

Quantifying scrambling in quantum neural networks

Roy J. Garcia, Kaifeng Bu and Arthur Jaffe

*Department of Physics, Harvard University,
Cambridge, Massachusetts 02138, U.S.A.*

E-mail: roygarcia@g.harvard.edu, kfbu@fas.harvard.edu,
arthur_jaffe@harvard.edu

ABSTRACT: We quantify the role of scrambling in quantum machine learning. We characterize a quantum neural network's (QNNs) error in terms of the network's scrambling properties via the out-of-time-ordered correlator (OTOC). A network can be trained by minimizing a loss function. We show that the loss function can be bounded by the OTOC. We prove that the gradient of the loss function can be bounded by the gradient of the OTOC. This demonstrates that the OTOC landscape regulates the trainability of a QNN. We show numerically that this landscape is flat for maximally scrambling QNNs, which can pose a challenge to training. Our results pave the way for the exploration of quantum chaos in quantum neural networks.

KEYWORDS: Random Systems, Stochastic Processes

ARXIV EPRINT: [2112.01440](https://arxiv.org/abs/2112.01440)

Contents

1	Introduction	1
1.1	Background on scrambling	2
2	Preliminaries	2
3	Main results	4
3.1	Error bounds	5
3.2	Gradient of loss function	7
4	Numerical simulations	8
5	Discussion	9
A	Training with cost functions	10
B	Supplementary numerical simulations	11
C	Generalizing true error	11
D	Diagrammatic formalism	12
E	Properties of the twirling channel	12
F	Calculus identity	13
G	Proof of Proposition 1	14
G.1	Computing the loss function	14
G.2	Computing true error	16
H	Proof of Corollary 1	18
I	Proof of Proposition 2	20
J	Proof of Proposition 3	23
J.1	Lipschitz constant	23
J.2	Maximally scrambling unitaries	26
K	Proof of Theorem 1	27
K.1	Lipschitz constant	27
K.2	Maximally scrambling unitaries	30
K.3	Vanishing gradient	31
L	Proof of Lemma 1	32
M	Proof of Proposition 4	33

1 Introduction

A quantum neural network (QNN) [1–6] is a quantum generalization of a classical neural network [7–9] used to learn or optimize functions. QNNs are of growing interest because of their potential to provide a quantum speed-up [10] and they present a promising application for near-term intermediate scale quantum devices [11]. The role of classical chaos in classical neural networks has long been established [12–15], with more recent results linking chaos to expressivity [16].

As the body of literature on QNNs grows, one question remains: what role does quantum chaos play in QNNs? Bridging these currently disjoint fields is necessary to understand the capabilities of QNNs in learning the chaotic properties of many-body systems, as recently demonstrated in [17]. The connection between chaos and the generalization capability of QNNs has been explored in [18], and the relation between chaos and approximation properties of QNNs has been studied via the Loschmidt echo in [19]. There is a growing interest in quantifying the role of quantum chaos in QNNs and further investigation is needed to rigorously establish this connection.

In this work, we relate chaos to QNNs by establishing upper and lower bounds on training error in terms of quantum scrambling. Scrambling measures the delocalization of quantum information arising from chaotic evolution and is hence a measure of quantum chaos [20–26]. It was recently shown that QNNs encounter barren plateaus, exponentially vanishing gradients in the cost function, when learning scrambling unitaries [27]. Hence, scrambling plays an important role in training. However, QNNs themselves may also have chaotic properties which characterize their learning ability. These properties have been investigated through scrambling measures such as the tripartite mutual information [28] and operator size [29]. Numerical evidence correlating the tripartite mutual information to the network’s empirical training error has been demonstrated in [28]. One contribution of our work involves relating the tripartite mutual information to the network’s true error via an inequality. The scrambling ability of QNNs has also been numerically linked to the design of efficient network architectures [29]. However, much needed analytic relations between scrambling measures and QNN training error largely remain unestablished. Our contribution is to establish a number of inequalities relating the two.

We demonstrate that training error can be bounded by the out-of-time-ordered correlator (OTOC), defined in the following section. This correlator is an essential tool in the study of chaos, as it can characterize fast scramblers [30–34] and has even been used to decode the Hayden-Preskill protocol [35, 36]. In our context, we use the OTOC to quantify how well a QNN architecture scrambles information. We show that learning a unitary requires learning its scrambling properties.

Our main result, given in Theorem 1, shows that training is regulated by the gradient of the OTOC. In other words, trainability is regulated by the OTOC landscape. Hence, training depends on how a network’s scrambling ability changes as its training parameters are perturbed. We provide numerical simulations to support the relevance of our analytic bounds. We show that when the QNN is maximally scrambling, the OTOC landscape is flat, which can pose a challenge to training.

1.1 Background on scrambling

Here, we introduce the out-of-time-ordered correlator as a scrambling measure. A common definition of scrambling is the growth of the Hilbert-Schmidt norm of the commutator between two local Pauli strings that initially commute, as one operator evolves under the action of the unitary Heisenberg group $U(t)$ [37–40]. A Pauli string is the tensor product of local Pauli operators.

Let O_A and O_D be two commuting, local Pauli strings on systems A and D , respectively. Define $O_D(t) = U^\dagger(t)O_D U(t)$. The Hilbert-Schmidt norm of the commutator can be expressed as

$$\| [O_D(t), O_A] \|_{\text{HS}} = \sqrt{2d_{\text{tot}} (1 - \langle O_D(t)O_A O_D(t)O_A \rangle)}. \tag{1.1}$$

The expectation value $\langle \cdot \rangle$ is taken with respect to the N -qubit maximally mixed state $\frac{I}{d_{\text{tot}}}$, and $d_{\text{tot}} = 2^N$. The quantity $\langle O_D(t)O_A O_D(t)O_A \rangle$ in eq. (1.1) is an out-of-time-ordered correlator. Although troublesome to measure, protocols to do this have been constructed [41–45].

To simplify notation, we suppress the variable t and write U for $U(t)$. We remove the dependence on the choice of O_A and O_D by redefining the out-of-time-ordered correlator as the average

$$\overline{\text{OTOC}}(U) = \mathbb{E}_{O_A} \mathbb{E}_{O_D} \langle U O_A U^\dagger O_D U O_A U^\dagger O_D \rangle, \tag{1.2}$$

where (unless otherwise state) \mathbb{E}_{O_A} (or \mathbb{E}_{O_D}) denotes an average over the Pauli group on system A (or D). A signature of chaos is that the OTOC decays to a floor value and the Hilbert-Schmidt norm reaches a maximum at large time; see Corollary 1. We define U to be maximally scrambling if the OTOC decays to this floor value.

2 Preliminaries

In this section, we briefly review some basics of QNNs. A QNN is a parameterized quantum circuit with unitary $U(\boldsymbol{\theta})$ and parameters $\boldsymbol{\theta}$. There are two disjoint input subsystems A, B and two disjoint output subsystems C, D each with $N_{S'}$ qubits and Hilbert space dimension $d_{S'} = 2^{N_{S'}}$, where $S' \in \{A, B, C, D\}$ (see figure 1). A Pauli string acting on system S' is denoted as $O_{S'}$. N denotes the total number of qubits in the system, and $d_{\text{tot}} = 2^N$ is the corresponding Hilbert space dimension.

The parameters of the QNN are tuned to train the network to either learn a target unitary, U_S , or optimize a cost function. We focus on unitary learning in this section and refer to appendix A for a cost function treatment. We train the QNN with data $S = \{ \rho_d, y_d \}_{d=1}^{n_s}$, where d denotes each data point and n_s is the total number of points. State ρ_d encodes the input data and y_d is the corresponding target function. $U(\boldsymbol{\theta})$ approximates y_d by computing the output function \tilde{y}_d . We use the standard notation for the target function and output function:

$$y_d = \text{Tr} \{ U_S \rho_d U_S^\dagger O_C \}, \quad \tilde{y}_d = \text{Tr} \{ U \rho_d U^\dagger O_C \}. \tag{2.1}$$

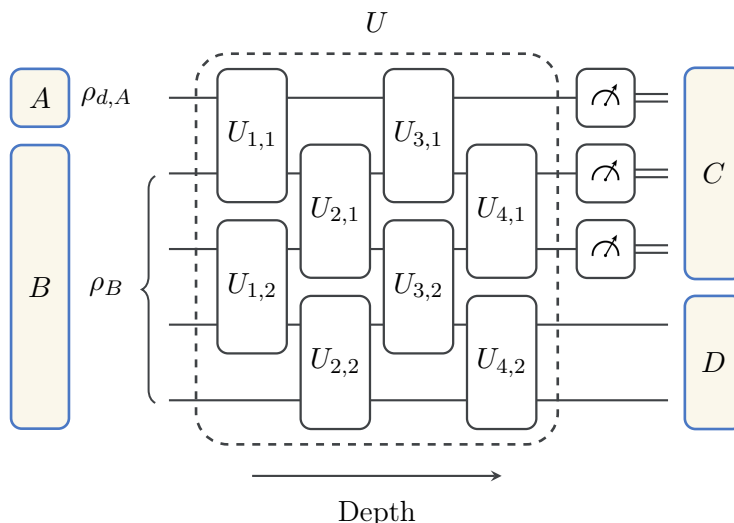


Figure 1. Quantum neural network with brick-wall architecture. Data is encoded in state $\rho_{d,A}$, while system B is initialized in the maximally mixed state, ρ_B . The QNN unitary U is composed of 2-qubit, parameterized unitaries, $U_{i,j}$. Output system C is measured after evolution by U .

The functions are expectation values of Pauli string O_C on system C with respect to the input state evolved with either U_S or U , respectively.

To determine the accuracy with which U approximates U_S , we define the loss function:

$$L_d = \mathbb{E}_{O_C} |\tilde{y}_d - y_d|^2. \quad (2.2)$$

Although it is common to measure the loss function with respect to one observable O_C , we take the average over all Pauli strings, as this will help establish a connection between L_d and the OTOC. The case where the loss function is defined with respect to one observable is considered in appendix C. The true error L is defined as the average of the loss function over the data set S :

$$L = \mathbb{E}_d L_d. \quad (2.3)$$

The empirical error is defined as the average of the loss function over a sample set, which is a finite subset of S . Practically, the QNN is trained by optimizing its parameters to minimize the empirical error. We assume L is approximated sufficiently well by the empirical error.

Assume ρ_d is local in system A , such that

$$\rho_d = \rho_{d,A} \otimes \rho_B. \quad (2.4)$$

The input data is encoded in pure state $\rho_{d,A} = |\psi_{d,A}\rangle \langle \psi_{d,A}| = U_A |0\rangle \langle 0|_A U_A^\dagger$ using unitary U_A , while system B is prepared in the maximally mixed state, $\rho_B = \frac{I_B}{d_B}$. Take U_A to be a Haar random unitary sampled from the unitary group on A . The true error becomes an average over all uniformly distributed input states, $\rho_{d,A}$: $L = \int_{\text{Haar}} dU_A L_d$. Since only a 2-design is required, the average over U_A can be taken over the Clifford group, which forms a 3-design [46, 47].

In this work, we quantify the role of quantum chaos in QNNs. Chaos in QNNs has been explored through the fidelity OTOC [19], which has the general form $|\langle \psi | U^\dagger M U | \psi \rangle|^2$ [22, 48, 49]. However, it was recently proposed that higher-point correlators can reveal the finer-grained dynamics of chaos [45, 50–52]. Since the fidelity OTOC carries the same information as the 2-point correlator $|\langle \psi | U^\dagger M U | \psi \rangle|$, it may not reveal the finer scrambling dynamics available to the 4-point OTOC in eq. (1.2). Therefore, we rely on $\overline{\text{OTOC}}(U)$ to study scrambling in QNNs.

3 Main results

Let us first introduce the connection between the training error and the out-of-time-ordered correlator. In the following proposition, we write both the loss function and true error in terms of out-of-time-ordered correlators.

Proposition 1 *The loss function can be written as*

$$L_d = d_D^2 (C_d(U, U) + C_d(U_S, U_S) - 2C_d(U, U_S)), \quad (3.1)$$

where $C_d(U_1, U_2)$ is

$$C_d(U_1, U_2) = \mathbb{E}_{O_D} \langle U_1 \rho_d U_1^\dagger O_D^\dagger U_2 \rho_d U_2^\dagger O_D \rangle. \quad (3.2)$$

It follows that the true error can be written in terms of OTOCs:

$$L = G \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\text{OP}(U, U_S) \right], \quad (3.3)$$

where $\text{OP}(U, U_S)$ and G are defined by

$$\text{OP}(U, U_S) = \mathbb{E}_{O_A} \mathbb{E}_{O_D} \langle U O_A U^\dagger O_D U_S O_A U_S^\dagger O_D \rangle, \quad (3.4)$$

$$G = \frac{d_A^2}{(d_A + 1)d_C^2}. \quad (3.5)$$

We give the proof of this proposition in appendix G. The expressions $C_d(U, U)$ and $C_d(U_S, U_S)$ have the form of out-of-time-ordered correlators, but they are sub-optimal scrambling measures since ρ_d is non-unitary. Hence, we will later rely on the trusted scrambling measure $\overline{\text{OTOC}}(U)$ to bound L_d in Proposition 3. The first two terms in eq. (3.3) depend only on the scrambling ability of U and U_S , respectively. The *optimization correlator*, $\text{OP}(U, U_S)$, reflects the optimization of U with respect to U_S when learning. In Proposition 2, we will establish bounds on L which are independent of $\text{OP}(U, U_S)$.

In the following corollary, we focus on the special case where the target unitary is maximally scrambling, as this is physically relevant when the QNN learns the large-time dynamics of a chaotic many-body system. As U_S becomes more scrambling, the true error approaches the value obtained by integrating U_S over the Haar measure on the unitary group [52], denoted as $L_{\text{scram}} = \int_{\text{Haar}} dU_S L$. When both U and U_S are maximally scrambling and independent of each other, the true error reaches a floor value of $L_{\text{floor}} = \int_{\text{Haar}} dU dU_S L$. We prove the following corollary in appendix H.

