Generalized free cumulants for quantum chaotic systems

Siddharth Jindal 🆔 **and Pavan Hosur** 🆔

Department of Physics and Texas Center for Superconductivity, University of Houston,
4302 University Dr, Houston, TX 77004, U.S.A.

E-mail: siddharth@sidjindal.dev, phosur@uh.edu

ABSTRACT: The eigenstate thermalization hypothesis (ETH) is the leading conjecture for the emergence of statistical mechanics in generic isolated quantum systems and is formulated in terms of the matrix elements of operators. An analog known as the ergodic bipartition (EB) describes entanglement and locality and is formulated in terms of the components of eigenstates. In this paper, we significantly generalize the EB and unify it with the ETH, extending the EB to study higher correlations and systems out of equilibrium. Our main result is a diagrammatic formalism that computes arbitrary correlations between eigenstates and operators based on a recently uncovered connection between the ETH and free probability theory. We refer to the connected components of our diagrams as generalized free cumulants. We apply our formalism in several ways. First, we focus on chaotic eigenstates and establish the so-called subsystem ETH and the Page curve as consequences of our construction. We also improve known calculations for thermal reduced density matrices and comment on an inherently free probabilistic aspect of the replica approach to entanglement entropy previously noticed in a calculation for the Page curve of an evaporating black hole. Next, we turn to chaotic quantum dynamics and demonstrate the ETH as a sufficient mechanism for thermalization, in general. In particular, we show that reduced density matrices relax to their equilibrium form and that systems obey the Page curve at late times. We also demonstrate that the different phases of entanglement growth are encoded in higher correlations of the EB. Lastly, we examine the chaotic structure of eigenstates and operators together and reveal previously overlooked correlations between them. Crucially, these correlations encode butterfly velocities, a well-known dynamical property of interacting quantum systems.

Contents

1 Introduction

There has been recent interest in the physics of thermalization in quantum many-body systems. Thermalization was historically established by Boltzmann's ergodic hypothesis, which states that systems uniformly sample the entire phase that is consistent with their macroscopic

symmetries [1]. Despite its remarkable predictive power, the ergodic hypothesis is manifestly inconsistent with unitarity. A similar issue arises in the study of the black hole information paradox, where general relativity appears inconsistent with unitarity near an event horizon [2]. Nonetheless, quantum systems do thermalize [3–6], and understanding how they thermalize has led to a deeper understanding of quantum and statistical mechanics [7, 8]. Since the field of quantum chaos was born partly to resolve the inconsistency between quantum and statistical mechanics, it is fitting that it has returned to prominence in attempts to solve the inconsistency between quantum mechanics and general relativity [9, 10].

Roughly 30 years ago, Deutsch and Srednicki introduced the eigenstate thermalization hypothesis (ETH) to justify the emergence of statistical mechanics for generic isolated quantum systems [4, 5]. Since then, the ETH has been implicated in a variety of physical phenomena including holography [11], quantum error correction [12], scrambling [13, 14], and transport [15]. Furthermore, the failure of the ETH has led to the discovery of a fascinating phase of matter known as many-body localization [16]. Six years ago, motivated by contemporaneous work in the string theory community [17], Foini and Kurchan reformulated the ETH to account for higher order correlations [13], and last year, along with Pappalardi, reinterpreted their results in terms of free probability theory [18].

The ETH purports that the emergence of statistical mechanics and thermalization in isolated quantum systems is the result of pseudorandomness in the matrix elements of operators. Formally [13],

$$\overline{X_{i_1 i_2} \cdots X_{i_n i_1}} = e^{-(n-1)S(\bar{E})} f_n(\bar{E}; \vec{\omega}), \quad i_p \neq i_q \text{ for } p \neq q, \tag{1.1}$$

where X is a local operator,[1] $\{|i_m\rangle\}$ are eigenstates of a chaotic Hamiltonian, S is the microcanonical entropy, f is an $\mathcal{O}(1)$ spectral function, and the overline denotes arithmetic averaging over a narrow energy band. More than justifying the results of equilibrium statistical mechanics, the ETH ensures that thermalization occurs in real time. Though systems which obey the ETH are considered *chaotic*, it is not clear how the ETH interacts with other principles of quantum chaos. For instance, eq. (1.1) does not fundamentally harbor a notion of locality except via an implicit but vague restriction that it only applies to "simple" operators. It also does not obviously capture the non-local entanglement structure of the Page curve.

A recent, state-based idea that addresses this issue is the ergodic bipartition (EB), an ansatz on the pseudorandom structure of chaotic eigenstates when split over two subsystems. Formally [19–21],

$$\langle i|ab\rangle \equiv c_{ab}^i, \quad \overline{|c_{ab}^i|^2} = e^{-S(E_i)} F(E_i - E_a - E_b) \tag{1.2}$$

where $|i\rangle$ is an eigenstate of the full Hamiltonian $H = H_A + H_B + H_{AB}$, $|ab\rangle$ is a product of eigenstates of subsystem Hamiltonians H_A and H_B, and H_{AB} is the interacting term that couples subsystems A and B. $e^{S(E)}$ is the density of states of the system at energy E and F is a window function that ensures $E_i \approx E_a + E_b$. The moments of F are roughly those of H_{AB},

---

[1]One may consider correlations between n distinct operators, too. For any sequence of operators or "word", W, constructed from individual operators or "letters", there is a unique associated f_W, up to cyclic permutations of the letters. For example, for operators X, Y, Z and word $YZYX$, we have $\overline{Y_{i_1 i_2} Z_{i_2 i_3} Y_{i_3 i_4} X_{i_4 i_1}} = e^{-3S(\bar{E})} f_{YZYX}(\bar{E}; \vec{\omega})$ for distinct i_1, i_2, i_3, i_4.

$\int_\omega F(\omega)\omega^n \approx \langle H_{AB}^n \rangle_i$ (see appendix C.2). Therefore, if H_{AB} lives on the boundary between A and B, it may be regarded as a subextensive perturbation and F is sharply peaked in its argument. The EB implies that the reduced density matrices of eigenstates are consistent with those of the microcanonical ensemble and obey a Page curve. Additionally, the tensor product structure allows it to distinguish local systems from nonlocal ones. Structurally, the EB is formulated in analogy with the ETH, but presently is only capable of computing static quantities for systems prepared in eigenstates.

Both the ETH and the EB are avatars of Berry's conjecture [22]: the hypothesis that most eigenfunctions of chaotic potentials are, in essence, Gaussian random waves. Berry's conjecture is predated by Von Neumann's quantum ergodic theorem [3], which states that the overwhelming majority of pure states in a many-body system approximate the local properties of microcanonical ensemble arbitrarily well. It is then natural to conjecture that the eigenstates of many-body systems also retain the properties of the microcanonical ensemble, if they can be treated as random waves. In this light, Berry's conjecture would be a manifestly quantum version of Boltzmann's ergodic hypothesis where uniform sampling of phase space has been supplanted by random vectors.

In systems with few degrees of freedom, not all eigenstates can be treated as random [22]. However, as one takes the thermodynamic limit, $\mathcal{V} \to \infty$, provided it is well-defined, two things happen simultaneously: (1) the level spacing of the system vanishes exponentially fast $\sim \mathcal{O}(e^{-S})$ and (2) the local physics of the system become insensitive to microscopic perturbations[2] $\ll \mathcal{O}(1)$ [23]. As such, any microscopic perturbation $\ll \mathcal{O}(1)$ added to the system can mix an exponentially large number of eigenstates $\sim \mathcal{O}(e^S)$ without modifying any physics. Thus, if the thermodynamic limit is to be well-defined, *all* pure states in a microscopically small energy window should have identical physical properties [15, 24].[3] This line of reasoning underlies early work on chaos in many-body systems [4, 26] and, more recently, investigations into *emergent rotational symmetry* at small frequencies, that is, an invariance of physics to arbitrary norm-preserving[4] linear transformations of eigenstates that are sufficiently close in energy [27, 28]. From it, we can conjecture that eigenstates of thermodynamically large systems retain the properties of random vectors and of the microcanonical ensemble.

In this paper, we formulate a many-body Berry's conjecture (MBBC): the hypothesis that chaotic eigenstates are essentially random vectors up to the symmetry constraints of the system, and show that it unifies the ETH and the EB. To our knowledge, this term was first used in ref. [20], which studied entanglement entropies of chaotic eigenstates. We show that free probability theory naturally emerges from the MBBC and build on this insight to develop a diagrammatic formalism for calculating general correlations of eigenstates and operators, greatly generalizing the EB, in particular. With this formalism we are able to justify or

---

[2]This cannot strictly be true for eigenstates sufficiently close to the ground state of the system as both conditions are violated. An arbitrarily small relevant perturbation can modify IR physics and the density states is not exponentially large.

[3]In a system with multiple conservation laws, or in a symmetry broken phase [25], mixing states with distinct, non-energy, quantum numbers would generically require a perturbation that is $\sim \mathcal{O}(1)$ and capable of modifying some local physics. In *integrable* systems, which contain an extensive number of conservation laws, the number of states in the microscopic window that can be mixed by an integrability-preserving perturbation is not exponentially large, and the ETH cannot hold in the strong form we present [23].

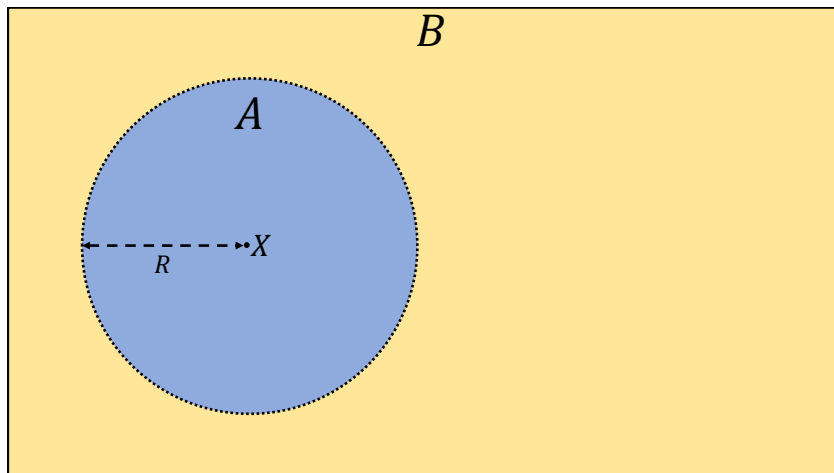[4]I.e. orthogonal, unitary, or symplectic.

Figure 1. The system is split into two subsystems A and B. An operator X lives within A, a distance R from the boundary.

improve several known results for chaotic eigenstates on a common footing. We show that chaotic eigenstates have thermal reduced density matrices consistent with the subsystem ETH conjecture and Von Neumann entanglement entropies consistent with the Page curve. We then show that both these properties are shared by non-equilibrium states at late times, indicating that chaotic systems thermalize in real time. We additionally study signatures of local dynamics: ballistic and diffusive entanglement growth and butterfly velocities.

In our setup, we consider a system split into two subsystems A and B and a local operator X living deep within A, see figure 1. The goal of doing so is to study the reduced density matrix on a single subsystem and the bipartite entanglement between the subsystems. However, our techniques should extend to any kind of partitioning of the system.

The outline of this paper is as follows. In section 2, we motivate our approach by studying the properties of random vectors without appealing to formal results of Haar measures or of free probability theory. The main result in this section, which illustrates the basic principle behind the MBBC, is that random vectors, and therefore chaotic eigenstates, have the same properties as equilibrium density matrices.

Section 3 is the technical backbone of the paper and contains our main results. It builds on section 2 to establish the framework of free probability theory and its connection to the EB and the ETH. Here, we introduce generalized free cumulants and describe the diagrammatic framework for computing arbitrary correlations between states and operators. This formalism is the main result of our paper. We compare our technique to analogous techniques in the literature and provide some numerical evidence for its validity. Furthermore, we connect the predictions of the MBBC to the predictions of rotationally invariant random matrix ensembles. We also discuss how the emergence of free probability in the ETH has unique consequences for the behavior of local observables over long times, connecting our observations to recent results in the literature. We argue that the ETH, eq. (1.1), the *quantum butterfly effect* characterized by the decay of *out-of-time-ordered correlators* (OTOCs) [17, 29], and the emergence of rotational symmetry at small frequencies [27, 28] are all, in essence, equivalent

definitions of chaos and avatars of free probability (we elaborate on these ideas in appendix A). To summarize, this section contains the mathematical and diagrammatic framework of free cumulants and a glimpse of implications that we explore in depth in the next two sections.

In section 4, we study the structure of chaotic eigenstates. Several previous works have been dedicated to this topic [19–21, 30], but our discussion is more than a review. We provide new derivations that show disparate results as consequences of the MBBC. First, in section 4.1, we show that the reduced density matrix of chaotic eigenstates takes the form,

$$\rho^i_{aa'} \equiv \langle a | \text{Tr}_B \left[|i\rangle \langle i| \right] |a'\rangle = e^{-S(E_i)+S_B(E_i-E_{aa'})} \left(\delta_{aa'} + \sqrt{e^{S_{\min}(E_{aa'})} \tilde{F}(E_{aa'};\omega_{aa'})} R_{aa'} \right) \quad (1.3)$$

where $E_{aa'} \equiv (E_a + E_{a'})/2$, $S_{\min} \equiv \min \left[S_A(E_{aa'}), S_B(E_i - E_{aa'}) \right]$, $\tilde{F}(E_{aa'}, \omega_{aa'})$ contains a $\omega_{aa'} \equiv E_a - E_{a'}$ dependence narrowly peaked around zero, and $R_{aa'}$ is a nearly Gaussian random matrix that encodes higher correlations in the density matrix. Eq. (1.3) improves the result of [21] which contained an unphysical suppression of off-diagonal elements and generalizes the *state-averaging ansatz* of ref. [31]. Next, we study the entanglement entropy of chaotic systems and show that we are able to reproduce the results of ref. [21] exactly. In particular, we show that the Von Neumann entropy obeys a Page curve,

$$S_1 = \min \left[S_A(E_{iA}), S_B(E_{iB}) \right] + \Delta S \quad (1.4)$$

where E_{iA} and E_{iB} are the microcanonically expected subsystem energies defined by $E_{iA} + E_{iB} = E_i$ and $S'_A(E_{iA}) = S'_B(E_{iB})$ and ΔS denotes subextensive corrections to the Page curve. Furthermore, we discuss an inherently free probabilistic structure in the replica calculation of entanglement entropy previously noticed in the gravitational path integral calculation of the Page curve for an evaporating black hole [32–34]. Finally, in section 4.3, we show that reduced density matrices of nearby eigenstates are exponentially close in trace distance, a hypothesis known as the subsystem ETH [35]. More precisely, we show,

$$\mathcal{O} \left(e^{S_A/2-S/2} \right) \lesssim ||\rho^i_A - \rho^j_A||_1 \lesssim \mathcal{O} \left(e^{S^{(\infty)}_{\min}/2+S_A/2-S/2} \right) \quad (1.5)$$

where $|| \cdots ||_1$ is the Schatten 1-norm[5] and $S^{(\infty)}_{\min}$ is the microcanonical entropy of the smaller subsystem at infinite temperature.[6]

In section 5, we then study a system prepared out-of-equilibrium with no entanglement between the subsystems. We first verify that all states with the same initial energy, E_0, relax towards that same equilibrium at late times, which shares its properties with nearby chaotic eigenstates.[7] Our approach reproduces the *equilibrated pure state* formalism of ref. [32]. Precisely, we show in section 5.1 that

$$\rho_{aa'}(t \to \infty) = e^{-S(E_0)+S_B(E_0-E_{aa'})} \left(\delta_{aa'} + \sqrt{e^{S_{\min}(E_{aa'})} \tilde{F}(E_{aa'};\omega_{aa'})} R_{aa'} \right) \quad (1.6)$$

---

[5]For an operator M, the Schatten p-norm is defined as $||M||_p \equiv \sqrt[p]{\sum_i |m_i|^p}$ where m_i are the eigenvalues or singular values of M.

[6]A stronger bound was derived in ref. [36] using similar methods.

[7]This statement holds for all properties except for fluctuations of approximately conserved operators, namely the fluctuations of subsystem energies. We discuss this caveat in appendix C.1 and thank Tarun Grover for bringing it to our attention.

and in section [5.2](#) that

$$S_1(t \to \infty) = \min\left[S_A(E_{0A}), S_B(E_{0B})\right] + \Delta S \tag{1.7}$$

where E_{0A}, E_{0B} are the microcanonical subsystem energies for total energy E_0. The key insight behind this result is that the time-independent partitions of the diagrams for time-dependent states map uniquely to those of eigenstates. In section [5.3](#), we study the different phases of entanglement growth [37–39] and find that distinct phases are encoded in distinct diagrams. Our results provide a novel organization to chaotic entanglement dynamics.

In section [6](#), we consider the correlations between operators and eigenstates. When an operator X lives deep within subsystem A, it may be natural to assume it becomes uncorrelated with c defined in eq. [(1.2)](#). In section [6.1](#), we show that this assumption strongly violates causality and that operator-state correlations are necessary for the ETH to apply in a system with many degrees of freedom. In essence, decorrelating the elements of X with those of c disconnects X from the time evolution operator e^{iHt} and trivializes the dynamics of the system. Thus, in section [6.2](#), we exploit causality to constrain operator-state correlations. For X living deep within subsystem A, time evolution under $H_A + H_B$ will be identical to that under H for times shorter than R/v_B, where v_B is the butterfly velocity of the system [40] and R is the distance to the boundary. Thus, local physics is necessarily encoded in eigenstate-operator correlations.

Some calculations and discussions in this paper are left to the appendices. Appendix [A](#) contains an informal introduction to the concepts of free probability theory we use in this paper and motivate why the subject plays a central role in quantum chaos. Appendix [B](#) provides an introduction to the use of saddle-point approximations for making thermal approximations with the ETH. Appendix [C](#) discuss the non-zero width of the F function that we mostly neglect in the body of this paper. Appendix [C.1](#) focuses on corrections to the energy fluctuations of subsystems that are a finite fraction of the whole system, which are able to distinguish different pure states with the same energy density, even in equilibrium. Appendix [C.2](#) discusses the form of F in eq. [(1.2)](#) in the context of previously conjectured Gaussian and Lorentzian forms for F [21, 30]. Our analysis complements appendix A of ref. [30]. Appendix [D](#) derives eq. [(1.1)](#) from the MBBC.

2 Many-body Berry's conjecture

Our fundamental postulate is that eigenstates of a chaotic Hamiltonian behave *for all intents and purposes* as random vectors up to the symmetry constraints of the system. This idea is what we refer to as the many-body Berry's conjecture and attempt to make precise in this section. We will assume conservation of energy and no other symmetries going forward. The purpose of this section is to develop the properties of random vectors while sidestepping a formal discussion of Haar measures or of free probability.

First, consider a d-dimensional Hilbert space and sample two normalized vectors randomly, $|1\rangle, |2\rangle$ from it. What is the expected value of their squared overlap, $\langle 1|2\rangle\langle 2|1\rangle$? An easy way to calculate this overlap is to rotate into an orthonormal basis, $\{|1'\rangle\}$, that contains $|1\rangle$. Then, $|2\rangle$, being chosen independently of $|1\rangle$, on average will have overlap $1/\sqrt{d}$ with

each basis element. Thus,

$$\overline{\langle 1|2\rangle\langle 2|1\rangle} = \frac{1}{d} \tag{2.1}$$

We may sample three vectors, $|1\rangle$, $|2\rangle$ and $|3\rangle$, and want to know the expected value of $\langle 1|2\rangle\langle 2|3\rangle\langle 3|1\rangle$. One may be forgiven for guessing that each overlap contributes a factor of $d^{-\frac{1}{2}}$ yielding $d^{-\frac{3}{2}}$, but that is incorrect. The *fluctuating* part of a single amplitude is indeed $d^{-\frac{3}{2}}$, but the noncommutativity of the projection operators $|i\rangle\langle i|$ ensures that the 3 overlaps are correlated. Instead, consider again summing over the entire basis that contains $|1\rangle$. This procedure corresponds to

$$\sum_{1'} \langle 1'|2\rangle\langle 2|3\rangle\langle 3|1'\rangle = \text{Tr}\left[|2\rangle\langle 2|3\rangle\langle 3|\right] = \langle 3|2\rangle\langle 2|3\rangle. \tag{2.2}$$

Linearity of averages implies

$$\text{Tr}\left[\overline{|2\rangle\langle 2|3\rangle\langle 3|}\right] = \overline{\langle 3|2\rangle\langle 2|3\rangle}. \tag{2.3}$$

Lastly, since the basis $\{|1'\rangle\}$ is essentially arbitrary,

$$\overline{\langle 1|2\rangle\langle 2|3\rangle\langle 3|1\rangle} = \frac{1}{d}\sum_{1'}\langle 1'|\overline{|2\rangle\langle 2|3\rangle\langle 3|}|1'\rangle = \frac{1}{d}\text{Tr}\left[\overline{|2\rangle\langle 2|3\rangle\langle 3|}\right] = \frac{1}{d}\overline{\langle 3|2\rangle\langle 2|3\rangle} = d^{-2}. \tag{2.4}$$

Following this inductive argument, we can now assert:

$$\overline{\langle 1|2\rangle\langle 2|\cdots|n\rangle\langle n|1\rangle} = d^{-(n-1)}. \tag{2.5}$$

Eq. (2.5) applies to vectors sampled uniformly over their Hilbert space. Note that the mean $d^{-(n-1)}$ is smaller than the fluctuating part $d^{-n/2}$ for $n > 2$.

We now want to understand how eigenstates of chaotic Hamiltonians can be understood as random vectors. First consider a pair of Hamiltonians: H_1 and $H_2 = H_1 + \lambda X$, where X is some subextensive perturbation. Physical arguments [4, 26] and numerical evidence [41] indicate that even for a very small[8] λ the eigenstates of H_2 are nearly orthogonal to those of H_1. Specifically, we expect that for eigenstates $|i_1\rangle$ of H_1 and $|i_2\rangle$ of H_2 with $E_{i_1} \approx E_{i_2}$,

$$\overline{\langle i_1|i_2\rangle\langle i_2|i_1\rangle} \sim e^{-S}. \tag{2.6}$$

Furthermore, even for $\lambda \sim \mathcal{O}(1)$, λX cannot mix eigenstates of H_1 that are very distant in energy. Our precise requirement is that λX is small enough that H_1 and H_2 retain the same entropy functions, $S_2(E) \approx S_1(E) \equiv S(E)$, which is expected to hold for any subextensive perturbation [30]. Lastly, X may have some nontrivial energy dependence. Thus we hypothesize,

$$\overline{\langle i_1|i_2\rangle\langle i_2|i_1\rangle} = e^{-S(\bar{E})}F(\bar{E};\omega), \tag{2.7}$$

---

[8]The precise expectation is that for $\lambda X \gtrsim \mathcal{O}(\mathcal{V}^{-\gamma})$, eigenstates of H_2 and H_1 within a small energy window are related essentially by a random rotation and cannot be computed from one another to any order in perturbation theory where $\mathcal{V}$ is the volume of the system, and γ is an exponent that depends on the transport properties of the system [15, 42, 43].

where $\overline{E} = \frac{1}{2}\left(E_{i_1} + E_{i_2}\right)$, $\vec{\omega} = E_{i_1} - E_{i_2}$, and F serves as a cutoff function that encodes the nontrivial energy dependences. Analogously, considering n distinct Hamiltonians $H_1, \ldots, H_n$ and an overlap of n respective eigenstates $|i_1\rangle, \ldots, |i_n\rangle$, we hypothesize,

$$\overline{\langle i_1|i_2\rangle\langle i_2|\cdots|i_n\rangle\langle i_n|i_1\rangle} = e^{-(n-1)S(\bar{E})}F(\bar{E};\vec{\omega}) \tag{2.8}$$

where $\bar{E} = \frac{1}{n}\sum_m E_m$ and $\vec{\omega} = (E_{i_1} - E_{i_2}, \cdots, E_{i_{n-1}} - E_{i_n})$. The crux of the MBBC is that the structure of eq. (2.8) holds for the eigenstates of any set of chaotic Hamiltonians that are neither extremely close to one another nor extremely far from one another. In section 3.1 we consider the situation where non-adjacent eigenstates in eq. (2.8) are taken from the same Hamiltonian.

To reiterate, we can understand eigenstates of physical Hamiltonians as random vectors by incorporating two constraints: mutual orthogonality and fixed energy. To handle the first constraint, we asserted that consecutive projection operators in eq. (2.8) are eigenstates of distinct, but similar, Hamiltonians. For the second constraint, the Hilbert space dimension, d was replaced with the density of states, e^S. Lastly, we inserted a cutoff function, $F(\bar{E};\vec{\omega})$, that ensures correlations are local in energy.

It is useful to formulate the above argument in terms of the density matrix of the space from which the eigenstate was sampled. Consider, again, the overlap of n random vectors sampled from the entire Hilbert space. The corresponding density matrix is simply the infinite temperature state, $\frac{1}{d}\mathbb{1}$. Our result for random vectors can be stated as

$$\overline{\langle 1|2\rangle\langle 2|\cdots|n\rangle\langle n|1\rangle} = \mathrm{Tr}\left[\left(\frac{1}{d}\mathbb{1}\right)^n\right] = d \cdot d^{-n} = d^{-(n-1)} \tag{2.9}$$

while for eigenstates,

$$\overline{\langle i_1|i_2\rangle\langle i_2|\cdots|i_n\rangle\langle i_n|i_1\rangle} \sim \mathrm{Tr}[\rho(\bar{E})^n] \sim e^{-(n-1)S(\bar{E})} \tag{2.10}$$

where $\rho(\bar{E})$ is the microcanonical density matrix with energy $\bar{E}$. Hence, narrow band averaging imbues eigenstates with the properties of corresponding equilibrium ensembles.

