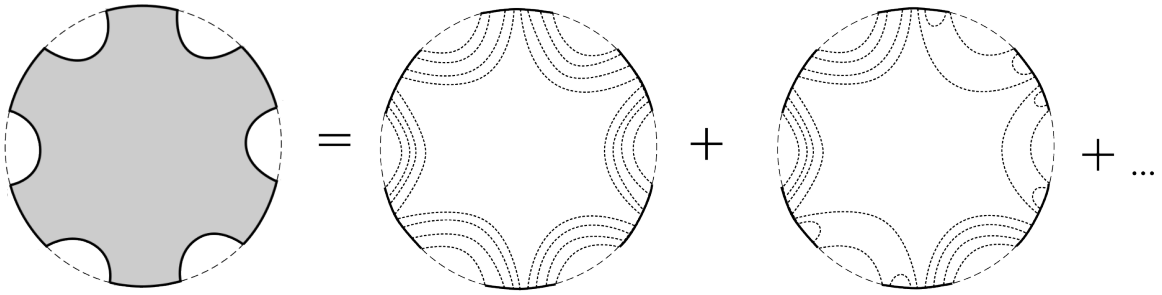


Figure 1. In our model, replica wormholes, in the sense of contributions to the Rényi entropies that connect the density matrices and restore unitarity, come from chaotic dynamics through the Haar averaging. The density matrices are quartic functions of the Hamiltonian-diagonalising unitary matrix, and taking the Haar average of the Rényi entropy is what connects the density matrices. The dashed lines represent unitary matrix index contractions.

The other $(N-1)$ states have all the energy in the environment and none in the system:

$$\mathrm{Tr}_{\mathsf{B}}(|\psi_2\rangle\langle\psi_2|) = \cdots = \mathrm{Tr}_{\mathsf{B}}(|\psi_N\rangle\langle\psi_N|) = |0\rangle\langle 0|_{\mathsf{A}}\,. \tag{1.2}$$

These represent the radiation states and they are entropically favoured within the micro-canonical ensemble. We take the initial state to be the special state $|\psi_1\rangle$, with the energy in the system, and this state represents the black hole or piece of coal.

A bit more informally, our Hilbert space therefore looks like

$$\mathcal{H} = \big(|\text{black hole}\rangle_{\mathsf{A}} \otimes |\text{empty}\rangle_{\mathsf{B}}\big) \oplus \big(|\text{empty}\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\text{rad},\mathsf{B}}\big), \tag{1.3}$$

with $\mathcal{H}_{\text{rad},\mathsf{B}}$ the $(N-1)$-dimensional space of radiation states in the B-system.

The matrix elements of the Hamiltonian within this narrow energy band, $H_{ij} = \langle\psi_i|\,H\,|\psi_j\rangle$, are taken to be random with GUE matrix statistics, with probability measure

$$\mu(H_{ij}) \propto \exp\left(-\frac{N}{2}|H_{ij}|^2\right). \tag{1.4}$$

Evolution with this random Hamiltonian generates a time-dependent distribution of pure states. For a given draw of a random Hamiltonian, the reduced density matrices of the system and environment, and their Rényi entropies, are functions of the Hamiltonian's diagonalising unitary matrix and its eigenvalues. Averaging over the unitary matrices, with the Haar measure, connects matrix indices. The Rényi entropies have contributions that connect the density matrices and disconnected components. In figure 1 we diagrammatically represent some of the matrix contractions from the Haar averaging. As for replica wormholes, the contributions that restore unitarity at late times connect the density matrices.

The result for the n'th Rényi entropy of the averaged environment density matrix ρ_{B}, in the large N approximation, is

$$S^{(n)}\left(\overline{\rho_{\mathsf{B}}}(t)\right) = \frac{1}{1-n}\log\left(\frac{1}{N^{n-1}}\left(1-\overline{g}(t)\right)^n + \overline{g}^n(t)\right), \tag{1.5}$$

JHEP07(2024)023

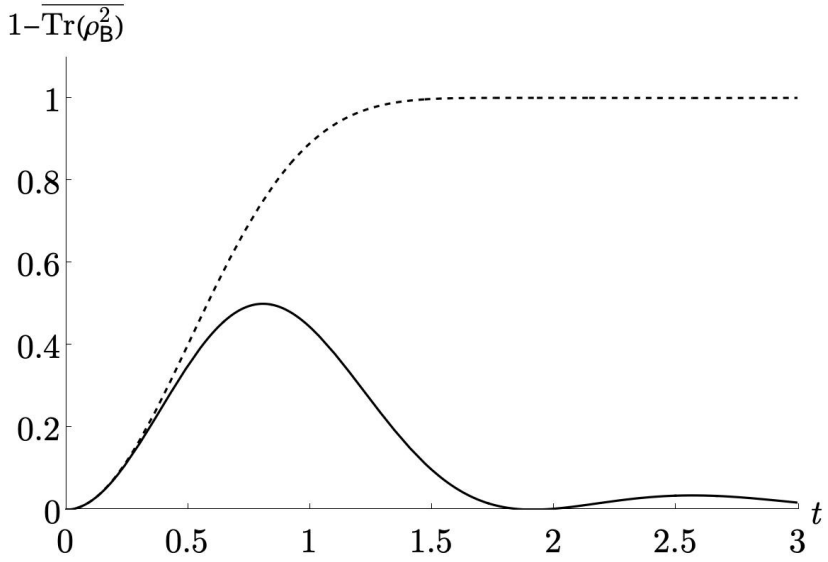


Figure 2. The averaged purity of the radiation in our model with GUE-random Hamiltonians. The dashed line is the result that neglects connected replica wormhole-like contributions from the Haar averaging, and shows non-unitary evolution, while the solid line does take into account these connected contributions, and shows unitary evolution.

where $\overline{g}(t)$ is the averaged normalised spectral form factor, while the averaged Rényi entropy is[3,4]

$$\overline{S^{(n)}(\rho_{\mathsf{B}}(t))} = \frac{1}{1-n} \log\left((1-\overline{g}(t))^n + \overline{g}^n(t)\right). \tag{1.6}$$

In figure 2 we plot the purity of the radiation density matrix. We calculate the von Neumann entropies by analytically continuing our results for the Rényi entropies and we find both the unitary and non-unitary Page curves.

While there are many features of black hole evaporation that our simple model does not capture, our goal is only to capture one particular aspect: the qualitative shape of the unitary and non-unitary Page curves. Our model captures this, and the difference between the curves comes through the Haar averaging of the random dynamics from the subset of matrix index contractions that connects the density matrices in a replica wormhole-like way.

It is perhaps worth emphasizing two key aspects of our model. The first is that one needs to have a mechanism which guarantees that most of the information will be transferred to the radiation system B. In many examples including our toy model, this follows from fairly straightforward entropy and energy considerations as we review in section 2. The second key aspect is that the random dynamics maps an initial pure state to a *classical statistical* mixture of pure states at later times. Computations which involve only one replica of the system cannot distinguish a classical statistical mixture of pure states from a corresponding single

[3]Here we take the average before taking the logarithm. We thus compute the annealed instead of the quenched averaged Rényi entropies.

[4]We implicitly assume that $\overline{g^n}(t) = \overline{g}(t)^n$. We will come back to the validity of this assumption, but expect it to hold for time scales relevant to us.

JHEP07(2024)023

mixed state as both yield the same expectation values. However, replica computations can tell the difference between the two. In particular, the expectation value of the purity is one for classical statistical mixtures of pure states, while it is smaller than one for mixed states. The ability of replica computations to distinguish classical mixtures of pure states from mixed states is what allows one to recover a Page curve using only semi-classical computations. Note that this does not imply that semi-classical computations predict the precise final state, for this one would need the actual UV completion of the system.

Alternatively, one can also view the success of semi-classical replica computations in producing the correct Page curve as a confirmation of the validity of the statistical interpretation of semi-classical gravity.

The plan for the remainder of this paper is as follows. In section 2, we detail the set-up of our toy model of black hole evaporation, and calculate the density matrices, the Rényi entropies, and von Neumann entropies, with and without averaging with respect to the random Hamiltonian. In section 3, we discuss the features of black hole evaporation that our model does not capture, how the model could be modified and extended, and further applications. In section 4, we discuss the results and conceptual takeaways and give possible directions for future research.

2 Toy model of black hole evaporation

After a black hole has evaporated, or a piece of coal burned away, we expect the final state of the radiation to be pure. Why? To have a model where the unitary Page curve returns to zero, we need to understand this.

Suppose we have a system A coupled to its environment B. At late times, for the environment's reduced state to be pure, it is sufficient if [23]:

1. The final state of the system is a ground state.
2. The ground state of the system is non-degenerate.
3. The evolution of the system and environment is unitary.
4. The initial state of the system and environment is pure.

Conditions (1) and (2) imply that the system's final state is pure. Conditions (3) and (4) imply that the final state of the system and environment is pure. Together they imply that the final state of the environment is pure.

Let us assume conditions (2)-(4). For a given initial state, condition (1) is satisfied if energy is conserved and all of it leaves the system to the environment. Whether this happens depends on the dynamics, and whether microstates where the energy has left the system are entropically favoured. It also requires the system's spectrum to be gapped, because if there is a continuous spectrum above the ground state for example, then the energy of the system can be arbitrarily low, while the entropy could be arbitrarily high. Assuming energy conservation and ergodic evolution, the long-time average of the energy in A will equal the ensemble average. If nearly all the microstates in the ensemble have zero energy in A, then the ensemble average of the energy in A will be small. If the ensemble average is much smaller than the gap

JHEP07(2024)023

between the (non-degenerate) ground state and the first excited state in A, then the long-time average of the reduced density matrix on A will be very close to the pure ground state.

A somewhat more quantitative description is as follows. Suppose we have a pure state $|\psi\rangle$ in $\mathcal{H}_\mathsf{A} \otimes \mathcal{H}_\mathsf{B}$ where the Hilbert spaces have dimensions d_A and d_B with $d_\mathsf{A} \ll d_\mathsf{B}$. The entanglement entropy between A and B is then bounded by $\log d_\mathsf{A}$. Now suppose that there is an operator H_A acting on the A-system with lowest eigenvalue E_0 and next lowest eigenvalue $E_0 + E_{\text{gap}}$. If the expectation value of H_A in the state $|\psi\rangle$ (i.e. $\text{Tr}(\rho_\mathsf{A} H_\mathsf{A})$) equals E and $\epsilon = (E - E_0)/E_{\text{gap}}$ is much smaller than one, then the entanglement entropy for the state $|\psi\rangle$ is bounded by

$$S_\mathsf{A} \leq -(1-\epsilon)\log(1-\epsilon) - \epsilon \log \epsilon + \epsilon \log(d_\mathsf{A} - 1). \tag{2.1}$$

This is much smaller than one as long as $\epsilon \ll 1/\log(d_\mathsf{A} - 1)$ and the reduced density matrix is close to that of a pure state. If we view E as the energy in the A-system and we equipartition the total energy E_{tot} of an initial state over the A and B-system then the condition $\epsilon \ll 1/\log(d_\mathsf{A} - 1)$ becomes

$$\frac{\frac{d_\mathsf{A}}{d_\mathsf{A}+d_\mathsf{B}} E_{\text{tot}} - E_0}{E_{\text{gap}}} \ll \frac{1}{\log(d_\mathsf{A} - 1)}\,, \tag{2.2}$$

which can always be achieved by making $d_\mathsf{B}/d_\mathsf{A}$ sufficiently large.

The takeaway is that we want our model to have energy conservation, ergodicity, a gapped spectrum, and entropic dominance of microstates where the energy has left the system.

2.1 Set-up

Here we describe the set-up for our toy model. We want the model to be as simple as possible, but still capable of dynamically capturing the following semiclassical gravitational predictions qualitatively: (1) the mixed Hawking radiation state and the non-unitary Page curve, (2) replica wormholes, and (3) the unitary Page curve.

First, we describe the structure of the Hilbert space and the space of states available under time evolution. The full Hilbert space $\mathcal{H}$ is a tensor product of system and environment Hilbert spaces:

$$\mathcal{H} = \mathcal{H}_\mathsf{A} \otimes \mathcal{H}_\mathsf{B}. \tag{2.3}$$

In general, we can decompose Hilbert spaces into direct sums over mutually orthogonal eigenspaces of any compact self-adjoint operator, by the spectral theorem. Let us decompose our $\mathcal{H}$ into a direct sum of microcanonical Hilbert subspaces with energies binned into narrow energy windows:

$$\mathcal{H} = \bigoplus_E \mathcal{H}_{\text{micro};E\pm\delta E}. \tag{2.4}$$

Each $\mathcal{H}_{\text{micro};E\pm\delta E}$ is a microcanonical Hilbert subspace spanned by the energy eigenstates with energies in the range $[E - \delta E, E + \delta E]$.

We are decomposing the Hilbert space this way to allow us to account for energy conservation and because it is a reasonable assumption that our initial state, if it represents

JHEP07(2024)023

a black hole microstate, has support only within a high energy microcanonical subspace $\mathcal{H}_{\text{micro};E\pm\delta E}$. Since time evolution cannot take the states out of this subspace, if energy is conserved, only this subspace is important. This is a feature of our model that differentiates it from those which treat black hole evaporation as a random unitary on $\mathcal{H}_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B}}$ [3, 4], which does not conserve energy.

If the system and environment are weakly interacting, then product states $|E - E'\rangle_{\mathsf{A}} \otimes |E'\rangle_{\mathsf{B}}$, which are eigenstates of the decoupled Hamiltonian with energy E, will be approximate eigenstates of the coupled Hamiltonian, meaning that they are superpositions of states in $\mathcal{H}_{\text{micro};E\pm\delta E}$. Then, it is a reasonable approximation that $\mathcal{H}_{\text{micro};E\pm\delta E}$ is also spanned by product states like $|E - E'\rangle_{\mathsf{A}} \otimes |E'\rangle_{\mathsf{B}}$, so that

$$\mathcal{H}_{\text{micro};E\pm\delta E} \simeq \bigoplus_{E'} \left(\mathcal{H}_{\mathsf{A};E-E'} \otimes \mathcal{H}_{\mathsf{B};E'}\right). \tag{2.5}$$

A state in $\mathcal{H}_{\mathsf{A};E-E'} \otimes \mathcal{H}_{\mathsf{B};E'}$ is interpreted in our model as a partially evaporated black hole, or burning piece of coal, that had initial energy E and has lost energy E' to the environment. To give a reasonable ansatz for the sizes of the Hilbert spaces, for a small evaporating AdS_4 black hole, $\log|\mathcal{H}_{\mathsf{A};E}| = S_{\text{BH}}(E) \sim (El_p)^2$ and $\log|\mathcal{H}_{\mathsf{B};E}| = S_{\text{rad}}(E) \sim (El_{\text{AdS}})^{3/4}$.

Recall that for the system's final state to be pure, the essential ingredients are: no ground state degeneracy, a gap in the spectrum, and entropic dominance of microstates where all the energy is in the environment. To that end, and to further simplify the space of states that are energetically accessible, we assume that there is one black hole microstate in $\mathcal{H}_{\text{micro};E\pm\delta E}$, and $N-1$ energetically accessible radiation states:

$$\mathcal{H}_{\text{micro};E\pm\delta E} = (|E\rangle_{\mathsf{A}} \otimes |0\rangle_{\mathsf{B}}) \oplus (|0\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B};E}). \tag{2.6}$$

We take the microcanonical entropy to be large,

$$e^{S_{\text{micro}}(E)} := |\mathcal{H}_{\text{micro};E\pm\delta E}| = N \gg 1, \tag{2.7}$$

and all but one of the microstates have the energy in the environment, $|\mathcal{H}_{\mathsf{B};E}| = N - 1$.

Our goal is to capture the qualitative rather than quantitative features of radiation density matrices and Page curves in the simplest toy model. While (2.6) has no partially-evaporated states and only one black hole microstate, we will show that it is sufficient to capture the qualitative features.

We choose an orthonormal basis $|i\rangle_{\mathsf{B}}$ for states in $\mathcal{H}_{\mathsf{B};E}$, with $i = 2, \ldots, N$, which in turn fixes an orthonormal basis for $\mathcal{H}_{\text{micro};E\pm\delta E}$,

$$|\psi_i\rangle := \begin{cases} |E\rangle_{\mathsf{A}}\,|0\rangle_{\mathsf{B}} & \text{for } i = 1, \\ |0\rangle_{\mathsf{A}}\,|i\rangle_{\mathsf{B}} & \text{for } i = 2, \ldots, N. \end{cases} \tag{2.8}$$

with $\langle\psi_i|\psi_j\rangle = \delta_{i,j}$.