Corollary 1 *When U is untrained and hence independent of U_S , L_{scram} and L_{floor} satisfy the following:*

$$L_{\text{scram}} = G \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}_{\text{scram}} - \frac{2}{d_A^2} \right], \quad (3.6)$$

$$L_{\text{floor}} = 2G \left[\overline{\text{OTOC}}_{\text{scram}} - \frac{1}{d_A^2} \right]. \quad (3.7)$$

Here, the OTOC for a maximally scrambling unitary is

$$\overline{\text{OTOC}}_{\text{scram}} = \frac{1}{(d_{\text{tot}}^2 - 1)} \left[\frac{d_{\text{tot}}^2}{d_A^2} - 1 + d_C^2 \left(1 - \frac{1}{d_A^2} \right) \right]. \quad (3.8)$$

When d_{tot} is large and $d_A \gg 1$, $L_{\text{floor}} \rightarrow \frac{2}{d_B d_{\text{tot}}}$. For fixed N_B , the true error in this limit vanishes exponentially with the total number of qubits N . This is a relevant limit when learning many-body unitaries.

3.1 Error bounds

In this subsection, we bound the true error and the loss function using OTOCs.

Proposition 2 *The true error L can be bounded by OTOCs:*

$$L_- \leq L \leq L_+, \quad (3.9)$$

where

$$L_{\pm}(U, U_S) = G \left[\sqrt{\overline{\text{OTOC}}(U)} \pm \sqrt{\overline{\text{OTOC}}(U_S)} \right]^2. \quad (3.10)$$

Also

$$|L - L_{\pm}| \leq 4G \sqrt{\overline{\text{OTOC}}(U) \overline{\text{OTOC}}(U_S)}. \quad (3.11)$$

We prove Proposition 2 in appendix I. $\overline{\text{OTOC}}(U)$ decays as U becomes more scrambling. Hence, the upper bound L_+ decays as U or U_S become more scrambling. The lower bound L_- depends on the distance between the OTOCs of U and U_S . This implies that L_- vanishes as the QNN learns the scrambling properties of the target unitary. A mismatch in the scrambling abilities of U and U_S may therefore inhibit the optimization of L . Hence, learning a target unitary requires learning its scrambling properties. The bound in ineq. (3.11) decays with scrambling, causing L to approach L_{\pm} .

At the start of training, the QNN is initialized as a random parameterized quantum circuit (RPQC) with unitary U_0 . Sufficiently deep random circuits are scrambling [53, 54] and lead to OTOC decay with circuit depth [55]. Therefore, the initialized QNN becomes increasingly scrambling with circuit depth, causing $\overline{\text{OTOC}}(U_0)$ and $L_+(U_0, U_S)$ to decay. Random local quantum circuits form approximate polynomial-designs [56, 57]. Hence, when the initialized QNN becomes sufficiently deep, $\overline{\text{OTOC}}(U_0)$ tends to $\overline{\text{OTOC}}_{\text{scram}}$ in eq. (3.8), the value found by integrating U_0 over the Haar measure on the unitary group [52]; see appendix H. The corresponding true error bounds are found by setting $\overline{\text{OTOC}}(U_0) \rightarrow \overline{\text{OTOC}}_{\text{scram}}$ in eq. (3.10).

Aside from the OTOC, the second Rényi entropy and the tripartite mutual information are also common measures for scrambling. We define the Choi isomorphism of U as $\rho(U) = |U\rangle\langle U|$, where $|U\rangle = \frac{1}{\sqrt{d_{\text{tot}}}} \sum_{i=1}^{d_{\text{tot}}} |i\rangle \otimes U|i\rangle$. For a system S' and its complement S'_c in system $ABCD$, define the reduced density operator $\rho_{S'}(U) = \text{Tr}_{S'_c}\{\rho(U)\}$. For example, $\rho_{AC}(U) = \text{Tr}_{BD}\{\rho(U)\}$. The second Rényi entropy is defined as $S_{AC}^{(2)}(U) = -\log_2\{\rho_{AC}^2(U)\}$. The tripartite mutual information is defined as $I_{3,U}(A:C:D) = I_U(A:C) + I_U(A:D) - I_U(A:CD)$, where the mutual information is, for example, $I_U(A:C) = S_A(U) + S_C(U) - S_{AC}(U)$ and the entanglement entropy is $S_{AC}(U) = -\text{Tr}\{\rho_{AC}(U) \log_2 \rho_{AC}(U)\}$.

The negativity of $I_{3,U}(A:C:D)$ measures how much information in A is shared by C and D after evolution by U . As the system scrambles, the magnitude of the tripartite mutual information increases. A numerical correlation between the empirical error and the tripartite mutual information was found in [28]. In the following corollary, we relate the true error to the second Rényi entropy and the tripartite mutual information.

Corollary 2 *The true error bounds L_{\pm} can be expressed in terms of the second Rényi entropy $S_{AC}^{(2)}(U_i)$ with $U_i \in \{U, U_S\}$ as follows,*

$$L_{\pm} = 2^{N-N_A-N_D} G \left[2^{-S_{AC}^{(2)}(U)/2} \pm 2^{-S_{AC}^{(2)}(U_S)/2} \right]^2. \quad (3.12)$$

Moreover, the true error can be bounded by the tripartite mutual information $I_{3,U_i}(A:C:D)$:

$$L \geq G \left[2^{(I_{3,U}(A:C:D)-2N_A)/2} + 2^{(I_{3,U_S}(A:C:D)-2N_A)/2} - 2\text{OP}(U, U_S) \right]. \quad (3.13)$$

The OTOC for U is related to the second Rényi entropy, $S_{AC}^{(2)}(U)$, through $\overline{\text{OTOC}}(U) = 2^{N-N_A-N_D-S_{AC}^{(2)}(U)}$ [36, 58]. We use this relation along with eq. (3.10) to prove eq. (3.12). It has been shown that $\overline{\text{OTOC}}(U) \geq 2^{(I_{3,U}(A:C:D)-2N_A)/2}$ [58], which implies that the tripartite mutual information lower bounds the true error via eq. (3.3). This yields ineq. (3.13).

In Proposition 3, we establish a bound on the loss function in terms of $\overline{\text{OTOC}}(U)$. In the proof (see appendix J), we use Levy's lemma to construct a concentration inequality for the loss function. When working with Levy's lemma, it is useful to define the following function

$$f(\epsilon) = \sqrt{\frac{9\pi^3}{2d_A} \ln \frac{2}{\epsilon}}. \quad (3.14)$$

Proposition 3 *Let $|\psi_{d,A}\rangle$ be sampled from the Haar measure on the Hilbert space of system A . With probability at least $1 - \epsilon$, the loss function $L_d(|\psi_{d,A}\rangle)$ satisfies*

$$|L_d - L_{\pm}| \leq \eta f(\epsilon) + 4G \sqrt{\overline{\text{OTOC}}(U) \overline{\text{OTOC}}(U_S)}, \quad (3.15)$$

and

$$|L_d - L| \leq \eta f(\epsilon), \quad (3.16)$$

where the Lipschitz constant η is

$$\eta = 8\sqrt{d_A(d_A + 1)L_+}. \quad (3.17)$$

The Lipschitz constant decays as U or U_S become more scrambling, since L_+ is a function of OTOCs, which decay. This decay concentrates L_d about L_{\pm} and L via ineqs. (3.15) and (3.16), respectively. Concentration occurs as a RPQC QNN becomes deeper. In the case where U and U_S are maximally scrambling, it is shown in appendix J that for large d_{tot} and $d_D \gg d_A$, the right-hand side of ineq. (3.16) becomes $\eta f(\epsilon) \rightarrow \frac{16}{d_C} \sqrt{\frac{9\pi^3}{2} \ln \frac{2}{\epsilon}}$. This implies that performing measurements on a larger system C will concentrate L_d about L . Also assuming $d_A \gg 1$, $L \rightarrow \frac{2}{d_B d_{\text{tot}}}$ by eq. (3.7). This decays exponentially with N , so L_d can be made to concentrate near 0. The loss function when learning many-body unitaries can therefore vanish.

3.2 Gradient of loss function

We now introduce our principal result regarding training. We bound the gradient of the loss function with the gradient of the OTOC. Doing so helps in performing gradient descent [59, 60], one of the widely used methods for training neural networks. Let us consider a QNN as in [61] given by $U(\boldsymbol{\theta}) = \prod_{l=1}^L U_l(\theta_l) W_l$ where $U_l(\theta_l) = e^{-i\theta_l V_l}$, V_l is Hermitian, W_l is a constant unitary, and L is the circuit depth. An application of Levy's lemma yields the following theorem; see appendix K.

Theorem 1 *Let $|\psi_{d,A}\rangle$ be sampled from the Haar measure on the Hilbert space of system A . With probability at least $1 - \epsilon$, $\partial_{\theta_l} L_d(|\psi_{d,A}\rangle)$ satisfies*

$$|\partial_{\theta_l} L_d - \partial_{\theta_l} L| \leq \eta_g f(\epsilon), \quad (3.18)$$

where

$$\partial_{\theta_l} L = G \left[\partial_{\theta_l} \overline{\text{OTOC}}(U(\boldsymbol{\theta})) - 2\partial_{\theta_l} \text{OP}(U(\boldsymbol{\theta}), U_S) \right], \quad (3.19)$$

and the Lipschitz constant η_g is

$$\eta_g = 8 \|V_l\|_{\infty} \left(\sqrt{d_A(d_A + 1)L_+} + 2 \right). \quad (3.20)$$

For a maximally scrambling target unitary and an untrained U ,

$$\partial_{\theta_l} L_{\text{scram}} = G \partial_{\theta_l} \overline{\text{OTOC}}(U(\boldsymbol{\theta})). \quad (3.21)$$

The quantity $\partial_{\theta_l} L_d$ is probabilistically bounded by the OTOC and its gradient via ineq. (3.18). It is shown in appendix K that

$$|\partial_{\theta_l} L| \leq \frac{8d_A^2}{(d_A + 1)d_C^2} \|V_l\|_{\infty}. \quad (3.22)$$

This bound decays exponentially with N_C , indicating that non-local measurements can produce vanishing gradients. This is consistent with the presence of exponentially vanishing gradients of globally defined functions found in [62, 63].

The Lipschitz constant η_g decays as U or U_S become more scrambling. When U and U_S are maximally scrambling unitaries, ineq. (3.18) simplifies to

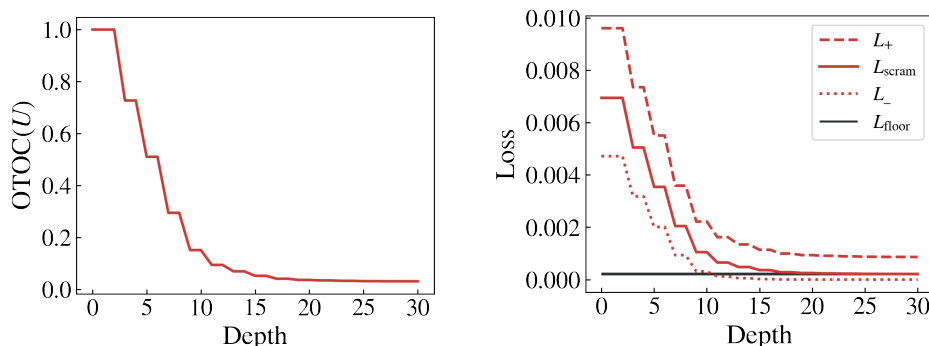


Figure 2. (Left) Plot of $\overline{\text{OTOC}}(U)$ against circuit depth at initialization for $N = 8$, $N_A = N_D = 3$. Systems A and D are disjoint (see figure 1 for example), so the commutator between any O_A and O_D vanishes, yielding $\overline{\text{OTOC}}(U) = 1$. The support of UO_AU^\dagger reaches D after three QNN layers, causing the OTOC to decay. The system becomes maximally scrambled once the OTOC reaches its floor value. (Right) The corresponding L_{scram} for a maximally scrambling target unitary is plotted along with L_+ , L_- , and L_{floor} .

$|\partial_{\theta_i} L_d - \partial_{\theta_i} L| \leq 16 \|V_i\|_\infty \left(\frac{1}{d_C} + \frac{1}{\sqrt{d_A}} \right) \sqrt{\frac{9\pi^3}{2} \ln \frac{2}{\epsilon}}$. This bound tightens as d_A and d_C increase, causing $\partial_{\theta_i} L_d$ to concentrate near $\partial_{\theta_i} L$. This corresponds to encoding the data using more qubits and performing measurements on a larger subsystem C .

Eq. (3.21) shows that the true error gradient is given by the gradient of the OTOC. The OTOC landscape therefore regulates trainability when implementing gradient descent. The OTOC reaches a floor value for a maximally scrambling QNN and no further chaotic evolution by U changes the OTOC appreciably. One therefore expects that an infinitesimal change in θ should not perturb the OTOC, and hence L_{scram} , from its fixed value. We present numerical simulations in section 4 to demonstrate this. This hints that barren plateaus, exponentially vanishing gradients [27, 61, 64], may potentially arise due to the flat OTOC landscape of maximally scrambling QNNs. This also suggests that weakly scrambling QNNs with θ -sensitive OTOCs may hold potential in avoiding barren plateaus. Indeed, shallow QNN architectures have been shown to circumvent barren plateaus [62, 65]; these architectures are typically weak scramblers.

4 Numerical simulations

For concreteness, we adopt the brick-wall network architecture (see figure 1) to numerically simulate the decay of the true error with circuit depth. At layer i in the QNN, a 2-qubit unitary $U_{i,j}$ is applied to each pair of neighboring qubits. Index j labels the unitaries in a given layer. Odd layers are staggered by one qubit with respect to even layers, forming a brick-wall geometry. Each unitary takes on the form $U_{i,j} = e^{-i\theta_{i,j}V_{i,j}}$, where $V_{i,j}$ is Hermitian and $\theta_{i,j}$ is a training parameter.

We numerically simulate the true error for a QNN at initialization when the target unitary is maximally scrambling. We choose each parameter $\theta_{i,j}$ randomly such that we can assume $U_{i,j}$ is effectively a Haar random unitary. L_{scram} is plotted in figure 2 (right) along with L_+ , L_- , and L_{floor} as a function of circuit depth. $\overline{\text{OTOC}}(U)$, and hence L_{scram} ,

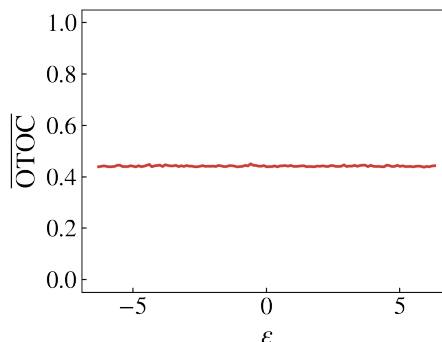


Figure 3. OTOC landscape with respect to perturbation parameter $\varepsilon \in [-2\pi, 2\pi]$. Each training parameter $\theta_{i,j}$ is randomly initialized. The OTOC is computed using updated parameters $\theta_{i,j} \rightarrow \theta_{i,j} + \varepsilon$. The QNN has a circuit depth of 30 layers and $N = 8, N_A = 1, N_C = N - 1$, ensuring the QNN is in the maximally scrambling regime.

decays with circuit depth. The bounds L_+ and L_- tighten with circuit depth; this is consistent with ineq. (3.11). L_{scram} decays to L_{floor} at sufficiently large depth, since the QNN becomes maximally scrambling.