We can also consider subsystems. Let the Hilbert space be a tensor product of subspaces A and B with dimensions d_A and $d_B = d/d_A$, respectively. We want to know how to compute the expected value of an arbitrary overlap between vectors sampled on the subspaces and the full space. If the vectors are sampled uniformly, we will again assume they can be replaced with the corresponding density matrix. For example, consider two vectors from each space $|1\rangle$, $|2\rangle$, $|1_A\rangle$, $|2_A\rangle$, $|1_B\rangle$, $|2_B\rangle$ and the following overlap,[9]

$$\begin{aligned}
\overline{\langle 1|1_A 1_B\rangle\langle 1_A 2_B|2\rangle\langle 2|2_A 2_B\rangle\langle 2_A 1_B|1\rangle} &= d^{-2}\,\mathrm{Tr}\left[\mathbb{1}\,|1_A 1_B\rangle\langle 1_A 2_B|\,\mathbb{1}\,|2_A 2_B\rangle\langle 2_A 1_B|\right] \\
&= d^{-2}\overline{\langle 2_A|1_A\rangle\langle 1_A|2_A\rangle} \\
&= d^{-2}d_A^{-1}.
\end{aligned} \tag{2.11}$$

An analogous overlap of eigenstates is expected to have weight e^{-2S-S_A}. We clarify this generalization in section 3.4 and provide numerical evidence for it in section 3.6.

---

[9]An analogous calculation was performed in appendix B of ref. [35].

Remarkably, eq. (2.8) has the same form as eq. (1.1). The fact that both changes-of-basis and local observables obey the same correlated structure may seem surprising and raise the question of precisely which objects are amenable to such an analysis. The answer is that eigenstates are the natural object of study of the ETH, where local operators and change-of-basis rotations serve equally well as scramblers (see appendix D for details). We claim that one can mix any combination of chaotic eigenstates, density matrices, and observables and retain such an expression. The remaining task of the section is to relate this observation to thermalization.

We study an *n*-point *cumulant* of an operator defined as the time-dependent sum over a product of operators that neglects repeated indices:

$$\sum_{[i_2 \cdots i_n]} e^{i\vec{\omega} \cdot \vec{t}} X_{i_1 i_2} \cdots X_{i_n i_1}, \quad \vec{t} \equiv (t_1, \ldots, t_{n-1}), \quad \vec{\omega} \equiv \left(\omega_{i_1 i_2}, \ldots, \omega_{i_{n-1} i_n} \right) \tag{2.12}$$

where the braces $[\cdots]$ indicate that indices are not to be repeated in summation (i.e. $i_m \neq i_{m'}$). We will replace sums over sufficiently smooth correlations with integrals via the substitution

$$\sum_i \longrightarrow \int_{E_i} e^{S(E_i)}. \tag{2.13}$$

Then we replace the right-hand-side of eq. (2.12) with an integral as

$$
\begin{aligned}
\sum_{[i_2 \cdots i_n]} e^{i\vec{\omega} \cdot \vec{t}} X_{i_1 i_2} \cdots X_{i_n i_1} &= \sum_{[i_2 \cdots i_n]} e^{i\vec{\omega} \cdot \vec{t} - (n-1)S(\bar{E})} f(\bar{E}; \vec{\omega}) \\
&= \int_{E_{i_2} \cdots E_{i_n}} e^{i\vec{\omega} \cdot \vec{t} + \sum_{m=2}^{n}(S(E_{i_m}) - S(\bar{E}))} f(\bar{E}; \vec{\omega}) \\
&= \int_{\vec{\omega}} e^{(i\vec{t} - \beta \vec{l}) \cdot \vec{\omega}} f(E_{i_1}; \vec{\omega}), \quad \vec{l} = \left(\frac{n-1}{n}, \ldots, \frac{1}{n} \right) \\
&\equiv f(E_{i_1}; \vec{t} + i\beta \vec{l})
\end{aligned}
\tag{2.14}
$$

where $\beta \equiv S'(E_{i_1})$ is the thermodynamic temperature. To go from the second to the third line in eq. (2.14), we have utilized the approximation $S(E_{i_m}) - S(E_{i_{m+1}}) \approx \beta \omega_{i_m i_{m+1}}$, which holds so long as the heat capacity of the system is extensive (see appendix B) and we have approximated $f(\bar{E}; \vec{\omega}) = f(E_{i_1} - \vec{l} \cdot \vec{\omega}; \vec{\omega}) \approx f(E_{i_1}; \vec{\omega})$ since f is a slow function of the total energy.

It is clear that the time-dependent cumulant is simply a mixed Fourier-Laplace transform of f. For the case of $n = 1$, the right-hand side of eq. (2.14) simply reduces to the time-independent microcanonical expectation value of X. For $n > 1$, we instead have a complete set of thermal correlation functions which must decay in order for the system to equilibrate. These cumulants are more precisely operator *free cumulants*, which we discuss in more detail in section 3.2.

One should wonder when replacing sums with integrals over smooth components, as we did in eq. (2.14), is safe. For example, the fluctuating component of $X_{i_1 i_2} \cdots X_{i_n i_1}$ is larger than its smooth component, $e^{-nS/2} \not< e^{-(n-1)S}$. The answer is that it depends on the number of sums being performed. Let the number of sums being performed be n^*. While the smooth parts will grow with $e^{n^* S}$, the fluctuating parts are uncorrelated and will only grow with $e^{n^* S/2}$. Thus, replacing the sums with integrals is valid when

$$e^{-nS/2 + n^* S/2} < e^{-(n-1)S + n^* S} \implies n^* > n - 2 \tag{2.15}$$

In eq. (2.14), we had $n^* = n - 1$ and were safe by one factor of e^S. If $n^* = n - 2$, then the fluctuating and smooth parts will have the same magnitude. If $n^* < n - 2$, then the smooth part will be washed away. In general, one may consider more complex situations with different entropy factors that will need to be compared.

Another difficulty with converting sums to integrals is the presence of *level repulsion*.[10] In chaotic systems, the eigenvalues of the Hamiltonian should be considered correlated (pseudo-)random variables [44]. Thus when converting a sum over multiple indices to a multi-variable integral, in principle, one would have to introduce a joint density of states, e.g.,

$$\sum_{ij} \longrightarrow \int_{E_i E_j} e^{S^{(2)}(E_i, E_j)} \tag{2.16}$$

for a pair of energy indices [45]. However, to leading order in e^{-S}, the joint density of states factorizes $e^{S^{(2)}(E_i, E_j)} \approx e^{S(E_i) + S(E_j)}$. Furthermore, level spacing corrections only become relevant at unphysical late times order of the Heisenberg timescale $\tau_H \sim e^S$ [6]. In principle, level spacing corrections may be computed if one has e.g. a *matrix model*[11] [46] for a system but a matrix model is not generally accessible in practice. More importantly, though we consider higher-order corrections in e^{-S}, we largely neglect the timescales necessary to observe level repulsion effects and will presume that joint densities of states factorize.

By the convolution theorem, the decay of time-dependent free cumulants, $f(\vec{t})$, is equivalent to the fine resolution behavior of their spectral free cumulants, $f(\vec{\omega})$. If the spectral free cumulant becomes smooth at over an energy scale ε, the associated time-dependent free cumulant will vanish to zero over a timescale ε^{-1}. In general, correlation functions of local operators are expected to have nontrivial behavior over short timescales that are associated with dissipation and scrambling and become trivial over longer timescales that are associated with transport and hydrodynamics [15]. Beyond those timescales, correlation functions are expected to vanish until the Heisenberg time when the level spacing of the system becomes relevant. Accordingly, a given spectral free cumulant should appear entirely smooth over energy scales that are smaller than those associated with hydrodynamics but are much larger than the level spacing of a system.

3 Generalized free cumulants

Recently, it was shown that the structure of the ETH, eq. (1.1), is intimately related to free probability theory [18], which describes the statistics of highly non-commutative random variables [47]. For correlated operators, eq. (1.1), turn out to be their *free cumulants*, analogs of the classical cumulants for maximally non-commutative variables. When classical (commutative) random variables are sampled independently, their classical cumulants[12]

---

[10]The suppressed likelihood of finding two eigenstates much closer in energy than the average level spacing in a chaotic system.

[11]A matrix model is an integral over matrices (e.g. a probability distribution). For example, the matrix model of ref. [46] is a distribution over all Hamiltonians consistent with a given set of correlators. Within the validity of saddle-point integration, (single trace) correlators computed from a single Hamiltonian sampled from a matrix model are identical to those averaged over that matrix model to all orders in e^{-S}. The ETH holds for a system if there is a matrix model in the sense of ref. [46] that generalizes its Hamiltonian.

[12]A cumulant of degree n is, in general, some polynomial in the moments of degree $m \le n$. Both free and classical cumulants can be defined by their respective moment-cumulant relationship.

vanish. Analogously, when large $N \times N$ random matrices are sampled independently, their free cumulants vanish to leading order in $1/N$. In the context of the ETH, we think of physics as being local in energy space so the $1/N$ expansion is replaced by an e^{-S} expansion and free probability becomes a good model in the thermodynamic limit. We discuss free probability theory and its role in quantum chaos in more detail in appendix A.

While ref. [18] clarifies the role of free cumulants for operators, an analogous structure is missing for states, even though the analysis of section 2 indicates that factors of e^{-S} show up for correlations of states in the same manner as they do for operators. In this section, we show that correlations between states are described by *generalized free cumulants* (GFCs) of which the EB, eq. (1.2), is a simple example. Our generalizations account for the partitioning of our system (figure 1) and correlations between eigenstates from distinct Hamiltonians. We introduce a diagrammatic representation of GFCs that will aid the investigation of state-based physics, such as evolution towards an equilibrium density matrix and the behavior of entanglement entropy without requiring a firm knowledge of free probability. In general, we are able to compute to arbitrary order in an e^{-S} expansion. This formalism is the main result of the paper.

To study ordinary free cumulants, refs. [13, 18, 48] utilize a similar diagrammatic formalism known as *cactus diagrams* while standard references on free probability utilize so-called *non-crossing partitions* [49] and ref. [46] utilizes *'t Hooft diagrams*. Existing diagrammatic approaches have substantial merit and we may have been able to modify them instead of introducing our own diagrams. However, we feel the diagrams we introduce will be more natural for our purposes.

3.1 Eigenstate correlations: perturbed Hamiltonian

To ease the introduction of GFCs, we first consider a Hamiltonian H_0 and a perturbed Hamiltonian H without necessarily bipartitioning the system. Later, we will take H_0 to be $H_A + H_B$ and H to be $H_A + H_B + H_{AB}$. We wish to compute, as an example, the diagonal elements with respect to H_0 of a system prepared in the canonical ensemble of H_0 but evolved under H. Let us again define the symbol $c^\mu_\nu \equiv \langle \mu | \nu \rangle$ for the change-of-basis tensor, $|I\rangle, |J\rangle$ as eigenstates of the initial Hamiltonian, and $|i\rangle, |j\rangle$ as eigenstates of the perturbed Hamiltonian. Then our expression is,

$$\rho_{II}(t) = \frac{1}{Z_\beta} \sum_{Jij} \langle I|i\rangle\langle i|J\rangle\langle J|j\rangle\langle j|I\rangle \, e^{-\beta E_J - i\omega_{ij}t} \equiv \frac{1}{Z_\beta} \sum_{Jij} c^I_i c^i_J c^J_j c^j_I e^{-\beta E_J - i\omega_{ij}t}. \tag{3.1}$$

One should expect that the typical magnitude of an element ρ_{II} is of order $\mathcal{O}(e^{-S})$. The right-hand expression in eq. (3.1) appears to sum e^{3S} terms while Z_β is the canonical partition function and contributes a factor of e^{-S}, so the summand should have average weight e^{-3S}. We need the overall expression to have magnitude e^{-S}. If we compare the summand to eq. (2.8), we see that a cyclic product of 4 overlaps has the necessary remaining factor of e^{-3S}.

However, we have neglected some terms. If $I = J$, the dominant contribution to the sum will be a product of 2 2-cycles, $\left(c^I_i c^i_I\right)\left(c^J_j c^j_I\right)$ and the summand will have magnitude $\left(e^{-(2-1)S}\right)^2 = e^{-2S}$. But we have also dropped a sum and are now only summing over e^{2S} terms, so our overall magnitude is the same. The same argument applies when $i = j$.

We may also have $I = J$ and $i = j$ simultaneously. Indeed, there are 4 distinct index partitions to consider:

$$
\begin{aligned}
Z_\beta \rho_{II}(t) &= \sum_{[Jij]} c_i^I c_J^i c_j^J c_I^j e^{-\beta E_J - i\omega_{ij} t} + \sum_{[ij]} c_i^I c_I^i c_j^I c_I^j e^{-\beta E_I - i\omega_{ij} t} \\
&\quad + \sum_{[Ji]} c_i^I c_J^i c_i^J c_I^i e^{-\beta E_J} + \sum_{[i]} c_i^I c_I^i c_i^I c_I^i e^{-\beta E_I} \\
&= \int_{\omega_1 \omega_2 \omega_3} e^{-it(\omega_2 + \omega_3) - \beta(E_I - \frac{1}{4}\omega_1 - \frac{1}{2}\omega_2 + \frac{1}{4}\omega_3)} F(\omega_1, \omega_2, \omega_3) \\
&\quad + \int_{\omega_1 \omega_3} e^{-it(-\omega_1 + \omega_3) - \beta(E_I + \frac{1}{2}\omega_1 + \frac{1}{2}\omega_3)} F_{(IJ)}(\omega_1, \omega_3) \\
&\quad + \int_{\omega_1 \omega_2} e^{-\beta E_I} F_{(ij)}(\omega_1, \omega_2) \\
&\quad + \int_{\omega_1} e^{-S(E_I) - \beta E_I} F_{(IJ)(ij)}(\omega_1)
\end{aligned}
\tag{3.2}
$$

where we have defined window functions for contracted partitions, $F_{(\cdots)}$, and used the symbol $(\cdots)$ to label relevant index contractions. We have also dropped the slow dependence of F on the average energy. Notice how the partition $(IJ)(ij)$ carries an extra factor of e^{-S} ensuring that its contributions will be suppressed. Furthermore, notice that the only other time-independent term is (ij). Therefore, if each F has a finite resolution, then the long time value of ρ_{II} will be entirely determined by the Laplace Transform of $F_{(ij)}$.

First, recognize that in figure 2(d), the suppression of $(IJ)(ij)$ is related to a double constraint imposed by the two contractions. In general, partitions may have any number of contractions, however, if any two contractions are connected by more than two indices topologically, they will be suppressed by a relevant factor of the density of states. What we refer to as a generalized free cumulant is a single connected partition, without contractions, before summing, in which no two indices are equal. In this case, the GFC's are individual loops. For example, a single, $n = 4$ generalized free cumulant is pictured in figure 2(a), whereas each of figures 2(b), 2(c), 2(d) depict partitions which are dominated by a product (or products) of $n = 2$ generalized free cumulants. GFCs are the irreducible building blocks of correlations in the e^{-S} expansion.

Each GFC has an associated window function. In the present case, we can use $F_2(\omega)$ to represent the window function for the $n = 2$ GFC and $F_4(\omega_1, \omega_2, \omega_3)$ for the $n = 4$ GFC.[13] More precisely,

$$
\begin{aligned}
\overline{c_j^I c_I^j} &= e^{-S(E_{Ij})} F_2(E_I - E_j) \\
\overline{c_i^I c_J^i c_j^J c_I^j} &= e^{-3S(E_{IJij})} F_4(E_I - E_i, E_i - E_J, E_J - E_j), \quad I \neq J, i \neq j.
\end{aligned}
\tag{3.3}
$$

Then, by comparison to eq. (3.2) or by inspecting figure 2, we can write down the window

---

[13]We should point out here the F depends, in general, on the Hamiltonians considered. Rather than name distinct functions for the various cases we consider throughout this paper, with some abuse of notation, we will reuse the label F for general window functions and let its arguments and context specify the specific function under consideration.
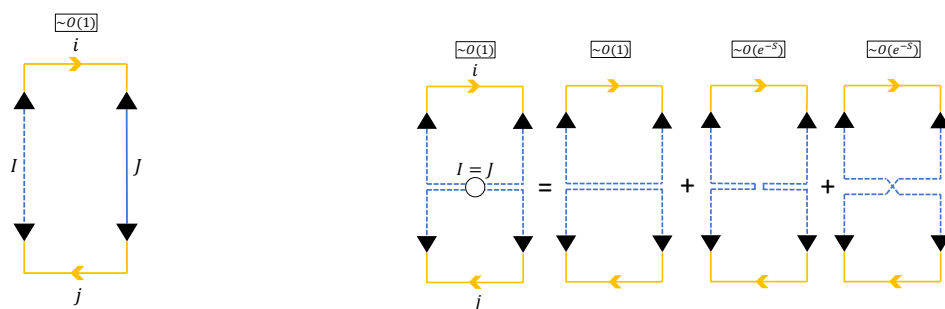
(a) The full partition has 1 4-loop and 3 sums, therefore the overall weight is $\mathcal{O}(e^{3S-3S}) \sim \mathcal{O}(1)$.

(b) Partition (IJ) is dominated by a product of 2 2-loops with 2 sums, therefore the overall weight is $\mathcal{O}(e^{2S-S-S}) \sim \mathcal{O}(1)$.
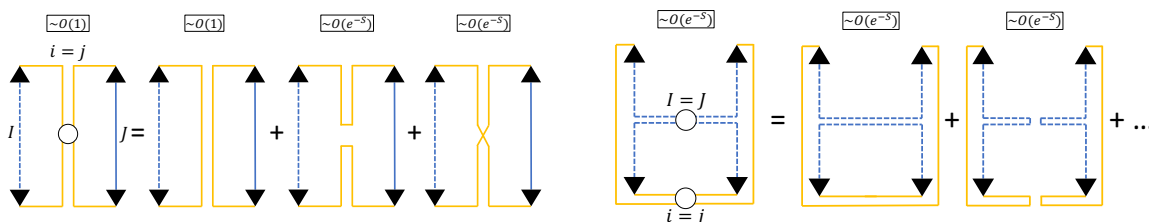
(c) Partition (ij) is dominated by a product of 2 2-loops with 2 sums, therefore the overall weight is $\mathcal{O}(e^{2S-S-S}) \sim \mathcal{O}(1)$.

(d) Partition $(IJ)(ij)$ is dominated by a product of 2 2-loops with 1 sum, therefore the overall weight is $\mathcal{O}(e^{S-S-S}) \sim \mathcal{O}(e^{-S})$.

Figure 2. The partitions in eq. (3.2). We have used color to indicate the eigenstates of different Hamiltonians. Unless explicitly stated otherwise in the figure, indices are taken to be distinct in summations. When a partition contains a contraction, it can be expressed as a sum of products of connected components obtained from different ways of pairing off indices. In (b) and (c), there are 3 pairings only 1 of which is dominant. In (d), there are $3 \times 3 = 9$ pairings, of which we have drawn 2 of 3 dominant terms. In (c) and (d), notice that a contraction of indices with opposite time-dependences yields partitions that are time-independent. *Note:* though we have removed the contraction symbol on the right-hand side of the above figures, the contractions still exist and need to be considered when counting the number of sums in a diagram.

functions for the contracted terms to leading order,

$$F(\omega_1, \omega_2, \omega_3) = F_4(\omega_1, \omega_2, \omega_3)$$

$$F_{(IJ)}(\omega_1, \omega_2) = F_2(\omega_1)F_2(\omega_2)$$

$$F_{(ij)}(\omega_1, \omega_2) = F_2(\omega_1)F_2(\omega_2)$$

$$F_{(IJ)(ij)}(\omega) = 3F_2(\omega)^2. \tag{3.4}$$

The general procedure for performing computations in this paper will be inspecting diagrams to acquire the correct factors of e^S and the correct form of the window function, then evaluating integrals via saddle-points by neglecting the finite width of the window functions. We have included a summary of notation we use in this paper in table 1.

	Notation	**Example**
(a)	$E_{ijk\cdots}$	$E_{ij} = \frac{1}{2}(E_i + E_j)$
(b)	$\bar{E}$	$\bar{E} = \frac{1}{n}\sum_{m=1}^{n} E_m$
(c)	ω_{ij}	$\omega_{ij} = E_i - E_j$
(d)	$(\cdots)$	eq. (3.2)
(e)	$[\cdots]$	$\sum_{[ij]} X_{ij}X_{ji} = \sum_i \sum_{j \neq i} X_{ij}X_{ji}$
(f)	f	$f_{XY}(E;\omega) = e^{-S(\bar{E})} \sum_{[ij]} X_{ij}Y_{ji}\delta(E_i - (E + \omega/2))\delta(E_j - (E - \omega/2))$
(g)	F	eq. (3.2)
(h)	$\langle \cdots \rangle_i$	$\langle X(t)Y(0)X(t)Y(0) \rangle_i = \langle i|X(t)Y(0)X(t)Y(0)|i \rangle$
(i)	$\langle \cdots \rangle_\beta$	$\langle X(t)Y(0)X(t)Y(0) \rangle_\beta = \sum_{ijkl} X_{ij}Y_{jk}X_{kl}Y_{li}e^{i(\omega_{ij}+\omega_{kl})t - \beta E_{ijkl}} / \sum_i e^{-\beta E_i}$

Table 1. Table of various notation. All notation holds analogously for various indices we will use (i, j, I, J, a, b). (a) We use subscripts of E to denote average energies over specified indices. (b) We use an overline to denote averaging the energy over an unspecified number of indices. (c) We use ω in general to denote energy differences, and subscripts to specify indices. (d) We use parentheses, $(\cdots)$, as a short hand to denote a set of indices that are contracted. Multiple contractions are written as products, $(\cdots)(\cdots)$. (e) Braces are used to indicate indices that are not equal to any other index in the expression. (f) We use the letter f to denote operator free cumulants in time and frequency space. (g) We used the letter F to denote state generalized free cumulants in time and frequency space. (h) We use $\langle \cdots \rangle_i$ to denote the ordinary expectation value evaluated on an individual eigenstate i. (i) We use $\langle \cdots \rangle_\beta$ to denote special thermally regulated correlators that are cyclically symmetric in their arguments and are natural objects in quantum chaos.

3.2 Operator correlations: OTOCs and freeness

In eq. (2.14), we showed that the sum over a product of operators neglecting repeated indices evaluated on an eigenstate reduced to the mixed Fourier-Laplace transform of the relevant f function. We referred to this object as a free cumulant. Instead of evaluating the free cumulant on an eigenstate, we can evaluate it via a special thermal regulator,

$$\frac{1}{Z_\beta} \sum_{[i_1 \cdots i_n]} X_{i_1 i_2} \cdots X_{i_n i_1} e^{i\vec{\omega} \cdot \vec{t} - \beta\bar{E}} = e^{\beta E_\beta - S(E_\beta)} \int_{\bar{E}\vec{\omega}} e^{i\vec{\omega}\cdot\vec{t} + (n-(n-1))S(\bar{E}) - \beta\bar{E}} f(\bar{E}; \vec{\omega})$$

$$= \int_{\vec{\omega}} e^{i\vec{\omega}\cdot\vec{t}} f(E_\beta; \vec{\omega})$$

$$= f_n(E_\beta; \vec{t}), \tag{3.5}$$

where the integral over $\bar{E}$ was evaluated via its saddle-point (see appendix B). The symmetry between indices has removed the shift vector $\vec{l}$ which forced us to work in complex time. Thus we find it is more natural to work with the thermally regulated correlator which is conventional in the study of quantum chaos [17, 50].

Here, we study an out-of-time-ordered correlator (OTOC),

$$\langle X(t)Y(0)X(t)Y(0) \rangle_\beta = \frac{1}{Z_\beta} \sum_{ijkl} X_{ij}Y_{jk}X_{kl}Y_{li} e^{i\omega_{ij}t + i\omega_{kl}t - \beta E_{ijkl}}. \tag{3.6}$$
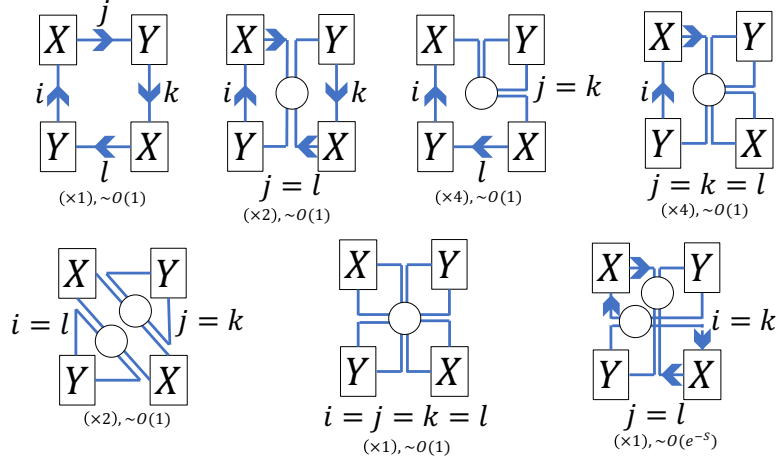
Figure 3. There are a total of 15 partitions from 7 different graphs. We have drawn one representative partition from each graph and labeled their multiplicities. Note that two different partitions with the same graph contribute differently, so the multiplicity is not a "symmetry factor". We have also labeled the overall weight of each partition divided by an implicit factor of Z_β.