Our initial state is $|\psi_1\rangle$. It is a non-stationary state within a particular microcanonical window $\mathcal{H}_{\text{micro};E\pm\delta E}$. It is an atypical state because all the energy is within the A subsystem, and its time evolution is towards typical microstates where all the energy is within the environment.

JHEP07(2024)023

Only the projection of the Hamiltonian onto the microcanonical window is important for the time evolution of $|\psi_1\rangle \in \mathcal{H}_{\text{micro},E\pm\delta E}$. By energy conservation, time evolution cannot take us out of this energy window, which implies that the projected Hamiltonian satisfies

$$H' := PHP = HP, \tag{2.9}$$

where P is the projection operator

$$P := \sum_{i=1}^{N} |\psi_i\rangle \langle\psi_i| . \tag{2.10}$$

As claimed, the time evolution only depends on the projected Hamiltonian:

$$\rho(t) := e^{iHt} |\psi_1\rangle \langle\psi_1| e^{-iHt} = e^{iH't} |\psi_1\rangle \langle\psi_1| e^{-iH't}. \tag{2.11}$$

H' still has a large diagonal component, $H' \approx E \cdot \mathbb{1}$, which we subtract off as it does not affect $\rho(t)$. The off-diagonal matrix elements of H' come from the coupling between A and B and vanish if the coupling is turned off. For small but finite coupling, for a many-body system, we cannot generally solve to find the exact eigenvalues of H' or the unitary matrix that diagonalises $\langle\psi_i| H' |\psi_j\rangle$. We do however expect $\langle\psi_i| H' |\psi_j\rangle$ to look like a $N \times N$ random matrix and its statistics to be predicted by random matrix theory (RMT).

Chaotic quantum many-body systems exhibit a strong form of universality [24–28]. RMT captures certain universal aspects of chaotic systems, such as the statistics of energy levels and matrix elements of observables [29–32]. The spectral statistics within sufficiently narrow energy bands of quantum systems with classically chaotic counterparts are believed to be captured by RMT [33]. This assumption of RMT behaviour within narrow energy bands is a weaker assumption than ETH [28]. High energy eigenstates in chaotic theories are of the form $\sum M_{ab} |E_a\rangle \otimes |E_b\rangle$ with M_{ab} a banded random matrix, with bandwidth controlled by the strength of the interaction [34, 35].

In our model, we replace time evolution with a single projected Hamiltonian H' by an ensemble of Hamiltonians with the same statistical properties as H', in particular, the symmetry class and the spectral distribution. In what follows, we will always assume that this ensemble is unitarily invariant, as this symmetry class is the least restrictive on the Hamiltonian.

When needed, for tractability and for explicit results, we will further assume that the ensemble is Gaussian. Gaussian ensembles have probability measures

$$\mu(H) \propto \exp\left(-\frac{1}{a^2} \operatorname{Tr} H^2\right), \tag{2.12}$$

where a sets the overall energy scale. When we specify ensembles of Hamiltonians H, like above, we will always be implicitly referring only to the distribution of the relevant matrix elements $\langle\psi_i| H |\psi_j\rangle$, which are the same as the projected Hamiltonian H'. Without loss of generality, we also choose ensembles whose mean is zero because the diagonal piece of $H' \approx E \cdot \mathbb{1}$ does not affect $\rho(t)$.

To recap our model, we have an initial state $|\psi_1\rangle$, and we are evolving with a unitary-invariant ensemble of $N \times N$ Hamiltonian matrices $\langle\psi_i| H |\psi_j\rangle$, with $|\psi_i\rangle$ defined in (2.8).

JHEP07(2024)023

We will show that the ensemble average of whatever quantity we are considering, density matrices or Rényi entropies, qualitatively equals the semiclassical gravitational approximation to the same. We will calculate reduced density matrices and entanglement measures before and after averaging, and connect the results to the unitary and non-unitary Page curves and replica wormholes. Evolution with an ensemble of Hamiltonians is fundamentally how non-unitarity is introduced into the model.

2.2 Exact density matrices

We will first calculate the density matrices, time-evolved with a single random H drawn from the ensemble. The Hamiltonian can be diagonalised to $H = U\Lambda U^\dagger$ and, since H is Hermitian, U is unitary and Λ is diagonal with real eigenvalues. When expressed in the $|\psi_i\rangle$ basis, the Hamiltonian is

$$\langle\psi_j| H |\psi_l\rangle = \sum_k \lambda_k U_{jk} U^\dagger_{kl}. \tag{2.13}$$

The probability distributions of the eigenvalues λ and unitary matrices U depend on the ensemble, and later we will specialise to the GUE.

The time-dependent density matrix can be found by acting with the time evolution operator e^{iHt} on the initial state, such that

$$\rho(t) = \sum_{i,j,k,l=1}^{N} U_{ik} U^\dagger_{k1} U_{1l} U^\dagger_{lj}\, e^{i(\lambda_k - \lambda_l)t} |\psi_i\rangle \langle\psi_j| . \tag{2.14}$$

Recall that the microcanonical Hilbert space (2.6) is embedded in the factorizable Hilbert space $\mathcal{H} = \mathcal{H}_\mathsf{A} \otimes \mathcal{H}_\mathsf{B}$, so partial traces are defined within this larger Hilbert space. Tracing out subsystem B, the only non-vanishing terms have $i = j$. The partial density matrix $\rho_\mathsf{A}(t)$ is thus given by

$$\rho_\mathsf{A}(t) = \sum_{k,l=1}^{N} e^{i(\lambda_k - \lambda_l)t} \left(U_{1k} U^\dagger_{k1} U_{1l} U^\dagger_{l1} |E\rangle \langle E|_\mathsf{A} + \sum_{i=2}^{N} U_{ik} U^\dagger_{k1} U_{1l} U^\dagger_{li} |0\rangle \langle 0|_\mathsf{A} \right). \tag{2.15}$$

This 2×2 density matrix is diagonal. If we trace out subsystem A instead, the partial density matrix of subsystem B is given by

$$\rho_\mathsf{B}(t) = \sum_{k,l=1}^{N} e^{i(\lambda_k - \lambda_l)t} \left(U_{1k} U^\dagger_{k1} U_{1l} U^\dagger_{l1} |0\rangle \langle 0|_\mathsf{B} + \sum_{i,j=2}^{N} U_{ik} U^\dagger_{k1} U_{1l} U^\dagger_{lj} |i\rangle \langle j|_\mathsf{B} \right). \tag{2.16}$$

Unlike $\rho_\mathsf{A}(t)$, the reduced density matrix in (2.16) is not diagonal.

2.3 Averaged density matrices

Next, we calculate the averaged density matrices and their purity and Rényi entropies. Our averaging is over random Hamiltonians with an appropriate probability measure: $\overline{f(H)} := \int d\mu(H) f(H)$. We will soon restrict the computation to the GUE measure, but for now let the measure be any that is invariant under unitary transformations: $\mu(H) = \mu(UHU^\dagger)$.

JHEP07(2024)023

One approach to calculating averaged Rényi entropyies is to first calculate the joint probability distribution function (PDF) of the entanglement spectra of the reduced density matrices. For us, since ρ_{A} is a 2×2 matrix, and $\mathrm{Tr}(\rho_{\mathsf{A}}) = 1$, there is only one independent eigenvalue in the entanglement spectrum between A and B. Nonetheless, we were unable to determine that PDF, so we have relegated the results to appendix A.

Our approach will be to average the density matrices over H: we first integrate over the diagonalising unitaries, then the eigenvalues. See appendix B for details on how to perform unitary integrals. The averaged reduced density matrix for A is

$$\overline{\rho_{\mathsf{A}}}(t) = \frac{N + \overline{W}(t)}{N(N+1)} |E\rangle \langle E|_{\mathsf{A}} + \frac{N^2 - \overline{W}(t)}{N(N+1)} |0\rangle \langle 0|_{\mathsf{A}} \,. \tag{2.17}$$

We have introduced the spectral form factor:

$$W(t) := \mathrm{Tr}\left(e^{iHt}\right) \mathrm{Tr}\left(e^{-iHt}\right) = \sum_{k,l=1}^{N} e^{i(\lambda_k - \lambda_l)t}. \tag{2.18}$$

Most of our results only require unitary invariance of the measure. The only place where the actual details of the measure come into play, are in the exact time dependence of this spectral form factor, which depends on the ensemble under consideration. Even without restricting to a specific ensemble, chaotic systems have some universal behavior of the spectral form factor. From its definition, $W(t = 0) = N^2$ and, in chaotic systems, the long-time average of W is order N.[5] To elucidate the leading order terms, we introduce the variable $g(t)$, defined as the ratio of $W(t)$ to N^2:

$$g(t) := \frac{W(t)}{N^2} \,. \tag{2.19}$$

At early times, $g(t)$ is of order one. However, as time progresses, it decays to a value of $1/N$ during late-time evolution. The spectral form factor is only a function of the Hamiltonian's eigenvalues and time and the precise function for $g(t)$ depends on the eigenvalue probability distribution of the ensemble. In what follows, we will sometimes omit the explicit t-dependence and just write g. The von Neumann entropy of the averaged density matrix in the large N approximation is

$$S(\overline{\rho_{\mathsf{A}}}(t)) = -\overline{g}(t) \log\left(\overline{g}(t)\right) - (1 - \overline{g}(t)) \log\left(1 - \overline{g}(t)\right) . \tag{2.20}$$

Thus, in the large N approximation, as depicted in figure 3 for the GUE, the entropy (2.20) starts at zero, increases to its maximum value $S_{\max} = \log 2$, and subsequently decreases to approximately zero at late times.

Let us now look at the behaviour of the von Neumann entropy of the average state in the environment, subsystem B. The averaged partial density matrix $\overline{\rho_{\mathsf{B}}}(t)$ is given by

$$\overline{\rho_{\mathsf{B}}}(t) = \frac{1 + \overline{g}(t)N}{N+1} |0\rangle \langle 0|_{\mathsf{B}} + \frac{N - \overline{g}(t)N}{N^2 - 1} \sum_{i=2}^{N} |i\rangle \langle i|_{\mathsf{B}} \,. \tag{2.21}$$

[5] The long time average of $\mathrm{Tr}(e^{iHt})$ is a sum of random phases from the uniform distribution, so W is the square of the absolute value of an N-step random walk in the complex plane.

JHEP07(2024)023

Calculating the entropy is straightforward, because the off-diagonal elements of ρ_{B} have averaged to zero, and, in the large N approximation, gives

$$S\left(\overline{\rho_{\mathsf{B}}}(t)\right) = -\overline{g}(t)\log\left(\overline{g}(t)\right) - (1-\overline{g}(t))\log\left(\frac{1-\overline{g}(t)}{N}\right), \tag{2.22}$$

At $t = 0$, the entropy is zero; $\rho_{\mathsf{B}}(0) = |0\rangle\langle 0|_{\mathsf{B}}$. However, as $\overline{g}(t)$ decays as a function of time, the von Neumann entropy keeps rising. Taking the long time average, for which $\overline{g} \sim 1/N$, the entropy of the averaged state limits to

$$\lim_{t\to\infty} S(\overline{\rho_{\mathsf{B}}}(t)) = \log N. \tag{2.23}$$

Contrary to the entropy of subsystem A, the von Neumann entropy of the averaged environment state does not return to zero at late times. This might seem contradictory, as we had taken time evolution to be unitary. The initial state of the combined system AB is pure, so the entropy of the system A and the environment B must be equal at all times. This is reminiscent of the classical information paradox, where the radiation left after the black hole has evaporated seems to be mixed, after unitarily evolving a pure state. Therefore, (2.22) is our analogue of the Hawking curve, where the radiation entropy grows and remains large at late times, signalling a mixed state. This apparent contradiction can be resolved by realising that we have not studied the average entropy $\overline{S_{\mathsf{A}}}(t)$ and $\overline{S_{\mathsf{B}}}(t)$, but rather the entropy of the averaged states.

2.3.1 GUE

If we specify which ensemble we are drawing our random Hamiltonians from, we can make more precise statements about the time evolution of the von Neumann entropy. We average over random Hamiltonians with the GUE probability measure, where we have taken (2.12) with $a = \sqrt{2/N}$:

$$\overline{f(H)} := \int dH \exp\left(-\frac{N}{2}\operatorname{Tr} H^2\right) f(H). \tag{2.24}$$

The choice of the prefactor a in the exponent controls the width of the eigenvalue spectrum. Our choice gives an N-independent width, $(\overline{H^2} - \overline{H}^2) \sim 1$, which is what we want for our microcanonical energy windows of fixed width.[6] Other Gaussian ensembles are the GOE and the GSE, where the integral is over the space of real symmetric matrices and symmetric quaternionic matrices respectively, and the pre-factor of the measure is slightly altered. We choose GUE, rather than GOE or GSE, because it does not assume time reversal symmetry and is the least restrictive symmetry class of the three Gaussian ensembles. We expect that all our results could be generalised to GOE or GSE if one wished to assume time reversal and/or rotational symmetry, and we could also consider quartic and higher order terms in the matrix potential.

[6] We could include an extra dimensionless parameter α by replacing N by $N\alpha^2$ in (2.24). This would be the same as rescaling time via $t \to t/\alpha$, and in this sense the unit of time can be chosen arbitrarily in our toy model.

JHEP07(2024)023

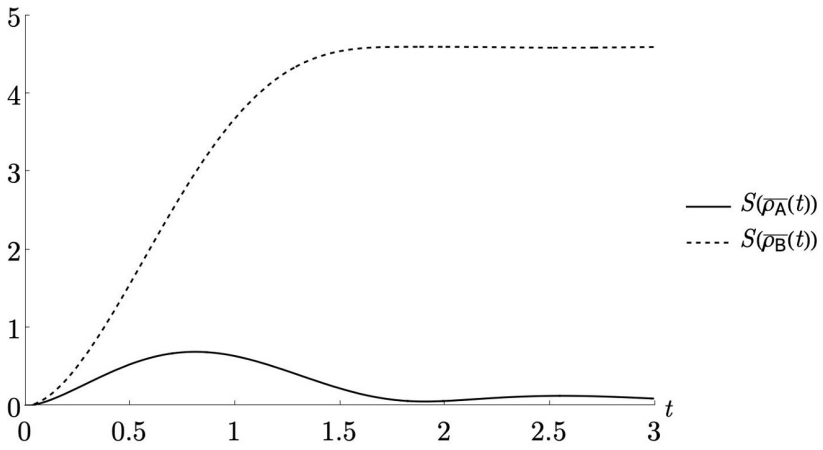


Figure 3. Entropy of the averaged system and environment density matrices. The solid line shows $S(\overline{\rho_{\mathsf{A}}}(t))$, while the dashed line shows $S(\overline{\rho_{\mathsf{B}}}(t))$. These are numerical plots that show the average over 10 Hamiltonians from the GUE with $N = 100$. We have used the analytic results for the Haar averaged density matrices, (2.17) and (2.21), and plotted using the numerical average of the spectral form factor, which also agrees with the analytically averaged result. The entropy for the subsystem A grows to a maximum value of $\log 2$ and then decreases until it is approximately zero. The entropy for the subsystem B rises monotonically and approaches $\log N$ at late times.

For the GUE, [36] studied the time evolution for the average spectral form factor, also using $a = \sqrt{2/N}$ as in (2.24). They find the approximate[7] result

$$g(t)N^2 = N^2 r_1(t)^2 - N r_2(t) + N, \tag{2.25}$$

where the functions $r_1(t)$ and $r_2(t)$ are defined as

$$r_1(t) \coloneqq \frac{J_1(2t)}{t}, \qquad r_2(t) \coloneqq \begin{cases} 1 - \frac{t}{2N}, & \text{for } t < 2N, \\ 0 & \text{for } t > 2N. \end{cases} \tag{2.26}$$

Here J_1 is a Bessel function of the first kind. For Hamiltonians of the GUE with $N = 100$, the time evolution of the von Neumann entropy of the averaged states is visualised in figure 3. It can be seen that indeed the entropy of the subsystem A increases until it reaches its maximum $\log 2$, and later drops and becomes close to zero again. On the other hand, the entropy of the subsystem B keeps increasing, approaching $\log N$.