In figure 3, we numerically simulate the OTOC landscape generated by varying each training parameter $\theta_{i,j}$ using perturbation parameter $\varepsilon \in [-2\pi, 2\pi]$. Each $\theta_{i,j}$ is randomly initialized. A circuit depth of 30 layers is used with $N = 8, N_A = 1, N_C = N - 1$. This choice of circuit parameters ensures that, at initialization, the QNN is maximally scrambling and the OTOC has attained its floor value. The flat landscape indicates that even after perturbing the parameters, the QNN remains maximally scrambling and the OTOC retains its floor value. From Theorem 1, we see that $\partial_{\theta_i} L_{\text{scram}} = G \partial_{\theta_i} \overline{\text{OTOC}}(U(\boldsymbol{\theta}))$. That is, the OTOC landscape determines the L_{scram} landscape. Hence, the flat OTOC landscapes arising from maximally scrambling QNNs may pose challenges to training via gradient descent algorithms.

5 Discussion

We have shown that training error is bounded by the OTOC, a scrambling measure. Our results demonstrate that learning a unitary necessitates learning its scrambling properties. Training of the QNN via gradient descent requires computing the gradient of a loss function. We establish an inequality relating this gradient to the gradient of the OTOC. As a result, training is regulated by the OTOC landscape. We show that maximally scrambling QNNs can produce flat OTOC landscapes, which may present a roadblock to training. An open question, which we pose, is to prove whether weakly scrambling QNN architectures can remove barren plateaus in the training landscape.

Acknowledgments

We thank Jordan Cotler for discussions about Levy’s lemma. We thank Liyuan Chen for discussions about Rényi entropy. This work was supported in part by the ARO Grant W911NF-19-1-0302 and the ARO MURI Grant W911NF-20-1-0082.

A Training with cost functions

In the main text, we considered the training of QNNs by examining the loss function and its gradient. This method is useful when training with a data set. However, QNNs can also be trained to optimize a problem-specific cost function, which does not involve a data set. Let $\rho_{d'}$ be a fixed, random state of the form eq. (2.4). We define the cost function as

$$\mathcal{C} = \mathbb{E}_{\mathcal{C}} \tilde{y}_{d'}^2, \quad (\text{A.1})$$

where $\tilde{y}_{d'}$ has the form of eq. (2.1). Eq. (A.1) does not include a target function, since the QNN is not learning a target unitary. It is shown in appendix L that the cost function and its gradient can be written in terms of the out-of-time-ordered correlator $C_{d'}(U, U)$ defined in eq. (3.2):

Lemma 1 *The cost function and its gradient can be expressed in terms of an out-of-time-ordered correlator, $C_{d'}(U, U)$, as follows*

$$\mathcal{C} = d_D^2 C_{d'}(U, U), \quad (\text{A.2})$$

$$\partial_{\theta_i} \mathcal{C} = d_D^2 \partial_{\theta_i} C_{d'}(U, U). \quad (\text{A.3})$$

To understand how the network's scrambling properties affect training via gradient descent algorithms, we bound the cost function and its gradient using the OTOC from eq. (1.2). We consider a QNN architecture as in section 3.2. The following proposition bounds \mathcal{C} and its gradient when $\rho_{d', A}$ is a random state (see appendix M for a proof).

Proposition 4 *Let $|\psi_{d', A}\rangle$ be sampled from the Haar measure on the Hilbert space of system A . With probability at least $1 - \epsilon$, $\mathcal{C}(|\psi_{d', A}\rangle)$ satisfies*

$$|\mathcal{C} - \mathcal{C}_{\text{av}}| \leq \eta_{\mathcal{C}} f(\epsilon), \quad (\text{A.4})$$

where the Lipschitz constant $\eta_{\mathcal{C}}$ is

$$\eta_{\mathcal{C}} = \frac{4\sqrt{d_A}d_A}{d_C} \sqrt{\overline{\text{OTOC}}(U)}, \quad (\text{A.5})$$

and the average cost function is

$$\mathcal{C}_{\text{av}} = \int_{\text{Haar}} dU_A \mathcal{C} = G \left[\frac{1}{d_A} + \overline{\text{OTOC}}(U) \right]. \quad (\text{A.6})$$

With probability at least $1 - \epsilon$, $\partial_{\theta_i} \mathcal{C}(|\psi_{d', A}\rangle)$ satisfies

$$|\partial_{\theta_i} \mathcal{C} - \partial_{\theta_i} \mathcal{C}_{\text{av}}| \leq \eta_{\mathcal{C}, g} f(\epsilon), \quad (\text{A.7})$$

where the Lipschitz constant $\eta_{\mathcal{C}, g}$ is

$$\eta_{\mathcal{C}, g} = 8 \|V_l\|_{\infty} \left(\frac{\sqrt{d_A}d_A}{d_C} \sqrt{\overline{\text{OTOC}}(U)} + 1 \right), \quad (\text{A.8})$$

and the gradient of \mathcal{C}_{av} is

$$\partial_{\theta_i} \mathcal{C}_{\text{av}} = G \partial_{\theta_i} \overline{\text{OTOC}}(U). \quad (\text{A.9})$$

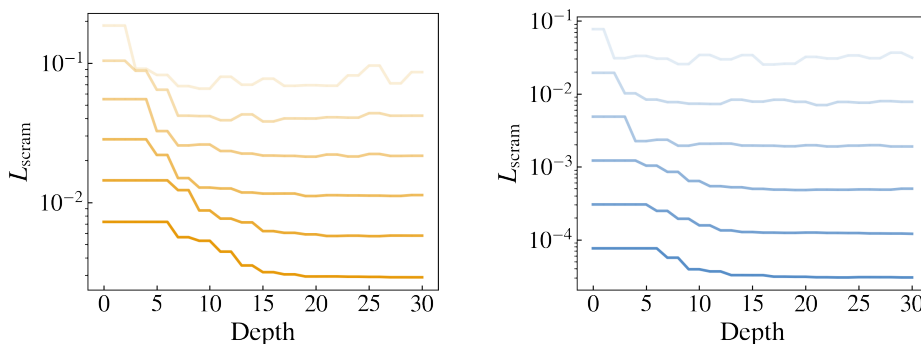


Figure 4. Plots of L_{scram} for a randomly initialized QNN of varying depth for a maximally scrambling target unitary. L_{scram} is plotted for $N = 3, 4, \dots, 8$ (darker colors indicate increasing N) for subsystem sizes: (left) $N_A = N - 1, N_C = 1$ and (right) $N_A = 1, N_C = N - 1$.

\mathcal{C} and $\partial_{\theta_i} \mathcal{C}$ are probabilistically bounded by $\overline{\text{OTOC}}(U)$. Importantly, ineq. (A.7) implies that $\partial_{\theta_i} \mathcal{C}$ can be made to concentrate about $\partial_{\theta_i} \mathcal{C}_{\text{av}}$, which depends only on the gradient of the OTOC, as shown in eq. (A.9). The OTOC landscape therefore regulates the network’s trainability when optimizing the cost function. Eqs. (A.3) and (A.9) suggest that maximally scrambling QNNs with flat out-of-time-ordered correlator landscapes may potentially account for barren plateaus.

B Supplementary numerical simulations

Figure 4 shows the decay of L_{scram} with respect to circuit depth for various N . In the left plot, the input data is encoded in a large state supported on $N_A = N - 1$ qubits and a single-qubit measurement on C is performed. In the right plot, the data is encoded in a single-qubit state and a measurement on $N_C = N - 1$ qubits is performed. For a given N , the right plot produces a smaller true error. This is consistent with the behavior of L_{floor} , which decreases as N_A decreases (i.e. as N_B increases) for a given N . Both plots indicate L_{scram} decays with circuit depth and width.

C Generalizing true error

Although L_d in eq. (2.2) is defined with respect to an average of O_C over \mathcal{P}_C , the Pauli group on C , variants of the true error may be defined with respect to a uniform average over a subset $S_C \subseteq \mathcal{P}_C$ with cardinality $|S_C|$. Some useful variants include:

$$\begin{aligned}
 L_{V1} &= \mathbb{E}_d \mathbb{E}_{O_C \in S_C} |\tilde{y}_d - y_d|^2, \\
 L_{V2} &= \mathbb{E}_d \left| \mathbb{E}_{O_C \in S_C} (\tilde{y}_d - y_d) \right|^2, \\
 L_{V3} &= \mathbb{E}_d \left| \mathbb{E}_{O_C \in S_C} (\tilde{y}_d - y_d) \right|,
 \end{aligned}
 \tag{C.1}$$

where \mathbb{E}_d implies the Haar average over U_A . Taking S_C to consist of a single Pauli observable gives a common definition of the loss function. It can be shown that

$$L_{V3}^2 \leq L_{V2} \leq L_{V1} \leq \frac{d_C^2}{|S_C|} L \leq \frac{d_C^2}{|S_C|} L_+. \quad (\text{C.2})$$

L_+ bounds a function of each variant. Hence, scrambling bounds a larger class of true error definitions. Variants of the cost function in eq. (A.1) can be defined in a similar fashion.

To prove ineq. (C.2), first recall $L \leq L_+$, as established in the main text. Also note that $\sum_{O_C \in S_C} |\tilde{y}_d - y_d|^2 \leq \sum_{O_C \in \mathcal{P}_C} |\tilde{y}_d - y_d|^2$. We can write

$$\mathbb{E}_{O_C \in S_C} |\tilde{y}_d - y_d|^2 = \frac{1}{|S_C|} \sum_{O_C \in S_C} |\tilde{y}_d - y_d|^2 \leq \frac{1}{|S_C|} \frac{d_C^2}{d_C^2} \sum_{O_C \in \mathcal{P}_C} |\tilde{y}_d - y_d|^2 = \frac{d_C^2}{|S_C|} \mathbb{E}_{O_C \in \mathcal{P}_C} |\tilde{y}_d - y_d|^2. \quad (\text{C.3})$$

This implies $L_{V1} \leq \frac{d_C^2}{|S_C|} L$. Since the variance is non-negative, $(\mathbb{E}_{O_C \in S_C} (\cdot))^2 \leq \mathbb{E}_{O_C \in S_C} ((\cdot)^2)$. This implies $L_{V2} \leq L_{V1}$. We also have $(\mathbb{E}_d (\cdot))^2 \leq \mathbb{E}_d ((\cdot)^2)$, which implies $L_{V3}^2 \leq L_{V2}$.

D Diagrammatic formalism

We can diagrammatically express unitary $U_i \in \{U, U_S\}$ as

$$U_i = \begin{array}{c} C \text{ --- } \boxed{U_i} \text{ --- } A \\ D \text{ --- } \boxed{U_i} \text{ --- } B \end{array}, \quad (\text{D.1})$$

where we label the input (A, B) and output (C, D) subsystems. Define the Bell state between systems S_1 and S_2 , each of dimension $d_{S'}$, as

$$|\text{Bell}\rangle_{S_1, S_2} = \frac{1}{\sqrt{d_{S'}}} \sum_{i=1}^{d_{S'}} |i\rangle_{S_1} \otimes |i\rangle_{S_2} = \frac{1}{\sqrt{d_{S'}}} \begin{array}{|} \hline S_1 \\ \hline S_2 \\ \hline \end{array}. \quad (\text{D.2})$$

The corresponding bra vector diagram is

$$\langle \text{Bell} |_{S_1, S_2} = \frac{1}{\sqrt{d_{S'}}} \begin{array}{|} \hline S_1 \\ \hline S_2 \\ \hline \end{array}. \quad (\text{D.3})$$

An identity relating an operator Q to its transpose is

$$\begin{array}{|} \hline \boxed{Q} \\ \hline \end{array} = \begin{array}{|} \hline \boxed{Q^T} \\ \hline \end{array}. \quad (\text{D.4})$$

E Properties of the twirling channel

Define the k -fold twirling channel for an operator Q on the k -copy Hilbert space as

$$\Phi_{\text{Haar}(d')}^{(k)}(Q) = \int_{\text{Haar}} dU U^\dagger \otimes^k Q U^{\otimes k}, \quad (\text{E.1})$$

where unitary U is sampled from Haar measure on the unitary group of dimension d' , $\mathcal{U}(d')$. This can be expanded in terms of permutation operators T_π , where $\pi \in \mathcal{S}_k$ and \mathcal{S}_k is the set of all permutations of the set $\{1, 2, \dots, k\}$:

$$\Phi_{\text{Haar}(d')}^{(k)}(Q) = \sum_{\pi, \sigma \in \mathcal{S}_k} C_{\pi, \sigma} T_\pi \text{Tr} \{T_\sigma Q\}. \quad (\text{E.2})$$

The coefficients $C_{\pi, \sigma}$ form the Weingarten matrix. In the case of $k = 1$,

$$\Phi_{\text{Haar}(d')}^{(1)}(Q) = \frac{1}{d'} I \text{Tr} \{Q\}. \quad (\text{E.3})$$

The case of $k = 2$ gives

$$\begin{aligned} \Phi_{\text{Haar}(d')}^{(2)}(Q) = \frac{1}{d'^2 - 1} & \left[I_{1,2} \text{Tr} \{I_{1,2} Q\} + \mathbb{S}_{1,2} \text{Tr} \{\mathbb{S}_{1,2} Q\} \right. \\ & \left. - \frac{1}{d'} I_{1,2} \text{Tr} \{\mathbb{S}_{1,2} Q\} - \frac{1}{d'} \mathbb{S}_{1,2} \text{Tr} \{I_{1,2} Q\} \right], \end{aligned} \quad (\text{E.4})$$

where $\mathbb{S}_{1,2}$ is the swap operator between the two Hilbert space copies. For a pure state $|\psi\rangle$, the following identity is useful

$$\Phi_{\text{Haar}(d')}^{(k)}(|\psi\rangle\langle\psi|^{\otimes k}) = \frac{\sum_{\pi \in \mathcal{S}_k} T_\pi}{d'(d'+1)\dots(d'+k-1)}. \quad (\text{E.5})$$

Permutation operators are invariant under the twirling channel:

$$\Phi_{\text{Haar}(d')}^{(k)}(T_\pi) = T_\pi. \quad (\text{E.6})$$

A k -design is a finite ensemble \mathcal{E} of unitaries which can replicate the first k moments of the Haar measure:

$$\mathbb{E}_{U \sim \mathcal{E}} [U^{\dagger \otimes k} Q U^{\otimes k}] = \int_{\text{Haar}} dU U^{\dagger \otimes k} Q U^{\otimes k}. \quad (\text{E.7})$$

The Pauli group forms a 1-design and the Clifford group forms a 3-design.

