Chaos and high temperature pure state thermalization

Yuri D. Lensky[a] **and Xiao-Liang Qi**[a,b]

[a]*Stanford Institute for Theoretical Physics, Stanford University,*
382 Via Pueblo Mall, Stanford, CA 94305, U.S.A.

[b]*Institute for Advanced Study,*
1 Einstein Drive, Princeton, NJ 08540, U.S.A.

E-mail: ydl@stanford.edu, xlqi@stanford.edu

ABSTRACT: Classical arguments for thermalization of isolated systems do not apply in a straightforward way to the quantum case. Recently, there has been interest in diagnostics of quantum chaos in many-body systems. In the classical case, chaos is a popular explanation for the legitimacy of the methods of statistical physics. In this work, we relate a previously proposed criteria of quantum chaos in the unitary time evolution operator to the entanglement entropy growth for a far-from-equilibrium initial pure state. By mapping the unitary time evolution operator to a doubled state, chaos can be characterized by suppression of mutual information between subsystems of the past and that of the future. We show that when this mutual information is small, a typical unentangled initial state will evolve to a highly entangled final state. Our result provides a more concrete connection between quantum chaos and thermalization in many-body systems.

Contents

1 Introduction

Empirically, there is a generic tendency towards entropy growth in many body systems. This "arrow of time" appears at odds with the fact that our physical models of these systems are often time-reversal symmetric. A common argument has been available from the time of Boltzmann: states are exponentially likely to evolve to states of higher entropy, simply due to the counting of states at given entropy (see for example [1]). In fact, in the presence of time-reversal symmetry this explanation is incomplete. For every state of a given entropy with entropy growth, there is a state with the same entropy but with entropy decay. The best we can actually hope for is to explain why some class of preferred states experiences entropy growth in some subsystem under specific dynamics.

This question is closely related to another aspect of quantum many body systems. The use of statistical ensembles to understand the long-time collective behavior of many degrees of freedom in terms of local microscopic interactions is one of the great simplifications and triumphs of modern physics. It is natural to ask when and why the use of these ensembles is justified; if all expectation values of interest for a given initial condition after time evolution can be computed to arbitrary accuracy in an ensemble depending only on macroscopic parameters, we will say that system thermalizes for those initial conditions and that time evolution. For isolated classical systems, dynamical chaos is a sufficient and generic condition for ergodicity in phase space, which explains the accuracy of the microcanonical ensemble and hence equilibrium statistical mechanics.

The quantum case is more complicated. It is important to note that there are classical systems with few degrees of freedom whose observables are well-described by the microcanonical ensemble, but upon quantization do not thermalize. Although there are

experimental probes of few-body quantum systems that *do* thermalize [2], generic thermalization in quantum systems appears to be inherently a many-body effect [3]. Quantum mechanics also supports a long-time behavior not present in classical systems, Many-Body Localization (MBL). Recent experiments detecting these phases [4] provide practical motivation to explain the mechanism of and conditions for thermalization of isolated systems. Finally, our understanding of quantum chaos and ergodicity is still incomplete, and leading justifications for quantum statistical mechanics are not as directly connected to quantum chaos as in the classical case.

There are two leading explanations for quantum thermalization. One is known as Canonical Typicality (CT) [5, 6], the statement that due to the exponentially large dimension of Hilbert space or the subspaces associated to finite energy windows, almost all pure states in the subspace will appear as if they were randomly chosen from that subspace, i.e. indistinguishable from the microcanonical ensemble, on any small subregion. Importantly, the CT approach can be extended to the dynamical result [7] that, under weak assumptions about the distribution of eigenvalues of the Hamiltonian, a subsystem interacting with a sufficiently large bath will spend most of its time close to its time average, independent of the initial state of the subsystem and for almost all initial states of the bath. One useful way to view CT is as an extension of the statistical argument for entropy growth: most states in a subspace are already close to maximum entanglement within the subspace. The other explanation is the Eigenstate Thermalization Hypothesis (ETH) [3, 8], which loosely stated is the conjecture that the high-energy eigenstates of quantized classically chaotic systems are indistinguishable from the microcanonical ensemble of the same system for local observables. This conjecture is well-supported numerically for a large class of systems, and gives a very clean description of quantum thermalization when it applies.

There are some deficiencies remaining in both approaches. There are no direct criteria to evaluate on a Hamiltonian to see if ETH holds, short of finding the eigenstates. On a related note, ETH has only been proven true for a small class of systems. Finally, although the ETH is inspired by ideas about classical chaos and ergodicity, there is no proof that chaos in dynamics implies ETH.[1] The conclusions of CT appear completely unrelated to whether a system is chaotic. The principle mechanism of CT is that typical states are close to maximally entangled, or already at equilibrium. The problem is that we would like information about highly atypical states (out-of-equilibrium low-entanglement states) that form a set of measure zero in most subspaces of high dimension. The dynamical extension [7] solves this problem for systems where a small subsystem can be highly atypical in a much larger typical bath. This is a reasonable assumption for a near-isolated quantum system interacting with the rest of the world, but is not useful when we wish to consider an even smaller class of states, where the entire system is far from equilibrium and has low entanglement. It is also a statement about time averages as opposed to instantaneous density matrices. Finally, like ETH, CT gives no criterion on the time evolution to distinguish thermalizing from non-thermalizing systems. On a related note, there is no explanation for the *mechanism* of thermalization, apart from the high dimension of Hilbert space.

---

[1]Some interesting progress along this direction is the relation between ETH and out-of-time-order correlation functions in [9].

In this work, we link the entropy growth of low-entanglement states under unitary time $U(t)$ to measures of quantum chaos associated with $U(t)$. In doing so, we begin to address the above deficiencies in explanations of quantum thermalization. More specifically, we consider the chaos criteria proposed in [10]. In this work, the unitary time evolution operator $U(t) = e^{-itH}$ is mapped to a doubled state. The doubled state is defined by considering two copies of the physical system and preparing maximally entangled EPR pairs between each site of the physical system and its doubling partner. Denoting this state by $|I\rangle$, the unitary $U(t)$ is mapped to the pure state $\mathbb{1} \otimes U(t)|I\rangle$. By construction, the two copies of systems (named as the past system and future system) are always maximally entangled, with $U(t)$ the Schmidt matrix of the wavefunction. Quantum chaos is characterized by the suppression of mutual information between subsystems of the future and past systems. A small mutual information between a region A in the past and a region B in the future tells us that operators in A mostly evolve to non-local operators exceeding the boundary of B, causing a suppression of local correlation functions. This criteria is shown to be related to another chaos criteria, the out-of-time-ordered correlation (OTOC) functions [11–16].

In this paper, we show that the mutual information criteria defined in [10] also controls the entropy growth for proper choices of low entropy initial states. More specifically, we consider a given partition of the system into multiple regions, and consider an ensemble of initial states that are tensor products of random states in each region. After time evolution by $U(t)$, we study the purity $\mathrm{Tr}\,\rho_A(t)^2$ of a subsystem A in the final state. The ensemble average of the second Renyi entropy is determined by a sum in which each term is controlled by the second Renyi mutual information in the doubled state. When the mutual information terms are sufficiently small, a typical product state at initial time evolves into a state with nearly maximal entropy. Therefore we have shown that chaos in the dynamics $U(t)$ implies thermalization, at least for the ensemble of unentangled initial states we define. Since the random product state has a high energy, the final state has maximal entropy and infinite temperature. We also discuss the generalization of our result to initial state ensembles with finite temperature.

The remainder of the paper is organized as follows. In section 2, we review the relevant aspects of quantum chaos and show how information theoretic quantities are linked to re-thermalization of the thermal ensemble. Sections 3 and 4 are the main results of this work, and demonstrate a connection between thermalization of product states, quantum information theory, and quantum chaos. Derivations of the main results are in appendix A.

In what follows, we denote density matrices of subsystems by $\rho_A = \mathrm{Tr}_{\bar{A}}\,\rho$, dimension of subsystem A by D_A, and operators that act by identity in $\bar{A}$ as O_A. As a reminder, the von Neumann entropy of a density matrix ρ is $S[\rho] = -\mathrm{Tr}[\rho \ln \rho]$, and the mutual information between two subsystems in ρ is $I[\rho; A, B] = S[\rho_A] + S[\rho_B] - S[\rho_{A \cup B}]$.