We have drawn all the possible contractions of eq. (3.6) in figure 3. In this case, since any index can contract with any other index, the GFCs of our diagrams, which are the cumulants mentioned above, are the celebrated free cumulants of the operators $X(t)$ and $Y(0)$ and are represented by single loop diagrams and are associated to the f functions.

We can isolate the free cumulants from eq. (3.6) using non-crossing *operator* partitions inline as follows. For convenience we will assume that all 1-point functions vanish, $\langle X \rangle_\beta = \langle Y \rangle_\beta = 0$. We define an operator partition as a symbol that represents the associated free cumulant of an ordered set of operators. A crossing of operator partitions results in an overwhelming suppression. Then,

$$
\begin{aligned}
\langle X(t)Y(0)X(t)Y(0)\rangle_\beta &\equiv \langle \overline{X(t)Y(0)X(t)Y}(0)\rangle_\beta \\
&\quad + \langle \overline{X(t)Y}(0)\overline{X(t)Y}(0)\rangle_\beta + \langle \overline{X(t)\overline{Y(0)X(t)}Y}(0)\rangle_\beta \\
&= f_{XYXY}(E_\beta; t, 0, t) + 2f_{XY}(E_\beta; t)^2 \\
&= \int_{\omega_1\omega_2\omega_3} f_{XYXY}(E_\beta; \omega_1, \omega_2, \omega_3)e^{it(\omega_1+\omega_3)} \\
&\quad + 2\left(\int_\omega f_{XY}(E_\beta; \omega)e^{i\omega t}\right)^2
\end{aligned}
\tag{3.7}
$$

where $E_\beta = \langle H \rangle_\beta$ and we have extracted 3 terms: a partition of all 4 operators, a product of the partition of the first two and last two operators, and a product of the partition of the first and last operator and the second and third operator. To a non-crossing partition of n operators is associated their n-point free cumulant. From eq. (3.7) we can see that the f-functions of the ETH are the regularized free cumulants that have a special significance in quantum chaos.

Partitions whose contractions cross exist, but are suppressed. In this case, we have one:

$$
\langle \overline{X(t)\overline{Y(0)X}(t)Y}(0)\rangle_\beta = \frac{1}{Z_\beta}\left(2f_{XY}(E_\beta; t)^2 + f_{XX}(E_\beta; 0)f_{YY}(E_\beta; 0)\right) \sim \mathcal{O}(e^{-S}).
\tag{3.8}
$$

One may wonder why 2 copies of $f_{XY}(E_\beta; t)^2$ appeared in eq. (3.8) when we appeared to have only contracted the pair of X's and the pair of Y's. One can find all 3 terms from the last partition in figure 3. But it is not obvious how to derive them directly from the left-hand expression in eq. (3.8). However, it is also natural to utilize non-crossing index partitions. In table 2, we compare our own diagrammatic approach to correlations to others in the literature, namely, cactus diagrams, non-crossing partitions of indices, and 't Hooft diagrams. All approaches fundamentally contain the same content and are united in their connection to free probability. Redoing the above calculations but taking into account 1- and 3-point functions will generate all 15 partitions discussed in figure 3. The full decomposition into non-crossing partitions is given in appendix A.

As discussed in section 2, the thermalization of a system follows from the decay of these free cumulants at late times. The equilibrium value of a correlation function can be obtained by dropping all time-dependent partitions or, equivalently, all diagrams that contain arrows. An equivalent formulation of this statement is that operators that satisfy the ETH become *freely independent*, or *free*, at long time separations. Free independence, or simply *freeness*, is the free probabilistic analog of classical independence [47]. For example, whereas a sum of many classically independent random variables acquires a Gaussian probability distribution function, a sum of many freely independent random variables acquires a semicircle spectrum, a famous signature of Gaussian matrix ensembles [45]. We can draw conclusions about physical operators from this observation.

Consider an operator X restricted to a narrow energy band,

$$X_{ij}^{(\omega_c)} \equiv X_{ij}\theta(\omega_c - |\omega_{ij}|). \tag{3.9}$$

Ref. [54] showed numerically on a spin chain that as $\omega_c \to 0$, $X^{(\omega_c)}$ obtains a semicircular spectrum and interpreted this observation as the onset of *random matrix theory*. We can interpret this observation as a consequence of free independence and, thus, a necessary consequence of the ETH. Consider the following identity,

$$X_{ij}^{(\omega_c)} \equiv X_{ij}\theta(\omega_c - |\omega_{ij}|) = \frac{\omega_c}{\pi} \int_{-\infty}^{\infty} X_{ij}(t)\operatorname{sinc}(\omega_c t)dt \tag{3.10}$$

which allows us to see $X^{(\omega_c)}$ as an average of $X(t)$ over a time period of ω_c^{-1}. If $X(t)$ becomes freely independent from $X(0)$ for t greater than some t_{free}, then for $\omega_c \ll t_{\text{free}}^{-1}$, $X^{(\omega_c)}$ must obtain a semicircular spectrum. This has been more recently interpreted in terms of the onset of an *emergent rotational symmetry* in the energy eigenspace of ETH satisfying systems at asymptotically small frequencies [28].

We can draw a more direct line between freeness and the quantum butterfly effect by defining the freeness of sublagebras. Consider two subalgebras, $\mathcal{X}$ and $\mathcal{Y}$, of the algebra of operators that act of the Hilbert space of our system. These subalgebras may be understood, for example, to be the Pauli algebras of two possibly identical sites in a spin chain at possibly different times. $\mathcal{X}$ and $\mathcal{Y}$ are freely independent if and only if for every $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$ with $\langle X_i \rangle = \langle Y_i \rangle = 0$,

$$\langle X_1 Y_1 \cdots X_q Y_q \rangle = 0, \, \forall q \geq 1. \tag{3.11}$$

Index	Cactus	Non-crossing	't Hooft

Table 2. Comparison of diagrammatic techniques. Column 1 contains two partitions from figure 3. Column 2 redraws these partitions using the formalism of refs. [13, 18]. Whether a diagram is leading order is equivalent to whether it resembles a cactus (essentially a tree with round branches) in the following sense. Each loop is a cactus pad; to any cactus pad, any number of other pads may be connected. However, to travel from any pad to any other by hopping to adjacent pads, there can only ever be one route without crossing the same contraction twice. This feature of cactus diagrams is shared by the loops in index diagrams so long as we are not considering subsystems (which we do in the next section). We also note that these are essentially the properties of tadpole diagrams which govern the Hartree approximation in large N vector systems [51]. Column 3 utilizes so-called non-crossing partitions of indices whose order is determined by the number of crossings. In eq. (3.7) we utilized non-crossing partitions of operators. Non-crossing partitions of operators and of indices enjoy a *dual* relationship and both function to describe free cumulants [18]. The black circumscribing circle represents the trace and the internal lines represent index contractions. The top row contains no crossings, while the bottom row contains 1. Column 4 depicts analogous 't Hooft diagrams, which play a role in large N gauge theories and whose order is closely related to the genus of the surface on which the diagram may be drawn without crossings [52]. As for non-crossing partitions, the black circumscribing circle represents the trace and the internal lines represent index contractions. The solid circles explicitly represent the appropriate free cumulants, often referred to in this context as *planar connected Green's functions* [52, 53], and as such the bottom row only encodes the first term on the right hand side of eq. (3.8) and not all terms of order e^{-S}. 't Hooft diagrams are the natural method of evaluating matrix models and main technology employed by ref. [46] to study the ETH. Both 't Hooft and index diagrams utilize a type of double-line notation, where the former uses open circle to denote index contractions, and the latter uses solid circles to denote operator contractions. In this way, index and 't Hooft diagrams share a dual relationship similar to the one shared by non-crossing partitions of operators and of indices.

Shifting $\mathcal{X}$ forward in time[14] $\mathcal{X} = e^{iHt}\mathcal{X}(0)e^{-iHt}$, we can see that eq. (3.11) contains the statement that all q-OTOCs of an element of $\mathcal{X}$ with an element of $\mathcal{Y}$ have decayed. Thus OTOCs are themselves a measure of freeness. This observation was first made in ref. [55], where a connection was drawn between OTOCs, freeness, and unitary designs. The idea we wish to convey is that satisfying the ETH, the quantum butterfly effect, and the emergence of rotational symmetry at small frequencies are, in essence, equivalent definitions of chaos and avatars of free probability (see appendix A for an elaboration on these ideas).

3.3 Random matrices, replicas, and higher moments

One may also be interested in partitions where some legs are left open, as these partitions represent fluctuations with a random tensor term, e.g. $\langle i|J\rangle = \sqrt{e^{-S(E_{iJ})}F(\omega_{iJ})}R_J^i$ or $X_{ij} = f_1(E_{ij})\delta_{ij} + \sqrt{e^{-S(E_{ij})}f_2(\omega_{ij})}R_{ij}$, where R represents random matrix terms with mean 0 and variance 1. The variance of such a term can be computed by replicating the diagram and connecting the open legs of the original to the open legs of its conjugate, e.g., $|\langle i|J\rangle|^2 \propto R_J^i R_i^J \sim 1$. In general, one may compute higher moments from suitably many replicas and apply the above-stated rules for the partitions that are generated. This procedure is key to our computations of entanglement entropy in sections 4.2 and 5.2.

A simple consequence of the rules we have presented is that random matrix terms are Gaussian distributed, to leading order in e^{-S}. For example, we can show that R_J^i is Gaussian distributed $\sim \mathcal{N}(0,1)$ by recognizing that (a) odd moments of R have open indices and vanish while (b) even moments will be dominated by the decompositions with the largest number of loops: products of pairs (see figure 4). Then, $\overline{|R_J^i|^{2p+1}} = 0$ while $\overline{|R_J^i|^{2p}} = \mu_{2p}(\overline{R_J^i R_i^J})^p + \mathcal{O}(e^{-S}) = (2p-1)!!$, where $\mu_{2p} = (2p-1)!!$ is the number of ways to pair up $2p$ copies of $|R_J^i|$ and are the central moments of the Gaussian distribution. An analogous argument was made for matrix elements of operators in [6] and follows concretely from (1.1). Another result which can readily obtained is that variance of the diagonal elements of operators is exactly twice the variance of nearby off-diagonal operators [27, 56], as the number of ways to pair indices of $2p$ copies of a diagonal element R_{ii} is instead $2^{2p}(2p-1)!!$. These properties are expected features of matrices sampled from rotationally invariant ensembles [45, 57].

The emergence of a Gaussian distribution originated from the decomposition of contractions into pairs of indices (rule 2). This derivation is essentially a converse of Isserlis' (Wick's) theorem. For the case of pure operator correlations, the validity of this decomposition is implicitly assumed throughout the literature and in key works [13, 18] and is supported by numerics on spin chains [58]. This assumption is justified so long as we imagine random matrix terms are consistent with a rotationally invariant random matrix ensemble. The matrix model of ref. [46] formalizes this idea for pure operator correlations, but heuristically it should hold for general eigenstate or operator-eigenstate correlations as well. To be more precise, as we discuss in appendix A, eq. (1.1) is an extension of the general form for the free cumulants of random matrices. But for the terms in eq. (1.1) to retain the combinatorial properties of free cumulants (i.e. generate non-crossing partitions), the pairwise decomposition must hold and as shown in ref. [46], eq. (1.1) implies that non-Gaussian terms in the corresponding

---

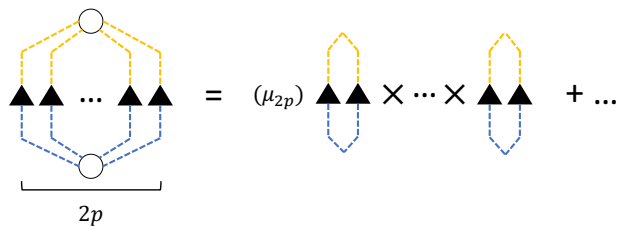[14]By time evolving each element of $\mathcal{X}$.

Figure 4. Decomposition of a contraction of $2p$ replicas into μ_{2p} products of p 2-loops. This decomposition is not special to the case shown but a general result on the moments of random matrix terms. Where the triangle is in the above figure, any arbitrarily complicated diagram with two external legs may be placed instead but the consequence of the replica calculation will be the same: random matrix terms are Gaussian distributed to leading order in e^{-S}.
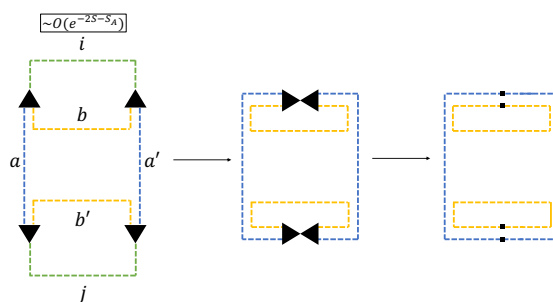

Figure 5. Schematic calculation of the weight associated with the uncontracted partition of eq. (3.12). Whole system indices are removed, each carrying a factor of e^{-S}, and then subsystem indices contribute a total of e^{-S_A} from what is left. Note that we are neglecting fluctuations which will vanish under summation.

matrix model are suppressed by appropriate density of states factors. Thus the suppression of non-Gaussianities should be understood as a generic feature of the ETH.

The procedure of using replicas to compute higher moments also holds for partitions without open legs. In general, computing moments will reduce to counting contractions between replicas. One will find that partitions with sufficiently many summed indices (in the sense of eq. (2.15)) will have higher central moments suppressed by factors of e^{-S}, as contractions inevitably cost sums without yielding sufficiently advantageous factorizations. Thus, these partitions will be very sharply peaked around their mean value. In contrast, partitions with very few summed indices will lose no sums from contractions but will gain advantageous factorizations and thus will be widely distributed from their means. The situation with open legs discussed in the previous paragraphs is a special case where random matrix terms acquire a Gaussian distribution with $O(1)$ variance about their means.

3.4 Eigenstate correlations: interacting subsystems

We now wish to generalize our formalism to the situation of two interacting subsystems A and B. We take $H_0 = H_A + H_B$ and $H = H_A + H_B + H_{AB}$. Take $|i\rangle$ and $|j\rangle$ as possibly identical eigenstates of H. We study, as an example of the various subtleties we will encounter, the
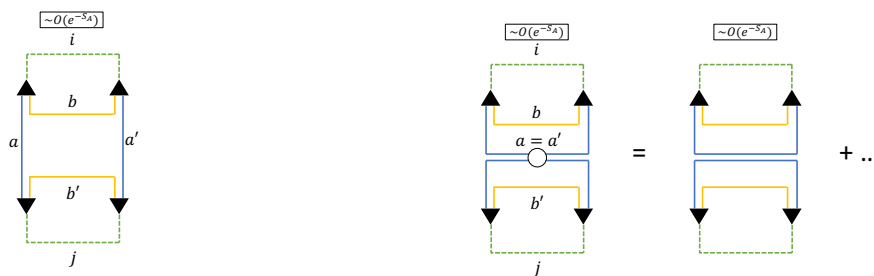
(a) The full partition from eq. (3.12) has overall weight $\mathcal{O}(e^{-S_A})$.

(b) Partition (aa') is dominated by a product of 2 2-cycles with 2 sums on subsystem B and 1 sum on subsystem a. Therefore the overall weight is $\mathcal{O}(e^{2S_B+S_A-S-S}) \sim \mathcal{O}(e^{-S_A})$.
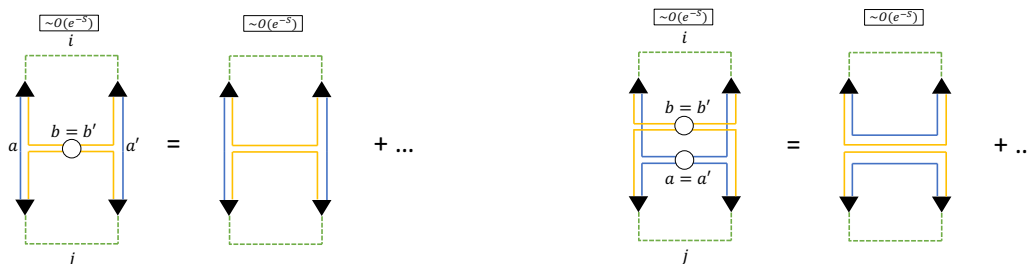
(c) Partition (bb') is dominated by the same partition as in figure 6(a), but with 1 fewer sum on subsystem B, so the overall weight is $\mathcal{O}(e^{-S_A-S_B}) = \mathcal{O}(e^{-S})$.

(d) Partition $(aa')(bb')$ is dominated by the same partition as in figure 6(b) but with 1 fewer sum on subsystem B, therefore the overall weight is $\mathcal{O}(e^{-S})$.

Figure 6. The partitions of eq. (3.12). For each diagram, leading order contributions are shown. It can be seen that the partition in figure 6(a) will share its window function with that of 6(c) and that the partition in figure 6(b) will share its window function with that of figure 6(d) while in each case having different weights.

Hilbert-Schmidt inner product of their reduced density matrices on A:

$$(\rho_A^i|\rho_A^j) \equiv \mathrm{Tr}_A\left[\rho_A^i\rho_A^j\right] = \sum_{aa'bb'} c_i^{ab}c_{a'b}^i c_j^{a'b'} c_{ab'}^j. \tag{3.12}$$

First, take i and j as distinct indices. We recognize that without sums, the term in eq. (3.12) is the eigenstate analog of equation (2.11) which we conjectured to have unsummed weight e^{-2S-S_A}. The spirit of the calculation in eq. (2.11) was that we can remove indices on the whole system and collect a factor of e^{-S} for each removed index until we are left with a product of partitions on the subsystem. We represent this calculation schematically in figure 5. The remaining subsystem partitions contribute their factors of the subsystem densities of states. In this case we get e^{-S} from each of i and j, and a factor of e^{-S_A} from the remaining 2 index cycles from a, a'. We neglect the orthogonality of $\langle a|a'\rangle$ so long as a and a' are not adjacent in the original diagram.

However, we can also read the correct weights off from the full diagram directly. Restricted to subsystem B, the left-hand-side of figure 5 appears as a product of 2 2-loops (yellow
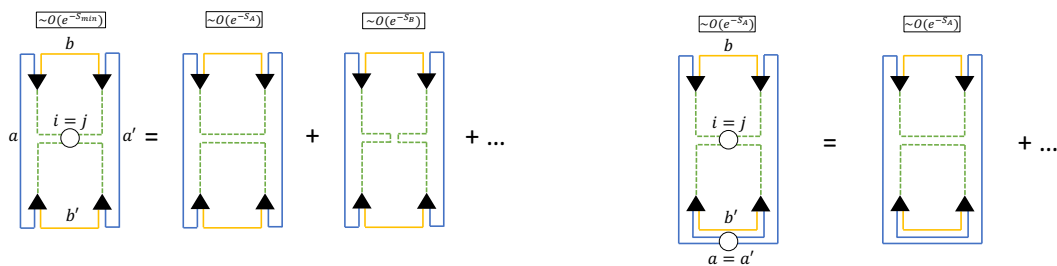
(a) The partition (ij) receives a contribution from two competing factors with respective weights e^{-S_A} and e^{-S_B}. We define $S_{\min}$ as $\min(S_A, S_B)$ in this context.

(b) Partition $(ij)(aa')$ is dominated by the same term as the diagrma in figure 6(b), which is 1 of 2 partitions which contribute to the diagram in figure 7(a), and carries the same weight, $\mathcal{O}(e^{-S_A})$.
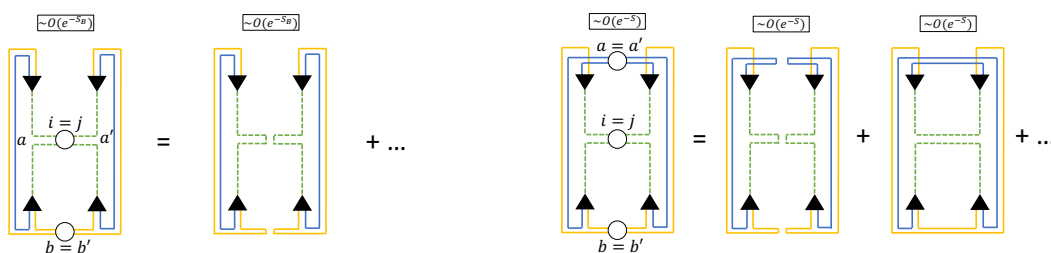
(c) Partition $(ij)(bb')$ is dominated by 1 of 2 partitions that contribute to the diagram in figure 7(a) and carries weight $\mathcal{O}(e^{-S_B})$.

(d) Partition $(ij)(aa')(bb')$ is dominated by the same partition as in figure 6(d) and one other enabled by the contraction (ij) and carries overall weight $\mathcal{O}(e^{-S})$.

Figure 7. The partitions of eq. (3.12) for $i = j$. Since there is no summation over i or j their contraction has not cost any factors of e^S, however, their contraction has enabled advantageous decompositions when $b = b'$ in figures 7(a), 7(c), and 7(d) not available in figures 6(a), 6(c), and 6(d).

and green lines), each of which would contribute a factor of e^{-S_B}. Restricted to subsystem A, it appears as a single 4-loop (blue and green lines) which would contribute e^{-3S_A}. For connected partitions that have potentially distinct behavior over different subsystems, the correct weight is acquired by combining terms of S_A and S_B as $S_A + S_B \to S$, leaving leftover factors of $S_{A(B)}$. In this case, we get $e^{-2S_B - 3S_A} \to e^{-2S - S_A}$.

Next, we discuss the consequences of index contractions. When summing over a, a', b, b', there will be terms where $a = a'$, or $b = b'$, or both. For $a = a'$, we can see from figure 6(b) that the diagram factorizes into two pieces, each with weight e^{-S}. However, for $b = b'$, we see from figure 6(c) that the contraction does not yield any advantageous factorization. Previously, in the case without subsystems, every contraction yielded an advantageous factorization. This is one way in which GFC's diverge from ordinary free cumulants. In the case of both $a = a'$ and $b = b'$, the best factorization is the same one as when $a = a'$ alone.

Now, take $i = j$, as in figure 7. First, note that unlike the case $i \neq j$, there is a choice in how to decompose the contraction (ij). This results in two possible weights from the partition, e^{-S_A} or e^{-S_B}, depending on the smaller of S_A and S_B (see figure 7(a)). In other words, the contraction (ij) has introduced a symmetry between the subsystems. Next, we consider again
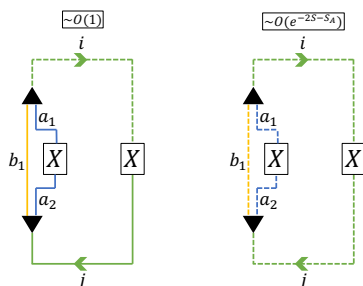
Figure 8. The summed and unsummed partitions representing eq. (3.13).

the effects of additional index contractions from $a = a'$ or $b = b'$. The contraction (aa') works as before factorizing the partition into 2 simple partitions (see figure 7(b)). The contraction of (bb'), however, can now take advantage of the new symmetry and factor the partition into 2 simple partitions as well (see figure 7(c)). The contraction $(aa')(bb')$ also receives an additional contribution (see figure 7(d)). Thus, we have seen that index contractions on partitions with subsystems can lead to more diverse behavior than on partitions without subsystems.

An analogous technique for calculating such partitions was utilized in ref. [36] which treated c_{ab}^i as a narrow banded Wishart matrix on its lower indices and then performed ensemble averages to calculate distinguishability measures via a version of 't Hooft double line notation.[15] Where both techniques apply, calculations are identical. However, our technique has the conceptual advantage of treating eigenstates of the full system on the same footing as eigenstates of the subsystem, while interpreting eigenstates as individual samples in the ensemble average. This conceptual advantage is particularly useful for introducing and discussing time-dependent physics as we do in sections 5 and 6.

3.5 Operator-eigenstate correlations

Lastly, we consider partitions which contain operators and changes-of-basis. Take an operator X deep within subsystem A. X can be understood in terms of its matrix elements in both the H_A and H eigenbases. We propose that either form of X can be used in GFC's. For instance, consider the 2-point correlator of X,

$$\langle X(t)X\rangle_i = \sum_{jaa'b} X_{ij} c_{ab}^j X_{aa'} c_i^{a'b} e^{i\omega_{ij}t}. \tag{3.13}$$

We have represented the right-hand side of eq. (3.13) in figure 8. One may wonder if there is any nontrivial relationship between the matrix elements $X_{aa'}$ and X_{ij}. We explore such a relationship in section 6.

3.6 Numerical evidence

In this section, we provide some numerical evidence for our conjectured scalings via exact diagonalization of a quantum spin chain. We consider the non-integrable Ising model for

---

[15]We thank Jonah Kudler-Flam for bring these calculations to our attention.
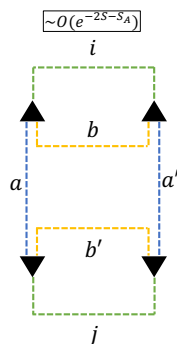
Figure 9. Diagrammatic representation of C (eq. (3.16)), up to averaging. Averaging ensures that fluctuations are suppressed.

N spins with periodic boundary conditions, augmented by unequal local fields on the first and the last site to break translational and reflection symmetries:

$$H = h_{z1}\sigma_1^z + h_{zN}\sigma_N^z + \sum_r \left[J\sigma_r^z\sigma_{r+1}^z + h_z\sigma_r^z + h_z\sigma_r^x\right] \tag{3.14}$$

We choose the parameters $J = 1.0$, $h_z = 0.5$, $h_x = -1.05$, $h_{z1} = -0.45$, $h_{zN} = 0.15$.