F Calculus identity

We derive a useful calculus identity. Define state $|\psi\rangle \in \mathbb{C}^{d_A}$ and real vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{d_A}$. Any state can be written as

$$|\psi\rangle = \mathbf{v}_1 + i\mathbf{v}_2. \quad (\text{F.1})$$

Define I_A as the $d_A \times d_A$ -dimensional identity matrix. We can write

$$\begin{aligned} |\psi\rangle &= \begin{bmatrix} I_A & iI_A \end{bmatrix} \mathbf{w} \\ \mathbf{w} &= \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}. \end{aligned} \quad (\text{F.2})$$

For a Hermitian matrix Q , we compute the following derivative:

$$\begin{aligned}
 \left\| \frac{d}{d\mathbf{w}} \langle \psi | Q | \psi \rangle \right\|_2 &= \left\| \frac{d}{d\mathbf{w}} \left(\mathbf{w}^T \begin{bmatrix} I_A \\ -iI_A \end{bmatrix} Q \begin{bmatrix} I_A & iI_A \end{bmatrix} \mathbf{w} \right) \right\|_2 \\
 &= \left\| \mathbf{w}^T \left(\begin{bmatrix} I_A \\ -iI_A \end{bmatrix} Q \begin{bmatrix} I_A & iI_A \end{bmatrix} + \begin{bmatrix} I_A \\ iI_A \end{bmatrix} Q^T \begin{bmatrix} I_A & -iI_A \end{bmatrix} \right) \right\|_2 \\
 &= \left\| \langle \psi | Q \begin{bmatrix} I_A & iI_A \end{bmatrix} + \langle \psi |^* Q^T \begin{bmatrix} I_A & -iI_A \end{bmatrix} \right\|_2 \\
 &= \left[\left(\langle \psi | Q \begin{bmatrix} I_A & iI_A \end{bmatrix} + \langle \psi |^* Q^T \begin{bmatrix} I_A & -iI_A \end{bmatrix} \right) \right. \\
 &\quad \left. \cdot \left(\begin{bmatrix} I_A \\ -iI_A \end{bmatrix} Q^\dagger | \psi \rangle + \begin{bmatrix} I_A \\ iI_A \end{bmatrix} Q^* | \psi \rangle^* \right) \right]^{1/2} \\
 &= \sqrt{\langle \psi | Q \begin{bmatrix} I_A & iI_A \end{bmatrix} \begin{bmatrix} I_A \\ -iI_A \end{bmatrix} Q^\dagger | \psi \rangle + \langle \psi |^* Q^T \begin{bmatrix} I_A & -iI_A \end{bmatrix} \begin{bmatrix} I_A \\ iI_A \end{bmatrix} Q^* | \psi \rangle^*} \\
 &= \sqrt{\langle \psi | Q(2I_A)Q^\dagger | \psi \rangle + \langle \psi |^* Q^T(2I_A)Q^* | \psi \rangle^*} \\
 &= \sqrt{2 \langle \psi | Q Q^\dagger | \psi \rangle + 2(\langle \psi |^* Q^T Q^* | \psi \rangle^*)^*} \\
 &= \sqrt{2 \langle \psi | Q^\dagger Q | \psi \rangle + 2 \langle \psi | Q^\dagger Q | \psi \rangle} \\
 &= 2 \|Q | \psi \rangle\|_2. \tag{F.3}
 \end{aligned}$$

The second line follows from a standard calculus identity.

G Proof of Proposition 1

G.1 Computing the loss function

We will show that the loss function can be written as

$$L_d = d_D^2 (C_d(U, U) + C_d(U_S, U_S) - 2C_d(U, U_S)), \tag{G.1}$$

where

$$C_d(U_1, U_2) = \mathbb{E}_{O_D} \langle U_1 \rho_d U_1^\dagger O_D^\dagger U_2 \rho_d U_2^\dagger O_D \rangle. \tag{G.2}$$

Let Q be an arbitrary operator on the d_{tot} -dimensional Hilbert space. We can define the more general correlator $C(Q, U_1, U_2)$ as

$$C(Q, U_1, U_2) = \mathbb{E}_{O_D} \langle U_1 Q U_1^\dagger O_D^\dagger U_2 Q U_2^\dagger O_D \rangle. \tag{G.3}$$

Note that $C_d(U_1, U_2) = C(\rho_d, U_1, U_2)$. We express $C(Q, U_1, U_2)$ diagrammatically:

$$C(Q, U_1, U_2) = \frac{1}{d_{\text{tot}} O_D} \mathbb{E} \left[\begin{array}{c} \text{Diagram: A sequence of operators } U_1, Q, U_1^\dagger, O_D^\dagger, U_2, Q, U_2^\dagger, O_D \text{ connected in a loop.} \end{array} \right]. \tag{G.4}$$

Using the transpose identity from eq. (D.4),

$$C(Q, U_1, U_2) = \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_D} \cdot \quad (\text{G.5})$$

Perform the average of O_D over the Pauli group (which forms a 1-design) using the identity from eq. (E.3): $\mathbb{E}_{O_D} O_D^\dagger Q' O_D = \frac{1}{d_D} I_D \text{Tr}_D\{Q'\}$ where $Q' = U_2 Q U_2^\dagger$. This identity produces

$$C(Q, U_1, U_2) = \frac{1}{d_{\text{tot}} d_D} \cdot \quad (\text{G.6})$$

Introduce a 1-design on system C by using the identity $\mathbb{E}_{O_C} O_C \otimes O_C^* = |\text{Bell}\rangle \langle \text{Bell}|_{C,C'}$,

$$C(Q, U_1, U_2) = \frac{d_C}{d_{\text{tot}} d_D} \mathbb{E}_{O_C} \cdot \quad (\text{G.7})$$

O_C is a Pauli string. Using $d_{\text{tot}} = d_C d_D$, we write this as

$$C(Q, U_1, U_2) = \frac{1}{d_D^2} \mathbb{E}_{O_C} \text{Tr} \{ U_1^* Q^T U_1^T O_C^* \} \text{Tr} \{ U_2 Q U_2^\dagger O_C \}. \quad (\text{G.8})$$

The transpose leaves the trace invariant, so

$$\mathrm{Tr} \left\{ U_1^* Q^T U_1^T O_C^* \right\} = \mathrm{Tr} \left\{ (U_1^* Q^T U_1^T O_C^*)^T \right\} = \mathrm{Tr} \left\{ O_C^\dagger U_1 Q U_1^\dagger \right\} = \mathrm{Tr} \left\{ U_1 Q U_1^\dagger O_C \right\}, \quad (\text{G.9})$$

where we used the Hermiticity of O_C . This produces

$$d_D^2 C(Q, U_1, U_2) = \mathbb{E}_{O_C} \mathrm{Tr} \left\{ U_1 Q U_1^\dagger O_C \right\} \mathrm{Tr} \left\{ U_2 Q U_2^\dagger O_C \right\}. \quad (\text{G.10})$$

Setting $Q = \rho_d$, we obtain an identity for $C_d(U_1, U_2)$,

$$d_D^2 C_d(U_1, U_2) = \mathbb{E}_{O_C} \mathrm{Tr} \left\{ U_1 \rho_d U_1^\dagger O_C \right\} \mathrm{Tr} \left\{ U_2 \rho_d U_2^\dagger O_C \right\}. \quad (\text{G.11})$$

L_d can be written as

$$L_d = \mathbb{E}_{O_C} \tilde{y}_d^2 + \mathbb{E}_{O_C} y_d^2 - 2 \mathbb{E}_{O_C} y_d \tilde{y}_d. \quad (\text{G.12})$$

From eq. (G.11), we readily see that

$$\begin{aligned} \mathbb{E}_{O_C} \tilde{y}_d^2 &= d_D^2 C_d(U, U), \\ \mathbb{E}_{O_C} y_d^2 &= d_D^2 C_d(U_S, U_S), \\ \mathbb{E}_{O_C} y_d \tilde{y}_d &= d_D^2 C_d(U, U_S). \end{aligned} \quad (\text{G.13})$$

This produces

$$L_d = d_D^2 (C_d(U, U) + C_d(U_S, U_S) - 2C_d(U, U_S)). \quad (\text{G.14})$$

G.2 Computing true error

We will prove that

$$L = \frac{d_A^2}{(d_A + 1)d_C^2} \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\text{OP}(U, U_S) \right]. \quad (\text{G.15})$$

Begin by writing L using eq. (3.1):

$$\begin{aligned} L &= \int_{\text{Haar}} dU_A L_d \\ &= d_D^2 \int_{\text{Haar}} dU_A (C_d(U, U) + C_d(U_S, U_S) - 2C_d(U, U_S)). \end{aligned} \quad (\text{G.16})$$

For unitaries $U_1, U_2 \in \{U, U_S\}$, we compute the following average

$$\begin{aligned} &\int_{\text{Haar}} dU_A C_d(U_1, U_2) \\ &= \frac{1}{d_{\text{tot}}^{O_D}} \mathbb{E} \int_{\text{Haar}} dU_A \mathrm{Tr} \left\{ U_1 \rho_d U_1^\dagger O_D U_2 \rho_d U_2^\dagger O_D \right\} \\ &= \frac{1}{d_{\text{tot}}^{O_D}} \mathbb{E} \int_{\text{Haar}} dU_A \mathrm{Tr} \left\{ U_1 (U_A |0\rangle \langle 0|_A U_A^\dagger \otimes \rho_B) U_1^\dagger O_D \right. \\ &\quad \left. \cdot U_2 (U_A |0\rangle \langle 0|_A U_A^\dagger \otimes \rho_B) U_2^\dagger O_D \right\} \\ &= \frac{1}{d_{\text{tot}}^{O_D}} \mathbb{E} \int_{\text{Haar}} dU_A \mathrm{Tr} \left\{ (U_1 \otimes U_2) (U_A |0\rangle \langle 0|_A U_A^\dagger \otimes \rho_B)^{\otimes 2} (U_1^\dagger \otimes U_2^\dagger) O_D^{\otimes 2} \mathbb{S}_{1,2} \right\} \\ &= \frac{1}{d_{\text{tot}}^{O_D}} \mathbb{E} \mathrm{Tr} \left\{ (U_1 \otimes U_2) \left[\int_{\text{Haar}} dU_A (U_A |0\rangle \langle 0|_A U_A^\dagger \otimes \rho_B)^{\otimes 2} \right] (U_1^\dagger \otimes U_2^\dagger) O_D^{\otimes 2} \mathbb{S}_{1,2} \right\}. \end{aligned} \quad (\text{G.17})$$

In the second line, we introduce $\rho_d = U_A |0\rangle \langle 0|_A U_A^\dagger \otimes \rho_B$. In the third line, we introduce the swap operator $\mathbb{S}_{1,2}$ on the doubled Hilbert space. Compute the average over U_A and keep track of the doubled systems (e.g. A_1 and A_2):

$$\begin{aligned}
 & \int_{\text{Haar}} dU_A (U_A |0\rangle \langle 0|_A U_A^\dagger \otimes \rho_B)^{\otimes 2} \\
 &= \int_{\text{Haar}} dU_A (U_A |0\rangle \langle 0|_{A_1} U_A^\dagger \otimes \rho_{B_1}) \otimes (U_A |0\rangle \langle 0|_{A_2} U_A^\dagger \otimes \rho_{B_2}) \\
 &= \int_{\text{Haar}} dU_A (U_A^{\otimes 2} (|0\rangle \langle 0|_{A_1} \otimes |0\rangle \langle 0|_{A_2}) U_A^{\dagger \otimes 2}) \otimes (\rho_{B_1} \otimes \rho_{B_2}) \\
 &= \Phi_{\text{Haar}(d_A)}^{(2)} (|0\rangle \langle 0|_{A_1} \otimes |0\rangle \langle 0|_{A_2}) \otimes (\rho_{B_1} \otimes \rho_{B_2}) \\
 &= \frac{1}{d_A(d_A+1)} \left[I_{A_1, A_2} + \mathbb{S}_{A_1, A_2} \right] \otimes (\rho_{B_1} \otimes \rho_{B_2}) \tag{G.18} \\
 &= \frac{1}{d_A(d_A+1)} \left[I_{A_1, A_2} + d_A \mathbb{E}_{O_A} O_A \otimes O_A^\dagger \right] \otimes (\rho_{B_1} \otimes \rho_{B_2}) \\
 &= \frac{1}{d_A(d_A+1)d_B^2} \left[I_{A, B} \otimes I_{A, B} + d_A \mathbb{E}_{O_A} ((O_A \otimes I_B) \otimes (O_A^\dagger \otimes I_B)) \right] \\
 &= \frac{1}{d_A(d_A+1)d_B^2} \left[I_{A, B}^{\otimes 2} + d_A \mathbb{E}_{O_A} (O_A \otimes O_A^\dagger) \right].
 \end{aligned}$$