2 Quantum chaos

We start by reviewing some recent results in understanding quantum chaos. In classical systems, one diagnostic of chaos is exponential sensitivity to initial conditions, quantified by the exponential growth of the Poisson bracket of some pair of phase space coordi-

nates $\{q(t), p(0)\} = \partial q(t)/\partial q(0) \sim e^{\lambda_L t}$. A natural generalization to the quantum case is $\langle -[O_A(t), O_B(0)]^2 \rangle_\rho$ [11]. Of the four terms in the expansion of this expression, the most interesting for our purpose is the out-of-time ordered four-point correlator (OTOC)

$$C_4(O_A(t), O_B(0))_\rho = \langle O_A(t) O_B(0) O_A(t) O_B(0) \rangle_\rho, \tag{2.1}$$

whose decay for thermal ρ has been interpreted as a signal of quantum chaos [10, 12, 13].

The decay of C_4 seems to be an operator-dependent statement, but is in fact related to an information-theoretic quantity, the second Renyi entropy $S^{(2)}$, computed from the time evolution $U(t) = e^{-iHt}$ [10]. Since $U(t) \in \mathscr{H} \otimes \mathscr{H}^*$, we can consider it as a normalized state in a Hilbert space with inner product $\langle A, B \rangle = \mathrm{Tr}[A^\dagger B]/D$ (here D is the dimension of $\mathscr{H}$). We can associate the copy of the Hilbert space corresponding to the future (past) with the left (right) tensor factor in $\mathscr{H} \otimes \mathscr{H}^*$. For intuition and computation, it can be useful to choose an isomorphism $\mathscr{H}^* \simeq \mathscr{H}$ that is compatible with the tensor factorization and think of $U(t)$ as an entangled state on two copies of the original system. Concretely, we consider the original Hilbert space $\mathscr{H}$ as a tensor product of small Hilbert spaces (for example on each site of a lattice): $\mathscr{H} = \bigotimes_x \mathscr{H}_x$. Denoting an orthonormal basis of $\mathscr{H}_x$ by $|\alpha_x\rangle, \alpha_x = 1, 2, \ldots, \dim \mathscr{H}_x$, one can define the maximally entangled state in the doubled Hilbert space as $|I\rangle = \bigotimes_x \left(D^{-1/2} \sum_{\alpha_x = \alpha_x^P = \alpha_x^F} |\alpha_x^P\rangle \otimes |\alpha_x^F\rangle \right)$. The state $|I\rangle$ encodes an isomorphism from operators (elements of $\mathscr{H}^* \otimes \mathscr{H}$) to states in a doubled system (elements of $\mathscr{H} \otimes \mathscr{H}$) by right action, so we can explicitly map the unitary operator $U(t)$ to the state $|U(t)\rangle = (\mathbb{1} \otimes U(t))|I\rangle$.[2] We denote the density matrix associated with the pure state $|U(t)\rangle$ as $\rho^{U(t)} = |U(t)\rangle\langle U(t)|$. The construction of $|U(t)\rangle$ and an example partial trace of $\rho^{U(t)}$ is illustrated in figure 1.

Correlations between the past and future copies of Hilbert space in $U(t)$ are related to chaos and scrambling. For example, the mutual information between a region A in the future and region B in the past bounds correlations in time:

$$I[\rho^{U(t)}; A_F, B_P] \geq \frac{1}{2} \left(\frac{\langle O_A(t) O_B(0) \rangle_{\beta=0} - \langle O_A(t) \rangle_{\beta=0} \langle O_B(0) \rangle_{\beta=0}}{\|O_A\| \|O_B\|} \right)^2. \tag{2.2}$$

We can already see a connection of the information content of $\rho^{U(t)}$ and thermalization in (2.2). If the mutual information between A in the future and B in the past is small in $\rho^{U(t)}$, the action of an operator in B in the past has no influence on the action of an operator in A in the future. This shows that small mutual information in $\rho^{U(t)}$ is sufficient for re-thermalization of the infinite temperature ensemble after perturbation. Thus in this case, we have the natural statement that information in $U(t)$ between regions in the future and past tells us how sensitive the future region is to the initial conditions in the past region. The main goal of this work is to extend this result to far-from-equilibrium pure states. More generally, (2.2) shows that we can think of the mutual information $I[\rho^{U(t)}; A_F, B_P]$ roughly as quantifying how much initial conditions in B determine the subsystem A after time evolution $U(t)$.

---

[2]It is important to note that $|U(t)\rangle$ depends on the basis choice used when defining $|I\rangle$, which determines the isomorphism $\mathscr{H} \simeq \mathscr{H}^*$. However, entanglement properties of $\rho^{U(t)}$ are completely independent from the basis choice.

(a) Time evolution $U(t)$.   (b) $U(t)$ as a state in $\mathscr{H} \otimes \mathscr{H}$.

(c) Density matrix $\rho_{A_F \cup B_P}^{U(t)}$; on the left, A has a white box around it, and B has a red box.
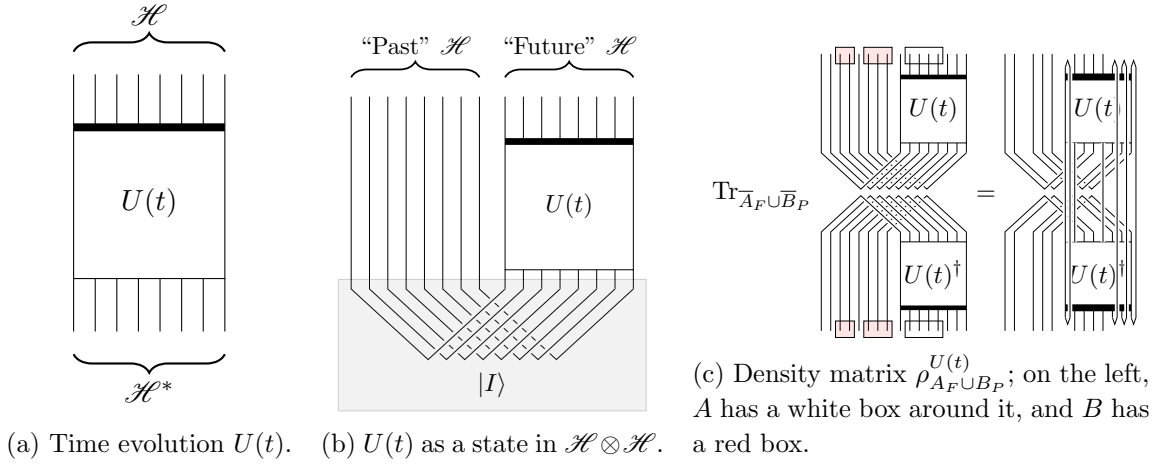
Figure 1. Pictorial representation and explicit construction of the mapping from time evolution operator $U(t)$ to the state $|U(t)\rangle \in \mathscr{H} \otimes \mathscr{H}$ and the associated density matrix $\rho^U(t)$. First, in 1a we introduce our notation and draw $U(t)$ as a tensor with "input" legs at the bottom and "output" legs at the top. To help keep track of the future and past, we draw the output edge of $U(t)$ as a bolded line. Each leg corresponds to a subsystem of $\mathscr{H}$ and denotes an index in the tensor, and contraction is represented by simply connecting "input" with "output" legs. A particular example of this operation is shown in 1b, where we depict action by $\mathbb{1} \otimes U(t)$ on the maximally entangled state $|I\rangle$, turning $U(t)$ into a state $|U(t)\rangle$ on a doubled system. In 1c we show $\rho_{A_F \cup B_P}^{U(t)} = \mathrm{Tr}_{\overline{A}_F \cup \overline{B}_P}|U(t)\rangle\langle U(t)|$. It is clear from this construction that $\rho^{U(t)}$ is maximally entangled between the past and future, so that for any region R exclusively in the past or future, $S[\rho_R^{U(t)}] = S^{(2)}[\rho_R^{U(t)}] = \ln D_R$.