We split the system into two subsystems A and B with N_A and $N_B = N - N_A$ spins, respectively, and define subsystem Hamiltonians,

$$H_A = h_{zN}\sigma_N^z + \sum_{r=N_B+1}^{N} \left[J\sigma_r^z\sigma_{r+1}^z + h_z\sigma_r^z + h_z\sigma_r^x\right],$$

$$H_B = h_{z1}\sigma_1^z + \sum_{r=1}^{N_B} \left[J\sigma_r^z\sigma_{r+1}^z + h_z\sigma_r^z + h_z\sigma_r^x\right]. \tag{3.15}$$

In this setup, we study the object we previously examined in figure 5 and redrawn in figure 9 for convenience. We had argued that C should scale as $\mathcal{O}(e^{-2S-S_A})$ based on correlations under the implicit overline averaging. Now, we explicitly implement the overline averaging via a Gaussian weight function over a narrow energy band of width ε, $\Delta_\varepsilon^{(iab)} \equiv e^{-\frac{(E_i - E_a - E_b)^2}{2\varepsilon^2}}$. Thus, we compute

$$C \equiv \overline{c_{ab}^i c_j^{ab'} c_{a'b'}^j c_i^{a'b}} \equiv \frac{\sum_{[ijaa'bb']} \Delta_\varepsilon^{(iab)} \Delta_\varepsilon^{(jab')} \Delta_\varepsilon^{(ja'b')} c_{ab}^i c_j^{ab'} c_{a'b'}^j c_i^{a'b}}{\sum_{[ijaa'bb']} \Delta_\varepsilon^{(iab)} \Delta_\varepsilon^{(jab')} \Delta_\varepsilon^{(ja'b')}} \tag{3.16}$$

Our results are presented in figure 10. We check the scaling with respect to N by fixing $N_A = 6$ and varying N from 10 to 14. We also check the scaling of N_A by fixing $N = 12$ and varying N_A from 5 to 7. In all cases, we find $\varepsilon = 0.4$ to be adequate. We are significantly limited by finite size corrections associated with the subsystem sizes which require us to discard data points for N_A, $N_B < 5$. Nonetheless, we find good agreement between the slopes of $\ln\left(d^{-2}d_A^{-1}\right) = (-2N - N_A)\ln(2)$ and $\ln(C)$ for accessible system sizes, which supports our analytical expectation of the scaling of C with N and N_A.
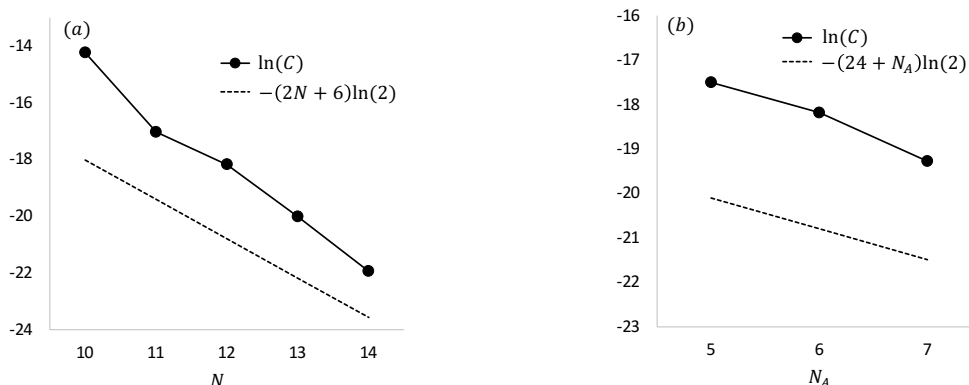
Figure 10. Scaling of C (eq. (3.16)) vs. (a) N for fixed $N_A = 6$, and (b) N_A for fixed $N = 12$. We compare are numerical calculations for an Ising spin chain (solid dots) to the asymptotic scaling associated with the Hilbert space dimensions (dashed line). We find qualitatively good agreement in both figures (a) and (b) for the slope of the curve.

4 Eigenstate correlations I: the structure of chaotic eigenstates

In this section, we consider an eigenstate $|i\rangle$ of the full system Hamiltonian H. We focus on two aspects of thermalization: the reduced density matrix and entanglement entropy. In section 4.1, we compute the on and off-diagonal elements of the reduced density matrix on subsystem A, improving the calculation of ref. [21] and generalizing that of ref. [31]. In section 4.2, we reproduce best-known results for the entanglement entropy of chaotic eigenstates from our formalism and discuss subextensive corrections. In doing so we remark on a qualitative resemblance of our calculation to the gravitational path integral calculation of the entanglement entropy of an evaporating black hole that is inherently free probabilistic [32–34]. Lastly, in section 4.3, we show that reduced density matrices of nearby eigenstates are exponentially close in trace distance, a hypothesis known as the subsystem ETH [35].

4.1 Reduced density matrix

We wish to study the matrix elements of the reduced density matrix of subsystem A,

$$\rho^i_{aa'} = \langle a|\mathrm{Tr}_B\left[|i\rangle\langle i|\right]|a'\rangle = \sum_b c_i^{ab} c_{a'b}^i. \tag{4.1}$$

Focusing on the diagonal elements first, we find a simple expression,

$$\rho^i_{aa} = \sum_b c_i^{ab} c_{ab}^i = \int_{E_b} e^{S_B(E_b)-S(E_i)} F(E_i - E_a - E_b) = e^{-S(E_i)+S_B(E_i-E_a)} \tag{4.2}$$

which reduces to a Gibbs state when A is much smaller than B. This term is represented in figure 11. We evaluated eq. (4.2) at the trivial saddle-point $E_i - E_a - E_b = 0$ associated with neglecting the finite width of F. If we neglect the width of window functions we are able to take the arguments of F functions (other than the overall energy) as saddle-points. However, including a nonzero width of F will slightly modify eq. (4.2) and the outcomes of the various saddle-point integrals we take in this paper. However, such a modification will
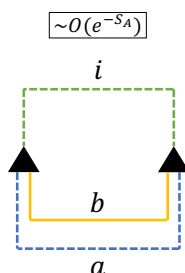
Figure 11. The diagonal elements of a reduced density matrix are represented by the simplest eigenstate cumulant for a bipartitioned system.

only be felt by the fluctuations of approximately conserved operators (i.e. the fluctuations of subsystem energies) and only when subsystem A is a finite fraction of the whole system (see appendix C.1).

For off-diagonal elements, we can immediately see that $\rho^i_{aa'}$ contains uncontracted indices and must have mean zero. We study instead, the variance,

$$\rho^i_{aa'}\rho^i_{a'a} = \sum_{bb'} c^{ab}_i c^i_{a'b} c^{a'b'}_i c^i_{ab'}. \tag{4.3}$$

We represent eq. (4.3) in figure 12. Notably, the leading order contribution to eq. (4.3) depends on the smaller of $S_A(E_a)$ and $S_B(E_i - E_a)$. We define $S_{\min}(E_a, E_b) \equiv \min\left[S_A(E_a), S_B(E_b)\right]$ and $S_{\min}(E_a) \equiv \min\left[S_A(E_a), S_B(E_i - E_a)\right]$. Then, we can read off leading order terms from figure 12 and evaluate via saddle-points,

$$\begin{aligned}
|\rho^i_{aa'}|^2 &= \sum_{[bb']} c^{ab}_i c^i_{a'b} c^{a'b'}_i c^i_{ab'} + \sum_b c^{ab}_i c^i_{a'b} c^{a'b}_i c^i_{ab} \\
&= \int_{E_b E_{b'}} e^{2S_B(E_{bb'}) - 2S(E) - S_{\min}(E_{aa'}, E_{bb'})} F(\cdots) + \int_{E_b} e^{S_B(E_{bb'}) - 2S(E)} F_{(bb')}(\cdots) \\
&= e^{-2S(E) + 2S_B(E_i - E_{aa'}) - S_{\min}(E_{aa'})} \tilde{F}(E_{aa'}; \omega_{aa'}) \tag{4.4}
\end{aligned}$$

where $\tilde{F}(E_{aa'}; \omega_{aa'})$ has been defined to capture a leftover $\omega_{aa'}$ dependence that suppresses correlations away from the diagonal. Combining eqs. (4.4) and (4.2), we obtain for the reduced density matrix,

$$\rho^i_{aa'} = e^{-S(E_i) + S_B(E_i - E_{aa'})} \left(\delta_{aa'} + \sqrt{e^{-S_{\min}(E_{aa'})} \tilde{F}(E_{aa'}; \omega_{aa'})} R_{aa'}\right) \tag{4.5}$$

where $R_{aa'}$ is an approximate Gaussian random matrix with mean zero and variance one that encodes higher correlations in the reduced density matrix. Eq. (4.5) is a generalization of the state-averaging ansatz of ref. [31] to the case of arbitrary subsystem sizes. Ref. [31] further discusses higher correlations in $R_{aa'}$ which we discuss implicitly in the next section in the context of entanglement entropy.

Eq. (4.5) also improves the result of ref. [21] which neglected the terms in figure 12(a) while keeping the term in figure 12(b). As a result, instead of our factor of $S_{\min}$, they had a factor of $S_B(E_i - E_{aa'})$ in our notation. Our improvement implies a much smaller suppression of off-diagonal elements of the reduced density matrix below the critical energy.
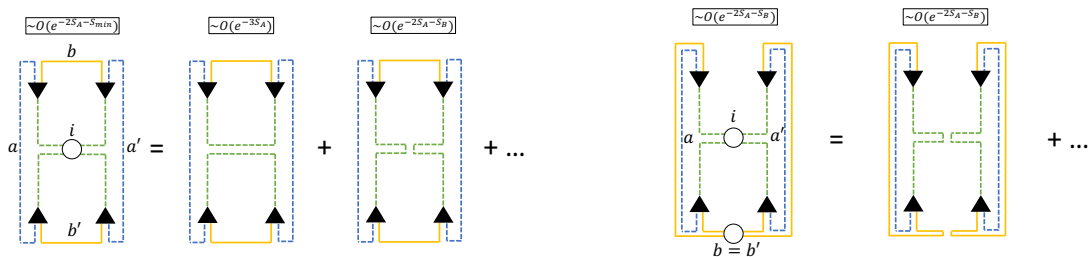
(a) A true 4-point contribution which captures sharp behavior in system size when $S_A \sim S_B$. Note the similarity to figure 7(a).

(b) Contribution that factors into 2-point functions and only contributes at leading order when $S_A \gtrsim S_B$. Note the similarity to figure 7(c).

Figure 12. Diagrams which contribute to the off-diagonal matrix elements of reduced density matrices.

This improvement is physically necessary on the following grounds. The suppression by e^{-S_B} would imply that the eigenstates of H_A are exponentially close to the eigenstates of ρ_A^i via $\langle a | (\rho_A^i)^2 | a \rangle = (\rho_{aa}^i)^2 \left(1 + \mathcal{O}(e^{-S_B}) \right)$. However, H_A is only defined up to an arbitrary area scaling term on the boundary of the subsystem and any two definitions of H_A would likely share no eigenstates. In contrast, ρ_A^i is defined unambiguously and cannot have eigenstates that are simultaneously exponentially close to eigenstates of all definitions of H_A. Another way of seeing the same problem is to consider preparing the system in an eigenstate of the total Hamiltonian H and then switching off the interaction H_{AB} thereby perturbing subsystem A along its boundary. If the off-diagonal elements of ρ_A^i are suppressed by e^{-S_B}, the timescale for diagonal elements of ρ_A^i to evolve is exponentially long $\sim \mathcal{O}(e^{S_B - S_A})$. Physically, however, we should only expect the timescale to be the time it takes for information from the boundary of A to reach the rest of the subsystem.

4.2 Entanglement entropies and the Page curve

To compute the entanglement entropies we first focus on the α-Renyi entropies for $\alpha \geq 2$,

$$S_\alpha \equiv \frac{1}{1-\alpha} \ln \left(\mathrm{Tr}_A \left[(\rho_A^i)^\alpha \right] \right) = \frac{1}{1-\alpha} \ln \left(\sum_{\{ab\}} c_{a_1 b_1}^i c_i^{a_2 b_1} \cdots c_{a_\alpha b_\alpha}^i c_i^{a_1 b_\alpha} \right). \qquad (4.6)$$

The diagrammatics for eq. (4.6) are represented in figure 13 for $\alpha = 2, 3, 4$. To leading order, we can ignore additional contractions in the sum and focus only on the full diagrams. The leading order terms will come from different ways of decomposing the contraction of i indices into pairs. When subsystem $A(B)$ is far smaller than its complement, there is a single leading partition with weight $e^{-\alpha S(E_i) - (\alpha - 1) S_{A(B)}(E_{a(b)})}$. However, when the subsystems are similar in size, there will be a critical energy $E_{A(B)}^*$ such that $S_{A(B)}(E_{A(B)}^*) = S_{B(A)}(E_i - E_{A(B)}^*)$ which denotes a crossover between leading partitions. Near $E_{A(B)}^*$, there will also be contributions from all other *non-crossing* pairings of i indices, and we draw examples of these partitions in figure 14. However, these partitions will, at best, contribute an area law term to entanglement entropy for $\alpha < 1$ and will be neglected other than to match terms to the time-dependent case we consider in section 5.
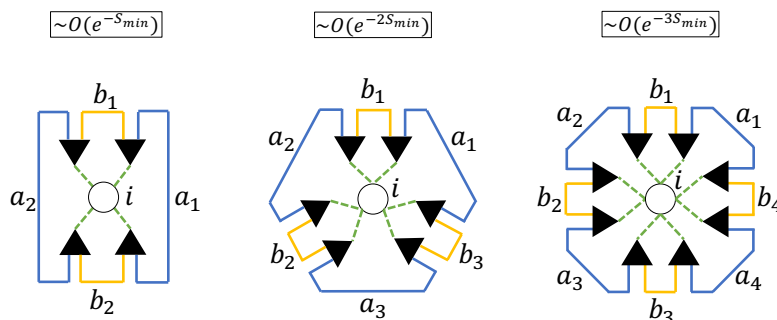
Figure 13. Diagrams which compute the second, third, and fourth moments of the reduced density matrix contained inside the log in eq. (4.6). For general Renyi index α, the diagram will appear as an α-fold rotationally symmetric version of the diagrams above.
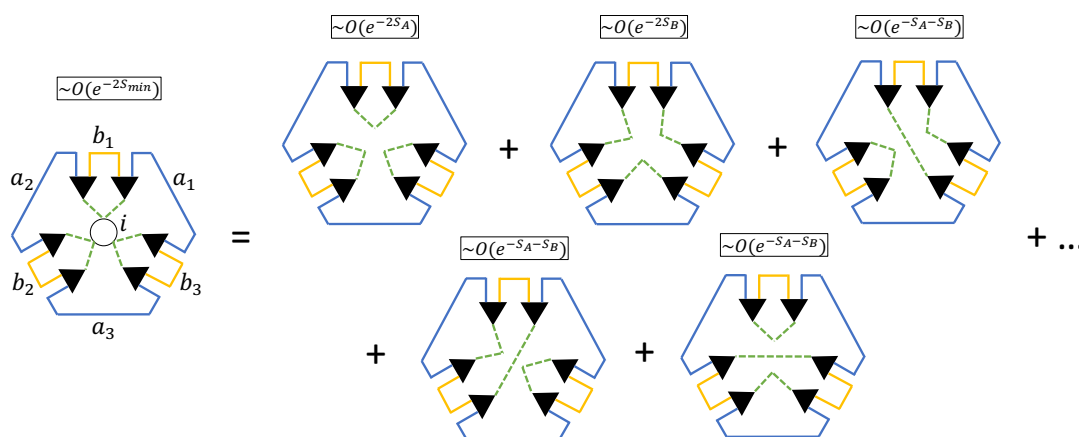

Figure 14. Pictured are the 5 five leading order contributions to eq. (4.6) for $\alpha = 3$, which come from the non-crossing pair partitions of the i index. There is be a term associated to each pair partition of 2α i indices. The non-crossing condition is not universally true of index contractions in index diagrams but are a consequence of the ordinary rules presented in section 3.1 applied to the Renyi entropy diagrams, figure 13. We focus only on the first 2 terms as the remaining 3 only contribute near the isolated point $S_A(E_a) = S_B(E_b)$. Not drawn are contractions of a and b indices that will also contribute near the critical energy so long as they do not create any additional crossings.

Then we can read off for general α,

$$S_\alpha = \frac{1}{1-\alpha} \ln\left(\int_{E_a E_b \{\omega\}} e^{\alpha S_A(E_a) + \alpha S_B(E_b) - \alpha S(E_i) - (\alpha-1)S_{\min}(E_a, E_b)} F(\cdots) \right) \qquad (4.7)$$

and taking $\alpha \to 1$,

$$S_1 = \int_{E_a E_b} e^{S_A(E_a) + S_B(E_b) - S(E_i)} S_{\min}(E_a, E_b) F(E_i - E_a - E_b). \qquad (4.8)$$

Eqs. (4.7) and (4.8) are equivalent to expressions derived in ref. [21] and our results are equivalent as well. In taking the $\alpha \to 1$ limit, we have neglected the implicit dependence of F on α. This neglect is valid as long as we also neglect the non-zero width of F.

We can then evaluate the integrals via saddle-points. The first saddle-point condition is given by F which will enforce that $E_a + E_b = E_i$. The second saddle-point condition is given by maximizing the exponents in eqs. (4.7) and (4.8). For $\alpha \neq 1$,

$$S_\alpha = \frac{1}{1 - \alpha} \left\{ \alpha \left[S_A(\mathcal{E}_A) + S_B(\mathcal{E}_B) - S(E_i) \right] - (\alpha - 1) S_{\min}(\mathcal{E}_A, \mathcal{E}_B) \right\} \qquad (4.9)$$

where $\mathcal{E}_A$ and $\mathcal{E}_B$ are the subsystem energies that dominantly contribute to the Renyi entropies and are defined by the saddle-point conditions of eq. (4.7):

$$\text{(I)} \quad \mathcal{E}_A + \mathcal{E}_B = E_i \quad \& \quad \begin{cases} \text{(II.1)} & S'_A(\mathcal{E}_A) = \alpha S'_B(\mathcal{E}_B) \\ \text{(II.2)} & S'_B(\mathcal{E}_B) = \alpha S'_A(\mathcal{E}_A) \\ \text{(II.3)} & S_A(\mathcal{E}_A) = S_B(\mathcal{E}_B) \end{cases} \qquad (4.10)$$

Here, condition (II) is determined by which of the three given saddle-points minimizes S_α. For $\alpha > 1$, the saddle point is given by condition (II.1) if $S_A < S_B$ and condition (II.2) if $S_A > S_B$. For $\alpha < 1$, the same result holds except at high temperatures and when the systems are of similar size, where condition (II.3) can hold, in which case $\mathcal{E}_A = E_A^*$. For $\alpha = 1$, we derive

$$S_1 = \min \left[S_A(E_{i,A}), S_B(E_{i,B}) \right] + \Delta S \qquad (4.11)$$

where the saddle-point conditions of (4.8) are given by,

$$\text{(I)} \quad E_{i,A} + E_{i,B} = E_i \quad \& \quad \text{(II)} \quad S'_A(E_{i,A}) = S'_B(E_{i,B}) \qquad (4.12)$$

and $\Delta S < 0$ denotes subextensive corrections to S_1 we will discuss shortly.

First, we note that eq. (4.11) recovers a finite temperature version of the Page curve, noted for its linear slopes as a function of subsystem size [59]. However, for $\alpha \neq 1$, eq. (4.9) does not obtain this form. In particular, $S_{\alpha>1}$ is superadditive, and $S_{\alpha<1}$ is subadditive and thus are convex and concave functions of system size, respectively. This convexity/concavity is referred to as the "failure of the Page curve" [20]. We can interpret this failure in the following manner. The subsystems of a chaotic eigenstate $|i\rangle$ obtain an effective thermodynamic inverse temperature from $S(E_i)$, $\beta \equiv S'(E_i)$, which determines all aspects of local physics in finite energy density states. However, for $\alpha \neq 1$, Renyi entropies of subsystems that are a finite fraction of the whole system have access to highly *nonlocal* information that encodes physics at different temperatures [60]. Thus, as the size of a small subsystem is increased, the Renyi entropies gain access to more information from different parts of the spectrum and bend accordingly. For $\alpha > 1$ ($\alpha < 1$), the Renyi entropies are dominated by low (high) temperature physics and as $\alpha \to \infty$ ($\alpha \to 0$) the entropy will approach that of the ground (infinite-temperature) state.

We now discuss the corrections to the Von Neumann entropy, ΔS. One generally expects area law corrections to entanglement entropy. To acquire them, one would need to carefully consider the structure of the F functions in the $\alpha \to 1$ limit. As the F functions are generally system-dependent, so are the area law corrections, but we do not preclude the possibility that some generic structure may exist. There is another correction from contributions away from the saddle point in eq. (4.8). These contributions are controlled by the heat capacities

of the subsystems and, in homogeneous systems, contribute a correction to entanglement entropy that is order square root in system size. This term is derived in ref. [21], and we direct readers to their calculation rather than repeat it. This is the leading order correction in $D < 2$ dimensional systems, whereas the area law correction generally is leading order in $D > 2$ dimensions. For $D = 2$ dimensions both terms have equal order.

Despite the fact that previous studies of eigenstate entanglement [20, 21, 35, 61, 62] did not have a general way to compute higher eigenstate correlations, refs. [20, 21] still successfully computed the Renyi and Von Neumann entropies.[16] Ref. [20] utilized similar assumptions to our own and computed the moments of the reduced density matrix averaged over all states consistent with the form in eq. (1.2). Ref. [21], on the other hand, computed entanglement entropies by guessing a distribution of eigenvalues of the reduced density matrix. Since the moments of the reduced density matrix are entirely determined by its eigenvalues, it is not necessarily surprising that this calculation could be done without direct computation of higher correlations.

The non-crossing constraint depicted in figure 14 is similar to one found in the gravitational path integral computation for the Page curve of an evaporating black hole in JT gravity and discussed in refs. [32–34]. A connection was drawn to the non-crossing partitions of free probability theory in ref. [33] and explicated in ref. [34] in terms of so-called Kreweras complements and free multiplicative convolution. To make the analogy precise, we model subsystem B as an evaporating black hole and subsystem A as its radiation. At early times, when B is much larger than A, the Page curve is dominated by the first partition on the right-hand side of figure 14 where pairings of i-indices disconnect the replicas of B. At late times, when most of B has evaporated, the Page curve is dominated by the second partition where pairings of i-indices connect all replicas of B together. Where in our calculation arise simple index contractions, the gravitational path integral predicts semiclassical wormholes that (dis)connect the black hole replicas (contrast our figures 13 and 14 with figure 2 of ref. [34]). Ref. [63] finds a similar analogy between index contractions and replica wormholes in a toy model for entanglement dynamics.

The common thread between all calculations is an implicit ensemble averaging over correlations of wavefunction overlaps that connect originally disconnected replicas and suppress crossings by an appropriate density-of-states factor. The validity of such an averaged calculation in systems without explicit ensemble averaging is referred to as the *factorization problem* [63]. The ETH may provide a solution. In our calculation, the ensemble average emerged as the smooth portion of the GFCs that dominate over fluctuating portions without any explicit ensemble average. Similar logic is summoned in ref. [31] to justify the appearance of semiclassical wormholes in evaluating gravitational path integrals.

4.3 Subsystem ETH

The expression for the density matrix given in eq. (4.5) holds up to polynomial corrections in the system size. In contrast, ref. [35], conjectured that reduced density matrices for a given

---

[16]Aside from the fact that [20] missed the possible saddle-point condition (II.3) in (4.10) and neglected subleading corrections.

subsystem of nearby[17] eigenstates of the full system should be *exponentially* close in trace distance when the subsystem is smaller than one-half the system. More precisely,

$$||\rho_A^i - \rho_A^j||_1 \equiv \text{Tr}_A \left[|\rho_A^i - \rho_A^j| \right] \sim \mathcal{O}(e^{-S/2}), \tag{4.13}$$

for $E_i = E_j$ up to irrelevant corrections of order the level spacing. The significance of this definition is that it implies the existence of an equilibrium reduced density matrix, ρ_A^{ETH} that specifies the thermal properties of subsystem A to a far greater degree than that of the canonical ensemble.