In the second line, we switch the order of the tensor product on systems B_1 and A_2 for ease of computation. We switch back to the correct order in line six. In the third line, we introduce the 2-fold twirling channel, $\Phi_{\text{Haar}(d_A)}^{(2)}(\cdot)$ (see appendix E). In the fourth line, we use eq. (E.5). In the fifth line we introduce the 1-design identity: $\mathbb{S}_{A_1, A_2} = d_A \mathbb{E}_{O_A} O_A \otimes O_A^\dagger$. In the sixth line, we use $\rho_B = \frac{1}{d_B} I_B$, reorganize the tensor product and drop the A_i and B_i labels. In line seven, we redefine notation $O_A \otimes I_B \rightarrow O_A$. Plugging into eq. (G.17),

$$\begin{aligned}
 & \int_{\text{Haar}} dU_A C_d(U_1, U_2) \\
 &= \frac{1}{d_{\text{tot}}} \frac{1}{d_A(d_A+1)d_B^2} \\
 & \quad \cdot \mathbb{E}_{O_D} \text{Tr} \left\{ (U_1 \otimes U_2) \left[I_{A, B}^{\otimes 2} + d_A \mathbb{E}_{O_A} (O_A \otimes O_A^\dagger) \right] (U_1^\dagger \otimes U_2^\dagger) O_D^{\otimes 2} \mathbb{S}_{1,2} \right\} \\
 &= \frac{1}{d_{\text{tot}}} \frac{1}{d_A(d_A+1)d_B^2} \\
 & \quad \cdot \mathbb{E}_{O_D} \text{Tr} \left\{ O_D^{\otimes 2} \mathbb{S}_{1,2} + d_A \mathbb{E}_{O_A} (U_1 \otimes U_2) (O_A \otimes O_A^\dagger) (U_1^\dagger \otimes U_2^\dagger) O_D^{\otimes 2} \mathbb{S}_{1,2} \right\} \tag{G.19} \\
 &= \frac{1}{d_{\text{tot}}} \frac{1}{d_A(d_A+1)d_B^2} \\
 & \quad \cdot \mathbb{E}_{O_D} \left[\text{Tr} \left\{ O_D^2 \right\} + d_A \mathbb{E}_{O_A} \text{Tr} \left\{ (U_1 O_A U_1^\dagger O_D \otimes U_2 O_A^\dagger U_2^\dagger O_D) \mathbb{S}_{1,2} \right\} \right] \\
 &= \frac{1}{d_{\text{tot}}} \frac{1}{d_A(d_A+1)d_B^2} \mathbb{E}_{O_D} \left[d_{\text{tot}} + d_A \mathbb{E}_{O_A} \text{Tr} \left\{ U_1 O_A U_1^\dagger O_D U_2 O_A U_2^\dagger O_D \right\} \right] \\
 &= \frac{1}{d_A(d_A+1)d_B^2} \left[1 + d_A F(U_1, U_2) \right].
 \end{aligned}$$

In the fourth line, we use the Hermiticity of O_A and $\text{Tr}\{O_D^2\} = \text{Tr}\{I_{C,D}\} = d_{\text{tot}}$. In line five, we define the correlation function

$$F(U_1, U_2) = \mathbb{E}_{O_A O_D} \mathbb{E} \langle U_1 O_A U_1^\dagger O_D U_2 O_A U_2^\dagger O_D \rangle. \quad (\text{G.20})$$

In the case where $U_1 = U_2$, $F(U_1, U_1) = \overline{\text{OTOC}}(U_1)$. In the case where $U_1 = U$ and $U_2 = U_S$, we retrieve the optimization correlator $F(U, U_S) = \text{OP}(U, U_S)$. From eq. (G.19), we readily find

$$\int_{\text{Haar}} dU_A C_d(U_1, U_1) = \frac{1}{d_A(d_A + 1)d_B^2} \left[1 + d_A \overline{\text{OTOC}}(U_1) \right], \quad (\text{G.21})$$

$$\int_{\text{Haar}} dU_A C_d(U, U_S) = \frac{1}{d_A(d_A + 1)d_B^2} \left[1 + d_A \text{OP}(U, U_S) \right]. \quad (\text{G.22})$$

Eq. (G.16) is then

$$\begin{aligned} L &= \frac{d_D^2}{d_A(d_A + 1)d_B^2} \left[(1 + d_A \overline{\text{OTOC}}(U)) + (1 + d_A \overline{\text{OTOC}}(U_S)) - 2(1 + d_A \text{OP}(U, U_S)) \right] \\ &= \frac{d_D^2 d_A}{d_A(d_A + 1)d_B^2} \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\text{OP}(U, U_S) \right] \\ &= G \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\text{OP}(U, U_S) \right]. \end{aligned} \quad (\text{G.23})$$

In the third line, we use $d_A d_B = d_C d_D$ and define $G = \frac{d_A^2}{(d_A + 1)d_C^2}$.

H Proof of Corollary 1

We show that the true error for a maximally scrambling target unitary is

$$L_{\text{scram}} = G \left(\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}_{\text{scram}} - \frac{2}{d_A^2} \right). \quad (\text{H.1})$$

This is computed from L by integrating U_S over the Haar measure on the unitary group, which is valid under the assumption U_S is scrambling:

$$\begin{aligned} L_{\text{scram}} &= \int_{\text{Haar}} dU_S L \\ &= G \int_{\text{Haar}} dU_S \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\text{OP}(U, U_S) \right]. \end{aligned} \quad (\text{H.2})$$

First compute the integral over the OTOC:

$$\begin{aligned} \int_{\text{Haar}} dU_S \overline{\text{OTOC}}(U_S) &= \int_{\text{Haar}} dU_S \mathbb{E}_{O_A O_D} \mathbb{E} \langle U_S O_A U_S^\dagger O_D U_S O_A U_S^\dagger O_D \rangle \\ &= \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_A O_D} \mathbb{E} \int_{\text{Haar}} dU_S \text{Tr} \left\{ U_S O_A U_S^\dagger O_D U_S O_A U_S^\dagger O_D \right\} \\ &= \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_A O_D} \text{Tr} \left\{ \left[\int_{\text{Haar}} dU_S U_S^{\otimes 2} O_A^{\otimes 2} U_S^{\dagger \otimes 2} \right] O_D^{\otimes 2} \mathbb{S}_{1,2} \right\} \\ &= \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_D} \text{Tr} \left\{ \mathbb{E}_{O_A} \Phi_{\text{Haar}(d_{\text{tot}})}^{(2)}(O_A^{\otimes 2}) O_D^{\otimes 2} \mathbb{S}_{1,2} \right\}. \end{aligned} \quad (\text{H.3})$$

The average over U_S is just the 2-fold twirling channel, $\Phi_{\text{Haar}(d_{\text{tot}})}^{(2)}(O_A^{\otimes 2})$ (see appendix E). The swap operator $\mathbb{S}_{1,2}$ acts over the doubled Hilbert space. Using eq. (E.4) we compute the average of the twirling channel,

$$\begin{aligned}
 \Phi_{\text{Haar}(d_{\text{tot}})}^{(2)}(O_A^{\otimes 2}) &= \frac{1}{d_{\text{tot}}^2 - 1} \left[I_{1,2} \text{Tr} \{ O_A^{\otimes 2} \} + \mathbb{S}_{1,2} \text{Tr} \{ O_A^{\otimes 2} \mathbb{S}_{1,2} \} \right. \\
 &\quad \left. - \frac{1}{d_{\text{tot}}} I_{1,2} \text{Tr} \{ O_A^{\otimes 2} \mathbb{S}_{1,2} \} - \frac{1}{d_{\text{tot}}} \mathbb{S}_{1,2} \text{Tr} \{ O_A^{\otimes 2} \} \right] \\
 &= \frac{1}{d_{\text{tot}}^2 - 1} \left[I_{1,2} \left(\text{Tr} \{ O_A \}^2 - \frac{1}{d_{\text{tot}}} \text{Tr} \{ O_A^2 \} \right) \right. \\
 &\quad \left. + \mathbb{S}_{1,2} \left(\text{Tr} \{ O_A^2 \} - \frac{1}{d_{\text{tot}}} \text{Tr} \{ O_A \}^2 \right) \right] \\
 &= \frac{1}{d_{\text{tot}}^2 - 1} \left[I_{1,2} \left(d_{\text{tot}}^2 \delta_{O_A, I_A} - 1 \right) + \mathbb{S}_{1,2} \left(d_{\text{tot}} - d_{\text{tot}} \delta_{O_A, I_A} \right) \right].
 \end{aligned} \tag{H.4}$$

In the above, O_A actually denotes $O_A \otimes I_B$, so $\text{Tr} \{ O_A \} = d_{\text{tot}} \delta_{O_A, I_A}$. We also use $\text{Tr} \{ O_A^2 \} = \text{Tr} \{ I_{A,B} \} = d_{\text{tot}}$. Now average the twirling channel over O_A , using $\mathbb{E}_{O_A} \delta_{O_A, I_A} = \frac{1}{d_A^2} \sum_{O_A} \delta_{O_A, I_A} = \frac{1}{d_A^2}$:

$$\begin{aligned}
 \mathbb{E}_{O_A} \Phi_{\text{Haar}(d_{\text{tot}})}^{(2)}(O_A^{\otimes 2}) &= \frac{1}{d_{\text{tot}}^2 - 1} \left[I_{1,2} \left(d_{\text{tot}}^2 \mathbb{E}_{O_A} \delta_{O_A, I_A} - 1 \right) + \mathbb{S}_{1,2} \left(d_{\text{tot}} - d_{\text{tot}} \mathbb{E}_{O_A} \delta_{O_A, I_A} \right) \right] \\
 &= \frac{1}{d_{\text{tot}}^2 - 1} \left[I_{1,2} \left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + \mathbb{S}_{1,2} \left(d_{\text{tot}} - \frac{d_{\text{tot}}}{d_A^2} \right) \right].
 \end{aligned} \tag{H.5}$$

Plug this into eq. (H.3)

$$\begin{aligned}
 &\int_{\text{Haar}} dU_S \overline{\text{OTOC}}(U_S) \\
 &= \frac{1}{d_{\text{tot}}(d_{\text{tot}}^2 - 1)} \mathbb{E}_{O_D} \text{Tr} \left\{ \left[I_{1,2} \left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + \mathbb{S}_{1,2} \left(d_{\text{tot}} - \frac{d_{\text{tot}}}{d_A^2} \right) \right] O_D^{\otimes 2} \mathbb{S}_{1,2} \right\} \\
 &= \frac{1}{d_{\text{tot}}(d_{\text{tot}}^2 - 1)} \mathbb{E}_{O_D} \left[\text{Tr} \{ O_D^{\otimes 2} \mathbb{S}_{1,2} \} \left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + \text{Tr} \{ \mathbb{S}_{1,2} O_D^{\otimes 2} \mathbb{S}_{1,2} \} \left(d_{\text{tot}} - \frac{d_{\text{tot}}}{d_A^2} \right) \right] \\
 &= \frac{1}{d_{\text{tot}}(d_{\text{tot}}^2 - 1)} \mathbb{E}_{O_D} \left[\text{Tr} \{ O_D^2 \} \left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + \text{Tr} \{ O_D \}^2 \left(d_{\text{tot}} - \frac{d_{\text{tot}}}{d_A^2} \right) \right] \\
 &= \frac{1}{d_{\text{tot}}(d_{\text{tot}}^2 - 1)} \mathbb{E}_{O_D} \left[d_{\text{tot}} \left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + d_{\text{tot}}^2 \delta_{O_D, I_D} \left(d_{\text{tot}} - \frac{d_{\text{tot}}}{d_A^2} \right) \right] \\
 &= \frac{1}{d_{\text{tot}}(d_{\text{tot}}^2 - 1)} \left[d_{\text{tot}} \left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + d_C^2 \left(d_{\text{tot}} - \frac{d_{\text{tot}}}{d_A^2} \right) \right] \\
 &= \overline{\text{OTOC}}_{\text{scram}},
 \end{aligned} \tag{H.6}$$

where we define

$$\overline{\text{OTOC}}_{\text{scram}} = \frac{1}{(d_{\text{tot}}^2 - 1)} \left[\left(\frac{d_{\text{tot}}^2}{d_A^2} - 1 \right) + d_C^2 \left(1 - \frac{1}{d_A^2} \right) \right]. \tag{H.7}$$

In the above, we use $\text{Tr}\{O_D\} = d_{\text{tot}}\delta_{O_D, I_D}$, $\text{Tr}\{O_D^2\} = d_{\text{tot}}$, and $\mathbb{E}_{O_D} \delta_{O_D, I_D} = \frac{1}{d_D^2}$. For large d_{tot} , $\overline{\text{OTOC}}_{\text{scram}} \rightarrow \frac{1}{d_A^2} + \frac{1}{d_D^2} - \frac{1}{d_A^2 d_D^2}$. This result was originally shown in [66].