2.4 Purity

2.4.1 Average purity of system and environment

The fact that we have found a different von Neumann entropy for the averaged state in subsystem A than subsystem B implies the total averaged state in the full system AB is no

[7] The approximation made to find this result is a short distance cutoff to regulate the divergence for closeby eigenvalues in the sine kernel. The divergence is an artefact of the expansion around infinite N. However, when comparing with numerics, [36] finds that the difference between this approximate function and the numerical solution does not decrease when N increases, meaning the approximation is not in effect a large N approximation. This discrepancy only appears in the ramp; the slope and plateau are well approximated by (2.25).

JHEP07(2024)023

longer pure. We explicitly compute this by averaging (2.14) over the unitary matrices:

$$\int_{U_N} \rho(t) = \frac{1+g(t)N}{N+1} |\psi_1\rangle \langle\psi_1| + \frac{N-g(t)N}{N^2-1} \sum_{i=2}^{N} |\psi_i\rangle \langle\psi_i| \,. \tag{2.27}$$

If we take the square of this density matrix and then take the trace, we find that the purity of the averaged state in the large N approximation is given by

$$\mathrm{Tr}\left(\overline{\rho}^2(t)\right) = \overline{g}(t)^2 + \frac{1}{N}\left(1-\overline{g}(t)\right)^2 . \tag{2.28}$$

Given the time evolution of the spectral form factor, we see that the purity of the averaged total state starts at 1, but then decreases and goes to $1/N$ in the large N approximation. However, under unitary time evolution, the density matrix remains pure at all times. Indeed, computing the expression for the purity without averaging the state gives

$$\mathrm{Tr}(\rho^2(t)) \equiv 1 \tag{2.29}$$

at all times. This follows from the properties of the unitary matrices and no averaging has come into play yet.

It is not immediately obvious what unitarity means when evolving with an ensemble of Hamiltonians. While evolution with a single Hamiltonian is unitary, the ensemble-averaged evolution of a given quantity may not be unitary in the sense of whether there exists a single unitary transformation that can give the same result as the ensemble average. The purity of $\overline{\rho}(t)$ that we calculated is not unitary in this sense.

2.4.2 Average purity of system

Let us now compute the averaged purity of the states reduced to A and B. First, to compare, the purities of the averaged density matrices, in the large N approximation, are

$$\mathrm{Tr}\left(\overline{\rho_{\mathsf{A}}}^2(t)\right) = \overline{g}(t)^2 + \left(1-\overline{g}(t)\right)^2 , \qquad \mathrm{Tr}\left(\overline{\rho_{\mathsf{B}}}^2(t)\right) = \overline{g}(t)^2 + \frac{1}{N}\left(1-\overline{g}(t)\right)^2 . \tag{2.30}$$

Therefore, at late times, the average of $\overline{\rho_{\mathsf{A}}}$ is pure, while $\overline{\rho_{\mathsf{B}}}$ has purity $\sim N^{-1}$. In contrast, at late times, the averaged purity of B should be approximately one. Assuming ergodicity, the long time average of the time-evolved state equals the microcanonical average and, since a large fraction $\frac{N-1}{N}$ of the states in $\mathcal{H}_{\mathrm{micro};E\pm\delta E}$ are in $|0\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B};E}$, the microcanonical average is approximately an unentangled product state with pure reduced density matrices. Therefore, we expect the average purity of both the system and environment at late times to be approximately one. This differs from the purity of the averaged state in B as seen in (2.30), which goes to zero at late times. We should then expect that, at late times,

$$\mathrm{Tr}\left(\overline{\rho_{\mathsf{B}}}^2\right) \neq \overline{\mathrm{Tr}\left(\rho_{\mathsf{B}}^2\right)}. \tag{2.31}$$

As we will show next, the difference between the averaged purity and the purity of the averaged state comes from, when averaging over the unitary group, whether the unitary matrix index contractions connect the reduced density matrices or not.

JHEP07(2024)023

2.4.3 Average purity of the environment

Let us now study the average purity of the state reduced B. Squaring the partial density matrix and taking the trace, we obtain

$$\begin{aligned}\mathrm{Tr}\left(\rho_{\mathsf{B}}^2(t)\right) = \sum_{k,l,p,q=1}^{N} e^{i(\lambda_k-\lambda_l+\lambda_p-\lambda_q)t}\Big(&U_{1k}U_{k1}^\dagger U_{1l}U_{l1}^\dagger\; U_{1p}U_{p1}^\dagger U_{1q}U_{q1}^\dagger \\ &+ \sum_{i,j=2}^{N} U_{ik}U_{k1}^\dagger U_{1l}U_{lj}^\dagger\; U_{jp}U_{p1}^\dagger U_{1q}U_{qi}^\dagger\Big).\end{aligned} \tag{2.32}$$

Note that both of these terms are very similar, where $i = j = 1$ for the first term, and for the second term both i and j can be anything but 1. For both terms, the first four unitaries come from the left copy of the density matrix, the last four unitary matrices come from the right one. Taking the trace and averaging over the unitary matrices, there are a priori many different contraction patterns that need to be taken into account.[8]

Large N approximation. Taking the Haar average of a product of unitary matrices contracts indices. To explain our contraction notation, we will draw the line above for contractions between unitary matrices and below for contractions between individual indices, so

$$\overline{U_{ab}U_{cd}^\dagger} \coloneqq \underbrace{U_{ab}U_{cd}^\dagger} \sim \frac{\delta_{ad}\delta_{bc}}{N}. \tag{2.33}$$

We will first work in the large N approximation, where only five contraction patterns of the indices of the unitary matrices turn out to be non-vanishing. The first contraction pattern that needs to be taken into account is

$$\overline{U_{ik}U_{k1}^\dagger}\,\overline{U_{1l}U_{lj}^\dagger}\quad\overline{U_{jp}U_{p1}^\dagger}\,\overline{U_{1q}U_{qi}^\dagger} \sim \frac{\delta_{i1}\delta_{j1}}{N^4} \longrightarrow \overline{g^2}(t). \tag{2.34}$$

There is a space left between the unitary matrices that belong to the first and second copy of ρ_{B}, to highlight whether the contraction connects the different copies of ρ_{B} or not. The delta functions give a contribution of $\overline{g^2}$ after evaluating the sums together with the exponential factor. Since g is of order one at early times, we indeed have to include this term in our leading order expansion for the purity. There are four more contractions that contribute to leading order to the purity:

$$\begin{aligned}
&\overline{U_{ik}U_{k1}^\dagger\,U_{1l}U_{lj}^\dagger\quad U_{jp}U_{p1}^\dagger\,U_{1q}U_{qi}^\dagger} \sim \frac{\delta_{kq}\delta_{lp}}{N^4} \longrightarrow 1,\\
&\overline{U_{ik}U_{k1}^\dagger}\,\underbrace{U_{1l}U_{lj}^\dagger\quad U_{jp}U_{p1}^\dagger}\,\overline{U_{1q}U_{q1}^\dagger} \sim \frac{-\delta_{kq}}{N^5} \longrightarrow -\overline{g}(t),\\
&\underbrace{U_{ik}U_{k1}^\dagger}\,\overline{U_{1l}U_{lj}^\dagger\quad U_{jp}U_{p1}^\dagger}\,\underbrace{U_{1q}U_{qi}^\dagger} \sim \frac{-\delta_{lp}}{N^5} \longrightarrow -\overline{g}(t),\\
&\underbrace{U_{ik}U_{k1}^\dagger\,U_{1l}U_{lj}^\dagger\quad U_{jp}U_{p1}^\dagger\,U_{1q}U_{qi}^\dagger} \sim \frac{1}{N^6} \longrightarrow \overline{g^2}(t).
\end{aligned} \tag{2.35}$$

[8] Averaging over the unitary matrices and taking the trace commute. The averaging is an integration and the trace is a sum which, because it is over a finite number of terms, commutes with the integration.

JHEP07(2024)023

Therefore we find that to leading order in the large N approximation, the purity of the environment is given by

$$\overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))} = 1 - 2\overline{g}(t) + 2\overline{g^2}(t)\,. \tag{2.36}$$

At $t = 0$, the purity is equal to one because $g(0) = 1$. At late times, in chaotic systems and in the large N approximation, g goes to zero at late times and the purity will become one again. We can disentangle the contributions to the purity into a disconnected component and a connected component, and find

$$\overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))}_{\text{disc.}} = \overline{g^2}(t)\,, \qquad \overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))}_{\text{conn.}} = \overline{(1-g(t))^2}\,. \tag{2.37}$$

The disconnected component of the purity as computed in (2.30) starts out at 1 but decreases to zero at late times. Therefore, the dominant contribution to the purity changes. At early times, the disconnected component dominates. However, at some intermediate time, the connected component starts to dominate. For a fixed ensemble, one can actually compute the time evolution of the purity. For the GUE, the purity and its decomposition into disconnected and connected components is shown in figure 4.

Our model captures the replica wormhole story as there is a dynamical transition in dominance: the connected contribution grows larger than the disconnected contribution. This connected contribution contracts the density matrix indices exactly how we would expect it to in the replica wormhole:

$$\rho_{b_1 b_2}\rho_{b_2 b_1}\,. \tag{2.38}$$

Let us calculate the time at which the averaged purity of ρ_{B} is minimal, which is an analogue of a black hole's Page time in our model. The turnaround is when $\overline{g} = \frac{1}{2}$, where the purity has dropped to $\frac{1}{2}$.[9] For the GUE, taking $\overline{g}(t) \approx e^{-t^2/2}$ which is valid at early times, it can be seen that the transition time is

$$t \approx \sqrt{\ln 2}. \tag{2.39}$$

This time is independent of N. In figure 4 it can be seen that this agrees with the numerical result since $\sqrt{\ln 2} \approx 0.83$. To compare, the Page time of a 4d asymptotically flat Schwarzschild black hole of initial energy E is

$$t_{\text{Page}} \sim G_N^2 E^3. \tag{2.40}$$

Our model has no parameters that directly represent the Planck scale, or E, the total energy of the system and environment, so cannot capture the dependence of the black hole entropy or the Page time on these quantities. Nonetheless, it does capture the qualitative shapes of the unitary and non-unitary Page curves. That said, in section 3 we will explain how our model can be modified to get a more physical Page time like (2.40).

[9] For this, we assume that $\overline{g^2} \approx \overline{g}^2$ for $t \ll \sqrt{N}$. Spectral form factors in chaotic theories are characterised by their dip, ramp and plateau, and they are self-averaging before their dip time [37], but not after. For the GUE ensemble, $t \sim \sqrt{N}$ is the dip time. Numerical evidence to justify our assumption is shown in figure 6.

JHEP07(2024)023

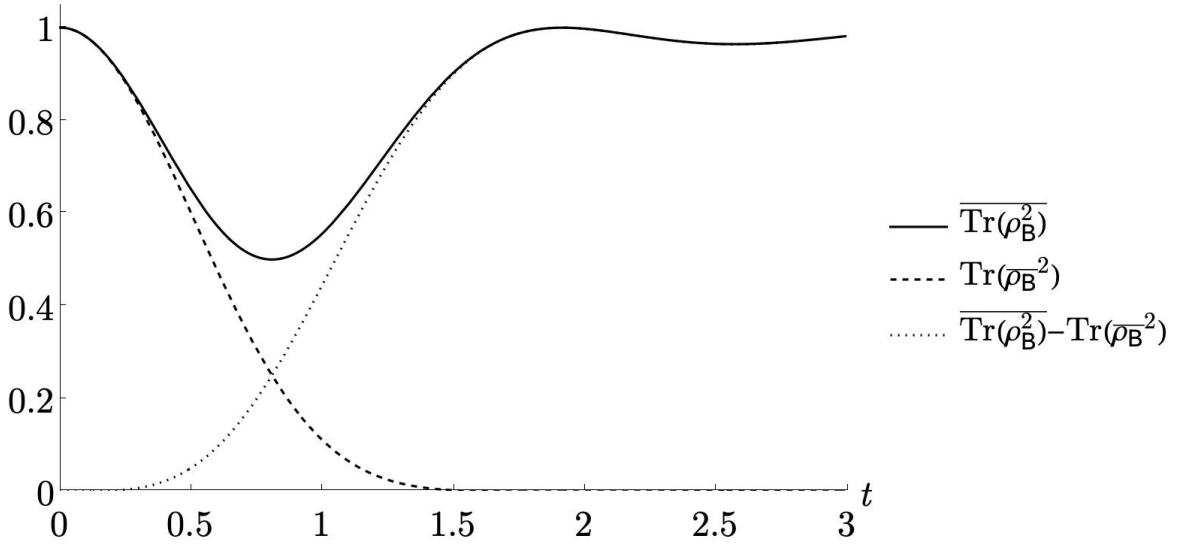


Figure 4. The purity of the environment B as a function of time. The solid black line is a numerical plot of the large N approximation of $\overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))}$ as defined in (2.36) averaged over 10 Hamiltonians drawn from the GUE with $N = 100$. The dashed line is the disconnected component of $\overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))}$ and the dotted line is the connected component, as defined in (2.37). At early times, the disconnected component dominates, while at late times the connected component dominates, resulting in a final state that is close to pure. This is closely analogous to the replica wormhole story, where unitarity in the Page curve is restored by an exchange of dominance to a saddle that connects the density matrices when calculating the Rényi entropies.

Exact in N. One can also compute an expression for the averaged purity that is exact in N. To do so, one needs the exact Weingarten functions for 4-cycles, which are listed in appendix B. The result is

$$\overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))} = \frac{N^4\left(1-2\overline{g}+2\overline{g^2}\right)+N^3\left(4-2\overline{g}+2\overline{g_3}+2\overline{g_3^*}\right)+N^2\left(5+4\overline{g}+2\overline{g_2}\right)+6N}{N(N+1)(N+2)(N+3)}\,. \tag{2.41}$$

Here $g_2(t)$ and $g_3(t)$ are new functions that are defined as

$$g_2(t) \coloneqq \frac{\mathrm{Tr}\left(e^{2iHt}\right)\mathrm{Tr}\left(e^{-2iHt}\right)}{N^2}\,, \qquad g_3(t) \coloneqq \frac{\mathrm{Tr}\left(e^{2iHt}\right)\mathrm{Tr}\left(e^{-iHt}\right)^2}{N^3}\,. \tag{2.42}$$

Both of these are of order one at early times, which is also their maximal value. Note that $g_2(t)$ is equal to $g(2t)$. It is straightforward to see that the large N approximation of (2.41) indeed reduces to (2.36).

2.4.4 Average purity of the system

Let us next calculate the purity of the system A. Squaring ρ_{A} and taking the trace yields

$$\begin{aligned}\mathrm{Tr}\,\rho_{\mathsf{A}}^2(t) = \sum_{k,l,p,q=1}^{N} e^{i(\lambda_k-\lambda_l+\lambda_p-\lambda_q)t}\Big(&U_{1k}U_{k1}^\dagger U_{1l}U_{l1}^\dagger\; U_{1p}U_{p1}^\dagger U_{1q}U_{q1}^\dagger \\ &+\sum_{i,j=2}^{N} U_{ik}U_{k1}^\dagger U_{1l}U_{lj}^\dagger\; U_{jp}U_{p1}^\dagger U_{1q}U_{qi}^\dagger\Big).\end{aligned} \tag{2.43}$$

JHEP07(2024)023

The unitaries in (2.43) are exactly the same as the unitaries that appeared in the purity of the environment. Note that $\mathrm{Tr}(\overline{\rho_{\mathsf{A}}}^2(t)) = \overline{\mathrm{Tr}\,(\rho_{\mathsf{A}}^2(t))}$, in the large N approximation, because the system A evolves towards its pure vacuum state for typical draws from the Hamiltonian ensemble, so its purity is self-averaging. In contrast to the environment B, no replica wormholes are necessary for unitarity in the purity of A. A small consequence of $\mathrm{Tr}(\overline{\rho_{\mathsf{A}}}^2(t)) = \overline{\mathrm{Tr}\,(\rho_{\mathsf{A}}^2(t))}$ is that $\mathrm{Tr}(\overline{\rho_{\mathsf{A}}}^2(t)) = \overline{\mathrm{Tr}(\rho_{\mathsf{B}}^2(t))}$ and we will return to this peculiarity in section 3.

2.5 Averaged Rényi and von Neumann entropies

In the previous subsection, we calculated and discussed the evolution of the purities of the system A and its environment B. Next, we would like to extend the analysis to von Neumann and the higher Rényi entropies. It would be interesting to see whether the results for the purity extend to higher n, in particular whether the analogue of the dominant replica wormhole in [15] is also the late time dominant contribution in our model and if so, when this exchange in dominance takes place.