Our diagrammatic formalism does not directly compute the Schatten 1-norm $|| \cdots ||_1$. However, we can compute the Schatten 2-norm and utilize the following bound,[18]

$$||M||_2 \leq ||M||_1 \leq \sqrt{\text{rank}(M)}||M||_2 \tag{4.14}$$

which holds for an arbitrary operator M and is a corollary of the Cauchy-Schwarz inequality. First, let us expand,

$$
\begin{aligned}
||\rho_A^i - \rho_A^j||_2^2 \equiv \text{Tr}_A \left[\left(\rho_A^i - \rho_A^j \right)^2 \right] &= -2 \, \text{Tr}_A \left[\rho_A^i \rho_A^j \right] + \text{Tr}_A \left[(\rho_A^i)^2 \right] + \text{Tr}_A \left[(\rho_A^j)^2 \right] \\
&\approx 2 \left(\text{Tr}_A \left[(\rho_A^i)^2 \right] - \text{Tr}_A \left[\rho_A^i \rho_A^j \right] \right) \\
&\equiv 2 \left((\rho_A^i | \rho_A^i) - (\rho_A^i | \rho_A^j) \right)
\end{aligned} \tag{4.15}
$$

where we the error in the approximation $\text{Tr}_A \left[(\rho_A^i)^2 \right] \approx \text{Tr}_A \left[(\rho_A^j)^2 \right]$ is suppressed by e^{-S} when $|i\rangle$ and $|j\rangle$ are nearby eigenstates.[19]

We can now recognize that the diagrammatics of $(\rho_A^i | \rho_A^i)$ and $(\rho_A^i | \rho_A^j)$ are given in figures 7 and 6, respectively. By inspection, the leading order difference comes from the terms in figures 7(a) and 7(c) that are absent in 6(a) and 6(c), where the contraction of index i allowed diagrams of order e^{-S_B} to contribute. Thus,

$$
\begin{aligned}
(\rho_A^i | \rho_A^i) - (\rho_A^i | \rho_A^j) &= \int_{E_b E_{b'} E_a E_{a'}} e^{2S_A(E_{aa'}) + S_B(E_{bb'}) - 2S(E_i)} F_{(ij)}(\cdots) \\
&\quad + \int_{E_b E_a E_{a'}} e^{2S_A(E_{aa'}) + S_B(E_b) - 2S(E_i)} F_{(ij)(bb')}(\cdots) \\
&\sim e^{-S(E_i) + S_A(E_{iA})}.
\end{aligned} \tag{4.16}
$$

Lastly, recognizing that $\ln[\text{rank}(\rho_A)]$ is the 0-Renyi entropy, per eq. (4.9), we can recognize that the reduced density matrices are full rank (up to a polynomial correction in subsystem sizes): $\text{rank}(\rho_A^{i(j)}) \sim e^{S_{\min}^{(\infty)}}$, where $e^{S_{\min}^{(\infty)}}$ is the microcanonical entropy of the smaller subsystem at infinite temperature. Then, $\text{rank}(\rho_A^i - \rho_A^j) \lesssim e^{S_{\min}^{(\infty)}}$ and eq. (4.14) becomes

$$\mathcal{O}(e^{S_A/2 - S/2}) \lesssim ||\rho_A^i - \rho_A^j||_1 \lesssim \mathcal{O}(e^{S_{\min}^{(\infty)}/2 + S_A/2 - S/2}) \tag{4.17}$$

which establishes the subsystem ETH as a consequence of the MBBC by its e^S dependence. The remaining factor crucially dictates the subsystem sizes for which the trace distance is suppressed [35, 36]. However, this factor remains to be determined and we leave a direct computation of the trace distance to future work.

---

[17]Defined as having a difference in energy at the order of the level spacing.

[18]It was brought to our attention that under the same assumptions a replica calculation of fidelities yields a stronger bound on the trace distance. This calculation was performed in ref. [36] (eq. (197)).

[19]The given suppression can be justified by recognizing the gradual dependence of the 2-Renyi entropy on total energy and its small fluctuations between nearby eigenstates.

5 Eigenstate correlations II: thermalization of a non-equilibrium initial state

In this section, we will show how a system that is not prepared in an eigenstate reaches thermal equilibrium. The general result reproduces the *equilibrated pure state* formalism of ref. [32]. Specifically, we show that reduced density matrices relax towards a form associated with their energy density, and entanglement entropies will relax towards their equilibrium values. The key insight that provides our general result in this section is that (a) time-dependent partitions vanish and (b) the time-independent partitions factor into a product of an "outer" partition that is identical to that of a system prepared in equilibrium and an "inner" partition that integrates out. In the special cases we consider, this factorization will reproduce the partitions we saw in section 4.

We initialize our system in a product of subsystem eigenstates $|ab\rangle$ s.t. $H_A^n H_B^m |ab\rangle = E_a^n E_b^m |ab\rangle$, but we stress that our main results do not depend on this choice. Our state has initial energy $E_0 = E_a + E_b$, and all states with initial energy E_0 will relax to the same equilibrium so long as we neglect the width of F. In appendix C.1, we show how fluctuations of the subsystem energy are sensitive to the specific form of F and do not exactly obtain the properties of eigenstates, even in equilibrium, when subsystem A is a finite fraction of the system. Ultimately, this is because the energy of a thermodynamically large subsystem is an approximately conserved quantity.

Since dynamics are generally system-dependent, the specific forms of time-dependent partitions that encode dynamics will be system-dependent as well. However, that does not preclude generic features in the time-dependent partitions. Indeed, in section 5.3, we find that the growth of entanglement entropy has an intriguing diagrammatic organization. We find that distinct diagrams contain the ballistic growth, diffusive growth, and saturation of entanglement entropy.

5.1 Reduced density matrix

Consider the matrix elements of the reduced density matrix on subsystem A over time given the initial state $|ab\rangle$:

$$\rho_{a'a''}(t) = \langle a'|\text{Tr}\left[e^{-iHt}|ab\rangle\langle ab|e^{iHt}\right]|a''\rangle \tag{5.1}$$

$$= \sum_{b'ij} \langle a'b'|i\rangle\langle i|ab\rangle\langle ab|j\rangle\langle j|a''b'\rangle \, e^{-i(E_i-E_j)t} \tag{5.2}$$

$$\equiv \sum_{b'ij} c_i^{a'b'} c_{ab}^i c_j^{ab} c_{a''b'}^j e^{-i\omega_{ij}t}. \tag{5.3}$$

We will once again be interested in both the diagonal elements $\rho_{a'a'}$ and the variance of off-diagonal elements $|\rho_{a'a''}|^2$. Let us first consider the diagonal elements,

$$\rho_{a'a'}(t) = \sum_{b'ij} c_i^{a'b'} c_{ab}^i c_j^{ab} c_{a'b'}^j e^{-i\omega_{ij}t}. \tag{5.4}$$

For ρ to relax to its equilibrium form in general, two things must be true: (1) the time-independent part of ρ must equal its equilibrium form, and (2) the time-dependent part
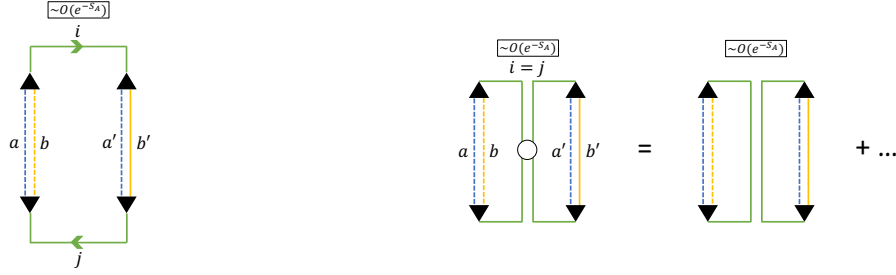
(a) Uncontracted partition. Several contractions exist, however, only ones which cancel that time dependences on i and j will be time-independent.

(b) Time independent contribution to the reduced density matrix. Note, there is only one sum over index i.

Figure 15. In (a) we have placed the uncontracted partition corresponding to eq. (5.4) for clarity. Time dependence ensures that it will decay as long as its corresponding F-function is smooth. In (b), we have placed the leading order time-independent partition. Time-independence is achieved by contracting indices with opposite time dependences. Further time-independent partitions can be generated by contracting $a = a'$ or $b = b'$. However, the former case will only contribute a factor of 2 for the matrix element ρ_{aa} while the latter case will suppressed by a factor of e^{-S_B}, so we neglect both cases. Focusing on the right hand side of the equation in (b) on the left, we have a partition that fully integrates out when summing over index i; we refer to this as the "inner" partition. On the right, we have an "outer" partition that is identical to the partition pictured in figure 11 and describes the diagonal elements of an equilibrium reduced density matrix. When we consider higher correlations the "inner" and "outer" labels will become more visually apparent.

must vanish. We show the former in this section. The latter condition is contained in the smoothness assumption of the ETH — the belief that there is a finite energy scale below which no structure can be seen in matrix elements. Given these conditions, in the $t \to \infty$ limit the oscillatory terms vanish and the time-independent value of ρ is obtained.

Diagrammatically, time-independent partitions are found by contracting indices with opposite time-dependences, thereby removing oscillatory terms. The expression in eq. (5.4) has several unique partitions, but only 1 is both nontrivial and time-independent, so we will restrict our attention to it (see figure 15). We compute,

$$
\begin{aligned}
\rho_{a'a'}(\infty) &= \sum_{b'i} \left| c_{a'b'}^i \right|^2 \left| c_{ab}^i \right|^2 \\
&= \int_{E_i E_{b'}} e^{S(E_i) + S_B(E_{b'}) - 2S(E_i)} F(E_i - E_a - E_b) F(E_i - E_{a'} - E_{b'}) \\
&= e^{-S(E_0) + S_B(E_0 - E_{a'})}
\end{aligned}
\tag{5.5}
$$

which is what we expect from the equilibrium case considered previously, though the second F function will modify the subsystem energy variance expression (see appendix C.1). We can see from this example that the recovery of the equilibrium result came from a factorization of the time-independent partition into a partition equivalent to the one considered for eigenstates, and a partition that was integrated out in eq. (5.5). This correspondence between time-independent partitions and equilibrium partitions is general.
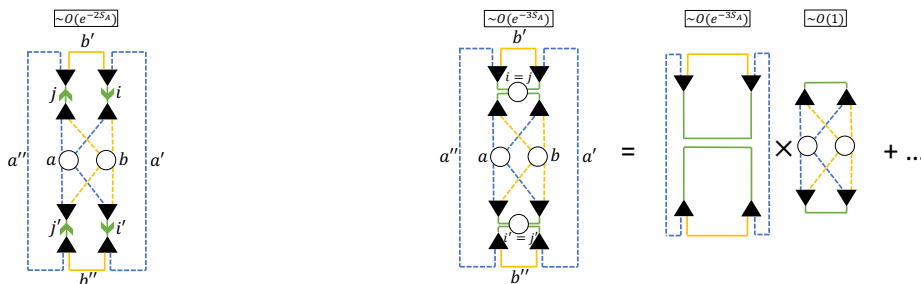
(a) Full partition representing the time-dependent off-diagonal elements of the reduced density matrix on subsystem A.

(b) Time-independent partition $(ij)(i'j')$ that contributes to the off-diagonal elements of the reduced density matrix on subsystem A.
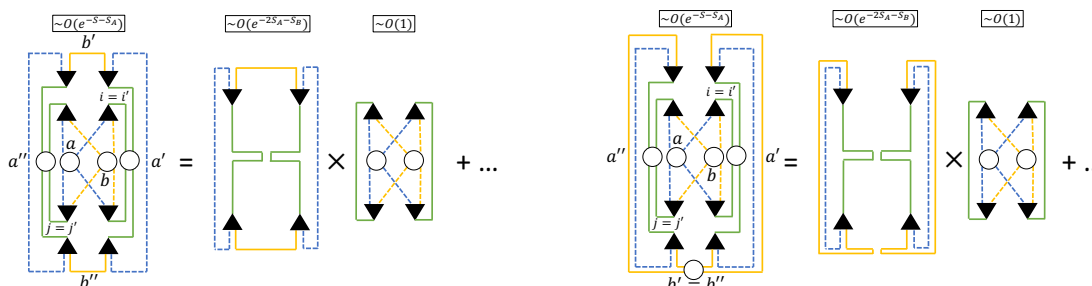
(c) Time-independent partition $(ii')(jj')$ that contributes to the off-diagonal elements of the reduced density matrix on subsystem A.

(d) Time-independent partition $(ii')(jj')(b'b'')$ that contributes to the off-diagonal elements of the reduced density matrix on subsystem A.

Figure 16. Diagrams that contribute to the off-diagonal matrix elements of the reduced density matrix on subsystem A. Subfigures (b-d) picture the time-independent terms and the factorizations that dominantly contribute. Each factorization contains an inner partition that integrates out and an outer partition that reduces to the partitions in figure 12. Note that one should not overcount the number of integrals on the right-hand side of each diagrammatic equation.

Next, we address the off-diagonal elements,

$$\rho_{a'a''}(t)\rho_{a''a'}(t) = \sum_{b'b''iji'j'} c_i^{a'b'} c_{ab}^i c_j^{ab} c_{a''b'}^j c_{i'}^{a''b''} c_{ab}^{i'} c_{j'}^{ab} c_{a'b''}^{j'} e^{-it(E_i - E_j - E_{i'} + E_{j'})}. \tag{5.6}$$

There are several dozen independent partitions; however, we will focus on just 4: the full partition and the 3 largest time-independent partitions (see figure 16). The full partition drawn in figure 16(a), is given by

$$F(E_{a'a''}; \omega_{a'a''}; t) = \sum_{[b'b''iji'j']} c_i^{a'b'} c_{ab}^i c_j^{ab} c_{a''b'}^j c_{i'}^{a''b''} c_{ab}^{i'} c_{j'}^{ab} c_{a'b''}^{j'} e^{-it(E_i - E_j - E_{i'} + E_{j'})}$$

$$= e^{-2S(E_a + E_b) + 2S_B(E_a + E_b - E_{a'a''})} F(\omega_{a'a''}; t) \tag{5.7}$$

where $F(\omega_{a'a''}; t)$ is a smooth term fated to die off in t, whatever its precise form. There are no other partitions that contribute at the same order. This partition will be studied in the context of the 2-Renyi Entropy in section 5.3.

Moving on to the time-independent partitions, $(ij)(i'j')$, $(ii')(jj')$, $(iji'j')$, we find that the partitions in figures 16(b), 16(c), 16(d) factor into inner partitions that integrate out and outer partitions that resemble those in figures 12(a) and 12(b). Then again, we see that time-independent partitions reduce to equilibrium partitions,

$$
\begin{aligned}
|\rho_{a'a''}(\infty)|^2 &= \int_{E_b' E_{b''} E_i E_{i'}} e^{2S_B(E_{b'b''})-2S(E_0)-S_A(E_{a'a''})} F_{(ij)(i'j')}(\cdots) \\
&+ \int_{E_{b'} E_{b''} E_i E_{i'}} e^{2S_B(E_{b'b''})-2S(E_0)-S_B(E_{b'b''})} F_{(ii')(jj')}(\cdots) \\
&+ \int_{E_{b'} E_{b''} E_i E_{i'}} e^{S_B(E_{b'b''})-2S(E_0)} F_{(ii')(jj')(b'b'')}(\cdots) + \cdots \\
&= e^{-2S(E_0)+2S_B(E_0-E_{a'a''})-S_{\min}(E_{a'a''})} \tilde{F}(\omega_{a'a''})
\end{aligned}
\tag{5.8}
$$

which is the same expression derived in eq. (4.4).

5.2 Entanglement entropies and the Page curve

Once again, initializing in $|ab\rangle$, we repeat our analysis for the Renyi entropies. The α-Renyi entropy is

$$
\begin{aligned}
S_\alpha(t) &\equiv \frac{1}{1-\alpha} \ln\left(\text{Tr}_A\left[\left(\text{Tr}_B\left[e^{-iHt}|ab\rangle\langle ab| e^{iHt} \right] \right)^\alpha \right] \right) \\
&= \frac{1}{1-\alpha} \ln\left(\sum_{a'_m b'_m i_m j_m} c_{i_1}^{a'_1 b'_1} c_{j_1}^{i_1} c_{j_1}^{ab} c_{b'_1 a'_2}^{j_1} \cdots c_{i_\alpha}^{a'_\alpha b'_\alpha} c_{j_\alpha}^{i_\alpha} c_{j_\alpha}^{ab} c_{b'_\alpha a'_1}^{j_\alpha} e^{-it\sum_m \omega_{i_m j_m}} \right).
\end{aligned}
\tag{5.9}
$$

Time-independent partitions are those for which $\sum_m \omega_{i_m j_m}$ vanishes, which are obtained by contracting i indices with j. We have drawn the minimally contracted and the largest time-independent partitions for subsystem A much smaller than B for the 2-, 3-, and 4-Renyi entropies in figure 17.

When $S_B > S_A$, the largest time-independent partition is $(i_m j_m)$, while for $S_A > S_B$, the largest is $(i_{m+1} j_m)$. When $S_A \approx S_B$, any pairing of i indices with j indices that is non-crossing, as depicted previously in figure 14, contributes. Once again, we can see that the time-independent partitions factor into an inner partition that integrates out and outer partitions that are identical to those considered in the equilibrium case. Since the result is general, we do not repeat our calculations from section 4.2.

5.3 Entanglement growth

From an initially unentangled state, the dynamics of Renyi entropies $\alpha > 1$ generically exhibit 4 regimes: (i) local equilibration, (ii) ballistic growth, (iii) diffusive growth, and (iv) saturation. At late times, the equilibrium partitions once again map to eigenstate partitions and compute a Page curve. In the previous subsection we associated saturation to equilibrium partitions. In this subsection we will provide similar diagrammatic interpretations to the different regimes of entanglement growth by counting entropic factors (see figure 18).

Initially, entanglement growth is limited by fidelity to the initial state [30],

$$
S_\alpha(t \approx 0) \lesssim \frac{1}{1-\alpha} \ln |F_R(t)|^{2\alpha}
\tag{5.10}
$$

where $F_R(t)$ is the return amplitude $F_R(t) \equiv \langle ab|e^{-iHt}|ab\rangle$ which arises under the contraction $(a_1 \ldots a_\alpha a)(b_1 \ldots b_\alpha b)$, represented for $\alpha = 2$ in figure 18(a). We discuss functional forms
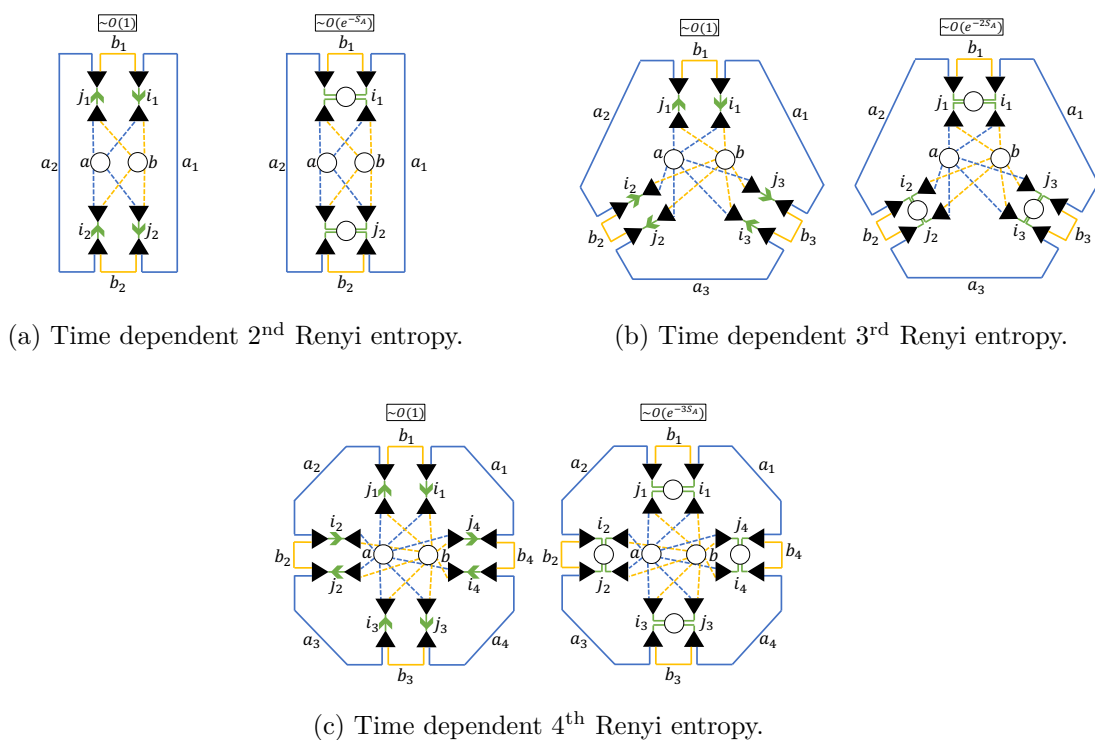
(a) Time dependent 2nd Renyi entropy.

(b) Time dependent 3rd Renyi entropy.

(c) Time dependent 4th Renyi entropy.

Figure 17. Time-dependent Renyi entropies adjacent to their leading time-independent counterpart. The contractions between i and j indices map one-to-one with the pairings of i for the equilibrium case, represented for the 3-Renyi entropy in figure 14.

for $F_{\mathrm{R}}(t)$ in appendix C.2. At a later time, $t_{\mathrm{eq}} \sim \mathcal{O}(1)$, the system will have reached local equilibrium [64] and will transition to ballistic entanglement growth. However, as we discuss later in this section, eq. (5.10) does not provide an adequate accounting of entanglement growth for $t \lesssim t_{\mathrm{eq}}$ and further work is needed to clarify the local equilibration timescale. The problem of modeling the propagator at short times in chaotic systems is closely related to the motivation for the *maximum entropy* approach to the properties of chaotic eigenstates presented in ref. [65]. Such an approach may be able to shed light on entanglement growth at early times.

At intermediate times $t_{\mathrm{eq}} \lesssim t \lesssim t_{\mathrm{bal}} \equiv \frac{R}{v_E^{(\alpha)}}$, entanglement growth is expected to be ballistic. For an effective 1D "strip" geometry (see figure 19), this expectation implies a form,

$$S_\alpha(t \lesssim t_{\mathrm{bal}}) \approx \frac{v_E^{(\alpha)} t}{R} S_\alpha(\infty) \tag{5.11}$$

where R is the radius of the strip. In this regime, $e^{(1-\alpha)S_\alpha}$ is initially $\sim \mathcal{O}(1)$. Whichever partition(s) dominates this regime of entanglement growth will be the slowest decaying partition that contributes at $\mathcal{O}(1)$. The remaining partitions that may contribute at $\mathcal{O}(1)$ are the full, uncontracted partition and partitions generated by contractions of any summed a_l- and b_l-indices with a and b. Performing all such contractions yielded the term composed solely of powers of the return ampltidue, eq. (5.10). Any partial such contraction will composed of other metrics of fidelity to the initial state such as $\left\langle \rho_{A(B)}(t) \right\rangle_{a(b)}$, $\left\langle \rho_{A(B)}(t)^2 \right\rangle_{a(b)}$, or other
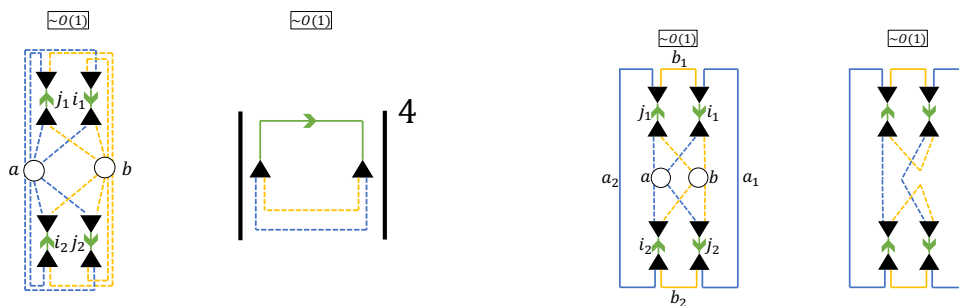
(a) The fourth power of the (modulus) return amplitude is obtained by contracting all summed a- and b-indices with the initial state. This is the only term which contributes at $t = 0$.

(b) The full, uncontracted partition is expected to govern the ballistic regime of entanglement growth, $t_{\text{eq}} < t < t_{\text{bal}}$, and encode the entanglement velocity.
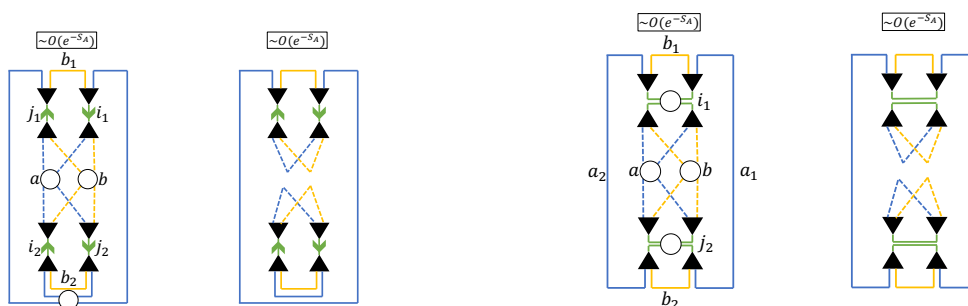
(c) At late times prior to saturation diffusive spreading of energy limits the grwoth of $\alpha > 1$-Renyi entropies. This is captured in the relaxation of the diagonal elements of reduced density matrices. Hence, we connect this behavior to the partition that contracts all external (summed) a-indices.

(d) At late times $t > t_{\text{sat}}$, phases cancel against each leaving only partitions that contract forward propogators with backwards propogators.