Now we compute the integral over the optimization correlator, assuming U is independent of U_S (i.e. U is not yet trained):

$$\begin{aligned}
\int_{\text{Haar}} dU_S \text{OP}(U, U_S) &= \frac{1}{d_{\text{tot}}} \int_{\text{Haar}} dU_S \mathbb{E}_{O_A O_D} \mathbb{E} \text{Tr} \left\{ U O_A U^\dagger O_D U_S O_A U_S^\dagger O_D \right\} \\
&= \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_A O_D} \mathbb{E} \text{Tr} \left\{ U O_A U^\dagger O_D \Phi_{\text{Haar}(d_{\text{tot}})}^{(1)}(O_A) O_D \right\} \\
&= \frac{1}{d_{\text{tot}}^2} \mathbb{E}_{O_A O_D} \mathbb{E} \text{Tr} \left\{ U O_A U^\dagger O_D^2 \right\} \text{Tr} \{O_A\} \\
&= \frac{1}{d_{\text{tot}}^2} \mathbb{E}_{O_A O_D} \mathbb{E} \text{Tr} \{O_A\}^2 \\
&= \frac{1}{d_{\text{tot}}^2} \mathbb{E}_{O_A O_D} \mathbb{E} (d_{\text{tot}}^2 \delta_{O_A, I_A}) \\
&= \frac{1}{d_A^2}.
\end{aligned} \tag{H.8}$$

L_{scram} becomes

$$L_{\text{scram}} = G \left(\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}_{\text{scram}} - \frac{2}{d_A^2} \right). \tag{H.9}$$

In the case where U is maximally scrambling, we obtain the floor value

$$L_{\text{floor}} = 2G \left(\overline{\text{OTOC}}_{\text{scram}} - \frac{1}{d_A^2} \right). \tag{H.10}$$

In the large d_{tot} limit, this floor value becomes

$$\begin{aligned}
\lim_{d_{\text{tot}} \rightarrow \infty} L_{\text{floor}} &= 2G \left(\lim_{d_{\text{tot}} \rightarrow \infty} \overline{\text{OTOC}}_{\text{scram}} - \frac{1}{d_A^2} \right) \\
&= 2G \left(\frac{1}{d_A^2} + \frac{1}{d_D^2} - \frac{1}{d_A^2 d_D^2} - \frac{1}{d_A^2} \right) \\
&= \frac{2d_A^2}{(d_A + 1)d_C^2} \left(\frac{1}{d_D^2} - \frac{1}{d_A^2 d_D^2} \right) \\
&= \frac{2}{(d_A + 1)d_B^2} \left(1 - \frac{1}{d_A^2} \right).
\end{aligned} \tag{H.11}$$

For $d_A \gg 1$,

$$\lim_{d_{\text{tot}} \rightarrow \infty} L_{\text{floor}} = \frac{2}{d_B d_{\text{tot}}}. \tag{H.12}$$

I Proof of Proposition 2

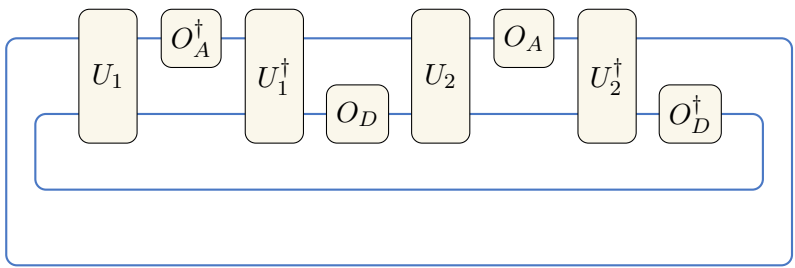
We will prove that L is bounded by

$$L_{\pm} = G \left[\sqrt{\overline{\text{OTOC}}(U)} \pm \sqrt{\overline{\text{OTOC}}(U_S)} \right]^2. \tag{I.1}$$

By noting L is non-negative, we apply the triangle inequality to the true error in eq. (3.3):

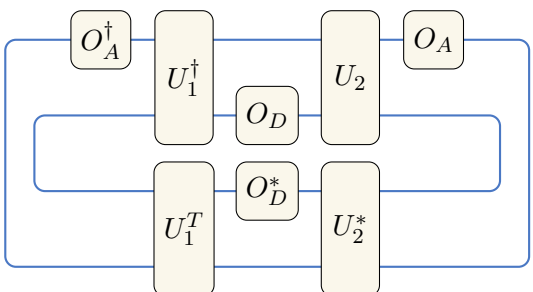
$$L \leq G \left[\left| \overline{\text{OTOC}}(U) \right| + \left| \overline{\text{OTOC}}(U_S) \right| + 2|\text{OP}(U, U_S)| \right]. \quad (\text{I.2})$$

We now bound the third term. Begin by rewriting $F(U_1, U_2)$ from eq. (G.20) diagrammatically, taking $O_A = O_A^\dagger$ and $O_D = O_D^\dagger$ wherever convenient:

$$F(U_1, U_2) = \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_A} \mathbb{E}_{O_D} \cdot$$


(I.3)

Using the transpose identity from eq. (D.4),

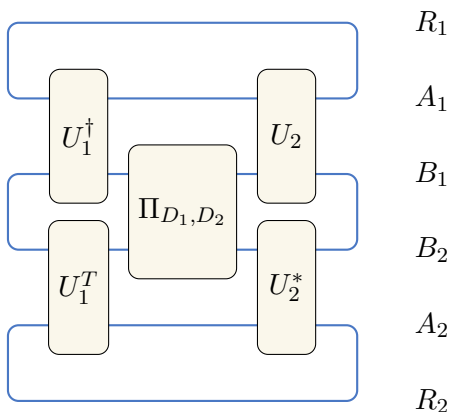
$$F(U_1, U_2) = \frac{1}{d_{\text{tot}}} \mathbb{E}_{O_A} \mathbb{E}_{O_D} \cdot$$


(I.4)

Let Π_{D_1, D_2} be the projection onto the Bell state between systems D_1 and D_2

$$\Pi_{D_1, D_2} = |\text{Bell}\rangle \langle \text{Bell}|_{D_1, D_2}. \quad (\text{I.5})$$

We average over O_A and O_D , and use the 1-design identity $\mathbb{E}_{O_D} O_D \otimes O_D^* = \Pi_{D_1, D_2}$:

$$F(U_1, U_2) = \frac{1}{d_{\text{tot}} d_A} \cdot$$


(I.6)

The labels on the right-hand side refer to the systems in the tensor network. Reference system R_i has dimension d_A , where $i \in \{1, 2\}$.

Define the states

$$\begin{aligned} |\psi(U_i)\rangle &= I_{R_1} \otimes U_i \otimes U_i^* \otimes I_{R_2} |\psi_0\rangle, \\ |\psi_0\rangle &= |\text{Bell}\rangle_{R_1 A_1} \otimes |\text{Bell}\rangle_{B_1 B_2} \otimes |\text{Bell}\rangle_{A_2 R_2}. \end{aligned} \quad (\text{I.7})$$

Eq. (I.6) can be written as

$$F(U_1, U_2) = \langle \psi(U_1) | \Pi_{D_1, D_2} | \psi(U_2) \rangle. \quad (\text{I.8})$$

Π_{D_1, D_2} actually denotes $I_{R_1, A_1, A_2, R_2} \otimes \Pi_{D_1, D_2}$, but the identity operator is omitted for convenience. Setting $U_1 = U_2$ yields $F(U_1, U_1) = \overline{\text{OTOC}}(U_1)$, as shown in appendix G.2. $F(U_1, U_1)$ also gives the probability $P(U_1)$ of projecting a Bell state onto $D_1 D_2$ while in state $|\psi(U_1)\rangle$ [66]. We have the relation

$$\overline{\text{OTOC}}(U_1) = \langle \psi(U_1) | \Pi_{D_1, D_2} | \psi(U_1) \rangle = P(U_1). \quad (\text{I.9})$$

Since the OTOC represents a probability, $0 \leq \overline{\text{OTOC}}(U_1) \leq 1$. In the case where $U_1 = U$ and $U_2 = U_S$, $F(U, U_S) = \text{OP}(U, U_S)$, which produces

$$\text{OP}(U, U_S) = \langle \psi(U) | \Pi_{D_1, D_2} | \psi(U_S) \rangle. \quad (\text{I.10})$$

We now bound $|\text{OP}(U, U_S)|$. Begin by writing

$$\begin{aligned} |\text{OP}(U, U_S)| &= \left| \langle \psi(U) | \Pi_{D_1, D_2}^2 | \psi(U_S) \rangle \right| \\ &= \left| \langle \psi(U, \Pi_{D_1, D_2}) | \psi(U_S, \Pi_{D_1, D_2}) \rangle \right|, \end{aligned} \quad (\text{I.11})$$

where we define the unnormalized state $|\psi(U_i, \Pi_{D_1, D_2})\rangle = \Pi_{D_1, D_2} |\psi(U_i)\rangle$ and use the fact $\Pi_{D_1, D_2} = \Pi_{D_1, D_2}^2$. Apply the Cauchy-Schwarz inequality,

$$\begin{aligned} |\text{OP}(U, U_S)| &\leq \sqrt{\langle \psi(U, \Pi_{D_1, D_2}) | \psi(U, \Pi_{D_1, D_2}) \rangle \langle \psi(U_S, \Pi_{D_1, D_2}) | \psi(U_S, \Pi_{D_1, D_2}) \rangle} \\ &= \sqrt{\langle \psi(U) | \Pi_{D_1, D_2}^2 | \psi(U) \rangle \langle \psi(U_S) | \Pi_{D_1, D_2}^2 | \psi(U_S) \rangle} \\ &= \sqrt{\langle \psi(U) | \Pi_{D_1, D_2} | \psi(U) \rangle \langle \psi(U_S) | \Pi_{D_1, D_2} | \psi(U_S) \rangle} \\ &= \sqrt{\overline{\text{OTOC}}(U) \overline{\text{OTOC}}(U_S)}. \end{aligned} \quad (\text{I.12})$$

Ineq. (I.2) becomes

$$L \leq G \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) + 2\sqrt{\overline{\text{OTOC}}(U) \overline{\text{OTOC}}(U_S)} \right]. \quad (\text{I.13})$$

The right-hand side is defined as the upper bound on L :

$$L_+ = G \left[\sqrt{\overline{\text{OTOC}}(U)} + \sqrt{\overline{\text{OTOC}}(U_S)} \right]^2. \quad (\text{I.14})$$

We also construct a lower bound on L . We first bound $\text{OP}(U, U_S)$:

$$\begin{aligned} \text{OP}(U, U_S) &\leq |\text{OP}(U, U_S)| \\ &\leq \sqrt{\overline{\text{OTOC}}(U) \overline{\text{OTOC}}(U_S)}. \end{aligned} \quad (\text{I.15})$$

We now bound the true error:

$$\begin{aligned} L &= G \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\text{OP}(U, U_S) \right] \\ &\geq G \left[\overline{\text{OTOC}}(U) + \overline{\text{OTOC}}(U_S) - 2\sqrt{\overline{\text{OTOC}}(U)\overline{\text{OTOC}}(U_S)} \right]. \end{aligned} \quad (\text{I.16})$$

We define the right-hand side as the lower bound on L :

$$L_- = G \left[\sqrt{\overline{\text{OTOC}}(U)} - \sqrt{\overline{\text{OTOC}}(U_S)} \right]^2. \quad (\text{I.17})$$

Using $|L - L_{\pm}| \leq L_+ - L_-$, we arrive at the bound

$$|L - L_{\pm}| \leq 4G\sqrt{\overline{\text{OTOC}}(U)\overline{\text{OTOC}}(U_S)}. \quad (\text{I.18})$$

J Proof of Proposition 3

We derive a concentration inequality for the loss function by using Levy's lemma. Assume input state ρ_d is given by eq. (2.4). It will be convenient to introduce the notation $L_d = L_d(|\psi_{d,A}\rangle)$. Levy's lemma in this context states:

Lemma 2 (Levy's lemma) *Let $L_d : S^{2d_{\text{tot}}-1} \rightarrow \mathbb{R}$ satisfy $|L_d(|\psi_{1,A}\rangle) - L_d(|\psi_{2,A}\rangle)| \leq \eta \| |\psi_{1,A}\rangle - |\psi_{2,A}\rangle \|_2$. Then $\forall \delta \geq 0$,*

$$\text{Prob}_{|\psi_{d,A}\rangle \sim \text{Haar}(d_A)} [|L_d(|\psi_{d,A}\rangle) - L| \geq \delta] \leq \epsilon, \quad (\text{J.1})$$

where $\epsilon = 2\exp\left(-\frac{2d_A\delta^2}{9\pi^3\eta^2}\right)$ and η is the Lipschitz constant. By definition, $L = \mathbb{E}_{|\psi_{d,A}\rangle \sim \text{Haar}(d_A)} [L_d(|\psi_{d,A}\rangle)]$. The average is over the uniform distribution on the Hilbert space of system A .

Parameter δ can be written in terms of ϵ : $\delta = \eta f(\epsilon)$, where we define $f(\epsilon) = \sqrt{\frac{9\pi^3}{2d_A} \ln \frac{2}{\epsilon}}$, as in eq. (3.14). Levy's lemma implies that with probability at least $1 - \epsilon$,

$$|L_d(|\psi_{d,A}\rangle) - L| \leq \eta f(\epsilon). \quad (\text{J.2})$$

With probability at least $1 - \epsilon$, we can also construct an inequality for $|L_d(|\psi_{d,A}\rangle) - L_{\pm}|$:

$$\begin{aligned} |L_d(|\psi_{d,A}\rangle) - L_{\pm}| &= |L_d(|\psi_{d,A}\rangle) - L + L - L_{\pm}| \\ &\leq |L_d(|\psi_{d,A}\rangle) - L| + |L - L_{\pm}| \\ &\leq \eta f(\epsilon) + |L - L_{\pm}| \\ &\leq \eta f(\epsilon) + 4G\sqrt{\overline{\text{OTOC}}(U)\overline{\text{OTOC}}(U_S)}. \end{aligned} \quad (\text{J.3})$$

J.1 Lipschitz constant

We now compute the Lipschitz constant. We can write state $|\psi_{d,A}\rangle$ in terms of real vectors $\mathbf{v}_1^d, \mathbf{v}_2^d \in \mathbb{R}^{d_A}$:

$$|\psi_{d,A}\rangle = \mathbf{v}_1^d + i\mathbf{v}_2^d. \quad (\text{J.4})$$

The state can also be written in the form:

$$\begin{aligned} |\psi_{d,A}\rangle &= [I_A \ iI_A] \mathbf{w}_d \\ \mathbf{w}_d &= \begin{bmatrix} \mathbf{v}_1^d \\ \mathbf{v}_2^d \end{bmatrix}. \end{aligned} \tag{J.5}$$

It is straightforward to show that

$$\| |\psi_{1,A}\rangle - |\psi_{2,A}\rangle \|_2 = \| \mathbf{w}_1 - \mathbf{w}_2 \|_2. \tag{J.6}$$

Therefore, the Lipschitz continuity condition for $L_d(|\psi_{d,A}\rangle)$ can be written as

$$|L_d(|\psi_{1,A}\rangle) - L_d(|\psi_{2,A}\rangle)| \leq \eta \| \mathbf{w}_1 - \mathbf{w}_2 \|_2. \tag{J.7}$$

We can therefore compute a Lipschitz constant by finding any η such that

$$\left\| \frac{d}{d\mathbf{w}_d} L_d(|\psi_{d,A}\rangle) \right\|_2 \leq \eta. \tag{J.8}$$

We compute

$$\begin{aligned} \left\| \frac{d}{d\mathbf{w}_d} L_d(|\psi_{d,A}\rangle) \right\|_2 &= \left\| \frac{d}{d\mathbf{w}_d} \mathbb{E}_{\mathcal{O}_C} |\tilde{y}_d - y_d|^2 \right\|_2 \\ &= \left\| \mathbb{E}_{\mathcal{O}_C} 2(\tilde{y}_d - y_d) \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left\| (\tilde{y}_d - y_d) \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 \\ &= 2 \mathbb{E}_{\mathcal{O}_C} |\tilde{y}_d - y_d| \left\| \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} (|\tilde{y}_d| + |y_d|) \left\| \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 \\ &\leq 4 \mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 \\ &\leq 4 \mathbb{E}_{\mathcal{O}_C} \left[\left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 + \left\| \frac{d}{d\mathbf{w}_d} y_d \right\|_2 \right]. \end{aligned} \tag{J.9}$$