A main result of [10] is a more explicit connection of past-future mutual information to chaos: the average of the OTOC (equation (2.1)) over operators in subsystem A in the future and B in the past is proportional to $S^{(2)}[\rho_{A_F \cup \bar{B}_P}^{U(t)}]$, where $\bar{B}_P$ is the complement of B_P in the past system. By average of operators on a subsystem A we mean a weighted sum over the D_A^2 Hermitian operators in a complete, orthonormal basis (under the above inner product on operators). We will write these operators in a script font, as $\mathcal{O}_A$, with the average implied wherever they appear. The second Renyi entropy is defined as

$$S^{(2)}[\rho] = -\log \mathrm{Tr}\, \rho^2 \tag{2.3}$$

and is a measure of uncertainty in ρ: for pure states, $S^{(2)} = 0$, while for maximally mixed states $S^{(2)} = S = \log D$ where S is the von Neumann entropy and D is the dimension of the Hilbert space. There are some other properties of $S^{(2)}[\rho_{A_F \cup B_P}^{U(t)}]$ that will be important in what follows. It can be seen from Jensen's inequality that $S^{(2)}[\rho] \leq S[\rho]$. Thus when $S^{(2)}$ is near-maximal, so is S. We also have the bounds[3] $\ln D_A + \ln D_B \geq S^{(2)}[\rho_{A_F \cup B_P}^{U(t)}] \geq |\ln D_A - \ln D_B|$. Thus for B much larger than A, $S^{(2)}$ is large for "kinematic" reasons, independent of the time evolution. The more interesting quantity in this case is a version of the mutual information adopted for Renyi entropy, $I^{(2)}[\rho^{U(t)}; A_F, B_P] =$

---

[3]As can be seen by writing $e^{-S^{(2)}[\rho_{A_F \cup B_P}^{U(t)}]} = \mathrm{Tr}[\pi_A(t)\pi_B(0)]/D_A D_B$ where π_A is the projector onto Hermitian operators in A in the operator Hilbert space defined above, or by the positivity of $I^{(2)}$.

$S^{(2)}[\rho_{A_F}^{U(t)}] + S^{(2)}[\rho_{B_P}^{U(t)}] - S^{(2)}[\rho_{A_F \cup B_P}^{U(t)}]$, which is non-negative in our state since $\rho^{U(t)}$ is maximally entangled between the future and past. $I^{(2)}[\rho^{U(t)}; A_F, B_P]$ captures fluctuations of $S^{(2)}[\rho_{A_F \cup B_P}^{U(t)}]$ about its kinematic value, and bounds the corresponding mutual information $I[\rho^{U(t)}; A_F, B_P]$ from above.

We can then write (as shown in [10])

$$D_A^2 C_4(\mathcal{O}_A(t), \mathcal{O}_{\bar{B}}(0))_{\beta=0} = \exp(I^{(2)}[\rho^{U(t)}; A_F, B_P]). \tag{2.4}$$

Thus scrambling in $U(t)$ as quantified by $S^{(2)}$ is directly related to chaos. A generically small four-point correlator means a large Renyi entropy or a small $I^{(2)}$. For A and B small, the expression (2.4) is actually in terms of non-local operators on $\bar{B}$. In this case, there is a more natural expression in terms of local operators,

$$(D_A D_B \langle \mathcal{O}_A(t) \mathcal{O}_B(0) \rangle_{\beta=0})^2 = \exp(I^{(2)}[\rho^{U(t)}; A_F, B_P]). \tag{2.5}$$

The two expressions (2.4) and (2.5) emphasize the important point that the second Renyi mutual information characterizes the behavior of both two-point functions and OTOC. $I^{(2)}[\rho^{U(t)}; A_F, B_P]$ between two small regions A and B is governed by two-point functions of operators supported on A and B, while that between a small region A and a big region B (bigger than half system size) is governed by the OTOC of operators supported on A and the smaller region $\bar{B}$. This is also consistent with the fact that the decay of the OTOC implies a stronger scrambling of information than simply the decay of two-point functions. As we will see below, one utility of the point of view of information is a unified treatment of the two- and four-point functions.

In [10], the relationship (2.4) is used to show that a four-point correlator decaying to some value less than ϵ in any region implies that the sum of mutual informations $I[\rho^{U(t)}; A_F, B_P] + I[\rho^{U(t)}; A_F, \bar{B}_P] \leq I^{(2)}[\rho^{U(t)}; A_F, B_P] + I^{(2)}[\rho^{U(t)}; A_F, \bar{B}_P] < 4 \ln D_A + 2 \ln \epsilon$. In principle, ϵ can be so small that this sum is arbitrarily close to zero. In more realistic models, we can expect that the OTOC will decay as some polynomial of the logarithm of the *total* Hilbert space dimension. This sum $I[\rho^{U(t)}; A_F, B_P] + I[\rho^{U(t)}; A_F, \bar{B}_P]$ minus $I[\rho^{U(t)}; A_F, B_P \bar{B}_P] \equiv 2 \ln D_A$ is called tripartite information and its negativity is proposed as a measure of "scrambling" due to unitary time evolution; then quantum chaos as measured by the decay of C_4 implies scrambling.

We would like to make a side remark at the end of this section. We treat $I^{(2)}$ and I for $\rho^{U(t)}$ as operator-independent diagnostics of chaos. It is clear from the discussion above that if the OTOC and two-point functions decay generically, $I^{(2)}$ will be small, which implies I is small as well. Although it is most direct from the discussion above to treat $I^{(2)}$ as the intrinsic measure of chaos and I simply as a quantity *also* small in chaotic systems only because it is bounded by $I^{(2)}$, the true mutual information I is more natural in many other contexts and it is intuitive that small mutual information of $\rho^{U(t)}$ should *imply* chaos. To that end, using a bound on von Neumann entropy in terms of Renyi entropy [17] (see appendix B), we can show that

$$\exp(I^{(2)}[\rho^{U(t)}; A_F, B_P]) \leq 1 + \left(1 + \frac{1}{\ln(D_A D_B)}\right)(D_A D_B - 1)\frac{I[\rho^{U(t)}; A_F, B_P]}{\ln(D_A D_B)}. \tag{2.6}$$

Thus a sufficiently small mutual information implies small $I^{(2)}$, which in turn implies chaos according to the OTOC. In the remainder of the work, we will focus on $I^{(2)}$, but (2.6) should be kept in mind as a way to bound $I^{(2)}$ in terms of the true mutual information.

3 Thermalization of completely random product states

Our goal is to understand how entropy growth and thermalization is related to quantum chaos as defined above. As discussed in section 1, entropy growth is a state-dependent statement and can only be true for specific classes of states, for example initial states with small entanglement. The most naive choice of initial state ensemble is product states of some fixed granularity. More precisely, we consider a partition of the initial system into regions R_s such that $\cup_{s=1}^{S} R_s = P$ is the whole system. Correspondingly, each region R_s has a Hilbert space $\mathscr{H}_s$, and the Hilbert space of the whole system can be written as a tensor product of subsystems $\mathscr{H} = \bigotimes_s \mathscr{H}_s$. We consider states of the form $|\psi(0)\rangle = \bigotimes_s |a_s\rangle$, with $|a_s\rangle$ a random pure state in $\mathscr{H}_s$. An example of one of these states, along with its time evolution, is shown in figure 2a. There is no change in the global entropy of a density matrix under unitary evolution, but there *can* be changes in subsystems. Thus we consider the second Renyi entropy of the density matrix corresponding to an initial low-entanglement pure state in some subsystem A[4] after time evolution by $U(t)$, averaged uniformly (according to the Haar measure) over initial states of the form $|\psi(0)\rangle$. Denoting the average $\mathbb{E}_H[f(\bigotimes_s |a_s\rangle)] = \int \prod_s dU_s f(\bigotimes_s U_s |a_s\rangle)$, where integrals are done over the Haar measure, and $\rho^\psi(0) = |\psi(0)\rangle\langle\psi(0)|$, we find our main result

$$\mathbb{E}_H\left[\exp(-S^{(2)}[\rho_A^\psi(t)])\right] = \frac{1}{\prod_s(1+1/D_s)} \frac{1}{D_A} \left(1 + \sum_{\substack{R\in\mathcal{P}(\{R_s\}) \\ R\neq\varnothing}} \frac{\exp(I^{(2)}[\rho^{U(t)}; A_F, R_P])}{D_R}\right)$$

(3.1)

where the sum runs over all nontrivial subregions $R = R_{i_1} \cup R_{i_2} \cup \cdots \cup R_{i_n}$ that are unions of some of the building blocks R_s. $\mathcal{P}(\{R_s\})$ denotes the set of all such R's, i.e. the powerset of $\{R_1, R_2, \ldots, R_S\}$. A similar relation has recently been studied in the context of random dynamics in [18]. A representation of a typical term in the sum is shown in figure 2b. We give some examples of this formula below, and present a derivation in appendix A.