Figure 18. Four partitions associated with the growth of 2-Renyi entropy adjacent to their leading factorization under the assumption that subsystem A is smaller than subsystem B. When the subsystems are of similar size, there will more contributions.

powers of the return amplitude. Since each of these terms are highly nonlocal, we expect they should decay at faster rates than any entanglement velocity. Hence, our expectation is that for each α-Renyi entropy, only the full uncontracted partition governs the whole regime of ballistic entanglement growth, while other terms $\sim \mathcal{O}(1)$ at best contribute during local equilibration. We've included the case $\alpha = 2$ in figure 18(b). Then we assert,

$$
\begin{aligned}
p(t/t_{\text{eq}})e^{(1-\alpha)\frac{v_E^{(\alpha)}t}{R}S_\alpha(\infty)} &\approx \sum_{[\cdots]} c_{i_1}^{a_1'b_1'} c_{ab}^{i_1} c_{j_1}^{ab} c_{b_1'a_2'}^{j_1} \cdots c_{i_\alpha}^{a_\alpha'b_\alpha'} c_{ab}^{i_\alpha} c_{j_\alpha}^{ab} c_{b_\alpha'a_1'}^{j_\alpha} e^{-it\sum_m \omega_{i_m j_m}} \\
&= \int_{\cdots} e^{2\alpha S(\bar{E}_i) + \alpha S_A(\bar{E}_a) + \alpha S_B(\bar{E}_b) - 3\alpha S(E_a + E_b)} F(\cdots) e^{-it\sum_m \omega_{i_m j_m}} \\
&= \int_{\{\omega\}} F(\omega_{i_1 j_1}, \cdots) e^{-it\sum_m \omega_{i_m j_m}}
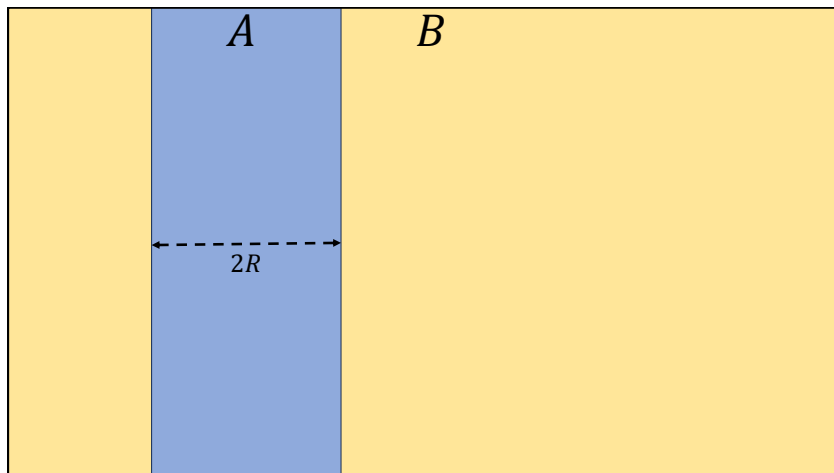\end{aligned}
\tag{5.12}
$$

– 36 –

Figure 19. An effective 1D system split into two subsystems A and B. The width of A is $2R$.

where the left-hand side contains the exponential decay with a short time suppression, $p(t/t_{\rm eq})$, that ensures the expression vanishes at $t \ll t_{\rm eq}$ and the right-hand side contains the only relevant index partition. For convenience, we define $\omega_E^{(\alpha)} \equiv (\alpha - 1)v_E^{(\alpha)}S_\alpha(\infty)/R$ and study the spectral function,

$$
\begin{aligned}
\mathcal{F}(\omega) &\equiv \int_{\omega_{i_1 j_1} \cdots} F(\omega_{i_1 j_1}, \cdots)\delta\left(\omega - \sum_m \omega_{i_m j_m}\right) \\
&\approx \int_{-\infty}^{\infty} p(t/t_{\rm eq})e^{i(\omega + i\omega_E^{(\alpha)})t}dt.
\end{aligned}
\tag{5.13}
$$

The exponential decay ensures that $\mathcal{F}(\omega)$ analytic within the strip $-\omega_E^{(\alpha)} \leq \mathrm{Im}[\omega] \leq \omega_E^{(\alpha)}$. Furthermore, we expect $p(t/t_{\rm eq})e^{-\omega_E^{(\alpha)}t}$ to maximize during the crossover from the local equilibration regime to the ballistic regime, $t \approx t_{\rm eq}$, which requires $\mathcal{F}(\omega)$ to oscillate over all scales larger than $t_{\rm eq}^{-1}$. Lastly, general arguments based on locality [66] enforce a sharp cutoff at large frequencies $\mathcal{F}(\omega) \lesssim e^{-\frac{|\omega| - V_{AB}}{g}}$, where V_{AB} is an area scaling term determined by the higher moments of H_{AB} and g is an $\mathcal{O}(1)$ effective coupling constant.

At late times still prior to saturation the Renyi entropies for $\alpha > 1$ will transition from ballistic to diffusive growth [38, 39] in systems with conservation laws (i.e. energy conservation). Close to saturation, diffusive growth implies a form

$$
S_\alpha(t \lesssim t_{\rm sat}) \propto \sqrt{Dt}
\tag{5.14}
$$

where D is the diffusivity of the system, and $t_{\rm sat}$ is the saturation time for the entanglement entropy. Since we just argued that the Renyi entropies are dominated by the full index partition in the entropic factors, naively, one may expect the same partition to contain both ballistic and diffusive regimes as was suggested in ref. [30]. However, at late times, the exponential decay of the full partition will diminish its magnitude by the entropy of the smaller subsystem. Then, any other partitions that were only suppressed by the entropy of the smaller subsystem can contribute provided they decay sufficiently slowly. Taking A

to be the smaller subsystem, we find precisely one partition to be suitable: the contraction of all summed a indices, $(a_1 a_2 \ldots a_\alpha)$.[20] Defining $\rho_A(E_a, t) \equiv \langle a| \mathrm{Tr}_A \left[e^{-iHt} |ab\rangle\langle ab| e^{iHt} \right] |a\rangle$ as the diagonal elements of the reduced density matrix on subsystem A in the subsystem energy eigenbasis, we can write,

$$F_{(a_1 a_2 \ldots a_\alpha)}(t) = \int_{E_a} e^{S_A(E_a)} \left(\rho_A(E_a, t) - \rho_A(E_a, \infty) \right)^\alpha + \cdots \sim \mathcal{O}(e^{(1-\alpha)S_A}) \qquad (5.15)$$

where the error in eq. (5.15) consists of partitions associated with the now negligible return amplitude (5.10) and partitions that are too suppressed to contribute at leading order in any regime. From eq. (5.15), it's clear why $(a_1 a_2 \ldots a_\alpha)$ is the right partition to describe diffusive relaxation: $\rho_A(E_a, t)$ encodes the dynamics of the approximately conserved subsystem energy. From an arbitrary unentangled inital state, the diffusive dynamics depend not only on the geometry of the subsystem, but on the temperature dependence of the diffusivity and heat capacity as well. We can conclude, however, that since $F_{(a_1 a_2 \ldots a_\alpha)}(t)$ decays slower than any exponential, but faster than any polynomial, its spectral function $\mathcal{F}_{(a_1 a_2 \ldots a_\alpha)}(\omega)$ is smooth, but non-analytic on the real line.[21]

We now discuss the $\alpha \to 1$ limit. Since the saturation regime shares its structure with pure eigenstates, the $\alpha \to 1$ limit will be the same as in section 4.2. Interestingly, S_1 does not exhibit a diffusive regime and so we should expect $F_{(a_1 a_2 \ldots a_\alpha)}(t)$ to vanish in this limit. Indeed, $\int_{E_a} e^{S_A(E_a)} \left(\rho_A(E_a, t) - \rho_A(E_a, \infty) \right)^1 = 1 - 1 = 0$. For the ballistic regime, eq. (5.11), the analytic continuation $\alpha \to 1$ holds by construction if $v_E^{(\alpha)}$ and $S_\alpha(\infty)$ can be regarded as analytic functions of α predicting a ballistic growth for S_1 where one is seen for $S_{\alpha>1}$. In general, for the analytic continuation to exist,[22] $\alpha = 1 + \epsilon$ must imply that $\mathrm{Tr}[\rho_A(t)^\alpha] = 1 - S_1(t)\epsilon + \mathcal{O}(\epsilon^2)$. However, eq. (5.10) clearly does not have this property, except for $t = 0$. Since $|F_R(t > 0)|^{2\alpha}$ does not converge to 1 with α, there must exist one or more partitions that exactly cancel the decay of $F_R(t)$ for all $t > 0$. We are forced to conclude that decay of the return amplitude does not provide an adequate picture of the local equilibration regime, which in our view remains a mysterious aspect of entanglement growth. Otherwise, our diagrams provide an intriguing structure to the phases of entanglement growth.

6 Operator-eigenstate correlations

In the preceding sections, we assume that both subsystems A and B are thermodynamically large, as parameterized by expansions in $e^{-S_{A(B)}}$. It follows that an operator X obeys the ETH, eq. (1.1), with respect to eigenstates both of H_A and of H. Furthermore, if X is deep within subsystem A it should have the same properties on the subsystem and the full system, at least over short times. This question was previously considered in ref. [30] where it was remarked that correlations between X and c are in principle unnecessary to satisfy these conditions based on a calculation that matched factors of the density-of-states. In

---

[20] Analogously for subsystem B. When the subsystems are a similar size, there will be a contribution from all non-crossing contractions of summed a and b indices.

[21] As an example of an appropriate non-analytic function, the Fourier transform of the standard bump function, $\exp\left(\frac{1}{x^2-1}\right)$, $|x| \leq 1$, can be evaluated via saddle-points to get the desired asymptotic decay $\sim e^{-\sqrt{t}}$, up to dimensionful constants and other physical features.

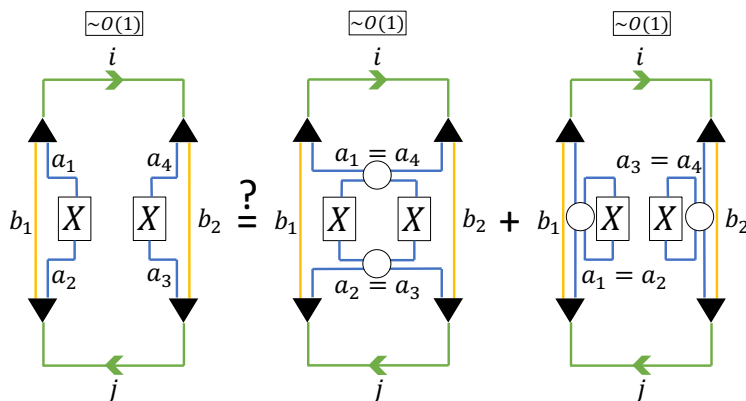[22] It must since ρ_A is a positive semi-definite operator.

Figure 20. This figure depicts the false hypothesis that operator-eigenstate correlations factor if operators are taken to be far from the boundary between the subsystems.

fact, we show that this is not true and nontrivial correlations between X and c are necessary for physical consistency. We additionally show that these correlations encode the time at which operator X receives information about subsystem B, and argue that this timescale is determined by the butterfly velocity.

6.1 The decay catastrophe

When X is deep within subsystem A it may be tempting to assume that its matrix elements become uncorrelated with those of c. Consider the 2-point cumulant of X,

$$f_2(t + i\beta/2) = \langle X(t)X \rangle_i - \langle X \rangle_i^2 = \sum_{[j]a_1a_2a_3a_4b_1b_2} c^i_{a_1b_1} X_{a_1a_2} c^{a_2b_1}_j c^j_{a_3b_2} X_{a_3a_4} c^{a_4b_2}_i e^{i\omega_{ij}t}. \quad (6.1)$$

If the elements of X and c are uncorrelated, their combined diagram must factor as in figure 20. There are two diagrams that contribute at leading order, but we just focus on one, $(a_1a_4)(a_2a_3)$, to illustrate the point. For $a_1 \neq a_2$, this partition relates the 2-point cumulant on the whole system, f_2, to the 2-point cumulant of subsystem A, $f_2^{(A)}$.

The factorization allows us to perform a direct computation,

$$\begin{aligned}
f_{(a_1a_4)(a_2a_3)}(t + i\beta/2) &= \sum_{[ja_1a_2b_1b_2]} c^i_{a_1b_1} c^j_{a_2b_2} X_{a_1a_2} X_{a_2a_1} e^{i\omega_{ij}t} \\
&= \int_{E_a E_{b_1} E_{b_2} \omega_{ij}\omega_a} F_{(a_1a_4)(a_2a_3)}(E_i; \cdots, \omega_{ij}, \omega_a) f_2^{(A)}(E_a; \omega_a) e^{i\omega_{ij}t - \beta\omega_{ij}/2} \\
&= F_{(a_1a_4)(a_2a_3)}(-t - i\beta/2) f_2^{(A)}(0)
\end{aligned} \quad (6.2)$$

where in the second line, we have defined F to be the relevant eigenstate cumulant and emphasized its dependence on ω_{ij} and $\omega_a = E_{a_1} - E_{a_2}$. In general, decorrelating c and X causes correlation functions to become convolutions between wavefunction partitions and operator partitions in frequency space and products in time.

Eq. (6.2) implies the obviously false conclusion that the entire dynamics of any local observable in subsystem A far from the boundary are contained in the dynamics of a single eigenstate cumulant that generally depends on the nature of the interaction between the subsystems. This outcome is not a bug associated with the partition we have chosen but an

inevitable consequence of decorrelating c with X. Thus, we are forced to conclude that X and c are correlated even at short times and far separations. Consequently, the elements of c contain non-local information about the whole system. In section 6.2, we will square this conclusion with causality.

6.2 The butterfly velocity

Existing literature discusses emergent causality in terms of the *butterfly velocity*, which bounds the time at which out-of-time-ordered-correlators (OTOCs) between distant operators can decay [40]. The simplest OTOC is $g(t) \equiv \langle W(t)VW(t)V \rangle$ for distinct operators V and W which commute at time $t = 0$. Let W and V be separated by a distance R. The decay of $g(t)$ measures the nontriviality of the commutator $[W(t), V]$. Until information of W has reached V, the commutator is approximately zero, and $g(t)$ cannot decay. The butterfly velocity v_B is defined such that the time at which decay first occurs is $t_B = R/v_B$.

We will study an analog of the OTOC between H_{AB} and X to bound when X has received information about subsystem B. Consider the Heisenberg evolution of our operator X located a distance R from the boundary between A and B. First, we will evolve X forwards in time under H_A for a time t, then switch on the interaction H_{AB} and reverse evolution back to time $t = 0$. For $t < t_B$, X should return approximately to its initial value. Concretely, we define

$$h(t) \equiv \left\langle e^{-iHt}e^{iH_At}Xe^{-iH_At}e^{iHt}X \right\rangle_i. \tag{6.3}$$

For convenience, we will take $\langle X \rangle_i = 0$ and $\langle X^2 \rangle_i = 1$.

We justify using $h(t)$ to study the butterfly velocity by defining an interaction propagator $U_I(t) \equiv e^{i(H_A+H_B)t}e^{-iHt}$, for which $U_I'(t) = -iH_{AB}(t)U_I(t)$, where we have defined the time dependence of operators in terms of $H_A + H_B$. With this definition, we have:

$$h(t) = \left\langle U_I^\dagger(-t)XU_I(-t)X \right\rangle_i,$$
$$h'(t) = -i\left\langle U_I^\dagger(-t)[H_{AB}(-t), X]U_I(-t)X \right\rangle_i. \tag{6.4}$$

We can understand $h'(t)$ as an inner product on the space of Hermitian operators between $-iU_I^\dagger(-t)[H_{AB}(-t), X]U_I(-t)$ and X and utilize Cauchy-Schwarz,

$$|h'(t)|^2 \leq -\left\langle U_I^\dagger(-t)[H_{AB}(-t), X]^2 U_I(-t) \right\rangle_i \left\langle X^2 \right\rangle_i = -\left\langle [H_{AB}, X(t)]^2 \right\rangle_i. \tag{6.5}$$

Thus, $h'(t)$ is bounded by the squared commutator of X and H_{AB}, separated by a time of t. This result can be readily adapted into a bound on the decay of $h(t)$ by assuming the growth of the commutator square is monotonic prior to t_B,

$$\begin{aligned}
|h(t) - h(0)| &\leq \int_0^t |h'(t)|\, dt \\
&\leq \int_0^t \sqrt{-\langle [H_{AB}, X(t)]^2 \rangle_i}\, dt \\
&\leq t\sqrt{-\langle [H_{AB}, X(t)]^2 \rangle_i},
\end{aligned} \tag{6.6}$$

hence, the decay of $h(t)$ is bounded by the nontriviality of the squared commutator.

We define t_d as the decay timescale of $h(t)$. When $t_B \gg t_d$, $h(t)$ will approach a step function. Consequently, for $\omega < t_d^{-1}$, $h(\omega)$ will behave as $\sim \mathrm{sinc}(R\omega/v_B)$, while for $\omega > t_d^{-1}$, $h(\omega)$ will have some cutoff that in general depends on the decay of $h(t)$. We want

to understand correlations in the energy eigenspace of our system, so let us write eq. (6.3) in terms of components of X and c,

$$h(t) = \sum_{ja_1a_2a_3a_4b_1b_2} c^i_{a_1b_1} X_{a_1a_2} c^{a_2b_1}_j c^j_{a_3b_2} X_{a_3a_4} c^{a_4b_2}_i e^{i(\omega_{a_1a_2} - \omega_{ij})t}. \tag{6.7}$$

Eq. (6.7) is nearly identical to eq. (6.1) and contains the same partitions. From the above considerations,

$$h(\omega) = 2\pi \sum_{ja_1a_2a_3a_4b_1b_2} c^i_{a_1b_1} X_{a_1a_2} c^{a_2b_1}_j c^j_{a_3b_2} X_{a_3a_4} c^{a_4b_2}_i \delta(\omega_{a_1a_2} - \omega_{ij} - \omega)$$

$$\sim \text{sinc}(R\omega/v_B)\Lambda_{t_d^{-1}}(\omega) \tag{6.8}$$

where $\Lambda_{t_d^{-1}}(\omega)$ is a cutoff function introduced by the finite rate of decay of $h(t)$.

Eqs. (6.7) and (6.8) ensure a nontrivial correlated structure between X and c that is relevant at arbitrarily short times and arbitrarily long distances. However, in contrast to previous quantities we have studied, it is not obvious that the right-hand sides of eqs. (6.7) and (6.8) have any particular interpretation in terms of individual generalized free cumulants indicating that our diagrammatic approach does not capture the full picture. Our analysis mirrors that of ref. [67], which considered the eigenbasis representation of the OTOC between spatially separated operators in the context of the ETH and found an analogous sinc-like universal form. More recently, ref. [65] was able to capture the physics of the butterfly velocity from a pure wavefunction perspective analogous to our own, but considering correlations between distinct partitionings of the system. It would be interesting to connect their approach to our own.

7 Conclusions and discussion

The eigenstate thermalization hypothesis was originally introduced to justify the application of statistical mechanics to quantum many-body systems. However, in recent years, especially given the connection to free probability theory, it has become clear that the ETH does more than justify statistical mechanics. Free probability theory has previously emerged in physics in the context of large N limits [53, 68]. Then, the ETH is perhaps best understood as a set of phenomena associated with an emergent e^{-S} expansion in quantum many-body systems. This interpretation of the ETH is automatic in the matrix model formulation developed in ref. [46]. In this work, we have presented the many-body Berry's conjecture as a reformulation of the ETH, which is traditionally understood in the sense of eq. (1.1). However, we may alternately view eq. (1.1), the many-body Berry's conjecture, and the various results in this paper and the literature as just aspects of a large e^S limit. A mature understanding of this limit remains to be developed.

In this paper, we have formulated the eigenstate thermalization hypothesis and the ergodic bipartition in terms of a many-body Berry's conjecture, the hypothesis that eigenstates of chaotic systems are random vectors up to the symmetry constraints of the system, and we have argued that the MBBC is the natural quantum generalization of the ergodic hypothesis. We showed how this approach naturally leads to a diagrammatic formalism developed in the language of free probability, which is our main result. We demonstrate the power of our formalism by showing that systems relax to a universal reduced density matrix and obey

the Page curve at late times, thus establishing irreversible thermalization under reversible unitary evolution. We also establish the subsystem ETH as a consequence of the MBBC. We also discuss the role that locality plays in the ETH and develop connections to butterfly velocities and entanglement growth.

Our considerations in this paper are generic, and our results should apply to any system where ergodicity is not explicitly broken, e.g., through many-body localization. Certainly much more can be done on a similarly generic footing. While we have largely ignored the functional forms of eigenstate cumulants (F functions), they are necessary to understand the time-dependent dynamics of thermalization and entanglement. There is some work that suggests eigenstate cumulants, in some cases, have universal forms that depend on the number of dimensions, locality of interactions, etc but not the model considered [30]. Additionally, while the operator cumulants (f functions) are generally model dependent, their high-frequency cutoffs appear connected to universal properties of scrambling, and future work may clarify this connection [14, 18]. Other work may clarify the connection between the butterfly velocity and spectral correlations and uncover correlated structures that exist beyond the ETH [65, 67]. We hope the formalism we have developed will aid in the classification of generic properties of quantum many-body systems.

On the other hand, future work may focus on concrete realizations of the above results. Exact functional forms of generalized free cumulants may be possible to derive within toy models, such as the SYK model. Applications of the ETH to conformal field theories and to holographic systems [11] have revealed important structure in those systems and provided insights into the structure of black holes [36]. Additionally, numerical evidence for our work is severely limited by the capabilities of exact diagonalization. Modern techniques and greater computational power could provide clean demonstrations of our results in spin chains or other toy models [69, 70].

Some techniques in this paper are reminiscent of others in the literature. In section 5, our means of imposing equilibrium on non-equilibrium partitions by matching opposing time dependences reproduces the equilibrated pure state formalism of ref. [32]. In another case, refs. [31, 65] study the statistical properties of eigenstates via an application of the *principle of maximum entropy*. Ref. [65] uses this principle to study propagators at early times and studies the behavior of OTOCs in Floquet quantum circuits. Ref. [31] argues that the various principles of the ETH may each be understood as consequences of a maximum entropy principle and derives a special case of our eq. (4.5) to study saddle-points in the gravitational path integral. In our view, the principle of maximum entropy is a more formal and fundamental approach to the heuristic arguments we present in sections 2 and 3 that may even extend the validity of our work.

Acknowledgments

A Free probability in quantum chaos

In this appendix, we discuss the principles of free probability theory and their emergence in quantum chaos. We do not present rigorous or technical arguments but instead discuss informally how key ideas, namely *freeness*, *free cumulants*, and *random matrices* arise, and obtain analogs in the study of quantum chaos. In particular, we wish to motivate why the decay of out-of-time-ordered correlators (OTOCs), the eigenstate thermalization hypothesis (ETH), and emergent rotational symmetry are, in essence, equivalent definitions of chaos and avatars of free probability theory. Except where another citation is provided, we direct readers to ref. [47] for background on free probability.

First, we discuss what it means to have a probability theory in a non-commutative setting.[23] The traditional foundation of classical probability theory starts with a sample space of possible events and a rule for assigning probabilities to subsets of events. Non-commutative probability theory instead starts with an algebra of observables and their expectation values. The rules of non-commutative probability theory generalize the familiar rules of quantum mechanics where we are given a (pure or mixed) density matrix and some (generically non-commuting) observables whose expectation values are of interest to us. Non-commutative probability also reduces to classical probability theory in the limit that observables of interest commute. Often in quantum systems, particularly when $\hbar$ can be considered small, operators are approximately commutative and principles of classical probability theory such as sample spaces and (classical) independence become useful emergent descriptions of statistical physics. Free probability theory concerns itself with, in some sense, an opposite limit in which observables are as non-commutative as possible. We will see that this notion of maximal non-commutativity can be made precise and that quantum mechanics indeed contains such a limit.