The sixth line follows from $|\tilde{y}_d|, |y_d| \leq 1$. The remaining inequalities follow by the triangle inequality. Let Q be a Hermitian operator on system AB . Compute the following norm:

$$\begin{aligned} \left\| \frac{d}{d\mathbf{w}_d} \text{Tr} \{ \rho_d Q \} \right\|_2 &= \left\| \frac{d}{d\mathbf{w}_d} \text{Tr} \{ (|\psi_{d,A}\rangle \langle \psi_{d,A}| \otimes \rho_B) Q \} \right\|_2 \\ &= \left\| \frac{1}{d_B} \frac{d}{d\mathbf{w}_d} \text{Tr} \{ (|\psi_{d,A}\rangle \langle \psi_{d,A}| \otimes I_B) Q \} \right\|_2 \\ &= \frac{1}{d_B} \left\| \frac{d}{d\mathbf{w}_d} \text{Tr}_A \{ |\psi_{d,A}\rangle \langle \psi_{d,A}| \text{Tr}_B \{ Q \} \} \right\|_2 \\ &= \frac{1}{d_B} \left\| \frac{d}{d\mathbf{w}_d} \langle \psi_{d,A}| \text{Tr}_B \{ Q \} |\psi_{d,A}\rangle \right\|_2 \\ &= \frac{2}{d_B} \| \text{Tr}_B \{ Q \} |\psi_{d,A}\rangle \|_2. \end{aligned} \tag{J.10}$$

$\text{Tr}_B\{\cdot\}$ denotes the partial trace. Line five follows from appendix F. Now we bound the average over the norm:

$$\begin{aligned} \mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} \text{Tr} \{ \rho_d Q \} \right\|_2 &\leq \frac{2}{d_B} \mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ Q \} \|_\infty \\ &\leq \frac{2}{d_B} \mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ Q \} \|_{\text{HS}} \\ &\leq \frac{2}{d_B} \sqrt{\mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ Q \} \|_{\text{HS}}^2}. \end{aligned} \quad (\text{J.11})$$

The second line follows from the Hilbert-Schmidt norm upper bounding the operator norm. The third line follows from the variance being non-negative: $(\mathbb{E}_{\mathcal{O}_C}(\cdot))^2 \leq \mathbb{E}_{\mathcal{O}_C}(\cdot)^2$. Letting $Q = U^\dagger O_C U$ gives

$$\mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 \leq \frac{2}{d_B} \sqrt{\mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ U^\dagger O_C U \} \|_{\text{HS}}^2}. \quad (\text{J.12})$$

Now we compute

$$\mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ U^\dagger O_C U \} \|_{\text{HS}}^2 = \mathbb{E}_{\mathcal{O}_C} \text{Tr}_A \left\{ \left| \text{Tr}_B \{ U^\dagger O_C U \} \right|^2 \right\}. \quad (\text{J.13})$$

Writing this diagrammatically,

$$\mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ U^\dagger O_C U \} \|_{\text{HS}}^2 = \mathbb{E}_{\mathcal{O}_C} \left[\text{Diagram} \right]. \quad (\text{J.14})$$

By introducing a 1-design, we can write this as

$$\begin{aligned} \mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ U^\dagger O_C U \} \|_{\text{HS}}^2 &= d_A \mathbb{E}_{\mathcal{O}_A} \mathbb{E}_{\mathcal{O}_C} \left[\text{Diagram} \right]. \end{aligned} \quad (\text{J.15})$$

Using the Hermiticity of Pauli string O_A ,

$$\mathbb{E}_{\mathcal{O}_C} \| \text{Tr}_B \{ U^\dagger O_C U \} \|_{\text{HS}}^2 = d_A \mathbb{E}_{\mathcal{O}_A} \mathbb{E}_{\mathcal{O}_C} \text{Tr} \{ U O_A U^\dagger O_C \}^2. \quad (\text{J.16})$$

Retrieving the identity from eq. (G.10) and setting $Q = O_A, U_1 = U_2 = U$,

$$\mathbb{E}_{\mathcal{O}_C} \text{Tr} \{ U O_A U^\dagger O_C \}^2 = d_D^2 C(O_A, U, U) = d_D^2 \mathbb{E}_{\mathcal{O}_D} \langle U O_A U^\dagger O_D U O_A U^\dagger O_D \rangle. \quad (\text{J.17})$$

This yields,

$$\begin{aligned} \mathbb{E}_{O_C} \left\| \text{Tr}_B \left\{ U^\dagger O_C U \right\} \right\|_{\text{HS}}^2 &= d_A d_D^2 \mathbb{E}_{O_A O_D} \mathbb{E} \langle U O_A U^\dagger O_D U O_A U^\dagger O_D \rangle \\ &= d_A d_D^2 \overline{\text{OTOC}}(U). \end{aligned} \quad (\text{J.18})$$

Ineq. (J.12) becomes

$$\begin{aligned} \mathbb{E}_{O_C} \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 &\leq \frac{2}{d_B} \sqrt{d_A d_D^2 \overline{\text{OTOC}}(U)} \\ &= \frac{2\sqrt{d_A} d_D}{d_B} \sqrt{\overline{\text{OTOC}}(U)}. \end{aligned} \quad (\text{J.19})$$

A similar inequality holds for $\tilde{y}_d \rightarrow y_d$. Ineq. (J.9) becomes

$$\begin{aligned} \left\| \frac{d}{d\mathbf{w}_d} L_d(|\psi_{d,A}\rangle) \right\|_2 &\leq \frac{8\sqrt{d_A} d_D}{d_B} \left[\sqrt{\overline{\text{OTOC}}(U)} + \sqrt{\overline{\text{OTOC}}(U_S)} \right] \\ &= \frac{8\sqrt{d_A} d_D}{d_B} \left[\frac{d_C}{d_A} \sqrt{(d_A + 1)L_+} \right] \\ &= 8\sqrt{d_A(d_A + 1)L_+}. \end{aligned} \quad (\text{J.20})$$

The second line follows from the definition of L_+ in eq. (3.10). The Lipschitz constant is therefore

$$\eta = 8\sqrt{d_A(d_A + 1)L_+}. \quad (\text{J.21})$$

J.2 Maximally scrambling unitaries

Consider the case where $U_i \in \{U, U_S\}$ is maximally scrambling. In the large d_{tot} limit,

$$\overline{\text{OTOC}}(U_i) = \frac{1}{d_A^2} + \frac{1}{d_D^2} - \frac{1}{d_A^2 d_D^2}. \quad (\text{J.22})$$

Taking the case where $d_D \gg d_A$, $\overline{\text{OTOC}}(U_i) = \frac{1}{d_A^2}$. The Lipschitz constant becomes

$$\eta = \frac{8\sqrt{d_A} d_D}{d_B} \left[2\sqrt{\frac{1}{d_A^2}} \right] = \frac{16\sqrt{d_A} d_D}{d_B d_A} = \frac{16\sqrt{d_A}}{d_C}. \quad (\text{J.23})$$

Therefore,

$$\begin{aligned} \eta f(\epsilon) &= \frac{16\sqrt{d_A}}{d_C} \sqrt{\frac{9\pi^3}{2d_A} \ln \frac{2}{\epsilon}} \\ &= \frac{16}{d_C} \sqrt{\frac{9\pi^3}{2} \ln \frac{2}{\epsilon}}. \end{aligned} \quad (\text{J.24})$$

Recalling that with probability at least $1 - \epsilon$, $|L_d(|\psi_{d,A}\rangle) - L| \leq \eta f(\epsilon)$, L_d concentrates near L as d_C increases.

K Proof of Theorem 1

Similar to appendix J, we can use Levy's lemma to compute a concentration inequality for $\partial_{\theta_i} L_d$. Levy's lemma in this context reads

Lemma 3 (Levy's lemma) *Let $\partial_{\theta_i} L_d : S^{2d_{\text{tot}}-1} \rightarrow \mathbb{R}$ satisfy $|\partial_{\theta_i} L_d(|\psi_{1,A}\rangle) - \partial_{\theta_i} L_d(|\psi_{2,A}\rangle)| \leq \eta_g \|\psi_{1,A} - \psi_{2,A}\|_2$. Then $\forall \delta \geq 0$,*

$$\text{Prob}_{|\psi_{d,A}\rangle \sim \text{Haar}(d_A)} [|\partial_{\theta_i} L_d(|\psi_{d,A}\rangle) - \partial_{\theta_i} L| \geq \delta] \leq \epsilon, \quad (\text{K.1})$$

where $\epsilon = 2 \exp\left(-\frac{2d_A \delta^2}{9\pi^3 \eta_g^2}\right)$ and η_g is the Lipschitz constant.

In the above, $\partial_{\theta_i} L = \mathbb{E}_{|\psi_{d,A}\rangle \sim \text{Haar}(d_A)} [\partial_{\theta_i} L_d(|\psi_{d,A}\rangle)]$. Levy's lemma implies that with probability at least $1 - \epsilon$,

$$|\partial_{\theta_i} L_d(|\psi_{d,A}\rangle) - \partial_{\theta_i} L| \leq \eta_g f(\epsilon), \quad (\text{K.2})$$

where $f(\epsilon)$ is defined as in eq. (3.14).

K.1 Lipschitz constant

We compute the Lipschitz constant. As in appendix J, we define $|\psi_{d,A}\rangle$ using eq. (J.5) so that η_g can be found through the inequality: $\left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} L_d(|\psi_{d,A}\rangle) \right\|_2 \leq \eta_g$. We compute the bound:

$$\begin{aligned} \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} L_d(|\psi_{d,A}\rangle) \right\|_2 &= \left\| \frac{d}{d\mathbf{w}_d} (\partial_{\theta_i} \mathbb{E}_{\mathcal{O}_C} |\tilde{y}_d - y_d|^2) \right\|_2 \\ &= \left\| 2 \frac{d}{d\mathbf{w}_d} \mathbb{E}_{\mathcal{O}_C} (\tilde{y}_d - y_d) \partial_{\theta_i} \tilde{y}_d \right\|_2 \\ &= 2 \left\| \mathbb{E}_{\mathcal{O}_C} \left(\partial_{\theta_i} \tilde{y}_d \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) + (\tilde{y}_d - y_d) \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right) \right\|_2 \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left\| \partial_{\theta_i} \tilde{y}_d \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) + (\tilde{y}_d - y_d) \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(\left\| \partial_{\theta_i} \tilde{y}_d \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 + \left\| (\tilde{y}_d - y_d) \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \right) \\ &= 2 \mathbb{E}_{\mathcal{O}_C} \left(|\partial_{\theta_i} \tilde{y}_d| \left\| \frac{d}{d\mathbf{w}_d} (\tilde{y}_d - y_d) \right\|_2 + |\tilde{y}_d - y_d| \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \right) \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(|\partial_{\theta_i} \tilde{y}_d| \left(\left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 + \left\| \frac{d}{d\mathbf{w}_d} y_d \right\|_2 \right) + 2 \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \right). \end{aligned} \quad (\text{K.3})$$

Before computing $\partial_{\theta_l} \tilde{y}_d$, it will be useful to compute $\partial_{\theta_l} U$:

$$\begin{aligned}
 \partial_{\theta_l} U &= \partial_{\theta_l} \prod_{j=1}^L U_j(\theta_j) W_j \\
 &= \left[\prod_{k=l+1}^L U_k(\theta_k) W_k \right] \partial_{\theta_l} U_l(\theta_l) W_l \left[\prod_{j=1}^{l-1} U_j(\theta_j) W_j \right] \\
 &= \left[\prod_{k=l+1}^L U_k(\theta_k) W_k \right] (-iV_l) U(\theta_l) W_l \left[\prod_{j=1}^{l-1} U_j(\theta_j) W_j \right] \\
 &= \left[\prod_{k=l+1}^L U_k(\theta_k) W_k \right] (-iV_l) \left[\prod_{j=1}^l U_j(\theta_j) W_j \right] \\
 &= U_+ (-iV_l) U_-.
 \end{aligned} \tag{K.4}$$

We define $U_- = \prod_{j=1}^l U_j(\theta_j) W_j$ and $U_+ = \prod_{k=l+1}^L U_k(\theta_k) W_k$. Similarly, $\partial_{\theta_l} U^\dagger = U_-^\dagger (iV_l) U_+^\dagger$.