Apart from bounding von Neumann entropy from below, the utility of computing $S^{(2)}$ is that it can be used to bound the one-norm difference of density matrices. Recall that $|\langle O\rangle_\rho - \langle O\rangle_{\rho'}|/\|O\| \leq \|\rho - \rho'\|_1$, so the one-norm is the natural distance for density matrices. By Jensen's inequality,

$$\mathbb{E}_H\left[\|\rho_A^\psi(t) - \rho_{A\beta=0}\|_1\right] \leq \sqrt{D_A \mathbb{E}_H e^{-S^{(2)}[\rho_A^\psi(t)]} - 1}.$$

(3.2)

Thus as long as the deviation from maximal entanglement is sufficiently small, we can say a density matrix thermalizes in the one-norm in expectation. As we will see below, it turns out that thermalizing in expectation (at infinite temperature) sufficiently well is

---

[4]Note that A need not factorize through the $\{\mathscr{H}_s\}$.

(a) Pictorial representation of the final state $|\psi(t)\rangle$ obtained from time evolution of the initial product state $|\psi(0)\rangle$.

(b) An example of region R involved in the sum of eq. (3.1).

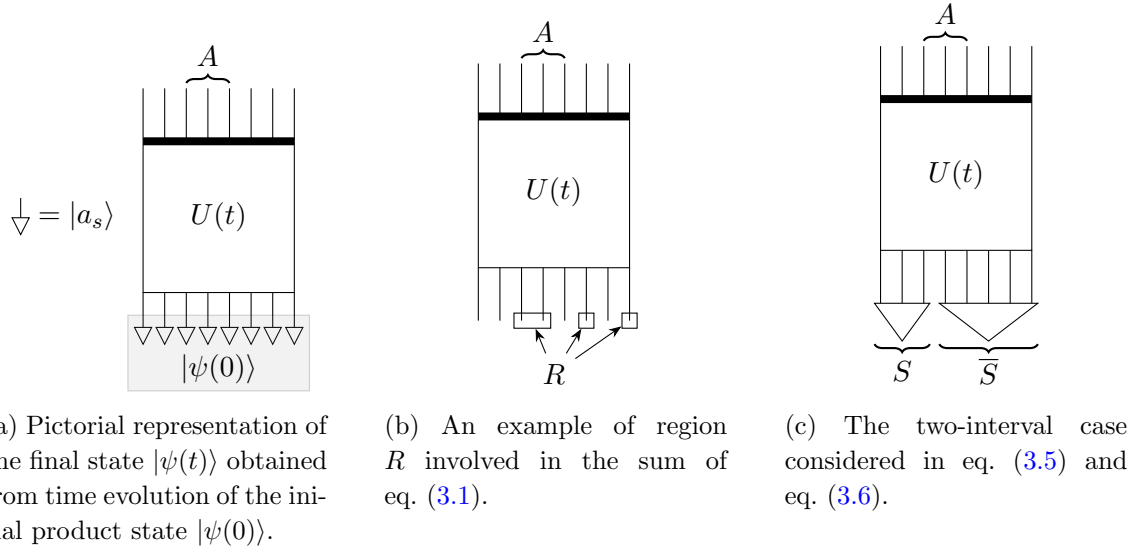(c) The two-interval case considered in eq. (3.5) and eq. (3.6).

Figure 2. Depiction of our setup. In figure 2a we show an example of a state from the ensemble we find the average entropy of in (3.1). Note in general each state $|a_s\rangle$ is different despite being drawn using the same symbol $\downvarrow$. In figure 2b, we show a typical region R from the powerset $\mathcal{P}(\{R_s\})$. Figure 2c illustrates the special case with bipartition of the past system.

sufficient for most states to thermalize. Note that the infinite temperature ensemble is the appropriate choice here, since for typical Hamiltonians most states will be infinite temperature states (cf. [19]), and we always have $\mathbb{E}_H \langle H \rangle_{\rho^\psi} = \langle H \rangle_{\beta=0}$.

To get an intuition for the implications of (3.1), it is illuminating to consider some special cases. First, we consider a trivial partition with only one region $R_1 = P$ equal to the whole system. In this case the ensemble is that of random pure states on the whole system. Our formula reduces to

$$\mathbb{E}_H \left[\exp(-S^{(2)}[\rho_A^\psi(t)]) \right] = \frac{1}{1 + 1/D} \frac{1}{D_A} \left(1 + \frac{D_A^2}{D} \right), \tag{3.3}$$

which is completely independent of dynamics. In the limit of large system size, as long as the subsystem A is less than half the system and grows at most linearly with system size, D_A^2/D decays exponentially with system size. Then (3.3) is the familiar statement that to exponential accuracy, a random pure state is close to maximally entangled in any small subsystem. This result is expected, as the typical pure state is indeed close to maximally entangled in any subsystem [20–22], and a random state evolves to another random state under *any* dynamics. In fact, the result of [20], derived by explicit integration on S^{2D-1}, is a special case of (3.3) for trivial evolution $U(t) = \mathbb{1}$. We can see the relationship to a more traditional measure of entanglement, the von Neumann entropy, by Jensen's inequality:

$$\mathbb{E}_H[S[\rho_A^\psi(t)]] \geq \mathbb{E}_H[S^{(2)}[\rho_A^\psi(t)]] \geq \ln D_A - \left(\ln(1 + D_A^2/D) - \ln(1 + 1/D) \right). \tag{3.4}$$

We move to the case of two initial subsystems, $\mathscr{H} = \mathscr{H}_S \otimes \mathscr{H}_{\bar{S}}$ where we take $1 \ll D_S \leq D_{\bar{S}}$. A typical state and its time evolution in this setup is shown in figure 2c. The

expression (3.1) becomes

$$
-\ln \mathbb{E}_H\left[\exp(-S^{(2)}[\rho_A^\psi(t)])\right] = \ln D_A - \ln\left(1 + \frac{e^{I^{(2)}[\rho^{U(t)};A_F,S_P]}}{D_S} + \frac{e^{I^{(2)}[\rho^{U(t)};A_F,\bar{S}_P]}}{D_{\bar{S}}} + \frac{D_A^2}{D}\right)
$$
$$
+ \ln(1 + (1 + D_S + D_{\bar{S}})/D). \tag{3.5}
$$

Already in this next-to-simplest case dynamics play a central role. First, if region S is large (and $\overline{S}$ is even larger), all the terms in $(1 + D_S + D_{\bar{S}})/D$ are exponentially small. Regardless, this contribution serves to *increase* $S^{(2)}$. For A smaller than half system size, D_A^2/D is exponentially small. The only decrease from maximal entanglement that can survive in the large system size limit is then due to the terms involving $I^{(2)}[\rho^{U(t)}; A_F, R_P]$. Thus small $I^{(2)}$ between A and both S and $\bar{S}$, equivalent respectively to the generic decay of two- and four-point correlators between A and S, is necessary and sufficient for the expectation of $S^{(2)}[\rho_A^\psi(t)]$ to be near the maximal (equivalently thermal at infinite temperature) value for initial product states in S and $\bar{S}$. It is important to note that "small" $I^{(2)}$ depends on our choice of S, as terms have the form $e^{I^{(2)}[\rho^{U(t)};A_F,R_P]}/D_R$. If we want this contribution less than ϵ_R, we only require the condition that $I^{(2)}[\rho^{U(t)}; A_F, R_P] < \ln(D_R \epsilon_R)$.