In $\mathrm{Tr}(\rho_{\mathsf{X}}^n)$, whether for $\rho_{\mathsf{X}} = \rho_{\mathsf{A}}, \rho_{\mathsf{B}}$ or ρ, each density matrix contributes four unitary matrices, see (2.14)–(2.16), so the Haar average of $\mathrm{Tr}(\rho_{\mathsf{X}}^n)$ is an integral over $4n$ unitary matrices with in total $8n$ indices. After averaging, all contraction patterns that are non-vanishing in the large N approximation have their $8n$ indices contracted in pairs.[10] These pairwise contractions either connect different density matrices or not. A key result in this subsection that we will show is how connected contributions to the Haar average of $\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)$ restores unitarity in the Page curve.

2.5.1 Disconnected contributions

We can calculate the averaged von Neumann entropy through[11]

$$\overline{S(\rho_{\mathsf{B}}(t))} = -\partial_n \overline{\mathrm{Tr}\,(\rho_{\mathsf{B}}^n(t))}\Big|_{n=1}. \tag{2.44}$$

The expression for $\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)$ can be compactly written as

$$\begin{aligned}\mathrm{Tr}\,(\rho_{\mathsf{B}}^n) = \sum_{k_1=1}^{N}\cdots\sum_{k_{2n}=1}^{N}\sum_{b_1=1}^{N}\cdots\sum_{b_n=1}^{N}\left(\prod_{m=1}^{n}\delta_{1b_m} + \prod_{m=1}^{n}(1-\delta_{1b_m})\right)\times \\ \times\prod_{i=1}^{n} e^{i(\lambda_{k_{2i-1}}-\lambda_{k_{2i}})t}U_{b_i k_{2i-1}}U^{\dagger}_{k_{2i-1}1}U_{1k_{2i}}U^{\dagger}_{k_{2i}b_{i+1}},\end{aligned} \tag{2.45}$$

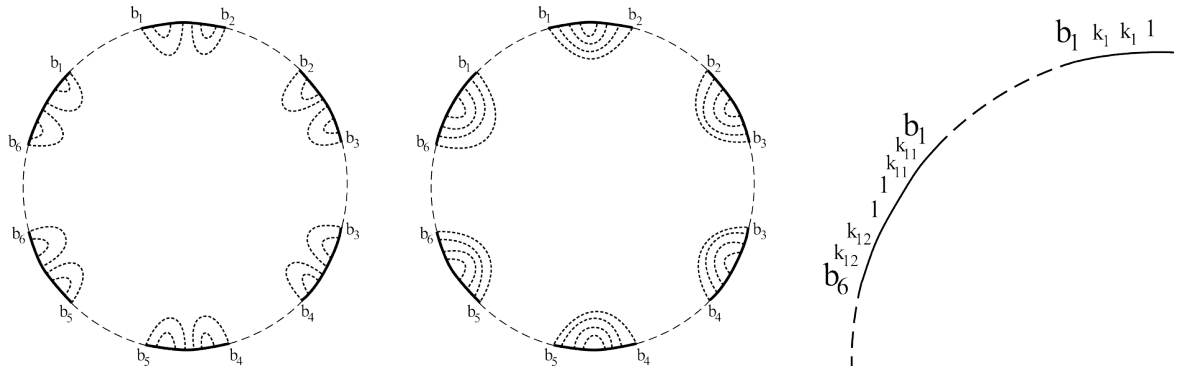
where $b_{n+1} = b_1$. Note that there are two distinct terms in (2.45): the first term has $b_\ell = 1$ for all ℓ, while the second term has $b_\ell \neq 1$ for all ℓ.

We will first study the disconnected contribution to $\overline{\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)}$, before we also take into account the connected components. Each disconnected density matrix ρ_{B} has four possibilities

[10] By contracted in pairs we mean the following. Each summed over index i appears twice. In the leading order contractions, each i is either contracted to itself, or both i's are contracted with the two occurrences of the same index j.

[11] Strictly speaking, we should average after taking the derivative, rather than the other way around. We will assume that these are the same; this seems like a mild assumption because ρ_{B} is a finite size matrix with eigenvalues in the compact interval $[0, 1]$, so we expect nice convergence and analyticity properties.

JHEP07(2024)023



(a) (b) (c)

Figure 5. Diagrammatics of disconnected contributions to the Rényi entropy, whose sum gives (2.47). The black lines on the edges of the first two circles represent the density matrices in $\mathrm{Tr}(\rho_{\mathsf{B}}^n(t))$, and the density matrix indices are labelled. In 5(a) and 5(b), two different contraction patterns of the unitary matrices contributing to (2.47) are visualised for $n = 6$. The contraction pattern in 5(a) corresponds to the contribution in (2.48) that gives $\overline{g^6}(t)$. At early times, this is 1, but at late times it decreases to $\sim N^{-6}$. The contraction in 5(b) contributes N^{-5}. At early times, this is subleading, but at late times it dominates over 5(a). In 5(c), a zoomed-in version on the upper left corner shows which lines correspond to which indices.

to contract its indices, to leading order in the large N approximation given by

$$\overline{U_{b_i k_{2i-1}} U^\dagger_{k_{2i-1}1} U_{1k_{2i}} U^\dagger_{k_{2i}b_{i+1}}} = \frac{\delta_{1b_i}\delta_{1b_{i+1}}}{N^2}\left(1-\frac{\delta_{k_{2i-1}k_{2i}}}{N}\right)+\frac{\delta_{b_i b_{i+1}}}{N^2}\left(\delta_{k_{2i-1}k_{2i}}-\frac{1}{N}\right). \tag{2.46}$$

Therefore, we find that the leading order contribution of the disconnected component to $\overline{\mathrm{Tr}\left(\rho_{\mathsf{B}}^n\right)}$ yields

$$\overline{\mathrm{Tr}\left(\rho_{\mathsf{B}}^n\right)}_{\text{disc.}} = \overline{g^n}(t) + \frac{\overline{(1-g(t))^n}}{N^{n-1}}\,. \tag{2.47}$$

The first term in (2.47) corresponds to the contraction

$$\prod_{i=1}^{n}\frac{\delta_{1b_i}}{N^2}\,. \tag{2.48}$$

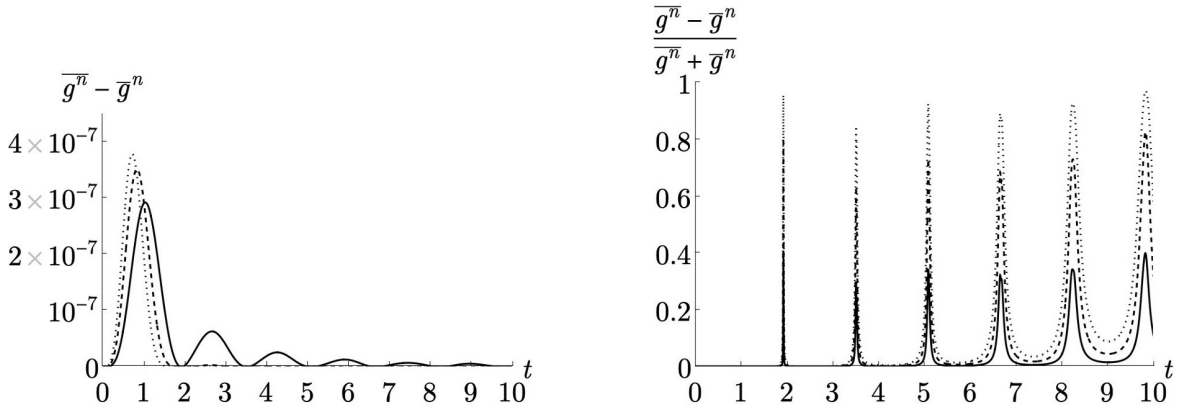
At late times, $\overline{g} \to 1/N$. Therefore, the contribution of this term is of order N^{-n} at late times. This contraction is visualised in figure 5(a) for $n = 6$. The second term in (2.47) corresponds to the contractions

$$\prod_{i=1}^{n}\frac{\delta_{b_i b_{i+1}}}{N^2}\left(\delta_{k_{2i-1}k_{2i}}-\frac{1}{N}\right). \tag{2.49}$$

At late times, this is of order N^{-n+1} and thus dominates over (2.48). The contraction that survives the large N approximation at late times and thus contributes N^{-n+1} is visualised in figure 5(b) for $n = 6$.

Spectral form factors are self-averaging before their dip times [37]. For GUE, the dip time is $t \sim \sqrt{N}$, which is long after the Page time of our model, so we can assume that g

JHEP07(2024)023



(a) (b)

Figure 6. In 6(a), the error $\overline{g^n} - \overline{g}^n$ is shown for $n = 2, 3, 4$. The solid lines correspond to $n = 2$, the dashed lines to $n = 3$ and the dotted lines to $n = 4$. The average is taken over 100 Hamiltonians drawn from the GUE with $N = 1000$. In 6(b), the relative error $\frac{\overline{g^n} - \overline{g}^n}{\overline{g^n} + \overline{g}^n}$ is visualised. It can be seen that the relative error has large spikes, centred around the zeroes of $J_1(2t)$ where $\overline{g}$ is expected to be very small before the dip time. We have numerical evidence that as N increases, the width of these spikes decreases.

is self-averaging and use the approximation

$$\overline{g^n}(t) \approx \overline{g}^n(t). \tag{2.50}$$

In figure 6 we give a numerical plot to show both the absolute and the relative error, to give further evidence for the approximation.

The leading order von Neumann entropy, only taking into account the disconnected components, can now be computed from (2.47) to be

$$\overline{S(\rho_{\mathsf{B}}(t))}_{\text{disc.}} \approx -\overline{g}(t)\log\left(\overline{g}(t)\right) - (1-\overline{g}(t))\log\left(1-\overline{g}(t)\right) + (1-\overline{g}(t))\log N\,. \tag{2.51}$$

At late times, this approaches $\log N$. This is the same as we found for $S(\overline{\rho_{\mathsf{B}}}(t))$ in (2.22) and thus at times when the assumption (2.50) is appropriate,

$$\overline{S(\rho_{\mathsf{B}})}_{\text{disc.}} \approx S(\overline{\rho_{\mathsf{B}}}). \tag{2.52}$$

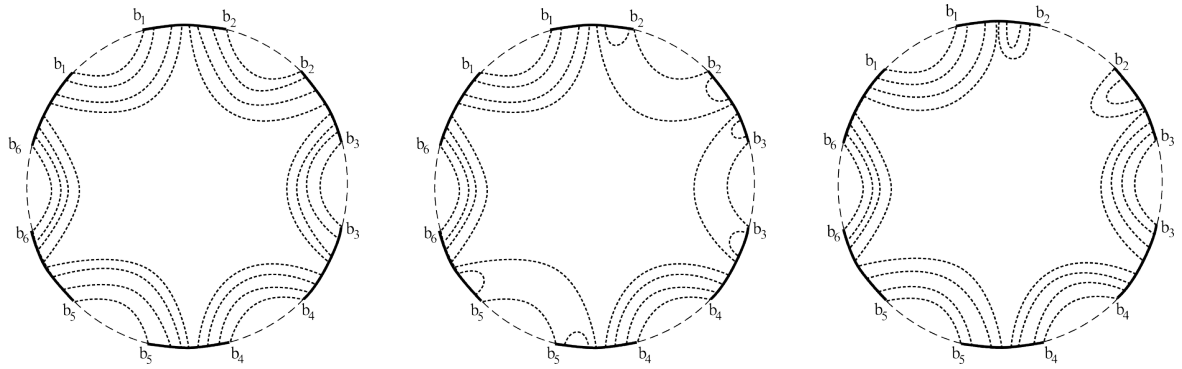
The assumption in (2.50) is known to be valid up to the dip time, therefore the analytic continuation to the von Neumann entropy is only valid up until the dip time as well. At this point, we are ignorant about the behaviour after the dip time, both for the n'th Rényi entropies and the von Neumann entropy.

2.5.2 Connected contributions

Now we return to the computation of $\overline{\operatorname{Tr}\rho_{\mathsf{B}}^n}$, also taking into account the connected components. This yields

$$\overline{\operatorname{Tr}\left(\rho_{\mathsf{B}}^n(t)\right)} = (1-\overline{g}(t))^n + \overline{g}^n(t). \tag{2.53}$$

JHEP07(2024)023



(a) (b) (c)

Figure 7. Connected contributions to the Rényi entropy of the unitary matrix contractions. In 7(a), 7(b) and 7(c), three contraction patterns contributing to (2.57) are visualised. The contraction in 7(a) contributes 1, the contraction in 7(b) contributes $-\overline{g^3}(t)$. The contraction in 7(c) contributes $\frac{\overline{g}(t)}{N^5}$ and is always subleading with respect to the others. At early times, 7(a) and 7(b) are of the same order, but at late times the only contraction pattern of the unitary matrices that is non-vanishing is the one in 7(a).

Here we have again used the approximation (2.50). The computation of (2.53) can be found in appendix C. The first term is the leading order contribution of the connected components, the second term corresponds to the disconnected component. We have dropped the terms corresponding to (2.49), as they are always subleading with respect to the connected components for $n > 1$. Three contributions to the connected components are visualised in figure 7. We note that at $t = 0$, the disconnected component is 1 and the sum of the connected components is zero. At late times, the disconnected component goes to zero, while the sum of the connected components approaches 1. From (2.53), the large N approximation of the n'th Rényi entropy is given by[12]

$$\overline{S^{(n)}(\rho_{\mathsf{B}}(t))} = \frac{\log\left((1-\overline{g}(t))^n + \overline{g}^n(t)\right)}{1-n}\,, \qquad n > 1. \tag{2.54}$$

2.6 Relation to replica wormholes

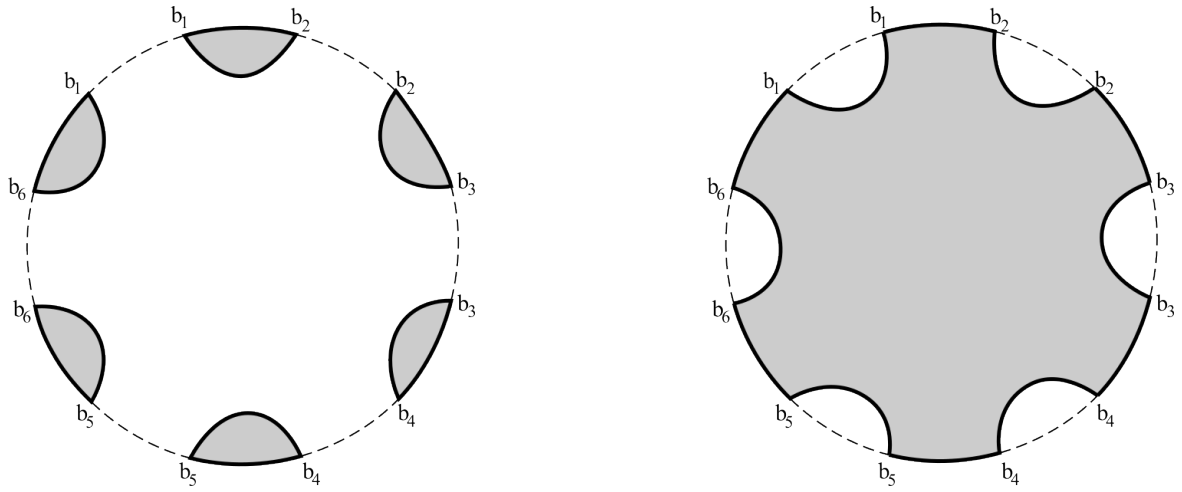
In order to make contact with the replica wormholes as found in [15], we have to study the contractions of the density matrices ρ_{B}. Each of these density matrices has two indices. We can thus write

$$\mathrm{Tr}\left(\rho_{\mathsf{B}}^n\right) = \rho_{b_1 b_2}\rho_{b_2 b_3}\rho_{b_3 b_4}\ldots\rho_{b_n b_1}. \tag{2.55}$$

To avoid clutter, we have suppressed the B-index of the partial density matrix. Two examples of contributions to the disconnected components were given in figure 5. In terms of the contraction pattern of the density matrices, we can visualise the sum of these unitary

[12]The averaged n'th Rényi entropy is proportional to $\overline{\log \mathrm{Tr}\left(\rho_{\mathsf{B}}^n\right)}$ but, as is well-known in the disorder-averaging literature, averaging after taking the logarithm is difficult so we instead average first and then take the logarithm. We thus are computing the annealed version instead of the quenched version of the Rényi entropies.