Consider two algebras of observables, $\mathcal{X}$ and $\mathcal{Y}$, and a state $\langle \cdots \rangle$ that satisfies $\langle ZW \rangle = \langle WZ \rangle$. As an example, one can take $\mathcal{Y}$ to be the algebra generated by Pauli operators on a single site of a spin chain, $\mathcal{X}$ to be the algebra generated by the Pauli operators on a different site evolved far forward in time, and the state to be the conventional thermal regulator discussed in row (i) of table 1. We will return to this example. $\mathcal{X}$ and $\mathcal{Y}$ are considered *freely independent*, or *free*, if

$$\langle X_1 Y_1 \cdots X_n Y_n \rangle = 0 \quad \text{for all} \quad X_i \in \mathcal{X},\, Y_i \in \mathcal{Y} \quad \text{such that} \quad \langle X_i \rangle = \langle Y_i \rangle = 0. \qquad \text{(A.1)}$$

The expression $\langle X_1 Y_1 \cdots X_n Y_n \rangle$ is known as an alternating moment and so long as $\mathcal{X}$ and $\mathcal{Y}$ are closed algebras any mixed moment of operators between them reduces to such an expression. It is not obvious, but eq. (A.1) encodes maximal non-commutativity between $\mathcal{X}$ and $\mathcal{Y}$. If there were any nontrivial algebraic relations between elements of $\mathcal{X}$ and $\mathcal{Y}$, e.g. $XY = YX$ for some X, Y, then the definition (A.1) would imply that at least one of X and Y is a scalar and thus that the algebraic relation is trivial, in contradiction with the assumption of nontriviality. We sketch this proof just for commutators. Assume that $\mathcal{X}$ and $\mathcal{Y}$ satisfy eq. (A.1) but that $XY = YX$ for some non-scalar X, Y. Consider the

---

[23]For an elaboration on the ideas in this paragraph, we direct readers to ref. [71].

fluctuations $\delta X \equiv X - \langle X \rangle$, $\delta Y \equiv Y - \langle Y \rangle$. By (A.1),

$$\left\langle \left(\delta X \delta Y \right)^n \left(\delta Y \right)^{m-n} \right\rangle = \left\langle \delta X \delta Y \cdots \delta X \delta Y^{m-n+1} \right\rangle = 0. \tag{A.2}$$

Since X and Y commute, eq. (A.2) reduces to,

$$\langle \delta X^n \delta Y^m \rangle = 0. \tag{A.3}$$

However, eq. (A.1) implies for $n = 1$,

$$\langle (X_1 - \langle X_1 \rangle)(Y_1 - \langle Y_1 \rangle) \rangle = 0 \quad \implies \quad \langle X_1 Y_1 \rangle = \langle X_1 \rangle \langle Y_1 \rangle \tag{A.4}$$

and taking $X_1 = \delta X^n$, $Y_1 = \delta Y^m$,

$$\langle \delta X^n \delta Y^m \rangle = \langle \delta X^n \rangle \langle \delta Y^m \rangle = 0. \tag{A.5}$$

However, eq. (A.5) states that the product of arbitrary central moments of X and Y vanish. This can only be the case if at least one of X, Y has vanishing central moments and thus is a scalar.[24] Hence, our hypothesis is contradicted, and no nontrivial element of $\mathcal{X}$ can commute with any nontrivial element of $\mathcal{Y}$. This argument can be extended to any algebraic relation between elements of $\mathcal{X}$ and $\mathcal{Y}$ and thus establishes the idea of maximal non-commutativity. The notion of freeness can also be extended from algebras to pairs of operators as freeness of the subalgebras they generate. X and Y are free if,

$$\langle (X - \langle X \rangle)^{p_1} (Y - \langle Y \rangle)^{q_1} \cdots (X - \langle X \rangle)^{p_n} (Y - \langle Y \rangle)^{q_n} \rangle = 0 \quad \text{for all} \quad n \geq 1, \tag{A.6}$$

for positive integer exponents p_m, q_m. A pair of algebras are free if and only if any pair of their operators are free.[25]

Freeness has an analogy in quantum chaos. A many-body system is considered chaotic if for simple, few-body observables X, Y,

$$\left\langle \left[\left(X(t) - \langle X(t) \rangle_\beta \right) \left(Y(0) - \langle Y(0) \rangle_\beta \right) \right]^q \right\rangle_\beta \to 0 \quad \text{for all} \quad q \geq 1 \tag{A.7}$$

for $t > t_{\text{scr}}$, where t_{scr} is known as the scrambling time of the system [72] and we have used the conventional thermal regulator. (A.7) is formulated to measure the non-commutativity of $X(t)$ with $Y(0)$. Eq. (A.7) is essentially the definition of the quantum butterfly effect, in which the decay of OTOCs characterizes the ability for the unitary evolution of a system to scramble quantum information [29]. In this light, the quantum butterfly effect is a consequence of the asymptotic freeness between simple few-body observables at long time separations. If the example of Pauli operators are to constitute freely independent algebras, then their OTOCs must decay.

The definition of freeness given in eq. (A.1) is the maximally non-commutative analog of the classical definition of independence between two variables. That is, $\mathcal{X}$ and $\mathcal{Y}$ are *classically independent*, or simply *independent*, if,

$$\langle X_1 Y_1 \cdots X_n Y_n \rangle = \langle X_1 \cdots X_n \rangle \langle Y_1 \cdots Y_n \rangle \quad \text{for all} \quad X_i \in \mathcal{X}, Y_i \in \mathcal{Y}. \tag{A.8}$$

---

[24]We assume faithfulness: $\langle X^2 \rangle = 0 \implies X = 0$. Freeness is defined with respect to a given state and cannot say much about operators for which the chosen state is not faithful.

[25]Freeness can be further extended to a collection of any number of algebras or operators.

Eqs. (A.1) and (A.8) are both rules for computing higher mixed moments from lower moments. Such a rule is obtained from eq. (A.1) by performing the binomial expansion of $\langle (X_1 - \langle X_1 \rangle)(Y_1 - \langle Y_1 \rangle) \cdots \rangle = 0$. The analogous rule from (A.8) is simply that mixed moments factorize. Whereas eq. (A.1) implied that elements of $\mathcal{X}$ and $\mathcal{Y}$ have no nontrivial algebraic relations, eq. (A.8) implies that all elements of $\mathcal{X}$ and $\mathcal{Y}$ satisfy the specific relation that they commute, i.e. $XY = YX$. This result is straightforward to prove. Consider two alternating moments,

$$\langle X_1 Y_1 \cdots X_{m-1} Y_{m-1} X_m Y_m X_{m+1} Y_{m+1} \cdots X_n Y_n \rangle = \langle X_1 \cdots X_n \rangle \langle Y_1 \cdots Y_n \rangle ,$$
$$\langle X_1 Y_1 \cdots X_{m-1} (Y_{m-1} Y_m)(X_m X_{m+1}) Y_{m+1} \cdots X_n Y_n \rangle = \langle X_1 \cdots X_n \rangle \langle Y_1 \cdots Y_n \rangle \qquad \text{(A.9)}$$

where both are equal per eq. (A.8). Then,

$$\langle \cdots X_{m-1} Y_{m-1} X_m Y_m X_{m+1} Y_{m+1} \cdots \rangle - \langle \cdots X_{m-1} (Y_{m-1} Y_m)(X_m X_{m+1}) Y_{m+1} \cdots \rangle = 0$$
$$\implies \quad \langle \cdots [X_m, Y_m] \cdots \rangle = 0. \qquad \text{(A.10)}$$

Since $X_1, Y_1 \cdots, X_{m-1}, Y_{m-1}, X_{m+1}, Y_{m+1}, \cdots, X_n, Y_n$ are arbitrary, eq. (A.10) can only hold if $[X_m, Y_m] = 0$ in general. Hence, eq. (A.8) enforces commutativity.

Many concepts of classical probability have exact analogs in free probability. Eq. (A.6) is analogous to the factorization of mixed moments of classically independent variables $\langle X^p Y^q \rangle = \langle X^p \rangle \langle Y^q \rangle$. The classical probability distribution of an operator is instead replaced with its spectrum. The convolution of probability distributions for sums of independent operators is replaced with a free additive convolution of spectra for sums of free operators. Of key importance in classical and free probability is the existence of cumulants, which to a physicist serve as the building blocks of a theory. Cumulants also compactly formulate the rules for computing higher moments from lower moments mentioned above for classically or freely independent variables. Whereas the vanishing of classical cumulants indicates independence, the vanishing of free cumulants indicates freeness. Furthermore, while classical cumulants admit a combinatorial interpretation in terms of set partitions, free cumulants can be interpreted in terms of *non-crossing* partitions. As an example we consider the cumulant decompositions of the moment $\langle XYXY \rangle$. First, in the classical case,

$$
\begin{aligned}
\langle XYXY \rangle \equiv &\; \langle \underline{XYXY} \rangle \\
&+ \langle \underline{XYX}Y \rangle + \langle \underline{XY}X\underline{Y} \rangle + \langle X\underline{YXY} \rangle + \langle \underline{X}Y\underline{XY} \rangle \\
&+ \langle \underline{XY}\,\underline{XY} \rangle + \langle \underline{XYX}\underline{Y} \rangle + \langle X\underline{YX}Y \rangle \\
&+ \langle XYXY \rangle \\
\equiv &\; k_{X^2 Y^2} \\
&+ 2 k_{X^2 Y} k_Y + 2 k_{XY^2} k_X \\
&+ k_{X^2} k_{Y^2} + 2 k_{XY}^2 \\
&+ k_X^2 k_Y^2 \qquad \text{(A.11)}
\end{aligned}
$$

where we've used k to denote classical cumulants. Then in the free case,

$$
\begin{aligned}
\langle XYXY \rangle \equiv {} & \langle \overbrace{XYXY} \rangle \\
& + \langle \overbrace{XY}\,X\!\overbrace{Y} \rangle + \langle \overbrace{XY\,XY} \rangle + \langle \overbrace{XY}\,\overbrace{XY} \rangle + \langle X\!\overbrace{Y\,XY} \rangle \\
& + \langle \overbrace{XY}\,\overbrace{XY} \rangle + \langle \overbrace{XY\,X}\,Y \rangle \\
& + \langle \underbrace{XY\,XY} \rangle \\
\equiv {} & f_{XYXY} \\
& + 2 f_{X^2Y} f_Y + 2 f_{XY^2} f_X \\
& + 2 f_{XY}^2 \\
& + f_X^2 f_Y^2
\end{aligned}
\tag{A.12}
$$

where we've used f to denote free cumulants and we've distinguished ordinary set partitions from non-crossing partitions by drawing the former below the line and the latter above it.

Hopefully the rules of the above partitions are clear, but there are a few aspects which should be emphasized. It is straightforward to verify that the vanishing of free cumulants, defined by the non-crossing partitions, implies freeness in the sense of eq. (A.6). Until 3$^{\text{rd}}$ order, free and classical cumulants are identical. At 4$^{\text{th}}$ order and above, classical cumulants become undefined for non-commutative operators while free cumulants become sensitive to order (up to cyclic permutations), i.e $f_{XYXY} = f_{YXYX} \neq f_{X^2Y^2} = f_{Y^2X^2}$. The difference is reflected in the fact that eqs. (A.11) and (A.12) differ by a single partition. As we discuss in section 3.2, free cumulants are crucial to the definition of the ETH. However, to justify the form given in eq. (1.1), we will turn to a concrete construction of free probability.

Random matrices are known to model free probability in their large N limit. A finite $N \times N$ hermitian matrix can be classically sampled from a distribution over its N real diagonal and $N(N-1)$ complex off-diagonal elements.[26] Of particular importance are rotationally invariant probability distributions. That is, for any matrix X and unitary rotation U,

$$
p(X) = p(U^\dagger X U)
\tag{A.13}
$$

where p is the probability density. Next, to take the large N limit one has to specify a sequence of $N \times N$ matrices $X^{(N)}$ sampled from a sequence of distributions $p^{(N)}$, $N \in \mathbb{Z}$, such that as $N \to \infty$, the spectrum of $X^{(N)}$ converges to a well-defined limit. Then consider two sequences of random matrices with large N limits, $X^{(N)}$ and $Y^{(N)}$. If for each N, the elements of $X^{(N)}$ and $Y^{(N)}$ are sampled independently and from rotationally invariant probability distributions, then,

$$
\lim_{N \to \infty} \left\langle \prod_{i=1}^n \left(X^{(N)} - \left\langle X^{(N)} \right\rangle \right)^{p_i} \left(Y^{(N)} - \left\langle Y^{(N)} \right\rangle \right)^{q_i} \right\rangle = 0 \quad \text{for all} \quad n \geq 1.
\tag{A.14}
$$

Eq. (A.14) is simply the freeness condition eq. (A.6) asymptotically in the large N limit. Since we have considered rotationally invariant distributions, the specific state $\langle \cdots \rangle$, is unimportant.

---

[26]One can similarly consider real symmetric matrices and orthogonal rotations or self-adjoint quaternionic matrices and symplectic rotations, but in this appendix we will stick to hermitian matrices and unitary rotations.

What we have seen is the remarkable fact that hermitian random matrices whose elements are sampled classically independently from a rotationally invariant ensemble, become freely independent in the large N limit. Rotational invariance is crucial. Asymptotic freeness in the large N limit holds if and only if the ensemble is rotationally invariant. Hence, where in eq. (A.7) we found that freeness occurs at long time separations for simple observables, we should expect that at small frequencies the matrix elements of those observables exhibit an emergent rotational invariance.

Lastly, we can introduce the expression for free cumulants in terms of matrix elements. For a sequence of matrices, $X^{(N)}$, that satisfies rotational invariance and has a well-defined limit spectrum, its free cumulants in the large N limit are given by the classical cumulants k_n of its matrix elements [57],

$$f_{X^n} = \lim_{N \to \infty} N^{n-1} k_n \left(X_{i_1 i_2}^{(N)}, X_{i_2 i_3}^{(N)}, \cdots, X_{i_n i_1}^{(N)} \right) \tag{A.15}$$

for any choice of distinct indices $i_1, \ldots, i_n$. Ref. [57] presents the result in terms of classical cumulants, but as discussed in ref. [13] the only products of matrix elements whose expectations are rotationally invariant are those whose indices are cyclic. Then, instead we have,

$$f_{X^n} = \lim_{N \to \infty} N^{n-1} \mathrm{E}(X_{i_1 i_2}^{(N)} X_{i_2 i_3}^{(N)} \cdots X_{i_n i_1}^{(N)}) \tag{A.16}$$

for any choice of distinct indices $i_1, \ldots, i_n$, where E represents the classical expectation value. It is clear that eq. (1.1) is the chaotic analogy to eq. (A.16) where the theoretical expectation value has simply been replaced with an empirical average and N with e^S.

Hence we have seen how key concepts of quantum chaos: the butterfly effect, the ETH, and emergent rotational invariance are each avatars of free probabilistic limit inherent in chaos of quantum many-body systems. Yet, there are many more tools in the free probability toolbox. We suspect broader awareness of the subject will help those tools find applications in physics.

B Saddle-points in the ETH

We briefly review the use of saddle-point integration in the context of the ETH. The basic principle of saddle-point integration is that the total integral of a rapidly varying integrand is sharply concentrated around its peak(s) or of a rapidly oscillating integrand around its point(s) of stationary phase. We refer to either peaks or stationary points as saddle-points. In the cases we consider, the contributions to integrals from regions away from the saddle-point will be suppressed by the system size, and thus, saddle-point integration becomes exact in the thermodynamic limit, at least within a logarithm.

We assume that our system has a Hamiltonian H with $\mathcal{V}$ degrees of freedom (volume). Extensive thermodynamic quantities such as the energy $E = \langle H \rangle$, microcanonical entropy $S(E) = \ln[\sum_i \delta(E_i - E)]$, and heat capacity $C(E) \equiv -\beta^2 [S''(E)]^{-1}$, all scale $\sim \mathcal{O}(\mathcal{V})$. Intensive thermodynamic quantities, such as the inverse temperature $\beta \equiv S'(E)$ and the expectation values of local operators $\langle X \rangle$ are $\sim \mathcal{O}(1)$.

The diagonal elements of ETH satisfying operators are equal to their microcanonical expectation value,

$$X_{ii} = e^{-S(E_i)} \sum_j \delta(E_j - E_i) X_j + \mathcal{O}(e^{-S}) \equiv f_1(E_i) + \mathcal{O}(e^{-S}). \tag{B.1}$$

Then eigenstates, by definition, obtain the same equilibrium properties as the microcanonical ensemble. We can also check the expectation values in the canonical ensemble,

$$\langle X \rangle_\beta \equiv \frac{\sum_i e^{-\beta E_i} X_{ii}}{\sum_i e^{-\beta E_i}} = \frac{\int_E e^{S(E)-\beta E} f_1(E) + \mathcal{O}(e^{-S/2})}{\int_E e^{S(E)-\beta E}} \tag{B.2}$$

where we have exploited (B.1) and the definition of microcanonical entropy to turn the sums into integrals. We find the saddle-point of the integral by setting the derivative of the integrand to zero and obtain

$$S'(E_\beta) - \beta + \frac{f_1'(E_\beta)}{f_1(E_\beta)} \implies S'(E_\beta) = \beta + \mathcal{O}(1/\mathcal{V}). \tag{B.3}$$

Taking only the saddle-point value of each integral we find,

$$\langle X \rangle_\beta = \frac{e^{S(E_\beta)-\beta E_\beta} f_1(E_\beta)}{e^{S(E_\beta)-\beta E_\beta}} = f_1(E_\beta). \tag{B.4}$$

Eq. (B.4) implies that the properties of the canonical ensemble with temperature β are equivalent to the properties of the microcanonical ensemble at a specific energy E_β. From eq. (B.3), we can realize that the saddle-point associated to the canonical inverse temperature β is equivalent to the microcanonical definition of temperature to leading order in $1/\mathcal{V}$. Thus, we establish ensemble equivalence in the ETH. This result extends readily to the non-equilibrium functions f_n for their slow dependence on the total system energy.

However, we have not considered the magnitude of corrections away from the saddle-point. To do so, we consider the Taylor expansion of the exponent about the saddle-point,

$$S(E) - \beta E = S(E_\beta) - \beta E_\beta - \frac{\beta^2 (E - E_\beta)^2}{2C} + \sum_{m=3}^{\infty} \frac{S^{(m)}(E_\beta)(E - E_\beta)^m}{m!}. \tag{B.5}$$

If we truncate this series to second order, we find that E is Gaussian distributed about E_β with a width $\sqrt{C/\beta^2}$ that scales $\sim \mathcal{O}(\mathcal{V}^{1/2})$. However, since $f_1'(E)$ is $\sim \mathcal{O}(1/\mathcal{V})$, contributions away from the saddle-point in the Gaussian approximation will only contribute at most at order $\sim \mathcal{O}(\mathcal{V}^{-1/2})$. If we assume that E stays close to E_β at any order in the expansion, self-consistently including higher derivatives of S results in an asymptotic series about $\mathcal{V} \to \infty$ which may be evaluated via Feynman diagrams. However, the higher order terms will be suppressed by further powers of $\mathcal{V}$, and going beyond zeroth order will be unnecessary for our purposes.

C Non-zero width of the window function

In this appendix we discuss the non-zero width of F. In C.1 we discuss how the width can affect some of the calculations in this paper by focusing on a specific case: subsystem energy fluctuations for both the eigenstates and non-eigenstates. In C.2 we discuss the form of F in eq. (1.2). This has been previously discussed in refs. [21] and [30] which predict a Gaussian and a Lorentzian form, respectively, in tension with one another. We shed some light on this disagreement.

C.1 Subsystem energy fluctuations

In this subappendix we consider an observable which can distinguish eigenstates from non-eigenstates: subsystem energy fluctuations. As we will see, these fluctuations gain an extra contribution from the energy uncertainty of a non-eigenstate when the subsystem is a finite fraction of the whole system.

To consider the effects of finite F width, we write

$$\ln\left[F(\omega)\right] \approx \ln\left[F(\omega_0)\right] - \frac{(\omega - \omega_0)^2}{2\Delta_0^2} + \dots \tag{C.1}$$

where Δ_0^2 represents the variance of F and ω_0 maximizes F. Eq. (C.1) is in essence a Gaussian approximation for F but the outcomes of convolutions will be exact so long as we are only computing variances.[27]

First, we consider the eigenstate case, eq. (4.2). The subsystem energy variance is given by,

$$\left\langle \Delta H_A^2 \right\rangle_i \equiv \left\langle (H_A - E_{iA})^2 \right\rangle_i = \int_{E_a E_b} (E_a - E_{iA})^2 e^{-S(E_i) + S_A(E_a) + S_B(E_b)} F(E_i - E_a - E_b). \tag{C.2}$$

Per appendix B, the density of states factors should be well approximated as Gaussians about the saddle-point of eq. (C.2). $e^{S_{A(B)}}$ obtains a variance of $C_{A(B)}T^2$, where $C_{A(B)}$ is the heat capacity of subsystem $A(B)$. Then, we can write,

$$\left\langle \Delta H_A^2 \right\rangle_i = \frac{1}{\mathcal{N}} \int_{E_a E_b} (E_a - E_{iA})^2 e^{-\frac{(E_a - E_{iA})^2}{2C_A T^2} - \frac{(E_b - E_{iB})^2}{2C_B T^2} - \frac{(E_i - E_a - E_b - \omega_0)^2}{2\Delta_0^2}} \tag{C.3}$$

where $\mathcal{N}$ represents an overall normalization that ensures $\langle i | i \rangle = 1$. The Gaussian integrals in eq. (C.3) can be directly evaluated to find,

$$\left\langle \Delta H_A^2 \right\rangle_i = T^2 \left(\frac{1}{C_A} + \frac{1}{C_B + \Delta_0^2/T^2} \right)^{-1}. \tag{C.4}$$

The variance of subsystem energies was considered in ref. [60] which contrasted the results for eigenstates and the canonical ensemble. For a Gibbs state on subsystem A, the variance is

$$\left\langle \Delta H_A^2 \right\rangle_{\text{Gibbs}} = C_A T^2. \tag{C.5}$$

For an eigenstate, one instead expects,[28]

$$\left\langle \Delta H_A^2 \right\rangle_{\text{Eigenstate}} = T^2 \left(\frac{1}{C_A} + \frac{1}{C_B} \right)^{-1}. \tag{C.6}$$

We can see that eq. (C.4) interpolates between the two cases. If for some reason $\Delta_0^2/T^2 \gg C_A$ eq. (C.4) reduces to eq. (C.5). This situation may be possible in models with long-range

---

[27]I.e. Bienaymé's identity.

[28]To connect our expression to that of ref. [60] one has to assume homogeneity. I.e. that $C_{A(B)} \propto \mathcal{V}_{A(B)}$ where $\mathcal{V}_{A(B)}$ is the volume of subsystem $A(B)$. Then one recovers $\left\langle \Delta H_A^2 \right\rangle_{\text{Eigenstate}} = C_A T^2 \left(1 - \mathcal{V}_A/\mathcal{V}\right)$ where $\mathcal{V}$ is the volume of the whole system.

interactions and would be interesting to explore further. Such systems may also exhibit some nontrivial scaling with temperature. If Δ_0^2 is vanishing, eq. (C.6) is recovered. This is more physical because for an eigenstate we should expect Δ_0^2 to scale as the area of the boundary between A and B. Thus, the finite width of F adds only an area scaling correction to the subsystem energy variance. Nonetheless, this correction is perceptible in the numerical calculations presented in figure 11 of ref. [60], where the graph of the subsystem energy variance shows a small enhancement with respect to eq. (C.6) for a single high energy eigenstate when A is larger than B.

Next, we consider an arbitrary initial state $|\psi\rangle$ with initial energy $E_\psi = \langle H \rangle_\psi$. We consider its eigenstate components,

$$|\psi\rangle \equiv \sum_i c_i^\psi |i\rangle \tag{C.7}$$

and define,

$$\overline{c_i^\psi c_\psi^i} \equiv e^{-S(E)} F_\psi(E_i - E_\psi) \tag{C.8}$$

and

$$\ln\left[F_\psi(\omega)\right] \approx \ln\left[F_\psi(\omega_\psi)\right] - \frac{(\omega - \omega_\psi)^2}{2\Delta_\psi^2} + \dots \tag{C.9}$$

where Δ_ψ is the energy variance of F_ψ and ω_ψ maximizes it. Generalizing eq. (5.5), in equilibrium, the subsystem energy variance will be given by,

$$\left\langle \Delta H_A^2(t \to \infty) \right\rangle_\psi = \int_{E_i E_a E_b} (E_a - E_{iA})^2 e^{-S(E_i)+S(E_a)+S(E_b)} F(E_i - E_a - E_b) F_\psi(E_\psi - E_i). \tag{C.10}$$

In this case we have two more variances to consider: CT^2 from e^S and Δ_ψ^2 from F_ψ. Once again evaluating the Gaussian integrals, we calculate,

$$\left\langle \Delta H_A^2(\infty) \right\rangle_\psi = T^2 \left(\frac{1}{C_A} + \frac{1}{C_B + \Delta_0^2/T^2 + (T^2/\Delta_\psi^2 + C^{-1})^{-1}} \right)^{-1}. \tag{C.11}$$

Thus we find some difference between the subsystem energy variance of a system prepared in an eigenstate and of a system that is prepared outside of an eigenstate but allowed to equilibrate. However, such a difference only shows up for a subsystem that is a finite fraction of the whole system.

C.2 Form of the window function

We have not yet discussed the form of F in eq. (1.2). Where in sections 5.3 and 6.2 we were able to relate forms of F in more complicated scenarios to known locality-related bounds, for this simplest case the literature provides different answers.

Ref. [21] conjectured that in this case F obtains an exactly Gaussian form in $D > 1$ dimensional systems,

$$F(\omega) \propto e^{-\frac{(\omega - \omega_0)^2}{2\Delta_0^2}} \tag{C.12}$$

where ω_0 maximizes $F(\omega)$ and Δ_0 is its variance. Their argument relies broadly on two claims:

1. $\int_\omega F(\omega)\omega^n \approx \langle H_{AB}^n\rangle_i$ up to subleading order corrections that are polynomially suppressed in the system size,

2. The moments of H_{AB} are those of a Gaussian distribution.

Claim 2 follows from the fact that H_{AB}, in $D > 1$ spatial dimensions, is a sum of an area scaling number of local terms in a system with a finite correlation length and thus its moments converge weakly towards those of a Gaussian distribution by central limit theorem. Claim 1 is more subtle. We take $H \equiv H_A + H_B + H_{AB}$, $H_0 \equiv H_A + H_B$ and $|i\rangle$, $|I\rangle$ to label their respective eigenstates. Starting from the identity[29]

$$\langle I|(H - E_I)^n|I\rangle = \sum_i |\langle i|I\rangle|^2 (E_i - E_I)^n = \int_\omega F(\omega)\omega^n \tag{C.13}$$

ref. [21] shows that,

$$\langle I|(H - E_I)^n|I\rangle = \langle I|([H_0 - E_I] + H_{AB})^n|I\rangle = \langle I|(H_{AB})^n|I\rangle + \mathcal{O}(n\mathcal{A}^{\frac{n-1}{2}}) \tag{C.14}$$

where $\mathcal{A}$ denotes the area of the boundary between the subsystems. Since the moments of H_{AB} are indeed Gaussian, ref. [21] concluded that F must be as well with mean $\omega_0 = \langle H_{AB}\rangle_I$ and variance $\Delta_0^2 = \Delta^2 \equiv \langle H_{AB}^2\rangle_I - \langle H_{AB}\rangle_I^2$. However, as pointed out in ref. [30], eq. (C.14) only bounds the first p central moments of $F(\omega)$ for some $p \sim \mathcal{O}\left(\sqrt{\mathcal{A}}\right)$ which may not be sufficient to determine the form of F.