It is also useful to rewrite (3.5) in terms of correlation functions

$$
-\ln \mathbb{E}_H\left[\exp(-S^{(2)}[\rho_A^\psi(t)])\right] = \ln D_A
$$
$$
- \ln\left(1 + D_A^2\left(\frac{1}{D} + D_S\langle \mathcal{O}_A(t)\mathcal{O}_S(0)\rangle_{\beta=0} + \frac{C_4(\mathcal{O}_A(t), \mathcal{O}_S(0))_{\beta=0}}{D_{\overline{S}}}\right)\right)
$$
$$
+ \ln(1 + (1 + D_S + D_{\bar{S}})/D). \tag{3.6}
$$

For large systems D_A^2/D is exponentially smaller than $D_A^2/D_{\overline{S}}$, so we can safely focus on the contributions due to correlators. It is clear that if the two-point functions decay and $D_A^2/D_{\overline{S}}$ is finite, the deviation from maximum entanglement will be dictated by the (strictly positive) four-point term $D_A^2 C_4(\mathcal{O}_A(t), \mathcal{O}_S(0))_{\beta=0}/D_{\overline{S}}$. Note that depending on the choice of A and S, even with both less than half system size and (for lattice models) $|A| < |S|$, $D_A^2/D_{\overline{S}}$ may be made of order, or even much greater than 1. As mentioned in section 2, $C_4(\mathcal{O}_A(t), \mathcal{O}_S(0))_{\beta=0}$ can be as small as an inverse polynomial in the logarithm of system dimension in chaotic systems, so as long as A and S are chosen so that $D_A^2/D_{\overline{S}}$ does not grow too quickly with system size, in chaotic systems random product states on S and $\overline{S}$ will evolve to look thermal in A. In the limit $D_A^2/D_{\bar{S}} \to 0$, only the local two-point function will contribute to deviations from $\ln D_A$. This is the case considered by CT, and shows that chaos in the OTOC sense is not necessary for thermalization into a much larger random bath. On the other hand, when $D_{\bar{S}}$ is finite, if four-point correlations do not decay sufficiently we can have significant corrections to thermal entropy.

This argument extends without significant modification to the case of S initial subsystems $\mathscr{H} = \bigotimes_{s=1}^S \mathscr{H}_s$, where S may grow linearly with system size. As long as two-point functions generically decay between A and subsystems up to half system size, the contribution from summands in (3.1) involving R less than half the system will be small. The decay of four-point correlators between A and subsystems up to half system size is necessary to

bound contributions from summands involving R greater than half system size. Concretely, we need, for regions R less than half system size, $\langle \mathcal{O}_A(t)\mathcal{O}_R(0)\rangle_{\beta=0} \ll 1/(D_A\sqrt{D_R})$ and $C_4(\mathcal{O}_A(t)\mathcal{O}_R(0))_{\beta=0} \ll D_{\overline{R}}/D_A^2$ for A to look maximally entangled on average. Note that in an integrable system, it is expected that $I^{(2)}$ will always be high for *some* subextensive[5] subregions R (for example numerics in [10]), although these regions may change in time as information propagates. Some subextensive region of initial conditions largely determines the density matrix in A. This demonstrates an obstacle to thermalization in non-chaotic isolated systems. In contrast, a chaotic system will scramble information about the initial conditions in each $\mathcal{H}_s$ across extensive regions of the system. Equivalently, extensive knowledge of initial conditions determines the density matrix in A. The only R for which $\exp(I^{(2)}[\rho^{U(t)}; A_F, R_P]) \gtrsim D_A$ have $D_R \sim D$, so that even if $I^{(2)}[\rho^{U(t)}; A_F, R_P]$ attains its maximum $2\ln D_A$, D_A^2/D_R decays exponentially in system size if A is chosen as in the examples above.

For systems with

$$\epsilon = \mathbb{E}_H\left[D_A e^{-S^{(2)}[\rho_A^\psi(t)]} - 1\right] = \frac{1}{\prod_s(1+1/D_s)} \sum_{\substack{R\in\mathcal{P}(\{R_s\}) \\ R\neq\varnothing}} \frac{e^{I^{(2)}[\rho^{U(t)}; A_F, R_P]}}{D_R}$$
$$+ \left(\frac{1}{\prod_s(1+1/D_s)} - 1\right) \tag{3.7}$$

small, we can meaningfully bound the number of states that do not thermalize by Markov's inequality:

$$\Pr\left(D_A e^{-S^{(2)}[\rho_A^\psi(t)]} - 1 > \delta\right) < \epsilon/\delta. \tag{3.8}$$

The conclusion is that if we find that states are expected to thermalize sufficiently well, then a particular state is likely to have the average behavior after long times. This bound is easily "weakened" to a statement about probabilities of significant deviation for local entropy. On the other hand, if states are not expected to thermalize, we do not expect to find such a bound on physical grounds; the long-time trajectory of non-thermalizing systems can depend sensitively on the details of initial conditions.

4 Finite temperature extension

As discussed above, the preceding results should be interpreted as statements about thermalization at infinite temperature. To get ensembles other than infinite temperature we must restrict the set of initial states we average over. One natural way is to still consider a partition into regions $R_s, s = 1, 2, \ldots, S$, but in each region we restrict the state into a subspace of Hilbert space $\mathcal{H}_s$, denoted as $\mathcal{H}_{Ms} \subset \mathcal{H}_s$. Physically, $\mathcal{H}_{Ms}$ is the subspace of states in an energy window $E_0 < E < E_0 + \Delta E$, when we define the energy with respect to the subsystem Hamiltonian of R_s, neglecting the boundary term contribution. We can define a "microcanonical" density matrix $\rho_{Ms} = \pi_{Ms}/D_{Ms}$ for each region, with π_{Ms} the projection

---

[5]By subextensive we do *not* mean not growing with system size, but we mean of dimension D_R small enough that D_A^2/D_R, with A chosen as in the examples above, is not exponentially small.

operator onto $\mathscr{H}_{Ms}$, and D_{Ms} the dimension of $\mathscr{H}_{Ms}$. Then we consider the initial state as pure states drawn from the ensemble $\rho_M = \otimes_s \rho_{Ms}$, which are states with zero entanglement entropy between different regions, and have a finite energy density. It will be convenient to change the normalization on the operator inner product to be $\langle A, B \rangle = \text{Tr}[A^\dagger B]$.

The above results suggest that entropy growth and thermalization of these product states should be related to entanglement properties of some state depending on $U(t)$; tentatively, call this state $|U_M(t)\rangle$ (in sections 2 and 3, the relevant state was isomorphic to the operator $U(t)$). As a first check that we have chosen a useful state, it is natural to require that some analog of (2.2) hold for $\rho^{U_M(t)}$. Such a result would suggest that correlations in $|U_M(t)\rangle$ are related in the same intuitive way to thermalization of states from the ensemble ρ_M as correlations in the state $|U(t)\rangle$ are to thermalization of states from the ensemble $\mathbb{1}/D$. A useful choice turns out to be $|U_M(t)\rangle = \mathbb{1} \otimes U(t)\rho_M^{1/2}|I\rangle$; since we have chosen ρ_M compatible with the tensor factorization of $\mathscr{H}$, to understand this state one can just put projectors on each input leg of $U(t)$ in figures 1 and 2. The case $\pi_M = \mathbb{1}$ has been described in section 2; it turns out this is a very special case, due to the fact that $\mathbb{1}$ commutes with everything. For general ρ_M and two regions A_F, B_P in the past and future systems, respectively, we have the bound

$$I[\rho^{U_M(t)}; A_F, B_P] \qquad (4.1)$$
$$\geq \frac{1}{2}\left(\frac{\left(\langle O_A(t)O_B(0)\rangle_{\rho_M} - \langle O_A(t)\rangle_{\rho_M}\langle O_B(0)\rangle_{\rho_M}\right) + \langle O_A(t)[\pi_M, O_B(0)]\rangle_{\rho_M}}{\|O_A\|\|O_B\|}\right)^2.$$

If the commutator $[\pi_M, O_B]$ is small, (4.1) becomes exactly analogous to (2.2). For example, suppose H has the form $H \equiv H_L = \sum_s H_s + \sum_{\partial s} H_{\partial s}$ where H_s act on disjoint subsystems $\mathscr{H}_s$, and the boundary terms $H_{\partial s}$ are allowed to couple "nearby" subsystems. If we then choose π_{Ms} to project onto some subsystem energy window (one where the eigenvalues of H_s lie in some fixed range) and take $\pi_M = \bigotimes_s \pi_{Ms}$, operators O_B that are local to subsystems and do not change the energy outside the energy window will have zero commutator with π_M. Another example is some local conserved quantity that we choose to concentrate in some subsystem $\mathscr{H}_s$ by choice of π_{Ms}; if O_B does not transport this charge across subsystems it will have zero commutator with π_M. If the above conditions are only met approximately (O_B has small matrix elements for bringing states out of and into $\mathscr{H}_M$), the commutator will be small. Of course, we can enforce a zero-commutator condition on O_B by simply taking it to be $\tilde{O}_B = \pi_M O_B \pi_M$. Then we can directly interpret (4.1) as the "re-equilibration" of ρ_M after acting by O_B; perturbing the state by O_B does not affect the action of O_A in the future. Of course, the case $\pi_M = \mathbb{1}$ reduces exactly to (2.2) for any choice of O_B.