JHEP07(2024)023



(a) (b)

Figure 8. The replica wormhole-like connectedness between density matrices that results from the unitary matrix contractions. In 8(a), the disconnected contraction pattern from (2.56) is visualised. This pattern dominates at early times where it is close to 1, but approaches zero at late times. In 8(b), the contraction pattern corresponding to (2.57) is visualised. This term is close to zero at early times, but is the dominant contribution and approaches 1 at late times.

contraction patterns as

$$\rho_{\underline{b_1}\underline{b_2}}\rho_{\underline{b_2}\underline{b_3}}\rho_{\underline{b_3}\underline{b_4}}\cdots\rho_{\underline{b_n}\underline{b_1}} = \overline{g^n}(t) + \frac{\overline{(1-g(t))^n}}{N^{n-1}}, \tag{2.56}$$

where each of the density matrices is contracted with itself. This contribution is 1 at $t = 0$, but decreases as time evolves. This contraction of the density matrices is visualised in figure 8(a) for $n = 6$.

On the other hand, in all of the leading order connected unitary contraction patterns, each b_i is contracted with itself. Two examples of these contraction patterns are visualised in figure 7(a) and figure 7(b). Although the contraction of the internal indices of the unitary matrices varies between these different contributions, on the level of the density matrices the sum of the leading order contractions to the connected components can be visualised as

$$\underbrace{\rho_{b_1 b_2}\rho_{b_2 b_3}\rho_{b_3 b_4}\cdots\rho_{b_n b_1}}_{\sqcup\sqcup\sqcup\sqcup\cdots\sqcup\sqcup} = \overline{(1-g(t))^n}, \tag{2.57}$$

and pictorially as in figure 8(b). This contribution is zero at $t = 0$, but increases as time evolves until it approaches 1 at late times. Note that there are other contraction patterns of the unitary matrices which may be harder to visualise as contraction patterns of the partial density matrices. An example of such a unitary contraction is given in figure 7(c). However, all the leading order contractions can all be visualised as in figure 8(b). This is reminiscent of the leading order replica wormhole contribution as found in [15].

Now one can compute the large N approximation of the von Neumann entropy of the environment B from (2.53) to be[13]

$$\overline{S(\rho_{\mathsf{B}}(t))} = -\overline{g}(t)\log\left(\overline{g}(t)\right) - \left(1-\overline{g}(t)\right)\log\left(1-\overline{g}(t)\right). \tag{2.58}$$

[13]To remind the reader, in deriving this equation we assume $\overline{g^n} \approx \overline{g}^n$. Furthermore, we average before taking the derivative of $\mathrm{Tr}(\rho_{\mathsf{B}}^n)$ instead of after.

JHEP07(2024)023

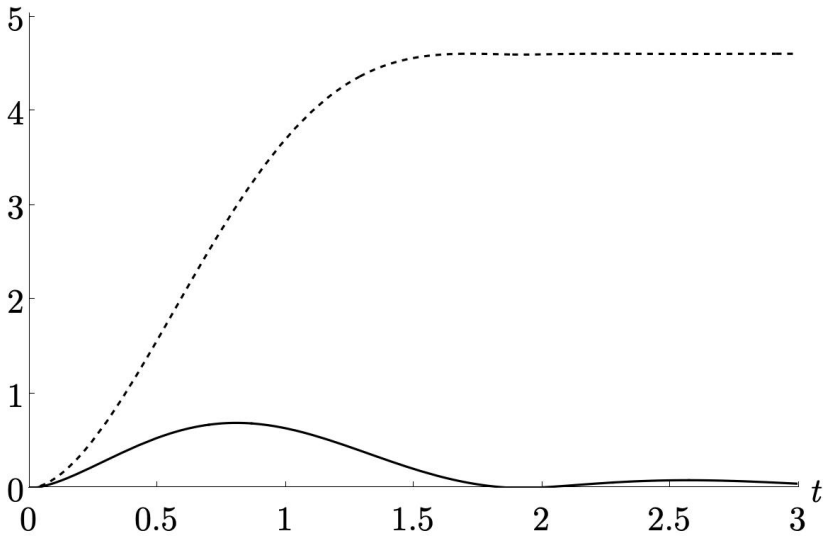


Figure 9. Time evolution of the large N limit of the von Neumann entropy of the environment (2.58), averaged over 10 Hamiltonians drawn from the GUE with $N = 100$. The dashed line is the von Neumann entropy of the disconnected components (2.51) and the solid line is the average von Neumann entropy (2.58).

The time evolution of the large N von Neumann entropy is visualised in figure 9. This is the unitary Page curve of our model.

It might be somewhat surprising to see that we have been able to drop $1/N$ subleading terms in the Rényi entropies and were able to retrieve the unitary result for the von Neumann entropy. It is not obvious that the large N approximation and the replica limit of the Rényi entropies commute, i.e. that

$$\lim_{N\to\infty}\lim_{n\to 1}\overline{S^{(n)}(\rho_{\mathsf{B}}(t))} = \lim_{n\to 1}\lim_{N\to\infty}\overline{S^{(n)}(\rho_{\mathsf{B}}(t))}. \tag{2.59}$$

It is clear that the l.h.s. should give the unitary result, as that is the large N limit of the analytic continuation of the exact result. However, it is not obviously trivial that the r.h.s. gives the same result; we are taking the leading order in $1/N$ for $n > 1$ and taking the analytic continuation. In a semiclassical computation of the replica geometries, one also neglects the higher order corrections, but the analytic continuation of this approximate result still gives the correct von Neumann entropy. The 'unreasonable effectiveness of the low energy EFT' in semiclassical gravity thus manifests itself through this mechanism in our model.

In our model, the multi-boundary replica wormhole shown in figure 8(b) is a representation of the class of unitary matrix index contractions that connect the density matrices in the way shown in figure 7(a) and 7(b). The connectedness arises from Haar averaging the unitary matrices coming from the random Hamiltonians. To summarise with a tongue-in-cheek slogan: ER = Haar.

3 Extensions and further applications of the model

In the previous section, we have seen how in a rather simple model we have captured several important qualitative features apparent in the black hole paradox, such as the naïve Hawking

JHEP07(2024)023

curve and the unitary Page curve, how replica wormholes are necessary to restore unitarity at late times and how the leading order contribution at late times is the cyclically connected replica wormhole. In this section, we will discuss some surprising features of our model, some improvements that can be made to make the model more realistic and we will look at the mutual information between the system, the environment and some reference system.

Difference with actual Page plot. In the standard Page computation, the Hawking and Page curves follow each other at early times and only start to deviate around the Page time, when the Hawking curve keeps growing and the Page curve drops back down. Our result is different in this aspect: we have

$$\overline{S(\rho_{\mathsf{B}}(t))} - S(\overline{\rho_{\mathsf{B}}}(t)) = (1 - \overline{g}(t)) \log N. \tag{3.1}$$

This means the two curves start deviating immediately, and around the Page time already have a difference of $\frac{1}{2} \log N$. A curve which follows this feature of the Page curve a bit more closely is the quantity

$$1 - \mathrm{Tr}(\rho_{\mathsf{B}}^2(t)). \tag{3.2}$$

The difference between the disconnected and connected component of this "impurity" is visualised in figure 2.

No need for replicas in the system. Another perhaps surprising feature of our model is that in the large N approximation,

$$\overline{S(\rho_{\mathsf{B}}(t))} = S(\overline{\rho_{\mathsf{A}}}(t)). \tag{3.3}$$

One way to see how this is the case is by noting that, for fixed Rényi index n, in the large N approximation,

$$\overline{\mathrm{Tr}(\rho_{\mathsf{A}}^n(t))} = \mathrm{Tr}(\overline{\rho_{\mathsf{A}}}^n(t)). \tag{3.4}$$

The advantage of connecting the different replica copies of ρ_{B} is to identify b_i in one copy of the density matrix to b_i in the next. In the end, after evaluating all the sums, there is then a factor

$$\left(\frac{1}{N}\right)^n + \left(\frac{N-1}{N}\right)^n \approx 1, \tag{3.5}$$

such that these terms are leading order. The reason that for $\rho_{\mathsf{A}}(t)$ these are subleading corrections is that by identifying a_i in one copy with a_i in the next, the accompanying factor is $\sim \left(\frac{1}{N}\right)^n \approx 0$.

From a statistical point of view, the above means that at any time, in the large N approximation, the classical statistical ensemble of reduced density matrices ρ_{A} is very sharply peaked around its expectation value with the variance being suppressed by factors $\sim 1/N$.

This feature is thus a consequence of the simplicity of our model, but it is not difficult to see that this still holds if we slightly generalise our model. If we for example consider the Hilbert space decomposition

$$\mathcal{H}_{\text{micro}} = (\mathcal{H}_{\mathsf{A},1} \otimes \mathcal{H}_{\mathsf{B},1}) \oplus (|0\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B},2}), \tag{3.6}$$

JHEP07(2024)023

with $|\mathcal{H}_{\mathsf{A},1}| = N_{\mathsf{A},1}$, $|\mathcal{H}_{\mathsf{B},1}| = N_{\mathsf{B},1}$ and $|\mathcal{H}_{\mathsf{B},2}| = N_{\mathsf{B},2}$, then the total Hilbert space dimension is

$$N = N_{\mathsf{A},1}N_{\mathsf{B},1} + N_{\mathsf{B},2}. \tag{3.7}$$

Under the assumption that $N_{\mathsf{A},1}N_{\mathsf{B},1} \ll N_{\mathsf{B},2}$, then it is not difficult to see that the connected contributions to $\overline{\mathrm{Tr}\,\rho_{\mathsf{A}}^n}$ will at most be of order

$$\left(\frac{N_{\mathsf{A},1}N_{\mathsf{B},1}}{N}\right)^n \ll 1. \tag{3.8}$$

Even if we extend the Hilbert space decomposition further as

$$\mathcal{H}_{\text{micro}} = \bigoplus_{E'} \mathcal{H}_{\mathsf{A};E-E'} \otimes \mathcal{H}_{\mathsf{B};E'}, \tag{3.9}$$

as long as $|\mathcal{H}_{\mathsf{A};0}| = 1$ and

$$|\mathcal{H}_{\mathsf{A};E-E'} \otimes \mathcal{H}_{\mathsf{B};E'}| \begin{cases} \ll |\mathcal{H}_{\mathsf{B};E}| & E' \neq E \\ \sim \mathcal{O}(N) & E' = E, \end{cases} \tag{3.10}$$

then the corrections that can come from taking the connected contributions into account similarly vanish in the large N limit. To recap, for the system A, in the large N approximation, it does not matter whether we average before or after taking the n'th power of the partial density matrix: $\mathrm{Tr}\,(\overline{\rho_{\mathsf{A}}}^n) = \overline{\mathrm{Tr}\,(\rho_{\mathsf{A}}^n)}$. The connections between the different density matrices are subleading corrections.

In contrast, for the environment B, the story is drastically different. The component that connects the neighbouring density matrices becomes the dominant contribution at late times. The disconnected component gives a result analogous to the naïve Hawking computation, with a final state where the radiation is maximally mixed. The connected component, analogous to the replica wormholes, purifies the radiation after the Page time, giving a dynamical entropy consistent with unitarity.

3.1 More realistic models of black hole evaporation

Our toy model captures both the unitary Page and non-unitary Hawking curves, which was the goal of this paper. If one wished to modify the model to make it a more realistic description of black hole evaporation, there are some properties that one would want to address:

1. Our model has only a single coal/black hole microstate within the microcanonical window.

2. The Planck scale l_p is not a tunable parameter. Our large N approximation is not a semiclassical approximation, as $\log N$ is the total microcanonical entropy which counts both black hole and radiation microstates.

3. Our Page time (2.39) is a constant that does not capture the dependence of a realistic black hole's Page time on initial energy or the Planck scale.

JHEP07(2024)023

These features are due to the simplicity of our model. There is only one dimensionless input parameter, N, while unitary Page curves need at least two: the maximum entropy, and the Page time in Planck units. Adding one more dimensionless parameter, the black hole entropy, to the model would likely be sufficient. These 'shortcomings' of our toy model are really a feature, as our goal was to come up with the simplest model that captures both the unitary and non-unitary Page curves and need not capture scales.

Larger number of microstates.

If we wished to make the model more realistic, one modification would be to increase the number of black hole/coal microstates from one. A more realistic model would take $e^{S_{\rm BH}}$ of the N microstates in $\mathcal{H}_{{\rm micro},E}$ to be black hole microstates so that now

$$\mathcal{H}_{{\rm micro};E} = (\mathcal{H}_{\mathsf{A};E} \otimes |0\rangle_{\mathsf{B}}) \oplus (|0\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B};E}), \tag{3.11}$$

where $|\mathcal{H}_{{\rm micro};E}| = N$ as before, but we now have a large number of black hole/coal microstates, with

$$N \gg |\mathcal{H}_{\mathsf{A},E}| = e^{S_{\rm BH}} \gg 1. \tag{3.12}$$

This modification increases the maximum entropy of the unitary Page curve from $\log 2$ to $S_{\rm max} \sim \log |\mathcal{H}_{\mathsf{A},E}|$, the initial black hole microcanonical entropy, but has a negligible effect on the Page time. The physical reason for the absence of an effect on the Page time is that our model still lacks intermediate states between black hole and no black hole; to get a Page time that depends on input parameters, we can introduce intermediate, half-evaporated states, and a notion of locality which captures the property that real black holes evaporate a few Hawking quanta at a time, and not all at once.

Locality. In the model, as it is right now, there is no notion of locality. To see this, note that we are equally likely to transition to any of the microstates in the microcanonical ensemble that form the basis in which our Hamiltonian is random. In particular, even with the modifications to the microcanonical Hilbert space structure discussed previously, where we increase the initial number of black hole microstates and add intermediate partially evaporated states, with a Hamiltonian that is Gaussian random for the set of states in the microcanonical window (2.5), the most likely transition from the initial state is to the entropically favoured radiation states in $|0\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B};E}$.

The eigenvalue statistics of quantum theories of which the classical counterpart is chaotic, are well described by random matrix statistics [33]. Furthermore, within the microcanonical window, the expectation is that the Hamiltonian is unitarily invariant.[14] This justified our model's assumption that H is GUE-random in the basis of states for $\mathcal{H}_{{\rm micro};E}$ in (2.5). In a theory with local interactions, the Hamiltonian will be of the form $H = H_{\mathsf{A}} + H_{\mathsf{B}} + H_{\rm int}$, where $H_{\rm int}$ couples nearby local operators across the shared boundary of the system and environment. We have chosen a basis of states that are tensor products of energy eigenstates of the subsystem Hamiltonians H_{A} and H_{B}. As the interaction between system and environment

[14] This is in the same spirit as the ETH conjecture: low energy observables are not able to distinguish between closely lying energy eigenstates.

JHEP07(2024)023

is turned off, in this basis the Hamiltonian becomes diagonal, so clearly it is a special one. Locality imposes more structure upon the Hamiltonian than we have assumed and, if we want to capture quantities like the Page time, then we have to account for that.

To model a theory with local interactions, we could make a different ansatz for the Hamiltonian in this basis. As explained above, for a weakly coupled local theory we expect H to be almost diagonal in the $|E'\rangle_{\mathsf{A}} \otimes |E - E'\rangle_{\mathsf{B}}$ basis. Instead of choosing the GUE, we could think of opting for a probability distribution like

$$\mu(H) \propto \exp\left(-\sum_{i,j} |H_{ij}|^2 c_{ij}\right), \qquad c_{ij} \sim e^{(i-j)^2\alpha}, \tag{3.13}$$

with $\alpha > 0$. In that case, if i and j are far from each other and H_{ij} is not very small, this Hamiltonian is heavily suppressed. This probability density function is by design no longer invariant under unitary transformations. Therefore, after the decomposition of the Hamiltonian into unitary matrices and eigenvalues, there will be a complicated measure over the unitary matrices alongside the Vandermonde determinant for the eigenvalues. As a consequence, the computation is much more complicated. In line with the ETH conjecture, we want to ensure that whatever measure we pick gives approximate unitary invariance in narrow energy bands with many states.