Ref. [30] instead argues for a Lorentzian form with an exponential cutoff at large frequencies in any number of dimensions. Their argument follows by approximating the solution to the characteristic equations for the eigenvalues of H after perturbing H_0 by H_{AB}. Preparing the system in a product state $|I\rangle$, we can decompose the Hilbert space into the direct sum of $|I\rangle$ and its complement, $\mathcal{H} = \mathcal{H}_c \oplus |I\rangle$. Diagonalizing H on the complement Hilbert space yields a set of $d-1$ eigenstates $|\epsilon_m\rangle$. Ref. [30] then mirrors a standard derivation of *Fermi's golden rule* (FGR) [73]. The key assumption is that the nonzero level spacing and finite bandwidth of the levels ϵ are irrelevant, that is,

$$\epsilon_{\max} - \epsilon_{\min} \gg \mathcal{O}(\mathcal{A}) \gg \delta\epsilon \tag{C.15}$$

where $\epsilon_{\max}$ ($\epsilon_{\min}$) is the largest (smallest) value of ϵ and $\delta\epsilon$ is the typical spacing between consecutive ϵ. Under these assumptions one can approximately solve a set of characteristic equations to yield,

$$F(E_i - E_I) \sim \frac{\Gamma_I^i}{(E_i - E_I)^2 + (\Gamma_I^i)^2} \tag{C.16}$$

where Γ_I^i gives the width of F. So long as Γ_I^i has a sufficiently slow dependence on $E_i - E_I$ eq. (C.16) is consistent with a Lorentzian form for F. As a consequence, the return amplitude

---

[29]Note that our normalization convention differs from that of ref. [21] but the difference can be absorbed into the definition of F.

of a system prepared in $|I\rangle$ and evolved under H exhibits an exponentially decaying return amplitude,

$$
\begin{aligned}
e^{iE_I t}\langle I|e^{-iHt}|I\rangle &= \sum_i |\langle i|I\rangle|^2 e^{-i(E_i - E_I)t} \\
&= \int_\omega F(\omega)e^{-i\omega t} \\
&\equiv F(t) \sim e^{-\Gamma_I^i t}
\end{aligned}
\tag{C.17}
$$

which reflects the physical picture of FGR. Recent work [74–76] has connected FGR and its breakdown to the *many-body localization* (MBL) to ETH crossover.[30] Ref. [76], in particular, argues that an exponential decay of the (modulus-squared) return amplitude is generic feature of ETH satisfying systems so long as the perturbation is not large enough to modify the entropy function. In contrast the argument of ref. [21] would suggest a Gaussian decay of the return amplitude. Various numerical calculations [30, 74, 76] corroborate a regime of exponential decay in 1 dimension, however, we are unaware of any numerical work in higher dimensions where the tension arises.

To understand the tension, it may be enlightening to study how the moment calculation eq. (C.14) influences the time dependence of $F(t)$.[31] Our analysis complements that in appendix A of ref. [30] by focusing on the decay of $F(t)$ where they focused on the decay of $F(\omega)$. At late times, the return amplitude saturates to the overlap of two microcanonically random vectors,

$$
e^{iE_I t}\langle I|e^{-iHt}|I\rangle \to e^{-S/2}.
\tag{C.18}
$$

The smooth F functions will not capture this saturation but determine what form the return amplitude takes prior to the saturation time. First we estimate the saturation time, t_{sat}, under the Gaussian form implied by eq. (C.12),

$$
e^{-\Delta^2 t_{\mathrm{sat}}^2/2} = e^{-S/2} \quad \implies \quad t_{\mathrm{sat}} = \sqrt{\frac{S}{\Delta^2}} \sim \mathcal{O}\left(\sqrt{\frac{\mathcal{V}}{\mathcal{A}}}\right)
\tag{C.19}
$$

where $\mathcal{V}$ is the volume of the system. Next, eq. (C.14) determines only the first p derivatives of $F(t)$ where $p \sim \mathcal{O}\left(\sqrt{\mathcal{A}}\right)$ is the first (even) central moment for which the error crosses the mean. Then we can estimate the error of the Gaussian description from its Taylor series,

$$
|F(t) - e^{-\Delta^2 t^2/2}| \approx \left|\frac{(p-1)!!\Delta^p t^p}{p!}\right| \approx \frac{1}{\sqrt{\pi}}\left|\frac{e\Delta^2 t^2}{p}\right|^{\frac{p}{2}}.
\tag{C.20}
$$

The error approximation derived in (C.20) implies that $e^{-\Delta^2 t^2/2}$ is only a reliable approximation to $F(t)$ for $t \lesssim t_{\mathrm{early}}$ defined as

$$
\begin{aligned}
e^{-\Delta^2 t_{\mathrm{early}}^2/2} &\approx \frac{1}{\epsilon\sqrt{\pi}}\left|\frac{e\Delta^2 t_{\mathrm{early}}^2}{p}\right|^{\frac{p}{2}} \\
\implies \frac{\Delta^2 t_{\mathrm{early}}^2}{p} &\approx W\left(e^{-1}\left(\epsilon^2\pi\right)^{\frac{1}{p}}\right) = 0.2785\cdots + \mathcal{O}\left(\frac{1}{p}\right) \\
\implies t_{\mathrm{early}} &\sim \mathcal{O}(\mathcal{A}^{-\frac{1}{4}})
\end{aligned}
\tag{C.21}
$$

---

[30]We are grateful to Philip Crowley for bringing this work to our attention.

[31]Neglecting phase factors throughout.

where ϵ is a small $\mathcal{O}(1)$ error threshold and W is the Lambert W function. Eq. (C.21) implies that increasing the size of $\mathcal{A}$ will actually shorten the validity of the moment calculation. This is physical since we are scaling p and Δ simultaneously and $F(t)$ grows sharp as $\Delta^{-1} \sim \mathcal{O}\left(\mathcal{A}^{-\frac{1}{2}}\right)$ which is faster than the constraint in eq. (C.21) grows tight. Thus, the Gaussian ansatz of ref. [21] can be expected to accurately capture the decay of $F(t)$ for a large number of periods Δ^{-1} yet only for a vanishing period of time in the thermodynamic limit. Furthermore, eq. (C.19) shows that t_{sat} diverges in the thermodynamic limit. We conclude that the argument of ref. [21] cannot determine the form of $F(t)$, or by extension $F(\omega)$, over all physically relevant scales.

Ref. [76] uses a statistical analysis of the Jacobi diagonalization algorithm applied to ETH satisfying systems to develop the following picture for the return amplitude. Initializing the system in an eigenstate of H_0 and perturbing it to H, the return amplitude is expected to experience an early time Gaussian decay, intermediate exponential decay, and a late time saturation, i.e.

$$e^{iH_0 t} \langle I | e^{-iHt} | I \rangle = \begin{cases} F_{\text{early}}(t) = e^{-J^2 t^2/2} \\ F_{\text{inter}}(t) \propto e^{-\Gamma t} \\ e^{-S/2} \quad \text{(saturation)} \end{cases} . \tag{C.22}$$

For a sufficiently weak perturbation a Gaussian description only holds at very early times with $J^2 \approx \Delta^2 \sim \Delta_0^2$. For a sufficiently strong volume-scaling perturbation the Gaussian decay may saturate early and entirely preempt the exponential regime with $J^2 \approx \Delta_0^2 \gg \Delta^2$ and. Then, inserting $F_{\text{early}}(t)$ for $F(t)$ into eqs. (C.20) and (C.21), we get,

$$e^{\left(J^2 - \Delta^2\right) t_{\text{early}}^2/2} - 1 = \epsilon \implies t_{\text{early}} = \sqrt{\frac{2\ln(1+\epsilon)}{J^2 - \Delta^2}}$$

$$\implies J^2 \approx \left(1 + \frac{2\epsilon}{W\left(e^{-1}(\epsilon^2 \pi)^{\frac{1}{p}}\right) p}\right) \Delta^2 = \Delta^2 + \mathcal{O}\left(\sqrt{\mathcal{A}}\right) \tag{C.23}$$

and our estimate of t_{early} indicates that the area-scaling perturbation, H_{AB}, should be considered weak with $J^2 \approx \Delta^2 \approx \Delta_0^2$. Returning to the frequency space picture, the exponential cut-off of $F(t)$ after t_{early} will constrain the form of $F(\omega)$ on scales below $t_{\text{early}}^{-1} \sim \mathcal{O}(\mathcal{A}^{1/4})$. However, the width of $F(\omega)$, $\Delta_0 \approx \Delta \sim \mathcal{O}(\mathcal{A}^{1/2})$ is parametrically larger than this scale in $D > 1$ dimensions. Thus, $F(\omega)$ should appear Gaussian in $D > 1$ dimensions over scales comparable to $\Delta \sim \mathcal{O}(\mathcal{A}^{1/2})$. $F(\omega)$ must also respect an $\mathcal{O}(1)$ cutoff at very large scales $\omega \sim \mathcal{O}(\mathcal{A})$ [66].

We conclude that the tension between the Gaussian and Lorentzian forms in $D > 1$ dimensions boils down to relevant timescales. The F function in eq. (1.2) captures the return amplitude of a system prepared in an eigenstate of $H_A + H_B$ and evolved under the Hamiltonian $H_A + H_B + H_{AB}$. We expect that the return amplitude will vanish rapidly as a Gaussian at very early times, but continue to decay as an exponential for a much longer period until saturation. Future work should clarify how the various arguments in the literature capture different aspects of this picture.

D Operator thermalization

We wish to show in this section that the ETH, eq. (1.1), is a consequence of the MBBC. Consider a local Hermitian operator X supported on a local Hilbert space of dimension d_X and an associated rotation,

$$U_X(s) \equiv e^{isX} = \sum_{m=0}^{d_X-1} p_m(s) X^m \qquad (D.1)$$

where the second equality states that U_X is a finite polynomial in X and follows from the Cayley-Hamilton theorem. For $|i\rangle$ an eigenstate of a Hamiltonian H, $|i(s)\rangle \equiv U_X(s)|i\rangle$ is an eigenstate of Hamiltonian $H(s) \equiv U_X(s)HU_X(-s)$. We can decompose $H(s)$ as

$$H(s) = H + (H(s) - H) \equiv H + V(s) \qquad (D.2)$$

where $V(s)$ is a small $O(1)$ perturbation so long as X is a local operator and H and sum of local terms. Then, by eq. (2.8) we can assert the following,

$$\overline{\langle i_1(s)|i_2(2s)\rangle\langle i_2(2s)|\cdots|i_n(ns)\rangle\langle i_n(ns)|i_1(s)\rangle} = e^{-(n-1)S(\bar{E})} F_X(s;\bar{E};\vec{\omega}) \qquad (D.3)$$

where we have considered n distinct eigenstates of H and rotated them each by a different angle s. Since distinct eigenstates are orthogonal, $F_X(0;\bar{E};\vec{\omega}) = 0$. In fact, the first $n-1$ derivatives of F_X with respect to s at $s=0$ must vanish as well,

$$e^{-(n-1)S(\bar{E})} F_X(s;\cdots) = \sum_{m=0}^{n(d_X-1)} \sum_{\sum_l m_l = m} q_{\{m\}}(s) \overline{(X^{m_1})_{i_1 i_2} \cdots (X^{m_n})_{i_n i_1}}$$
$$= 0 \cdot s^0 + \cdots + 0 \cdot s^{n-1} + q_{\{1\cdots 1\},n} s^n \overline{X_{i_1 i_2} \cdots X_{i_n i_1}} + \mathcal{O}(s^{n+1}) \quad (D.4)$$

where $q_{\{m\}}$ collects the sum of products of p_m and we have defined the expansion $q_{\{m\}}(s) \equiv \sum_{m'} q_{\{m\},m'} s^{m'}$. In general, the terms p_m and $q_{\{m\}}$ will depend only on the characteristic polynomial of X and are finite for any d_X. Thus we can extract correlations of X from the following non singular limit,

$$\overline{X_{i_1 i_2} \cdots X_{i_n i_1}} = \lim_{s \to 0} q_{\{1\cdots 1\},n}^{-1} s^{-n} e^{-(n-1)S(\bar{E})} F_X(s;\cdots) \equiv e^{-(n-1)S(\bar{E})} f(\bar{E};\vec{\omega}) \qquad (D.5)$$

and we arrive at our desired result. The approach of this section suggests a sufficient criteria for operators to satisfy the ETH in the sense of eq. (1.1): for a given operator X, $U_X(s)HU_X(-s)$ is close to H even as X mixes a large number of eigenstates.

References

[1] F. Schwabl, *Statistical Mechanics*, Springer Berlin Heidelberg (2006) [DOI:10.1007/3-540-36217-7].

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

[3] J. von Neumann, *Proof of the ergodic theorem and the H-theorem in quantum mechanics: Translation of: Beweis des Ergodensatzes und des H-Theorems in der neuen Mechanik*, *Eur. Phys. J. H* **35** (2010) 201.

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

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

[6] M. Srednicki, *The approach to thermal equilibrium in quantized chaotic systems*, *J. Phys. A* **32** (1999) 1163 [INSPIRE].

[7] S. Popescu, A.J. Short and A. Winter, *Entanglement and the foundations of statistical mechanics*, *Nature Phys.* **2** (2006) 754 [INSPIRE].

[8] N. Linden, S. Popescu, A.J. Short and A. Winter, *Quantum mechanical evolution towards thermal equilibrium*, *Phys. Rev. E* **79** (2009) 061103 [arXiv:0812.2385] [INSPIRE].

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

[10] S.H. Shenker and D. Stanford, *Black holes and the butterfly effect*, *JHEP* **03** (2014) 067 [arXiv:1306.0622] [INSPIRE].

[11] N. Lashkari, A. Dymarsky and H. Liu, *Eigenstate Thermalization Hypothesis in Conformal Field Theory*, *J. Stat. Mech.* **1803** (2018) 033101 [arXiv:1610.00302] [INSPIRE].

[12] N. Bao and N. Cheng, *Eigenstate Thermalization Hypothesis and Approximate Quantum Error Correction*, *JHEP* **08** (2019) 152 [arXiv:1906.03669] [INSPIRE].

[13] L. Foini and J. Kurchan, *Eigenstate thermalization hypothesis and out of time order correlators*, *Phys. Rev. E* **99** (2019) 042139 [arXiv:1803.10658] [INSPIRE].

[14] C. Murthy and M. Srednicki, *Bounds on chaos from the eigenstate thermalization hypothesis*, *Phys. Rev. Lett.* **123** (2019) 230606 [arXiv:1906.10808] [INSPIRE].

[15] A. Dymarsky, *Bound on Eigenstate Thermalization from Transport*, *Phys. Rev. Lett.* **128** (2022) 190601 [arXiv:1804.08626] [INSPIRE].

[16] R. Nandkishore and D.A. Huse, *Many body localization and thermalization in quantum statistical mechanics*, *Ann. Rev. Condens. Mat. Phys.* **6** (2015) 15 [arXiv:1404.0686] [INSPIRE].

[17] J. Maldacena, S.H. Shenker and D. Stanford, *A bound on chaos*, *JHEP* **08** (2016) 106 [arXiv:1503.01409] [INSPIRE].

[18] S. Pappalardi, L. Foini and J. Kurchan, *Eigenstate Thermalization Hypothesis and Free Probability*, *Phys. Rev. Lett.* **129** (2022) 170603 [arXiv:2204.11679] [INSPIRE].

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

[20] T.-C. Lu and T. Grover, *Rényi Entropy of Chaotic Eigenstates*, *Phys. Rev. E* **99** (2019) 032111 [arXiv:1709.08784] [INSPIRE].

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

[22] M.V. Berry, *Regular and irregular semiclassical wavefunctions*, *J. Phys. A* **10** (1977) 2083 [INSPIRE].

[23] J.M. Deutsch, *Eigenstate thermalization hypothesis*, *Rept. Prog. Phys.* **81** (2018) 082001 [INSPIRE].

[24] A. Dymarsky and H. Liu, *New characteristic of quantum many-body chaotic systems*, *Phys. Rev. E* **99** (2019) 010102 [arXiv:1702.07722] [INSPIRE].

[25] K.R. Fratus and M. Srednicki, *Eigenstate thermalization in systems with spontaneously broken symmetry*, *Phys. Rev. E* **92** (2015) 040103.

[26] M. Feingold and A. Peres, *Distribution of Matrix Elements of Chaotic Systems*, *Phys. Rev. A* **34** (1986) 591 [INSPIRE].

[27] L. Foini and J. Kurchan, *Eigenstate Thermalization and Rotational Invariance in Ergodic Quantum Systems*, *Phys. Rev. Lett.* **123** (2019) 260601 [arXiv:1906.01522] [INSPIRE].

[28] J. Wang et al., *Emergence of unitary symmetry of microcanonically truncated operators in chaotic quantum systems*, arXiv:2310.20264 [INSPIRE].

[29] P. Hosur, X.-L. Qi, D.A. Roberts and B. Yoshida, *Chaos in quantum channels*, *JHEP* **02** (2016) 004 [arXiv:1511.04021] [INSPIRE].

[30] Z.D. Shi, S. Vardhan and H. Liu, *Local dynamics and the structure of chaotic eigenstates*, *Phys. Rev. B* **108** (2023) 224305 [arXiv:2306.08032] [INSPIRE].

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

[32] 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].

[33] 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].

[34] J. Wang, *Beyond islands: a free probabilistic approach*, *JHEP* **10** (2023) 040 [arXiv:2209.10546] [INSPIRE].

[35] A. Dymarsky, N. Lashkari and H. Liu, *Subsystem eigenstate thermalization hypothesis*, *Phys. Rev. E* **97** (2018) 012140 [arXiv:1611.08764] [INSPIRE].

[36] J. Kudler-Flam, V. Narovlansky and S. Ryu, *Distinguishing Random and Black Hole Microstates*, *PRX Quantum* **2** (2021) 040340 [arXiv:2108.00011] [INSPIRE].

[37] M. Mezei and D. Stanford, *On entanglement spreading in chaotic systems*, *JHEP* **05** (2017) 065 [arXiv:1608.05101] [INSPIRE].

[38] T. Rakovszky, F. Pollmann and C.W. von Keyserlingk, *Sub-ballistic Growth of Rényi Entropies due to Diffusion*, *Phys. Rev. Lett.* **122** (2019) 250602 [arXiv:1901.10502] [INSPIRE].

[39] Y. Huang, *Dynamics of Rényi entanglement entropy in diffusive qudit systems*, arXiv:2008.00944 [DOI:10.1088/2633-1357/abd1e2] [INSPIRE].

[40] D.A. Roberts and B. Swingle, *Lieb-Robinson Bound and the Butterfly Effect in Quantum Field Theories*, *Phys. Rev. Lett.* **117** (2016) 091602 [arXiv:1603.09298] [INSPIRE].

[41] M. Rigol and L.F. Santos, *Quantum chaos and thermalization in gapped systems*, *Phys. Rev. A* **82** (2010) 011604.

[42] B. Georgeot and D.L. Shepelyansky, *Quantum chaos border for quantum computing*, *Phys. Rev. E* **62** (2000) 3504 [quant-ph/9909074] [INSPIRE].

[43] R. Modak and S. Mukerjee, *Finite size scaling in crossover among different random matrix ensembles in microscopic lattice models*, *New J. Phys.* **16** (2014) 093016.

[44] F.J. Dyson and M.L. Mehta, *Statistical Theory of the Energy Levels of Complex Systems. IV*, *J. Math. Phys.* **4** (1963) 701.

[45] M. Lal Mehta, *Random Matrices*, Elsevier Science & Technology, San Diego, U.S.A. (2004).

[46] D.L. Jafferis, D.K. Kolchmeyer, B. Mukhametzhanov and J. Sonner, *Matrix Models for Eigenstate Thermalization*, *Phys. Rev. X* **13** (2023) 031033 [arXiv:2209.02130] [INSPIRE].

[47] J.A. Mingo and R. Speicher, *Free Probability and Random Matrices*, Springer New York (2017) [DOI:10.1007/978-1-4939-6942-5].

[48] S. Pappalardi, F. Fritzsch and T. Prosen, *General Eigenstate Thermalization via Free Cumulants in Quantum Lattice Systems*, arXiv:2303.00713 [INSPIRE].

[49] R. Speicher, *Free Probability Theory and Non-Crossing Partitions*, *Sem. Lothar. Combin.* **B39c** (1997) 1.

[50] S. Pappalardi, L. Foini and J. Kurchan, *Microcanonical windows on quantum operators*, *Quantum* **8** (2024) 1227 [arXiv:2304.10948] [INSPIRE].

[51] E. Fradkin, *Quantum Field Theory: An Integrated Approach*, Princeton University Press (2021) [INSPIRE] [ISBN: 9780691149080].

[52] G. 't Hooft, *A Planar Diagram Theory for Strong Interactions*, *Nucl. Phys. B* **72** (1974) 461 [INSPIRE].

[53] R. Gopakumar and D.J. Gross, *Mastering the master field*, *Nucl. Phys. B* **451** (1995) 379 [hep-th/9411021] [INSPIRE].

[54] J. Richter, A. Dymarsky, R. Steinigeweg and J. Gemmer, *Eigenstate thermalization hypothesis beyond standard indicators: Emergence of random-matrix behavior at small frequencies*, *Phys. Rev. E* **102** (2020) 042127 [arXiv:2007.15070] [INSPIRE].

[55] M. Fava, J. Kurchan and S. Pappalardi, *Designs via Free Probability*, arXiv:2308.06200 [INSPIRE].

[56] R. Mondaini and M. Rigol, *Eigenstate thermalization in the two-dimensional transverse field Ising model. II. Off-diagonal matrix elements of observables*, *Phys. Rev. E* **96** (2017) 012157 [INSPIRE].

[57] B. Collins, J.A. Mingo, P. Sniady and R. Speicher, *Second Order Freeness and Fluctuations of Random Matrices, III. Higher order freeness and free cumulants*, math/0606431.

[58] R. Steinigeweg, J. Herbrych and P. Prelovšek, *Eigenstate thermalization within isolated spin-chain systems*, *Phys. Rev. E* **87** (2013) 012118.

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

[60] J.R. Garrison and T. Grover, *Does a single eigenstate encode the full Hamiltonian?*, *Phys. Rev. X* **8** (2018) 021026 [arXiv:1503.00729] [INSPIRE].

[61] Y. Huang, *Universal eigenstate entanglement of chaotic local Hamiltonians*, *Nucl. Phys. B* **938** (2019) 594 [arXiv:1708.08607] [INSPIRE].

[62] Y. Huang, *Universal entanglement of mid-spectrum eigenstates of chaotic local Hamiltonians*, *Nucl. Phys. B* **966** (2021) 115373.

[63] J. de Boer, J. Hollander and A. Rolph, *Page curves and replica wormholes from random dynamics*, *JHEP* **07** (2024) 023 [arXiv:2311.07655] [INSPIRE].

[64] H. Liu and S.J. Suh, *Entanglement Tsunami: Universal Scaling in Holographic Thermalization*, *Phys. Rev. Lett.* **112** (2014) 011601 [arXiv:1305.7244] [INSPIRE].

[65] D. Hahn, D.J. Luitz and J.T. Chalker, *The statistical properties of eigenstates in chaotic many-body quantum systems*, arXiv:2309.12982 [INSPIRE].

[66] I. Arad, T. Kuwahara and Z. Landau, *Connecting global and local energy distributions in quantum spin models on a lattice*, *J. Stat. Mech.* **1603** (2016) 033301 [arXiv:1406.3898] [INSPIRE].

[67] A. Chan, A. De Luca and J.T. Chalker, *Eigenstate Correlations, Thermalization and the Butterfly Effect*, *Phys. Rev. Lett.* **122** (2019) 220601 [arXiv:1810.11014] [INSPIRE].

[68] L. Hruza and D. Bernard, *Coherent Fluctuations in Noisy Mesoscopic Systems, the Open Quantum SSEP, and Free Probability*, *Phys. Rev. X* **13** (2023) 011045 [arXiv:2204.11680] [INSPIRE].

[69] P. Hosur, *Polynomial-time algorithm for studying physical observables in chaotic eigenstates*, *Phys. Rev. B* **103** (2021) 195159 [arXiv:2002.09514] [INSPIRE].

[70] M. Luo, R. Trivedi, M.C. Bañuls and J.I. Cirac, *Probing off-diagonal eigenstate thermalization with tensor networks*, *Phys. Rev. B* **109** (2024) 134304 [arXiv:2312.00736] [INSPIRE].

[71] T. Tao, *254A, Notes 5: Free probability*, (2010).

[72] D.A. Roberts and B. Yoshida, *Chaos and complexity by design*, *JHEP* **04** (2017) 121 [arXiv:1610.04903] [INSPIRE].

[73] S. Santra, B. Cruikshank, R. Balu and K. Jacobs, *Fermi's golden rule, the origin and breakdown of Markovian master equations, and the relationship between oscillator baths and the random matrix model*, *J. Phys. A* **50** (2017) 415302.

[74] P.J.D. Crowley and A. Chandran, *Partial thermalisation of a two-state system coupled to a finite quantum bath*, *SciPost Phys.* **12** (2022) 103 [arXiv:2104.03312] [INSPIRE].

[75] T. Micklitz, A. Morningstar, A. Altland and D.A. Huse, *Emergence of Fermi's Golden Rule*, *Phys. Rev. Lett.* **129** (2022) 140402 [arXiv:2206.00704] [INSPIRE].

[76] D.M. Long, D. Hahn, M. Bukov and A. Chandran, *Beyond Fermi's golden rule with the statistical Jacobi approximation*, *SciPost Phys.* **15** (2023) 251 [arXiv:2306.16457] [INSPIRE].