As mentioned, $\pi_M = \mathbb{1}$ is a very special case, and this gives rise to important differences when relating information measures to chaos and equilibration for generic ρ_M. The bound $S^{(2)} \leq S$ is always true, so $S^{(2)}$ of $\rho^{U_M(t)}$ is still a good measure of the "correlation" between the past and future. Important special cases are $S^{(2)}[\rho_{A_F}^{U_M(t)}] = S^{(2)}[\rho_{MA}(t)]$ and $S^{(2)}[\rho_{B_P}^{U_M(t)}] = S^{(2)}[\rho_{MB}(0)]$. On the other hand, since ρ_M can have non-trivial time evolution, the quantity $S^{(2)}[\rho_{A_F}^{U_M(t)}] + S^{(2)}[\rho_{B_P}^{U_M(t)}] - S^{(2)}[\rho_{A_F \cup B_P}^{U_M(t)}]$ can become negative (in

contrast to the $\pi_M = \mathbb{1}$ case), so $I^{(2)}$ as defined in section 2 is not as fundamental a quantity. It also does not bound the corresponding mutual information I. We *can* define a quantity that upper bounds I,

$$\tilde{I}[\rho^{U_M(t)}; A_F, B_P] = I[\rho^{U_M(t)}; A_F, B_P] + (S[\rho_{A_F \cup B_P}^{U_M(t)}] - S^{(2)}[\rho_{A_F \cup B_P}^{U_M(t)}])$$
$$= S[\rho^{U_M(t)}; A_F] + S[\rho^{U_M(t)}; B_P] - S^{(2)}[\rho_{A_F \cup B_P}^{U_M(t)}] \qquad (4.2)$$

For $\pi_M = \mathbb{1}$, $\tilde{I}[\rho^{U_M(t)}; A_F, B_P] = I^{(2)}[\rho^{U_M(t)}; A_F, B_P]$. There are equalities analogous to (2.4) and (2.5) relating $\tilde{I}$ to chaos:

$$e^{\tilde{I}[\rho^{U_M(t)}; A_F, R_P]} = \left(\frac{e^{S[\rho_{MA}(t)]}}{D_A} \right) \left(\frac{D_{\bar{R}}}{D_{M\bar{R}}} \right) D_A^2 C_4(\mathcal{O}_A(t), \tilde{\mathcal{O}}_{\bar{R}}(0))_{\rho_M} \qquad (4.3)$$

$$= \left(\frac{e^{S[\rho_{MA}(t)]}}{D_A} \right) \left(\frac{D_{MR}}{D_R} \right) (D_A D_R \langle \mathcal{O}_A(t) \mathcal{O}_R(0) \rangle_{\rho_M})^2, \qquad (4.4)$$

for regions $R \in \mathcal{P}(\{R_s\})$ (for other sorts of regions, factors of D_{MR} will be replaced by entropies), where $\tilde{\mathcal{O}}_R = \pi_M \mathcal{O}_R \pi_M$. This modification of $\mathcal{O}_R$ has a natural interpretation, paralleling the discussion of (4.1). The OTOCs in (4.3) are to be computed for operators in the past that do not move states out of $\mathscr{H}_M$ (and act by zero on states outside); in the examples following (4.1), $\tilde{\mathcal{O}}_R$ will conserve local energy density or subsystem charge, respectively. We also clearly have $[\pi_M, \tilde{\mathcal{O}}_R] = 0$, so according to (4.1) the mutual information bounds the effect of these operators in the most intuitive way. Note that as $\mathscr{H}_{Ms}$ becomes smaller, the four-point contributions (4.3) become more important.

With these preparations, we can extend the results of section 3 to the case of generic ρ_M. For product states $|\psi_M\rangle = \bigotimes_s |a_{Ms}\rangle$ with each $|a_{Ms}\rangle$ taken from $\mathscr{H}_{Ms}$, we obtain the result analogous to (3.1):

$$\mathbb{E}_H[\exp(-S^{(2)}[\rho_A^{\psi_M}(t)])] = \frac{1}{\prod_s (1 + 1/D_{Ms})}$$

$$\times e^{-S^{(2)}[\rho_{MA}(t)]} \left(1 + \sum_{\substack{R \in \mathcal{P}(\{R_s\}) \\ R \neq \varnothing}} e^{-(S[\rho_{MA}(t)] - S^{(2)}[\rho_{MA}(t)])} \frac{e^{\tilde{I}[\rho^{U_M(t)}; A_F, R_P]}}{D_{MR}} \right). \qquad (4.5)$$

The $\tilde{I}$ terms are the positive corrections for product states from the entropy computed from ρ_M. We can also generalize the discussion surrounding (3.2) to see that

$$\mathbb{E}_H \left[\|\rho_A^{\psi_M}(t) - \rho_{MA}(t)\|_1 \right] \leq \sqrt{D_A \left(\mathbb{E}_H e^{-S^{(2)}[\rho_A^{\psi_M}(t)]} - e^{-S^{(2)}[\rho_{MA}(t)]} \right)}, \qquad (4.6)$$

so states chosen from the ensemble given by ρ_M "equilibriate" (in the above sense of 1-norm) to ρ_M given small $\tilde{I}$ terms. The major difference is that ρ_M is generically not thermal. Interpreting small $\tilde{I}$ as chaos, this shows that chaos is sufficient to "scramble" initial conditions to the extent that the particular state within the initial ensemble is irrelevant, but we have not shown that the "unentangled microcanonical ensemble" ρ_M itself thermalizes.

That said, in a system with a local Hamiltonian and with a choice of the regions R_s of size much bigger than the thermal correlation length, the contribution of boundary terms to energy is small, and ρ_M has a volume law entropy that is close to the thermal value at the same energy expectation value. In other words, an initial pure state drawn from ρ_M has already almost thermalized when the reduced density matrix approaches that of ρ_M.

5   Conclusion

We have explored the consequences of small correlation (as computed in $U(t)$ and $U_M(t)$) between the past and future. For the density matrices $\rho_{\beta=0}$ and ρ_M the expressions (2.2) and (4.1) respectively show that certain types of perturbations to these density matrices in some region B are "forgotten" in some region A as long as the information between A in the past and B in the future for $U(t)$ or $U_M(t)$ has had time to decay. Next, (3.1) and (4.5) show that the decay of information between past and future regions corresponds to entropy growth for far-from-equilibrium pure states. Finally, using these expressions in combination with (3.2), (3.8), and (4.6), we have shown that this decay of information between past and future means the initial conditions of a particular pure state chosen from an ensemble are forgotten (although the ensemble itself is not). Although the discussion proceeds most naturally in terms of information, we can also conclude that quantum chaos as diagnosed by the OTOC and the decay of local two-point functions imply entropy growth and erasure of initial conditions ("equilibration") by relating the OTOC and two-point functions to information. Likewise, generic thermalization implies the contribution of $I^{(2)}$ or $\tilde{I}$ terms in (3.1) or (4.5) are small, so mutual information between local regions in the future and sub-extensive regions in the past is bounded above. Thus there is a sense in which quantum thermalization implies chaos.

The most important extension of this work is a deeper understanding of the finite temperature results. The same state may be a member of several ensembles ρ_M, but the formalism we developed does not identify a preferred density matrix. Such a preferred density matrix should be a time independent distribution, for example the Boltzmann distribution, when that pure state equilibriates. A first step may be to find conditions such that the "more thermal" (higher entanglement) ρ_M thermalize. It should also be possible to improve the factor D_A in (4.6) in the case that $\mathbb{E}_H e^{-S^{(2)}[\rho_A^{\psi_M}(t)]} \simeq e^{-S^{(2)}[\rho_{MA}(t)]}$ for locally thermalizing ρ_M. Finally, to make this work more practically applicable, it is important to show that either chaos as measured by the OTOC or decay of information between the future and past is generic for local Hamiltonians.