Another possibility more closely related to the GUE is that the probability density function will be dependent on a trace like in the GUE, but of the commutator of some fixed diagonal matrix M with H:

$$\mu(H) \propto \exp\left(-\operatorname{Tr}\left([M,H]^2\right)\right) = \exp\left(-\sum_{i,j}(m_i - m_j)^2 |H_{ij}|^2\right). \tag{3.14}$$

Yet another possibility is to manually add locality in the interaction Hamiltonian by giving it a small support only on the interaction Hilbert space:

$$\mathcal{H} = \bigoplus_{E_1,E_2,E_3} \mathcal{H}_{\mathsf{A}';E-E_1} \otimes \mathcal{H}_{\mathsf{A}_{\text{int}};E_2} \otimes \mathcal{H}_{\mathsf{B}_{\text{int}};E_3} \otimes \mathcal{H}_{\mathsf{B}';E_1-E_2-E_3}. \tag{3.15}$$

The interaction Hilbert space is given by

$$\mathcal{H}_{\text{int}} = \bigoplus_{E_2,E_3} \mathcal{H}_{\mathsf{A}_{\text{int}};E_2} \otimes \mathcal{H}_{\mathsf{B}_{\text{int}};E_3}\,, \tag{3.16}$$

and we take $|\mathcal{H}_{\text{int}}|$ to be small. Furthermore, we take $|\mathcal{H}_{\mathsf{B}';E}|$ to be much larger than any other subspace. The reason we decompose the interaction Hilbert space into an A_{int}- and B_{int}-part is because we still want to be able to compute the entropies in the entire A and B, which are the primed parts and interaction parts together.

The expectation is that the behaviour of the system will be similar to what was found here since moving all the energy into the environment is entropically favoured. However, the exact time dependence and the maximum of the entropies might change and take on dependence on the sizes of the subspaces.

Planckian and stringy corrections. Our model has one dimensionless parameter N, the number of states in the microcanonical window. If we wished to, how would we incorporate stringy or Planckian corrections? There are several other dimensionless parameters we could

JHEP07(2024)023

add like the number of black hole states $e^{S_{\rm BH}}$ and coefficients of higher order terms in the matrix model potential. Non-perturbative gravitational corrections of the form e^{-1/G_N} would appear as corrections in this new parameter. Part of the dynamics of a more realistic model could be accounted for by including higher-order terms in the matrix model potential, and the precise interpretation of these terms would depend on the way the toy model is embedded in a more realistic set-up.

3.2 Information transfer

Now we will study how the evaporation discussed in the toy model transfers information from the black hole to the radiation, similarly to the Hayden-Preskill process [38]. In the original process, Alice throws a diary into a black hole. This diary is maximally entangled with some sets of qubits in the hand of Charlie. While the black hole evaporates, Bob collects the Hawking radiation. In this process, Bob's final state is maximally entangled with Charlie's system, showing how the information has been transferred from the black hole to the Hawking radiation.

We will study a different process, as our evaporation is not imposed by moving qubits from the black hole Hilbert space to the radiation Hilbert space at discrete time steps but rather by assuming chaotic dynamics within a microcanonical energy window. The model will also differ in another aspect. In the Hayden-Preskill process, Alice throws her diary into an existing (either young or old) black hole where Bob has collected the radiation that has already come out of the black hole. Here Bob's radiation will be maximally entangled with the remainder of the black hole.

Our model is more analogous to a burning piece of coal (or a burning diary). We let Alice prepare two copies of the diary and maximally entangle them. Then she will burn one of the diaries in her system while giving the other diary to Charlie. We will take Bob's initial system, the environment, to be unentangled with the burning diary, but he collects the radiation coming off of the diary as it burns. We want to track where the information in the burning diary is as a function of time. When the diary has evaporated entirely, the information in it must have moved into the radiation. To quantify the information of the burning diary, we will compute the mutual information between the intact diary in Charlie's possession and the separate subsystems of Alice and Bob.

The effective Hilbert space is taken to be the tensor product between Charlie's Hilbert space $\mathcal{H}_{\mathsf{C}}$ and $\mathcal{H}_{\mathsf{AB};E}$; the combined microcanonical Hilbert space of Alice and Bob:

$$\mathcal{H}_{\rm eff} = \mathcal{H}_{\mathsf{AB};E} \otimes \mathcal{H}_{\mathsf{C}}, \qquad \mathcal{H}_{\mathsf{AB};E} = (\mathcal{H}_{\mathsf{A};E} \otimes |0\rangle_{\mathsf{B}}) \oplus (|0\rangle_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B};E})\,. \tag{3.17}$$

Here we take both the microcanonical Hilbert space $\mathcal{H}_{\mathsf{A};E}$ and $\mathcal{H}_{\mathsf{C}}$ to be m-dimensional and $\mathcal{H}_{\mathsf{B};E}$ to be $(N-m)$-dimensional,

$$\mathcal{H}_{\mathsf{A};E} = \bigoplus_{i=1}^{m} |a_i\rangle\,, \qquad \mathcal{H}_{\mathsf{B};E} = \bigoplus_{i=1}^{N-m} |b_i\rangle\,, \qquad \mathcal{H}_{\mathsf{C}} = \bigoplus_{i=1}^{m} |c_i\rangle\,. \tag{3.18}$$

We thus have $|\mathcal{H}_{\mathsf{AB};E}| = N$, and $|\mathcal{H}_{\rm eff}| = mN$. Furthermore, we assume $N \gg m$. We will take our initial density matrix to be of the form

$$\rho_0 = \rho_{\mathsf{AC}} \otimes \rho_{\mathsf{B}}. \tag{3.19}$$

JHEP07(2024)023

Here ρ_{AC} is a maximally entangled state between the diary in A and the auxiliary system in C, and we will take the environment to be initially in the ground state:

$$\rho_{\mathsf{AC}} = |\psi_{\mathsf{AC}}\rangle \langle\psi_{\mathsf{AC}}|\,, \qquad |\psi_{\mathsf{AC}}\rangle = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} |a_i\, c_i\rangle\,, \qquad \rho_{\mathsf{B}} = |0\rangle \langle 0|\,. \tag{3.20}$$

We will take the Hamiltonian to act trivially on $\mathcal{H}_{\mathsf{C}}$ and randomly on the N-dimensional subspace $\mathcal{H}_{\mathsf{AB};E}$:

$$H = H_{\mathrm{GUE}(N)} \otimes \mathbb{1}_{\mathsf{C}}. \tag{3.21}$$

That is, to model the chaotic evaporation of the diary in $\mathcal{H}_{\mathsf{AB};E}$, we take the Hamiltonian to be drawn randomly from the $N \times N$ Gaussian unitary ensemble (GUE). Based on the computation in section 2, the expectation is that now the diary will evaporate as time evolves, and the information in it will move into the radiation in the environment. The mutual information between the separate systems is defined as

$$I(\rho_{\mathsf{X}}, \rho_{\mathsf{Y}}, \rho_{\mathsf{XY}}) \coloneqq S(\rho_{\mathsf{X}}) + S(\rho_{\mathsf{Y}}) - S(\rho_{\mathsf{XY}}), \qquad S(\rho_{\mathsf{X}}) \coloneqq -\partial_n \operatorname{Tr} \rho_{\mathsf{X}}^n \Big|_{n=1}. \tag{3.22}$$

This quantity measures the reduction in uncertainty about the state in X after measuring the state in Y, and vice versa.

3.2.1 Information in the system

First, we want to compute the mutual information between the system A and the auxiliary system C. In order to compute this, we need the following quantities:

$$\overline{\operatorname{Tr}(\rho_{\mathsf{A}}^n)} = \frac{\overline{g}^n}{m^{n-1}} + (1-\overline{g})^n, \qquad \overline{\operatorname{Tr}(\rho_{\mathsf{AC}}^n)} = \frac{(1-\overline{g})^n}{m^{n-1}} + \overline{g}^n, \qquad \overline{\operatorname{Tr}(\rho_{\mathsf{C}}^n)} = \frac{1}{m^{n-1}}. \tag{3.23}$$

The computation of these quantities is done in a similar manner as in section 2, and we have again made the approximation $\overline{g^n} \approx \overline{g}^n$. We then find[15]

$$\begin{aligned} \overline{S(\rho_{\mathsf{A}})} &= \overline{g} \log m - \overline{g} \log \overline{g} - (1-\overline{g}) \log(1-\overline{g}), \\ \overline{S(\rho_{\mathsf{C}})} &= \log m, \\ \overline{S(\rho_{\mathsf{AC}})} &= (1-\overline{g}) \log m - \overline{g} \log \overline{g} - (1-\overline{g}) \log(1-\overline{g}). \end{aligned} \tag{3.24}$$

From here, the mutual information can be computed to be

$$\overline{I(\rho_{\mathsf{A}}, \rho_{\mathsf{C}}, \rho_{\mathsf{AC}})} = 2\overline{g} \log m. \tag{3.25}$$

Initially, the mutual information between A and C is given by $2 \log m$. At late times, it goes to 0. This means that, as expected, the information leaves the system A as time evolves.

[15] As before, we are computing the von Neumann entropy by averaging before taking the derivative instead of the other way around.

JHEP07(2024)023

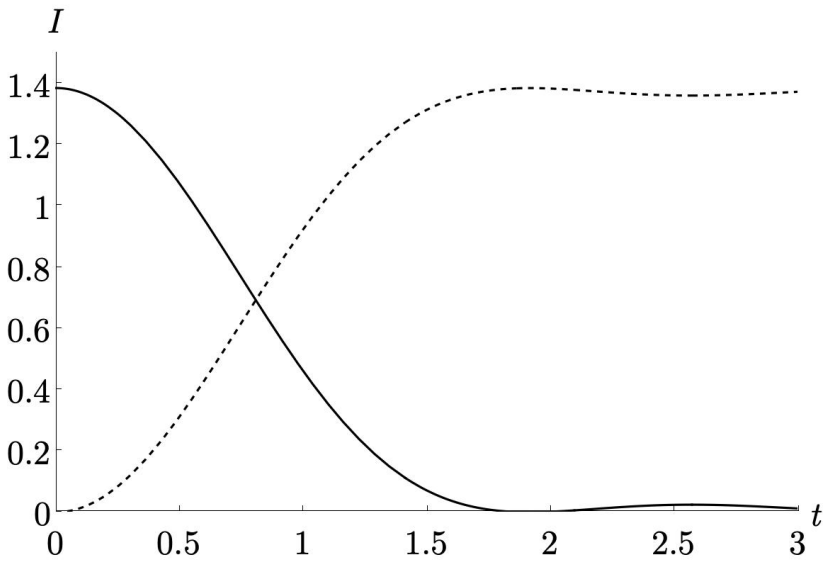


Figure 10. The mutual information I between the auxiliary system C and various systems is shown for $m = 2$. The solid line is the mutual information between A and C, the dashed line is the mutual information between B and C. It can be seen that the mutual information between A and C decreases from $2\log m$ at $t = 0$ to zero at late times, while the mutual information between B and C increases from 0 initially to $2\log m$ at late times. This is the average mutual information computed for 100 Hamiltonians drawn from the GUE with $N = 100$.

3.2.2 Information in the environment

Similarly, we can compute the mutual information between the environment B and the diary in C. Computing the relevant quantities gives

$$\overline{\mathrm{Tr}(\rho_{\mathsf{B}}^n)} = \frac{(1-\overline{g})^n}{m^{n-1}} + \overline{g}^n, \qquad \overline{\mathrm{Tr}(\rho_{\mathsf{BC}}^n)} = \frac{\overline{g}^n}{m^{n-1}} + (1-\overline{g})^n. \tag{3.26}$$

Note that these are exactly the quantities one would expect from unitarity. These give the entropies

$$\begin{aligned} \overline{S(\rho_{\mathsf{B}})} &= -\overline{g}\log\overline{g} - (1-\overline{g})\log(1-\overline{g}) + (1-\overline{g})\log m, \\ \overline{S(\rho_{\mathsf{BC}})} &= -\overline{g}\log\overline{g} - (1-\overline{g})\log(1-\overline{g}) + \overline{g}\log m. \end{aligned} \tag{3.27}$$

The mutual information between the environment and the intact diary in C is thus given by

$$\overline{I(\rho_{\mathsf{B}}, \rho_{\mathsf{C}}, \rho_{\mathsf{BC}})} = 2(1-\overline{g})\log m. \tag{3.28}$$

From here it is clear that the information gradually moves into the environment as time evolves. The time evolution of the mutual information is visualised in figure 10 for $m = 2$.

It is interesting to note what we would have obtained if we did not take into account the replica wormholes. The equations in (3.23) will be unaltered when first averaging and then taking the n'th power. However, the ones in (3.26) will differ:

$$\mathrm{Tr}(\overline{\rho_{\mathsf{B}}}^n) = \overline{g}^n + \frac{(1-\overline{g})^n}{N^{n-1}}, \qquad \mathrm{Tr}(\overline{\rho_{\mathsf{BC}}}^n) = \frac{\overline{g}^n}{m^{n-1}} + \frac{(1-\overline{g})^n}{(mN)^{n-1}}. \tag{3.29}$$

JHEP07(2024)023

This gives the entropies

$$\begin{aligned} S(\overline{\rho_{\mathsf{B}}}) &= -\overline{g}\log\overline{g} - (1-\overline{g})\log(1-\overline{g}) + (1-\overline{g})\log N, \\ S(\overline{\rho_{\mathsf{BC}}}) &= -\overline{g}\log\overline{g} - (1-\overline{g})\log(1-\overline{g}) - (1-\overline{g})\log N + \log m. \end{aligned} \tag{3.30}$$

This means that without taking into account the replica wormholes, mutual information between the radiation and the diary is

$$I(\overline{\rho_{\mathsf{B}}}, \overline{\rho_{\mathsf{C}}}, \overline{\rho_{\mathsf{BC}}}) = 0. \tag{3.31}$$

Naïvely, the information does not show up in the radiation at all.

We have thus seen how the information in a chaotically burning diary shows up in the radiation. Modelling the chaos is done by taking the time evolution of the diary and radiation to be drawn randomly from an ensemble of the GUE. Using the replica trick, we have computed the average von Neumann entropies and used those to quantify the mutual information between the intact diary in C and either the system with the burning diary A or the environment with the radiation B. The information enters the environment through the replica wormholes: if one does not take the connected components into account, the mutual information between the environment and the diary in C is zero at all times.

4 Discussion

In this paper, we have shown how the non-unitary Hawking curve and the unitary Page curve naturally appear in a simple toy model, assuming ergodicity, energy conservation and a gapped system. By observing the state of the environment B in the large N approximation, at late times one finds a maximally mixed density matrix for which the von Neumann entropy grows to $\log N$. This is our analogue of the thermal Hawking radiation, leading to the Hawking curve for the von Neumann entropy.

However, by more carefully probing the entropy through the replica trick and computing the quantities $\overline{\mathrm{Tr}\,\rho_{\mathsf{B}}^n}$ in the large N approximation, we have found a curve for the entropy consistent with unitarity, our analogue of the Page curve. Studying the contraction patterns of the density matrices, one finds two leading order contributions. The first one is a disconnected component, and if only taking this one into account, the result reduces to the Hawking curve. However, the second leading contribution cyclically connects the neighbouring density matrices, forming an analogue of the replica wormholes. This contribution starts to dominate at the Page time. By taking this term into account, the von Neumann entropy decays after the Page time. Our model clarifies how semi-classical replica wormholes are able uncover a Page curve consistent with microscopic unitarity. Gravitational computations involve some coarse graining but the coarse graining is such that they map pure initial states into classical statistical mixtures of pure states at later times. Semi-classical replica computations can diagnose whether the statistical mixture consists of pure states or not, but are unable to accurately identify the indidual pure states themselves. Our Hamiltonian ensembles are unitary-invariant. The relation between unitary matrix contractions and replica wormholes is a consequence of the Haar averaging. The eigenvalue distribution is important for the shapes of the radiation entropy curves $\overline{S(\rho_{\mathsf{B}})}$ and $S(\overline{\rho_{\mathsf{B}}})$, through the averaged spectral form

JHEP07(2024)023

factor $\overline{g}(t)$. A $\overline{g}(t)$ that returns to order 1 too soon, as may be expected in systems such as the simple harmonic oscillator, would give shapes that are inconsistent with the unitary and non-unitary Hawking radiation entropy curves.

Previous work on time evolution with random Hamiltonians from the GUE shows that while the GUE is a good approximation for late time physics of local quantum chaotic theories, at early times there is a discrepancy [36, 39]. The GUE is unitarily invariant, which makes computations tractable, but as a consequence, it only describes non-local physics. These theories scramble much faster than is to be expected for local theories of quantum chaos. However, at late times, all local operators in quantum chaotic systems have had time to spread and the notion of locality is lost and the useful information of the system is found in the spectrum. The GUE is a good description for the spectral properties of the Hamiltonian [33], which means the GUE is capable of accurately describing the late time physics. Even though the GUE thus fails to capture local physics at early times, the late-time conclusions of this work remain even for local chaotic systems. At late times, it is clear that the leading saddle is the connected contribution as in figure 8(b), whereas only considering the disconnected components will give an entropy $\sim \log N$. Initially, when the state is exactly pure in A and B separately, the disconnected component as in figure 8(a) will be the only contribution. Then at some point there will be an exchange in dominance between these two saddles. Therefore, the qualitative picture will not change; only the time-scales and the maximum of the entropy might be affected by introducing local physics. In particular, the recovery of a unitary Page curve only requires fairly general input; assuming a microcanonical window with chaotic dynamics is sufficient to find a unitary Page curve once the replica wormholes are included.

4.1 Comparison to other work

Toy models with dynamical unitary Page curves. There are many toy qubit models of black hole evaporation, see [40] and references therein, and quenched field theory models [41, 42]. The difficult aspect of any toy model is capturing both the non-unitarity of semiclassical black holes and showing how unitarity is restored.

An approach that is close to our own is looking for dynamical Page curves and replica wormholes in SYK-based models [43, 44]. It is close in part because their Hamiltonians are also drawn from an ensemble, and because of the attempt to capture replica wormholes, but it is different because of the specificity of SYK.

There have also been random unitary circuit models of black hole evaporation [45], and a model that attempts to incorporate energy conservation with Haar-random unitary evolution [46]. The "operator-gas approach" [47] finds a dynamical mechanism for the Page curve from quantum chaos without resorting to typicality, by studying the support of operators and modelling the probability of creating a void in this support. [48] computes the late-time entropy by projecting the states onto diagonal subspaces called equilibrated states. [49] models the black hole as a box leaking a gas of chaotic hard spheres. These models dynamically capture the unitary Page curve but their underlying mechanisms are fundamentally different from ours. None evolve within microcanonical windows with an ensemble of Hamiltonians.

Page's theorem and the West Coast model. In the original papers on the information paradox by Page [3, 4], the main assumption is that the total system is in a random pure

JHEP07(2024)023

state in the factorised Hilbert space $\mathcal{H} = \mathcal{H}_{\mathsf{A}} \otimes \mathcal{H}_{\mathsf{B}}$. By averaging over this random state with the Haar measure, the average entropy in the smaller subsystem is computed. Increasing the dimension of this smaller subsystem step-by-step, it can be found that the von Neumann entropy follows the well-known Page curve. The authors of [15] find a unitary Page curve from a toy model with random Hamiltonians. Besides time evolution with a random Hamiltonian, a set of random states is considered, similarly as in [4]. The average overlap between these states is what gives rise to the replica wormholes in this computation. Again, by manually increasing the dimension of the radiation Hilbert space, the entropy is shown to go up and then down again. Both [4] and [15] thus take the state to be a random pure state, modelling uncertainty in the exact microscopic details of the state. Assuming evaporation, the dimensionality of the radiation Hilbert space then grows and this enforces the unitary Page curve.

The main difference with our approach is that we instead take the Hamiltonian to be random in the microcanonical window, motivated by chaos RMT and modelling uncertainty in microscopic physics. Without assuming evaporation, we show how replica wormholes are able to restore unitarity in chaotic systems.

Entanglement and random dynamics. Time evolution with GUE-random Hamiltonians, and the resulting entanglement dynamics, have been previously studied in the condensed matter community [36, 39, 50–52]. One important result is that the smaller subsystem evolves to the maximally mixed state.

To compare, we take the Hamiltonian to be random only within microcanonical windows, to model energy conservation, and this ensures that the entanglement entropy is small at late times due to the argument outlined in section 2.

4.2 Future work

Extending the model. One possible direction for future work is to extend the existing model. For example, the model in section 3.2 might be extended to more closely resemble the actual Hayden-Preskill model, where the diary is thrown into an existing black hole and where Bob has already collected all of the early radiation. Another future direction could be to study the higher moments of the spectral form factor. The computation of the von Neumann entropy assumes the spectral form factor is self-averaging, which is only true before the dip time. After the dip time, the assumption $\overline{g^n} \approx \overline{g}^n$ is not valid anymore and we need additional information to be able to compute the $n \to 1$ limit of the Rényi entropies. The dip time is of order $\sqrt{N}$, so for large N this is at very late times. At these late times, $\overline{g^n}$ can grow close to 1 and thus give rise to large Rényi entropies. In any case, a better estimate for the late time value of the higher moments of the spectral form factor would be valuable for better understanding chaotic systems at late times.

(B)CFT models. In future work, we would like to return to one of our motivating questions: what are the necessary and sufficient conditions on a holographic (B)CFT to get a unitary Page curve? To investigate this, we need to upgrade our model. We could take a single holographic CFT, and consider the evolution of a pure high energy primary state. If the CFT's OPE coefficients are not exact, but drawn from an ensemble, such as in [16–20], can

JHEP07(2024)023

the von Neumann entropy of the averaged state still be zero at all times? At first sight the answer would seem to be yes, because this ensemble will map an initial pure state into a classical statistical mixture of pure states, just as in our toy model. To what extent can we introduce non-unitarities in the operator dimensions and OPE coefficients and still get a unitary Page curve? What is the set of CPTP maps that give the same unitary Page curve as ordinary time evolution? We can ask these questions in bottom-up or specific top-down holographic CFTs models, singly or doubly holographic, with or without radiation baths. For other work connecting statistical properties of CFTs with gravity, see [53–55].

Eigenstate thermalisation hypothesis. The eigenstate thermalisation hypothesis (ETH) for the matrix elements of simple operators in the energy eigenbasis is an ansatz that leads to thermalising behaviour in quantum systems. Because of ETH's connection to chaotic systems and random matrix theory, it is natural to ask whether it alone could capture both the non-unitary Page curve and replica wormhole-like unitarity-restoring corrections. One toy model would be to take the interaction Hamiltonian between system and environment to be $H_{\text{int}} = \lambda \mathcal{O}_{\mathsf{A}} \otimes \mathcal{O}_{\mathsf{B}}$, a coupling between simple local operators, with one or both of the operators satisfying the ETH ansatz.

Non-Markovian open quantum systems. We would like to connect our unitarity-restoring random matrix contractions with non-Markovian corrections to the Lindblad master equation. The radiation subsystem is an open quantum system, and Lindbladian evolution is a good approximation when the dynamics are Markovian. If the Lindblad operators are Hermitian, then the purity of the state decreases monotonically and the long-time average is maximally mixed [56]. Markovian dynamics are inconsistent with a unitary Page curve, and it would be interesting to see how and when non-Markovian corrections appear in our present and future models. One approach may be to extend the Lindblad equation to the dynamics of replicated open systems.

Factorisation problem. At the heart of both the factorisation problem and the replica wormhole story are Euclidean saddles that connect copies of gravitational systems. The factorisation problem is a tension between the calculation of $Z(\beta)^n$ in gravitational theories, which in some examples does not factorise due to connected geometries in the gravitational path integral, and the factorisation predicted from the boundary dual being a single CFT theory. The wormholes that connect multiple boundaries have a statistical interpretation, they quantify the correlations in the classical statistical ensemble which corresponds to semi-classical gravity [16–22]. The contribution of the wormholes can heuristically be expressed as

$$\int dH \mu(H) \left(\operatorname{Tr} e^{-\beta H}\right)^n - \left(\int dH \mu(H) \operatorname{Tr} e^{-\beta H}\right)^n , \tag{4.1}$$

where one integrates over all Hamiltonians which are semi-classically indistinguishable. It is interesting to notice that our toy model strongly suggests that replica wormholes appear to have a similar statistical origin. It would be interesting to pursue this analogy further and to explore to what extent the various issues associated with multi-boundary wormholes have a replica wormhole counterpart.

JHEP07(2024)023

Acknowledgments

We thank Ramesh Ammanamanchi, Luis Apolo, Igal Arav, Vladimir Gritsev, Diego Liška, Dominik Neuenfeld, Boris Post, Mark van Raamsdonk, Kamran Salehi Vaziri for useful discussions.

The work of JH is supported by the Dutch Black Hole Consortium with project number NWA.1292.19.202 of the research programme NWA which is (partly) financed by the Dutch Research Council (NWO). The work of AR is supported by the Stichting Nederlandse Wetenschappelijk Onderzoek Instituten (NWO-I) through the Scanning New Horizons project, by the Uitvoeringsinstituut Werknemersverzekeringen (UWV), by FWO-Vlaanderen project G012222N, and by the Vrije Universiteit Brussel through the Strategic Research Program High-Energy Physics. JdB is supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013), ERC Grant agreement ADG 834878.

A Entanglement spectrum probability distribution

The entanglement spectrum of our time evolved state, for a given draw of a random Hamiltonian, is specified by a single real number

$$|\lambda|^2 = (e^{iHt})_{11}(e^{-iHt})_{11}, \tag{A.1}$$

where H is a random matrix in the GUE ensemble; this simplicity of the entanglement spectrum is manifest in the reduced matrix on A which can be written as

$$\rho_{\mathsf{A}}(t) = |\lambda|^2 \rho_{\mathsf{A}}(0) + (1 - |\lambda|^2)\rho_{\mathsf{A},\text{vac.}}\,. \tag{A.2}$$

The Renyi entropies of A and B are simple functions of $|\lambda|^2$, so if we knew the probability distribution function (PDF) of $|\lambda|^2$ then we could calculate quantities like $\overline{S^{(n)}_{\mathsf{B}}(t)}$ directly by integrating over the PDF.

The PDF we want is

$$\begin{aligned} P(|\lambda|^2) &= \int d\mu_{\text{GUE}}(H)\delta\left(|\lambda|^2 - (e^{iHt})_{11}(e^{-iHt})_{11}\right) \\ &= \int d\mu(\vec{\lambda})d\mu(U)\delta\left(|\lambda|^2 - \left|\sum_i |U_{1i}|^2 e^{i\lambda_i t}\right|^2\right). \end{aligned} \tag{A.3}$$

In the second line, we have split the integral: $\mu(U)$ is the Haar measure on $\mathrm{U}(N)$ and $\mu(\lambda)$ is the joint PDF on eigenvalues of H.

While we can explicitly evaluate (A.3) for low values of N, we were not able to for general N and we have soft evidence that the problem isn't tractable: if we consider the apparently easier calculation that ignores the unitary matrix elements and integrals in (A.3), then we are calculating

$$P\left(\left|\sum_i e^{i\lambda_i t}\right|^2\right). \tag{A.4}$$

JHEP07(2024)023

We recognise the argument of this PDF as the spectral form factor (SFF). To our knowledge, the PDF of the SFF for the GUE, or other standard random ensembles, is not known; the mean and the variance are known [57], but not the full PDF. It would be of interest to the RMT community to pursue the calculation of the PDF of the SFF for Gaussian ensembles further.

B Unitary integrals and Weingarten functions

The unitary integrals can be evaluated using the Weingarten functions:

$$\int_{U_N} U_{i_1 j_1} \dots U_{i_q j_q} U^*_{i'_1 j'_1} \dots U^*_{i'_q j'_q} = \sum_{\sigma,\tau \in S_q} \delta_{i_1 i'_{\sigma(1)}} \dots \delta_{i_q i'_{\sigma(q)}} \delta_{j_1 j'_{\tau(1)}} \dots \delta_{j_q j'_{\tau(q)}} \mathsf{Wg}(\sigma\tau^{-1}, N). \tag{B.1}$$

The Weingarten function $\mathsf{Wg}(\sigma\tau^{-1}, N)$ only depends on the conjugacy class of $\sigma\tau^{-1}$ and N. We have listed below the values for $1 \le q \le 4$ [58]:

$$\begin{aligned}
q = 1: \quad & \mathsf{Wg}([1], N) = \frac{1}{N}, \\
q = 2: \quad & \mathsf{Wg}([1,1], N) = \frac{1}{N^2-1}, \\
& \mathsf{Wg}([2], N) = \frac{-1}{N(N^2-1)}, \\
q = 3: \quad & \mathsf{Wg}([1,1,1], N) = \frac{N^2-2}{N(N^2-1)(N^2-4)}, \\
& \mathsf{Wg}([2,1], N) = \frac{-1}{(N^2-1)(N^2-4)}, \\
& \mathsf{Wg}([3], N) = \frac{2}{N(N^2-1)(N^2-4)}, \\
q = 4: \quad & \mathsf{Wg}([1,1,1,1], N) = \frac{N^4-8N^2+6}{N^2(N^2-1)(N^2-4)(N^2-9)}, \\
& \mathsf{Wg}([2,1,1], N) = \frac{-1}{N(N^2-1)(N^2-9)}, \\
& \mathsf{Wg}([2,2], N) = \frac{N^2+6}{N^2(N^2-1)(N^2-4)(N^2-9)}, \\
& \mathsf{Wg}([3,1], N) = \frac{2N^2-3}{N^2(N^2-1)(N^2-4)(N^2-9)}, \\
& \mathsf{Wg}([4], N) = \frac{-5}{N(N^2-1)(N^2-4)(N^2-9)}.
\end{aligned} \tag{B.2}$$

For large N, the Weingarten functions have the asymptotic expression

$$\mathsf{Wg}(\sigma\tau^{-1}, N) = N^{-q-|\sigma\tau^{-1}|} \prod_i \left((-1)^{C_i - 1} c_{C_i - 1} \right) + O(N^{-q-|\sigma\tau^{-1}|-2}). \tag{B.3}$$

Here $|\sigma\tau^{-1}|$ is the minimal number of transpositions needed to write $\sigma\tau^{-1}$, $\sigma\tau^{-1}$ contains cycles of length C_i and c_n is a Catalan number:

$$c_n = \frac{(2n)!}{n!(n+1)!}. \tag{B.4}$$

JHEP07(2024)023

C Computation: higher Rényi entropies

We will work in the large N approximation. We will consider fixed n, such that combinatorial factors cannot cancel the $1/N$ suppression. We will compute the higher Rényi entropies, defined as

$$\overline{S^{(n)}(\rho_{\mathsf{B}})} = \frac{\log \overline{\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)}}{1-n}. \tag{C.1}$$

Note that we take the average before computing the logarithm. We are thus computing the annealed instead of the quenched Rényi entropies. The expression for $\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)$ reads

$$\begin{aligned}\mathrm{Tr}\,(\rho_{\mathsf{B}}^n) = \sum_{k_1=1}^{N}\cdots\sum_{k_{2n}=1}^{N}\sum_{b_1=1}^{N}\cdots\sum_{b_n=1}^{N}\left(\prod_{m=1}^{n}\delta_{1b_m}+\prod_{m=1}^{n}(1-\delta_{1b_m})\right)\times\\ \times\prod_{i=1}^{n}e^{i(\lambda_{k_{2i-1}}-\lambda_{k_{2i}})t}U_{b_ik_{2i-1}}U^{\dagger}_{k_{2i-1}1}U_{1k_{2i}}U^{\dagger}_{k_{2i}b_{i+1}},\end{aligned} \tag{C.2}$$

where $b_{n+1}=b_1$. We will compute the leading order contribution to the Rényi entropies for both the first term, with $b_i=1$ for all i, and the second term, with $2\le b_i\le N$ for all i, separately. It will be useful to consider what the maximal contribution from the separate components is. The maximum contribution from the Weingarten functions is $\sim N^{-2n}$,[16] where we used (B.3) with $q=2n$ and $|\sigma\tau^{-1}|\ge 0$. The maximal contribution from the b_i's is when they are all contracted to themselves, giving a contribution $\sim N^n$. The maximal contribution from the $2n$ different k_i's is when they are all contracted to themselves, giving $\sim W^n\sim N^{2n}$.