Acknowledgments

A Derivation of main results

We present the derivation of the results in section 4, which are a strict generalization of the results in sections 2 and 3. The main tool is the Schur-Weyl duality, which describes the combined action of the symmetric and unitary groups on tensor product spaces. On the vector space $V_n = (\mathbb{C}^D)^{\otimes^n}$, the symmetric group on n letters, $\mathbb{S}^n$, acts in a natural way by permuting the n factors, while the unitary group $U(D)$ acts by $U^{\otimes^n}$ for $U \in U(D)$.

Theorem 1 (Schur-Weyl Duality). *Under the combined natural actions of $\mathbb{S}^n$ and $U(D)$ on the vector space V_n (as defined above), V_n can be written as a direct sum*

$$V_n = \bigoplus_Y W_Y \otimes S_Y$$

where Y is an index running over Young diagrams with n boxes, and W_Y (S_Y) is an irreducible representation of $U(D)$ ($\mathbb{S}^n$) not isomorphic to any other representation appearing with different Y.

We will typically use this theorem to constrain operators that commute with the action of $U(D) \times \mathbb{S}^n$; since each irrep of the combined action appears only once in the decomposition of V_n, such an operator must act as multiplication by a constant on each irrep by Schur's lemma. Furthermore, the theorem tells us we can project onto each irrep by projecting onto an irrep of only $\mathbb{S}^n$, so each such operator can be written as a sum of projectors, each of which is in turn a sum of elements of $\mathbb{S}^n$.

As an intermediate result, we must compute the Haar integral $A_n = \int dU (U|\psi\rangle\langle\psi|U^\dagger)^{\otimes^n}$, which is clearly independent of the choice of $|\psi\rangle$. Furthermore, A_n commutes with the above actions of $U(D)$ and $\mathbb{S}^n$, so we take A_n to be a sum of $\sigma \in \mathbb{S}^n$. It is easy to check that in fact $\sigma A_n = A_n$ for all $\sigma \in \mathbb{S}^n$, so $A_n \propto \sum_{\sigma \in \mathbb{S}^n} \sigma$. To find the normalization factor, note that[6]

$$\mathrm{Tr}\left[\sum_{\sigma \in \mathbb{S}^n} \sigma\right] = \sum_{c=1}^n (\text{count of permutations with } c \text{ cycles})D^c = \prod_{m=0}^{n-1}(D+m).$$

This gives

$$\mathbb{E}_H[|\psi\rangle\langle\psi|] = \int dU(U|\psi\rangle\langle\psi|U^\dagger)^{\otimes^n} = \left[\prod_{m=1}^{n-1}(1+m/D)\right]^{-1}\frac{1}{D^n}\sum_{\sigma \in \mathbb{S}^n} \sigma.$$

We now compute the expectation of the second Renyi entropy in the setup of section 4. Call the non-identity element of $\mathbb{S}^2$, that swaps tensor factors, $\mathbb{X}$. If we have a Hilbert space $\mathscr{H}$ with a subsystem labelled A, there is a permutation group $\mathbb{S}^n_A$ that acts on $\mathscr{H}^{\otimes^n}$ by only permuting tensor factors corresponding to subsystem A between copies. We refer to these group elements by a subscript A, so for example $\mathbb{X}_A$. To compute the expectation

---

[6] The last equality is proved easily after noting that for a given permutation of $n-1$ elements, to form a permutation of n elements the nth element either forms a new cycle, contributing a factor of D, or can be put in an existing cycle in $n-1$ distinct ways (regardless of the permutation).

of the second Renyi entropy, we use the relation $\text{Tr}[\rho_A^2] = \text{Tr}[\rho^{\otimes 2}\mathbb{X}_A]$. As a reminder, the distribution on initial ρ^{ψ_M} in our case is fixed as follows. We are given a partition of Hilbert space into S subsystems, $\mathscr{H} = \bigotimes_s \mathscr{H}_s$, and in the vector space associated to each subsystem we choose a linear subspace $\mathscr{H}_{Ms}$ (with associated projector π_{Ms}). The distribution on ρ^{ψ_M} is independently Haar random on the subspace of each subsystem. From the above discussion on the Haar integral, it follows that

$$\mathbb{E}_H[\rho^{\psi_M}(0)^{\otimes 2}] = \frac{1}{\prod_s(1 + 1/D_{Ms})}\rho_M^{\otimes 2}\prod_s(\mathbb{1}_s + \mathbb{X}_s) = \frac{1}{\prod_s(1 + 1/D_{Ms})}\rho_M^{\otimes 2}\sum_{R \in \mathcal{P}(\{R_s\})}\mathbb{X}_R,$$

where D_{Ms} is the dimension of $\mathscr{H}_{Ms}$, $\rho_M = \prod_s \pi_{Ms}/D_{Ms}$, and $\mathcal{P}(\{R_s\})$ is the powerset of subsystems. The second equality comes from noting that the product has 2^S terms, based on a choice of $\mathbb{1}_s$ or $\mathbb{X}_s$ for each subsystem, and the included swaps combine to give a single swap of all included subsystems. We then have

$$\mathbb{E}_H\left[\text{Tr}[\rho^{\psi_M}(t)_A^2]\right] = \text{Tr}[U(t)^{\otimes 2}\mathbb{E}_H[\rho^{\psi_M}(0)^{\otimes 2}](U(t)^\dagger)^{\otimes 2}\mathbb{X}_A] \tag{A.1}$$

$$= \frac{1}{\prod_s(1 + 1/D_{Ms})}$$
$$\times \sum_{R \in \mathcal{P}(\{R_s\})}\text{Tr}[(U(t)\sqrt{\rho_M})^{\otimes 2}\mathbb{X}_R(U(t)\sqrt{\rho_M})^{\dagger\otimes 2}\mathbb{X}_A] \tag{A.2}$$

$$= \frac{1}{\prod_s(1 + 1/D_{Ms})}\sum_{R \in \mathcal{P}(\{R_s\})}e^{-S^{(2)}[\rho_{A_F \cup R_P}^{U_M(t)}]} \tag{A.3}$$

where $U_M(t) = U(t)\sqrt{\rho_M}$. To see the last equality, we refer to the explicit construction of the state corresponding to $U_M(t)$ following the procedure of figure 1 to check that the index contractions are correct, and the proportionality factor is correct since $\text{Tr}[U(t)\rho_M U(t)^\dagger] = 1$. By the same construction, we can compute $S^{(2)}[\rho_{R_P}^{U_M(t)}] = S[\rho_{R_P}^{U_M(t)}] = \ln D_{MR}$, $S^{(2)}[\rho_{A_F}^{U_M(t)}] = S^{(2)}[\rho_{MA}(t)]$, and $S[\rho_{A_F}^{U_M(t)}] = S[\rho_{MA}(t)]$, so that upon multiplication by appropriate factors of entropy, equations (3.1) and (4.5) follow from (A.3).

To connect (A.1), and more generally entropies of $\rho^{U_M(t)}$, to observables, we use the explicit form of projectors onto irreps of $\mathbb{S}^2$: $\pi_\pm = (\mathbb{1} \pm \mathbb{X})/2$. Then an operator A on $\mathscr{H}^{\otimes 2}$ that commutes with the joint action of $U(D) \times \mathbb{S}^2$ can explicitly be written as a sum of $\pi_\pm$, with coefficients $\text{Tr}[A\pi_\pm]/\text{Tr}[\pi_\pm]$. This gives

$$A = \frac{1}{D^2 - 1}\left((\text{Tr}[A] - \text{Tr}[A\mathbb{X}]/D)\mathbb{1} + (\text{Tr}[A\mathbb{X}] - \text{Tr}[A]/D)\mathbb{X}\right).$$

In particular, we can implement $\mathbb{X}$ in terms of any operators that commute with $U(D) \times \mathbb{S}^2$ and have $D\,\text{Tr}[A] = \text{Tr}[A\mathbb{X}]$. One example can be found by taking a complete basis of Hermitian operators orthonormal under the inner product $\langle A, B \rangle = \text{Tr}[A^\dagger B]/D$, $\mathcal{B}$ (we make this normalization choice so that eigenvalues of operators in the basis can be ± 1). Since conjugation by a unitary $U \in U(D)$ preserves Hermeticity and the inner product, conjugating every element of $\mathcal{B}$ gives a new, still orthonormal basis of the real vector space of Hermitian operators, so can be written as a real orthogonal matrix acting on the elements of $\mathcal{B}$. Thus the average over Hermitian operators $Y = D^{-2}\sum_{\mathcal{O} \in \mathcal{B}}\mathcal{O}^{\otimes 2}$ is invariant under

conjugation by a unitary (and is therefore independent of the particular choice of operator basis), and clearly commutes with the action of $\mathbb{S}^2$; in other words Y commutes with the joint action of $U(D) \times \mathbb{S}^2$. Taking a basis including the identity shows $\mathrm{Tr}[Y] = 1$, and the normalization condition gives $\mathrm{Tr}[Y\mathbb{X}] = D^{-2} \sum \mathrm{Tr}[\mathcal{O}^2] = D$, in other words $Y = \mathbb{X}/D$. We can then write (using (A.2) and the following discussion), assuming that R factors through the tensor factorization into $\mathscr{H}_s$ for convenience,

$$
\begin{aligned}
e^{-S^{(2)}[\rho_{A_F \cup R_P}^{U_M(t)}]} &= \mathrm{Tr}[U(t)^{\dagger \otimes 2} \mathbb{X}_A U(t)^{\otimes 2} \mathbb{X}_R \rho_M^{\otimes 2}] = D_A D_R \,\mathrm{Tr}[(\mathcal{O}_A(t)\mathcal{O}_R \rho_M)^{\otimes 2}] \\
&= D_A D_R \left(\langle \mathcal{O}_A(t)\mathcal{O}_R \rangle_{\rho_M} \right)^2 \\
&= D_A \,\mathrm{Tr}[\mathcal{O}_A(t)^{\otimes 2} \mathbb{X}_{\overline{R}} \rho_M^{\otimes 2} \mathbb{X}] \\
&= \frac{D_A D_{\overline{R}}}{D_M} \,\mathrm{Tr}[\mathcal{O}_A(t) \mathcal{O}_{\overline{R}} \pi_M \mathcal{O}_A(t) \mathcal{O}_{\overline{R}} \rho_M] \\
&= \frac{D_A D_{\overline{R}}}{D_M} \,\mathrm{Tr}[\mathcal{O}_A(t) \tilde{\mathcal{O}}_{\overline{R}} \mathcal{O}_A(t) \tilde{\mathcal{O}}_{\overline{R}} \rho_M] = \frac{D_A D_{\overline{R}}}{D_M} C_4(\mathcal{O}_A(t), \tilde{\mathcal{O}}_{\overline{R}}(0))_{\rho_M}
\end{aligned}
$$

where the first equality on the last line follows since $[\pi_M, \mathbb{X}_R] = 0$. These expressions, after multiplication by $D_{MR} e^{S[\rho_{MA}(t)]}$, give equations (2.4), (2.5), (4.3), and (4.4). Finally, equations (2.2) and (4.1) are direct consequences of the well-known inequality

$$
I[\rho; A, B] \geq \frac{1}{2} \left(\frac{\langle O_A O_B \rangle_\rho - \langle O_A \rangle_\rho \langle O_B \rangle_\rho}{\|O_A\| \|O_B\|} \right)^2
$$

and the definition of $\rho^{U_M(t)}$.

As a possible tool for computing higher Renyi entropies, we note that elements of $\mathbb{S}^n$ can be implemented in terms of averages of local operators for $n > 2$, despite the fact that we used a property special to $n = 2$ (invariance of $\sum \mathcal{O}^{\otimes 2}$ when summed over an orthonormal basis) above. The idea is to take some as-yet unchosen Hermitian operator A, and define $M = \int dU (UAU^\dagger)^{\otimes n}$, which commutes with $U(D) \times \mathbb{S}^n$, and so is a weighted sum of projectors onto irreps of $\mathbb{S}^n$. The numbers $\mathrm{Tr}[M\sigma]$ for $\sigma \in \mathbb{S}^n$ determine the weights; these are in turn products of traces of powers of A. Thus M depends only on the spectrum of A, and by tuning this spectrum we can tune the weights of projectors. For example, if we take $n = 2$, the Haar average over operators A with some fixed spectrum satisfying $D(\sum_i \lambda_i)^2 = \sum_i \lambda_i^2$ is proportional to $\mathbb{X}$.

B Derivation of equation (2.6)

We present a short derivation of (2.6), based on equation 23 of [17]. That equation, in our notation, is

$$
S[\rho_{A_F \cup B_P}^{U(t)}] \leq \ln(D_A D_B) \left(1 - \tau \left(\frac{e^{-S^{(2)}[\rho_{A_F \cup B_P}^{U(t)}]} - (D_A D_B)^{-1}}{1 - (D_A D_B)^{-1}} \right) \right),
$$

where another result of [17] is that $\tau \geq \ln(D_A D_B)/(\ln(D_A D_B) + 1)$. Then, noting that $S[\rho_{A_F}^{U(t)}] = S^{(2)}[\rho_{A_F}^{U(t)}] = \ln D_A$ and likewise for $\rho_{B_F}^{U(t)}$ as $\rho^{U(t)}$ is maximally entangled between past and future (see figure 1c), we can multiply both sides by $D_A D_B$ and use the bound on τ to obtain (2.6).

References

[1] L.D. Landau and E.M. Lifshitz, *Statistical physics. Volume* 5*: course of theoretical physics*, Pergamon Press, Oxford, U.K. (1969).

[2] C. Neill et al., *Ergodic dynamics and thermalization in an isolated quantum system*, [*Nature Phys.* **12** (2016) 1037](#) [[arXiv:1601.00600](#)].

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

[4] J.Y. Choi et al., *Exploring the many-body localization transition in two dimensions*, [*Science* **352** (2016) 1547](#) [[arXiv:1604.04178](#)].

[5] S. Goldstein, J.L. Lebowitz, R. Tumulka and N. Zanghi, *Canonical typicality*, [*Phys. Rev. Lett.* **96** (2006) 050403](#) [[cond-mat/0511091](#)] [[INSPIRE](#)].

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

[7] 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](#)].

[8] M. Srednicki, *Chaos and quantum thermalization*, [*Phys. Rev.* **E 50** (1994) 888](#) [[cond-mat/9403051](#)].

[9] Y. Huang, F.G. S.L. Brandao and Y.-L. Zhang, *Finite-size scaling of out-of-time-ordered correlators at late times*, [arXiv:1705.07597](#) [[INSPIRE](#)].

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

[11] A.I. Larkin and Yu.N. Ovchinnikov, *Quasiclassical method in the theory of superconductivity*, [*Sov. Phys. JETP* **28** (1969) 1200](#) [*Zh. Eksp. Teor. Fiz.* **55** (1969) 2262].

[12] D.A. Roberts and D. Stanford, *Two-dimensional conformal field theory and the butterfly effect*, [*Phys. Rev. Lett.* **115** (2015) 131603](#) [[arXiv:1412.5123](#)] [[INSPIRE](#)].

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

[14] A. Kitaev, *A simple model of quantum holography (part* 1*)*, talk at *KITP program: entanglement in strongly-correlated quantum matter*, [KITP](#), University of California, Santa Barbara, CA, U.S.A., 7 April 2015.

[15] A. Kitaev, *A simple model of quantum holography (part* 2*)*, talk at *KITP program: entanglement in strongly-correlated quantum matter*, [KITP](#), University of California, Santa Barbara, CA, U.S.A., 27 May 2015.

[16] A. Kitaev, *Hidden correlations in the Hawking radiation and thermal noise*, talk at 2015 *breakthrough prize in fundamental physics symposium*, [KITP](#), University of California, Santa Barbara, CA, U.S.A., 12 February 2015.

[17] P. Harremoës and F. Topsoe, *Inequalities between entropy and index of coincidence derived from information diagrams*, *IEEE Trans. Inform. Theory* **47** (2001) 2944.

[18] Y.-Z. You and Y. Gu, *Entanglement features of random hamiltonian dynamics*, *Phys. Rev.* **B 98** (2018) 014309 [`arXiv:1803.10425`] [ɪɴSPIRE].

[19] J.P. Keating, N. Linden and H.J. Wells, *Spectra and eigenstates of spin chain hamiltonians*, *Commun. Math. Phys.* **338** (2015) 81 [`arXiv:1403.1121`].

[20] E. Lubkin, *Entropy of an n-system from its correlation with a k-reservoir*, *J. Math. Phys.* **19** (1978) 1028.

[21] S.K. Foong and S. Kanno, *Proof of Page's conjecture on the average entropy of a subsystem*, *Phys. Rev. Lett.* **72** (1994) 1148.

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