C.1 $b_i = 1$

For the first term in (C.2) all b_i are set equal to 1. Since the Weingarten function is of order N^{-2n} or lower, we need a contribution of at least order $\sim N^{2n}$ from the k_i's for the contraction to survive in the large N limit. There is exactly one k_i-contraction pattern that gives this contribution, and that is when all the k_i's are contracted to themselves. Writing the unitaries in the notation of (B.1), we have

$$\begin{aligned}U_{b_ik_{2i-1}} &= U_{i_{2i-1}j_{2i-1}}\,, & U^{*}_{1k_{2i-1}} &= U^{*}_{i'_{2i-1}j'_{2i-1}}\,,\\ U_{1k_{2i}} &= U_{i_{2i}j_{2i}}\,, & U^{*}_{b_{i+1}k_{2i}} &= U^{*}_{i'_{2i}j'_{2i}}\,,\end{aligned} \tag{C.3}$$

where we set $b_i=b_{i+1}=1$ here. The only contraction pattern that survives the large N limit therefore is the one with $\tau=e$. In order to not pick up any further suppression from the Weingarten function, we need $\sigma\tau^{-1}=e$ and therefore the only possibility is $\sigma=e$. The corresponding contribution is, in the large N approximation,

$$\overline{\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)}\Big|_{b_i=1} = \overline{g^n}. \tag{C.4}$$

C.2 $b_i = 2,\ldots,N$

In this case, any of the b_i's can take $N-1$ values. The maximum contribution this gives is or order $\sim N^n$, when every b_i is contracted with itself. The maximum contribution from the k_i's however, is no longer $\sim N^{2n}$; we cannot contract any k_i with itself without making a

[16] By the symbol $\sim$ we mean here that this is the leading order in N.

JHEP07(2024)023

cross-contraction. As can be seen from (C.3), without a cross-contraction the corresponding b would be contracted with 1. This is the one value b is not allowed to have for the second term in (2.45). We thus have to take cross-contractions, where each cross-contraction involves a suppression by at least one factor of N. Then we would get a contribution (from two pairs of k's contracted to themselves) of order at most $\sim W/N = gN$, which at early times is of order N. If we don't contract a k_i with itself but pair it with another k_j, together it will contribute a factor N to the sum. Therefore, the maximum contribution from all k's together can be of order N^n. We can conclude that we need all of the b_i's to be contracted with themselves to be able to remain leading order.

Let us now examine the suppression of one of these cross-contractions. Writing again the unitaries in the notation of (B.1), we have

$$\begin{aligned} U_{1k_{2i}} &= U_{i_{2i}j_{2i}}\,, & U^*_{b_{i+1}k_{2i}} &= U^*_{i'_{2i}j'_{2i}}\,, \\ U_{b_{i+1}k_{2i+1}} &= U_{i_{2i+1}j_{2i+1}}\,, & U^*_{1k_{2i+1}} &= U^*_{i'_{2i+1}j'_{2i+1}}\,. \end{aligned} \tag{C.5}$$

Contracting b_{i+1} with itself corresponds to $\sigma(2i+1) = 2i$. If we want to contract k_{2i+1} with itself, we read off $\tau(2i+1) = 2i+1$. From here, we already see we will need at least one transposition in $\sigma\tau^{-1}$ because σ and τ act on $2i+1$ differently. Minimising the number of transpositions in $\sigma\tau^{-1}$, we see we thus have to take $\tau(2i) = 2i$ and $\sigma(2i) = 2i+1$. Then $\sigma\tau^{-1}$ has one extra transposition. After summing over the free variables with the exponential in front, this contributes a factor of $\frac{-W}{N} = -g$.

If we were to not contract k_{2i+1} with itself, we minimise the number of transpositions by taking $\tau(2i+1) = 2i$. By doing so, we have contracted k_{2i+1} with k_{2i}. To introduce no further contractions, we thus take $\sigma, \tau(2i) = 2i+1$. This means we have introduced no further transpositions to $\sigma\tau^{-1}$ and this term contributes N to the sum.

We thus see that for $b_i = 2, \ldots, N$, all leading order contributions have $\sigma = \prod_{i=1}^{n} \chi_i$ where χ_i is the transposition that contracts b_i with itself:

$$\chi_i = (2i-2, 2i-1). \tag{C.6}$$

Here $\chi_1 = (0,1) \equiv (1, 2n)$. Including χ_i in τ is equivalent to contracting k_{2i+1} and k_{2i} with themselves. For each χ_i, we can choose whether or not to include it in τ. In total, we thus have 2^n contraction patterns that contribute to leading order to $\overline{\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)}$ when $b_i = 2, \ldots, N$. We note that all of these contraction patterns couple different copies ρ_{B} to each other. Summing these contributions, we find that in the large N approximation,

$$\overline{\mathrm{Tr}\,(\rho_{\mathsf{B}}^n)}\Big|_{b_i=2,\ldots,N} = \sum_{\ell=0}^{n} \binom{n}{\ell} \overline{(-g)^\ell} = \overline{(1-g)^n}. \tag{C.7}$$

In total, there are $((2n)!)^2$ contraction patterns that contribute to the n'th Rényi entropy. However, because of the large N limit, only 2^n+1 of them contribute to leading order. In total, the n'th Rényi entropy of the environment for $n \geq 2$ in the large N approximation is given by

$$\overline{S(\rho_{\mathsf{B}})^{(n)}} = \frac{\log\left(\overline{(1-g)^n} + \overline{g^n}\right)}{1-n}. \tag{C.8}$$

JHEP07(2024)023

References

[1] S.W. Hawking, *Particle Creation by Black Holes*, *Commun. Math. Phys.* **43** (1975) 199 [*Erratum ibid.* **46** (1976) 206] [INSPIRE].

[2] S.W. Hawking, *Breakdown of Predictability in Gravitational Collapse*, *Phys. Rev. D* **14** (1976) 2460 [INSPIRE].

[3] D.N. Page, *Information in black hole radiation*, *Phys. Rev. Lett.* **71** (1993) 3743 [hep-th/9306083] [INSPIRE].

[4] D.N. Page, *Average entropy of a subsystem*, *Phys. Rev. Lett.* **71** (1993) 1291 [gr-qc/9305007] [INSPIRE].

[5] D.N. Page, *Time Dependence of Hawking Radiation Entropy*, *JCAP* **09** (2013) 028 [arXiv:1301.4995] [INSPIRE].

[6] G. Penington, *Entanglement Wedge Reconstruction and the Information Paradox*, *JHEP* **09** (2020) 002 [arXiv:1905.08255] [INSPIRE].

[7] A. Almheiri, N. Engelhardt, D. Marolf and H. Maxfield, *The entropy of bulk quantum fields and the entanglement wedge of an evaporating black hole*, *JHEP* **12** (2019) 063 [arXiv:1905.08762] [INSPIRE].

[8] A. Almheiri, R. Mahajan, J. Maldacena and Y. Zhao, *The Page curve of Hawking radiation from semiclassical geometry*, *JHEP* **03** (2020) 149 [arXiv:1908.10996] [INSPIRE].

[9] S. Ryu and T. Takayanagi, *Holographic derivation of entanglement entropy from AdS/CFT*, *Phys. Rev. Lett.* **96** (2006) 181602 [hep-th/0603001] [INSPIRE].

[10] V.E. Hubeny, M. Rangamani and T. Takayanagi, *A covariant holographic entanglement entropy proposal*, *JHEP* **07** (2007) 062 [arXiv:0705.0016] [INSPIRE].

[11] T. Faulkner, A. Lewkowycz and J. Maldacena, *Quantum corrections to holographic entanglement entropy*, *JHEP* **11** (2013) 074 [arXiv:1307.2892] [INSPIRE].

[12] N. Engelhardt and A.C. Wall, *Quantum Extremal Surfaces: Holographic Entanglement Entropy beyond the Classical Regime*, *JHEP* **01** (2015) 073 [arXiv:1408.3203] [INSPIRE].

[13] A. Rolph, *Quantum bit threads*, *SciPost Phys.* **14** (2023) 097 [arXiv:2105.08072] [INSPIRE].

[14] A. Almheiri et al., *Replica Wormholes and the Entropy of Hawking Radiation*, *JHEP* **05** (2020) 013 [arXiv:1911.12333] [INSPIRE].

[15] G. Penington, S.H. Shenker, D. Stanford and Z. Yang, *Replica wormholes and the black hole interior*, *JHEP* **03** (2022) 205 [arXiv:1911.11977] [INSPIRE].

[16] A. Belin and J. de Boer, *Random statistics of OPE coefficients and Euclidean wormholes*, *Class. Quant. Grav.* **38** (2021) 164001 [arXiv:2006.05499] [INSPIRE].

[17] A. Belin, J. de Boer, P. Nayak and J. Sonner, *Generalized spectral form factors and the statistics of heavy operators*, *JHEP* **11** (2022) 145 [arXiv:2111.06373] [INSPIRE].

[18] A. Belin, J. de Boer and D. Liška, *Non-Gaussianities in the statistical distribution of heavy OPE coefficients and wormholes*, *JHEP* **06** (2022) 116 [arXiv:2110.14649] [INSPIRE].

JHEP07(2024)023

[19] T. Anous, A. Belin, J. de Boer and D. Liška, *OPE statistics from higher-point crossing*, *JHEP* **06** (2022) 102 [arXiv:2112.09143] [INSPIRE].

[20] A. Belin et al., *Approximate CFTs and Random Tensor Models*, arXiv:2308.03829 [INSPIRE].

[21] J. de Boer, D. Liška, B. Post and M. Sasieta, *A principle of maximum ignorance for semiclassical gravity*, *JHEP* **02** (2024) 003 [arXiv:2311.08132] [INSPIRE].

[22] J. Pollack, M. Rozali, J. Sully and D. Wakeham, *Eigenstate Thermalization and Disorder Averaging in Gravity*, *Phys. Rev. Lett.* **125** (2020) 021601 [arXiv:2002.02971] [INSPIRE].

[23] D. Marolf, *The black Hole information problem: past, present, and future*, *Rept. Prog. Phys.* **80** (2017) 092001 [arXiv:1703.02143] [INSPIRE].

[24] M.V. Berry, *Semiclassical Theory of Spectral Rigidity*, *Proc. Roy. Soc. Lond. A* **400** (1985) 229.

[25] A. Altland and M.R. Zirnbauer, *Nonstandard symmetry classes in mesoscopic normal-superconducting hybrid structures*, *Phys. Rev. B* **55** (1997) 1142 [cond-mat/9602137] [INSPIRE].

[26] J.M. Deutsch, *Quantum statistical mechanics in a closed system*, *Phys. Rev. A* **43** (1991) 2046 [INSPIRE].

[27] M. Srednicki, *Chaos and Quantum Thermalization*, *Phys. Rev. E* **50** (1994) 888 [cond-mat/9403051] [INSPIRE].

[28] L. D'Alessio, Y. Kafri, A. Polkovnikov and M. Rigol, *From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics*, *Adv. Phys.* **65** (2016) 239 [arXiv:1509.06411] [INSPIRE].

[29] E.P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, *Annals Math.* **62** (1955) 548.

[30] E.P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions II*, *Annals Math.* **65** (1957) 203.

[31] E.P. Wigner, *On the distribution of the roots of certain symmetric matrices*, *Annals Math.* **67** (1958) 325.

[32] F.J. Dyson, *Statistical theory of the energy levels of complex systems. I*, *J. Math. Phys.* **3** (1962) 140 [INSPIRE].

[33] O. Bohigas, M.J. Giannoni and C. Schmit, *Characterization of chaotic quantum spectra and universality of level fluctuation laws*, *Phys. Rev. Lett.* **52** (1984) 1 [INSPIRE].

[34] J.M. Deutsch, *Thermodynamic entropy of a many-body energy eigenstate*, *New J. Phys.* **12** (2010) 075021.

[35] C. Murthy and M. Srednicki, *Structure of chaotic eigenstates and their entanglement entropy*, *Phys. Rev. E* **100** (2019) 022131 [arXiv:1906.04295] [INSPIRE].

[36] J. Cotler, N. Hunter-Jones, J. Liu and B. Yoshida, *Chaos, Complexity, and Random Matrices*, *JHEP* **11** (2017) 048 [arXiv:1706.05400] [INSPIRE].

[37] R.E. Prange, *The Spectral Form Factor Is Not Self-Averaging*, *Phys. Rev. Lett.* **78** (1997) 2280 [INSPIRE].

[38] P. Hayden and J. Preskill, *Black holes as mirrors: Quantum information in random subsystems*, *JHEP* **09** (2007) 120 [arXiv:0708.4025] [INSPIRE].

[39] S. Vijay and A. Vishwanath, *Finite-Temperature Scrambling of a Random Hamiltonian*, arXiv:1803.08483 [INSPIRE].

JHEP07(2024)023

[40] S.D. Mathur and C.J. Plumberg, *Correlations in Hawking radiation and the infall problem*, *JHEP* **09** (2011) 093 [arXiv:1101.4899] [INSPIRE].

[41] P. Dadras and A. Kitaev, *Perturbative calculations of entanglement entropy*, *JHEP* **03** (2021) 198 [*Erratum ibid.* **10** (2022) 201] [arXiv:2011.09622] [INSPIRE].

[42] P. Zhang, *Perturbative Page curve induced by external impulse*, *JHEP* **09** (2023) 056 [arXiv:2305.18329] [INSPIRE].

[43] K. Su, P. Zhang and H. Zhai, *Page curve from non-Markovianity*, *JHEP* **06** (2020) 156 [arXiv:2101.11238] [INSPIRE].

[44] H. Wang, C. Liu, P. Zhang and A.M. García-García, *Entanglement transition and replica wormholes in the dissipative Sachdev-Ye-Kitaev model*, *Phys. Rev. D* **109** (2024) 046005 [arXiv:2306.12571] [INSPIRE].

[45] L. Piroli, C. Sünderhauf and X.-L. Qi, *A Random Unitary Circuit Model for Black Hole Evaporation*, *JHEP* **04** (2020) 063 [arXiv:2002.09236] [INSPIRE].

[46] B. Yoshida, *Soft mode and interior operator in the Hayden-Preskill thought experiment*, *Phys. Rev. D* **100** (2019) 086001 [arXiv:1812.07353] [INSPIRE].

[47] H. Liu and S. Vardhan, *A dynamical mechanism for the Page curve from quantum chaos*, *JHEP* **03** (2021) 088 [arXiv:2002.05734] [INSPIRE].

[48] H. Liu and S. Vardhan, *Entanglement Entropies of Equilibrated Pure States in Quantum Many-Body Systems and Gravity*, *PRX Quantum* **2** (2021) 010344 [arXiv:2008.01089] [INSPIRE].

[49] C. Krishnan and V. Mohan, *Hints of gravitational ergodicity: Berry's ensemble and the universality of the semi-classical Page curve*, *JHEP* **05** (2021) 126 [arXiv:2102.07703] [INSPIRE].

[50] Vinayak and M. Žnidaric, *Subsystem dynamics under random Hamiltonian evolution*, arXiv:1107.6035 [DOI:10.1088/1751-8113/45/12/125204].

[51] D. Chernowitz and V. Gritsev, *Entanglement Dynamics of Random GUE Hamiltonians*, *SciPost Phys.* **10** (2021) 071 [arXiv:2001.00140] [INSPIRE].

[52] Y.-Z. You and Y. Gu, *Entanglement Features of Random Hamiltonian Dynamics*, *Phys. Rev. B* **98** (2018) 014309 [arXiv:1803.10425] [INSPIRE].

[53] A. Altland and J. Sonner, *Late time physics of holographic quantum chaos*, *SciPost Phys.* **11** (2021) 034 [arXiv:2008.02271] [INSPIRE].

[54] J. Chandra, S. Collier, T. Hartman and A. Maloney, *Semiclassical 3D gravity as an average of large-c CFTs*, *JHEP* **12** (2022) 069 [arXiv:2203.06511] [INSPIRE].

[55] J. Cotler and K. Jensen, *AdS_3 gravity and random CFT*, *JHEP* **04** (2021) 033 [arXiv:2006.08648] [INSPIRE].

[56] D. Manzano, *A short introduction to the Lindblad master equation*, *AIP Adv.* **10** (2020) 025106 [INSPIRE].

[57] J. Liu, *Spectral form factors and late time quantum chaos*, *Phys. Rev. D* **98** (2018) 086026 [arXiv:1806.05316] [INSPIRE].

[58] A. Mironov, A. Morozov and G.W. Semenoff, *Unitary matrix integrals in the framework of generalized Kontsevich model. I. Brezin-Gross-Witten model*, *Int. J. Mod. Phys. A* **11** (1996) 5031 [hep-th/9404005] [INSPIRE].

JHEP07(2024)023