We now compute $\partial_{\theta_l} \tilde{y}_d$:

$$\begin{aligned}
 \partial_{\theta_l} \tilde{y}_d &= \partial_{\theta_l} \text{Tr} \left\{ U \rho_d U^\dagger O_C \right\} \\
 &= \text{Tr} \left\{ (\partial_{\theta_l} U) \rho_d U^\dagger O_C \right\} + \text{Tr} \left\{ U \rho_d (\partial_{\theta_l} U^\dagger) O_C \right\} \\
 &= \text{Tr} \left\{ (U_+ (-iV_l) U_-) \rho_d U^\dagger O_C \right\} + \text{Tr} \left\{ U \rho_d (U_-^\dagger (iV_l) U_+^\dagger) O_C \right\} \\
 &= \text{Tr} \left\{ \rho_d U_-^\dagger U_+^\dagger O_C U_+ (-iV_l) U_- \right\} + \text{Tr} \left\{ \rho_d U_-^\dagger (iV_l) U_+^\dagger O_C U_+ U_- \right\} \\
 &= \text{Tr} \left\{ \rho_d U_-^\dagger [iV_l, U_+^\dagger O_C U_+] U_- \right\} \\
 &= \text{Tr} \left\{ \rho_d Q \right\}.
 \end{aligned} \tag{K.5}$$

In line six, we define the Hermitian operator $Q = U_-^\dagger [iV_l, U_+^\dagger O_C U_+] U_-$. Now compute the following norm:

$$\begin{aligned}
 \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_l} \tilde{y}_d \right\|_2 &= \left\| \frac{d}{d\mathbf{w}_d} \text{Tr} \left\{ \rho_d Q \right\} \right\|_2 \\
 &= \frac{2}{d_B} \left\| \text{Tr}_B \left\{ Q \right\} | \psi_{d,A} \right\|_2 \\
 &= \frac{2}{d_B} \sqrt{\langle \psi_{d,A} | (\text{Tr}_B \left\{ Q \right\})^2 | \psi_{d,A} \rangle} \\
 &= \frac{2}{d_B} \sqrt{\text{Tr} \left\{ \rho_{d,A} (\text{Tr}_B \left\{ Q \right\})^2 \right\}}.
 \end{aligned} \tag{K.6}$$

In the second line, we use eq. (J.10). Now diagrammatically write

$$\text{Tr} \left\{ \rho_{d,A} (\text{Tr}_B \left\{ Q \right\})^2 \right\} = \text{Diagram} \tag{K.7}$$

We introduce a 1-design:

$$\text{Tr} \left\{ \rho_{d,A} (\text{Tr}_B \{Q\})^2 \right\} = d_B \mathbb{E}_{O_B} \left[\text{Tr} \left\{ \rho_{d,A} \left(\text{Tr}_B \left\{ Q O_B Q \right\} \right)^2 \right\} \right]. \quad (\text{K.8})$$

Introduce the Bell state:

$$\text{Tr} \left\{ \rho_{d,A} (\text{Tr}_B \{Q\})^2 \right\} = d_B^2 \mathbb{E}_{O_B} \left[\text{Tr} \left\{ \rho_{d,A} \left(\text{Tr}_B \left\{ Q O_B Q \right\} \right)^2 \right\} \right]. \quad (\text{K.9})$$

We can now bound the trace:

$$\begin{aligned} \text{Tr} \left\{ \rho_{d,A} (\text{Tr}_B \{Q\})^2 \right\} &= d_B^2 \mathbb{E}_{O_B} \text{Tr} \left\{ (\rho_{d,A} \otimes \Pi_{B_1, B_2}) O_B^\dagger (Q \otimes I_B) O_B (Q \otimes I_B) \right\} \\ &= d_B^2 \left| \mathbb{E}_{O_B} \text{Tr} \left\{ (\rho_{d,A} \otimes \Pi_{B_1, B_2}) O_B^\dagger (Q \otimes I_B) O_B (Q \otimes I_B) \right\} \right| \\ &\leq d_B^2 \mathbb{E}_{O_B} \left| \text{Tr} \left\{ (\rho_{d,A} \otimes \Pi_{B_1, B_2}) O_B^\dagger (Q \otimes I_B) O_B (Q \otimes I_B) \right\} \right| \\ &\leq d_B^2 \mathbb{E}_{O_B} \left\| O_B^\dagger (Q \otimes I_B) O_B (Q \otimes I_B) \right\|_\infty \\ &\leq d_B^2 \mathbb{E}_{O_B} \left\| O_B^\dagger \right\|_\infty \left\| (Q \otimes I_B) \right\|_\infty \left\| O_B \right\|_\infty \left\| (Q \otimes I_B) \right\|_\infty \\ &= d_B^2 \mathbb{E}_{O_B} \|Q\|_\infty^2 \\ &= d_B^2 \|Q\|_\infty^2 \\ &\leq 4d_B^2 \|V_l\|_\infty^2. \end{aligned} \quad (\text{K.10})$$

In the first line, we take O_B to denote $I_A \otimes O_B \otimes I_B$. The second line follows from the trace on the left-hand side being non-negative. The third line follows from the triangle inequality. The fifth line follows from the sub-multiplicativity property of the norm. The sixth line uses the fact that Pauli strings have a maximum eigenvalue of 1. The last line follows from the following bound:

$$\begin{aligned} \|Q\|_\infty &= \left\| U_-^\dagger [iV_l, U_+^\dagger O_C U_+] U_- \right\|_\infty \\ &= \left\| [iV_l, U_+^\dagger O_C U_+] \right\|_\infty \\ &= \left\| iV_l U_+^\dagger O_C U_+ - U_+^\dagger O_C U_+ iV_l \right\|_\infty \\ &\leq \left\| iV_l U_+^\dagger O_C U_+ \right\|_\infty + \left\| U_+^\dagger O_C U_+ iV_l \right\|_\infty \\ &= 2 \|V_l\|_\infty. \end{aligned} \quad (\text{K.11})$$

The second and fifth lines follow from the unitary invariance of the Schatten norms. In the fourth line, we use the triangle inequality. We can now bound eq. (K.6):

$$\left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \leq \frac{2}{d_B} \sqrt{4d_B^2 \|V_l\|_\infty^2} = 4 \|V_l\|_\infty. \quad (\text{K.12})$$

Also, bound $|\partial_{\theta_i} \tilde{y}_d|$:

$$\begin{aligned} |\partial_{\theta_i} \tilde{y}_d| &= |\text{Tr} \{ \rho_d Q \}| \\ &\leq \|Q\|_\infty \\ &\leq 2 \|V_l\|_\infty. \end{aligned} \quad (\text{K.13})$$

Ineq. (K.3) becomes

$$\begin{aligned} &\left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} L_d(|\psi_{d,A}\rangle) \right\|_2 \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(|\partial_{\theta_i} \tilde{y}_d| \left(\left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 + \left\| \frac{d}{d\mathbf{w}_d} y_d \right\|_2 \right) + 2 \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \right) \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(2 \|V_l\|_\infty \left(\left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 + \left\| \frac{d}{d\mathbf{w}_d} y_d \right\|_2 \right) + 2 \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_d \right\|_2 \right) \\ &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(2 \|V_l\|_\infty \left(\left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_d \right\|_2 + \left\| \frac{d}{d\mathbf{w}_d} y_d \right\|_2 \right) + 8 \|V_l\|_\infty \right) \\ &\leq 2 \left(2 \|V_l\|_\infty \left(\frac{2\sqrt{d_A d_D}}{d_B} \sqrt{\text{OTOC}(U)} + \frac{2\sqrt{d_A d_D}}{d_B} \sqrt{\text{OTOC}(U_S)} \right) + 8 \|V_l\|_\infty \right) \\ &= 2 \left(\frac{4\sqrt{d_A d_D}}{d_B} \|V_l\|_\infty \left(\sqrt{\text{OTOC}(U)} + \sqrt{\text{OTOC}(U_S)} \right) + 8 \|V_l\|_\infty \right) \\ &= 8 \|V_l\|_\infty \left(\frac{\sqrt{d_A d_D}}{d_B} \left(\sqrt{\text{OTOC}(U)} + \sqrt{\text{OTOC}(U_S)} \right) + 2 \right) \\ &= 8 \|V_l\|_\infty \left(\frac{\sqrt{d_A d_D}}{d_B} \left(\frac{d_C}{d_A} \sqrt{(d_A + 1)L_+} \right) + 2 \right) \\ &= 8 \|V_l\|_\infty \left(\sqrt{d_A (d_A + 1)L_+} + 2 \right). \end{aligned} \quad (\text{K.14})$$

In line four, we use ineq. (J.19). In line seven, we use the definition of L_+ from eq. (3.10). The Lipschitz constant is therefore

$$\eta_g = 8 \|V_l\|_\infty \left(\sqrt{d_A (d_A + 1)L_+} + 2 \right). \quad (\text{K.15})$$

K.2 Maximally scrambling unitaries

Similar to appendix J.2, let U and U_S be maximally scrambling, let d_{tot} be large, and take $d_D \gg d_A$. Then $\overline{\text{OTOC}}(U) = \overline{\text{OTOC}}(U_S) = \frac{1}{d_A^2}$. The Lipschitz constant is

$$\begin{aligned} \eta_g &= 8 \|V_l\|_\infty \left(\frac{\sqrt{d_A d_D}}{d_B} \frac{2}{d_A} + 2 \right) \\ &= 16 \|V_l\|_\infty \left(\frac{\sqrt{d_A}}{d_C} + 1 \right). \end{aligned} \quad (\text{K.16})$$

Levy's lemma then gives the following concentration inequality:

$$\begin{aligned}
 |\partial_{\theta_l} L_d(|\psi_{d,A}) - \partial_{\theta_l} L| &\leq \eta_g f(\epsilon) \\
 &= 16 \|V_l\|_{\infty} \left(\frac{\sqrt{d_A}}{d_C} + 1 \right) \sqrt{\frac{9\pi^3}{2d_A} \ln \frac{2}{\epsilon}} \\
 &= 16 \|V_l\|_{\infty} \left(\frac{1}{d_C} + \frac{1}{\sqrt{d_A}} \right) \sqrt{\frac{9\pi^3}{2} \ln \frac{2}{\epsilon}}.
 \end{aligned} \tag{K.17}$$

K.3 Vanishing gradient

We bound $|\partial_{\theta_l} L|$:

$$\begin{aligned}
 |\partial_{\theta_l} L| &= G \left| \partial_{\theta_l} \overline{\text{OTOC}}(U(\boldsymbol{\theta})) - 2\partial_{\theta_l} \text{OP}(U(\boldsymbol{\theta}), U_S) \right| \\
 &\leq G \left[\left| \partial_{\theta_l} \overline{\text{OTOC}}(U(\boldsymbol{\theta})) \right| + 2 \left| \partial_{\theta_l} \text{OP}(U(\boldsymbol{\theta}), U_S) \right| \right].
 \end{aligned} \tag{K.18}$$

We bound the first term:

$$\begin{aligned}
 &|\partial_{\theta_l} \overline{\text{OTOC}}(U)| \\
 &= \left| \mathbb{E}_{O_A O_D} \mathbb{E} \partial_{\theta_l} \langle (U O_A U^\dagger O_D)^2 \rangle \right| \\
 &= 2 \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle (U O_A U^\dagger O_D) \partial_{\theta_l} (U O_A U^\dagger O_D) \rangle \right| \\
 &= 2 \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle (U O_A U^\dagger O_D) ((\partial_{\theta_l} U) O_A U^\dagger O_D + U O_A (\partial_{\theta_l} U^\dagger) O_D) \rangle \right| \\
 &= 2 \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle (U O_A U^\dagger O_D) (U_+ (-iV_l) U_- O_A U^\dagger O_D + U O_A U_-^\dagger (iV_l) U_+^\dagger O_D) \rangle \right| \\
 &= 2 \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle (U_+ U_- O_A U_-^\dagger U_+^\dagger O_D) (U_+ (-iV_l) U_- O_A U_-^\dagger U_+^\dagger O_D + U_+ U_- O_A U_-^\dagger (iV_l) U_+^\dagger O_D) \rangle \right| \\
 &= 2 \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle (U_- O_A U_-^\dagger U_+^\dagger O_D U_+) ((-iV_l) U_- O_A U_-^\dagger U_+^\dagger O_D U_+ + U_- O_A U_-^\dagger (iV_l) U_+^\dagger O_D U_+) \rangle \right| \\
 &= 2 \left| \mathbb{E}_{O_A O_D} \mathbb{E} (T_1 + T_2) \right| \\
 &\leq 2 \mathbb{E}_{O_A O_D} \mathbb{E} (|T_1| + |T_2|),
 \end{aligned} \tag{K.19}$$

where we define

$$\begin{aligned}
 T_1 &= \langle U_- O_A U_-^\dagger U_+^\dagger O_D U_+ (-iV_l) U_- O_A U_-^\dagger U_+^\dagger O_D U_+ \rangle, \\
 T_2 &= \langle U_- O_A U_-^\dagger U_+^\dagger O_D U_+ U_- O_A U_-^\dagger (iV_l) U_+^\dagger O_D U_+ \rangle.
 \end{aligned} \tag{K.20}$$

Now bound

$$\begin{aligned}
 |T_1| &= \left| \langle U_- O_A U_-^\dagger U_+^\dagger O_D U_+ (-iV_l) U_- O_A U_-^\dagger U_+^\dagger O_D U_+ \rangle \right| \\
 &\leq \left\| U_- O_A U_-^\dagger U_+^\dagger O_D U_+ (-iV_l) U_- O_A U_-^\dagger U_+^\dagger O_D U_+ \right\|_{\infty} \\
 &= \|V_l\|_{\infty}.
 \end{aligned} \tag{K.21}$$

The third line follows from the invariance of the Schatten norm under unitaries, and also noting that all operators but V_l are unitary. Similarly, $T_2 \leq \|V_l\|_\infty$. We therefore have

$$|\partial_{\theta_l} \overline{\text{OTOC}}(U)| \leq 4 \|V_l\|_\infty. \quad (\text{K.22})$$

We now bound

$$\begin{aligned} & |\partial_{\theta_l} \text{OP}(U(\boldsymbol{\theta}), U_S)| \\ &= \left| \mathbb{E}_{O_A O_D} \mathbb{E} \partial_{\theta_l} \langle U O_A U^\dagger O_D U_S O_A U_S^\dagger O_D \rangle \right| \\ &= \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle ((\partial_{\theta_l} U) O_A U^\dagger + U O_A (\partial_{\theta_l} U^\dagger)) O_D U_S O_A U_S^\dagger O_D \rangle \right| \\ &= \left| \mathbb{E}_{O_A O_D} \mathbb{E} \langle (U_+ (-iV_l) U_- O_A U_-^\dagger U_+^\dagger + U_+ U_- O_A U_-^\dagger (iV_l) U_+^\dagger) O_D U_S O_A U_S^\dagger O_D \rangle \right| \quad (\text{K.23}) \\ &= \left| \mathbb{E}_{O_A O_D} \mathbb{E} (\tilde{T}_1 + \tilde{T}_2) \right| \\ &\leq \mathbb{E}_{O_A O_D} \mathbb{E} (|\tilde{T}_1| + |\tilde{T}_2|), \end{aligned}$$

where

$$\begin{aligned} \tilde{T}_1 &= \langle U_+ (-iV_l) U_- O_A U_-^\dagger U_+^\dagger O_D U_S O_A U_S^\dagger O_D \rangle, \\ \tilde{T}_2 &= \langle U_+ U_- O_A U_-^\dagger (iV_l) U_+^\dagger O_D U_S O_A U_S^\dagger O_D \rangle. \end{aligned} \quad (\text{K.24})$$

We can bound $|\tilde{T}_1|, |\tilde{T}_2| \leq \|V_l\|_\infty$. This produces

$$|\partial_{\theta_l} \text{OP}(U(\boldsymbol{\theta}), U_S)| \leq 2 \|V_l\|_\infty. \quad (\text{K.25})$$

Ineq. (K.18) becomes

$$\begin{aligned} |\partial_{\theta_l} L| &\leq G(4 \|V_l\|_\infty + 2(2 \|V_l\|_\infty)) \\ &= \frac{8d_A^2}{(d_A + 1)d_C^2} \|V_l\|_\infty. \end{aligned} \quad (\text{K.26})$$

This upper bound vanishes as d_C increases.

L Proof of Lemma 1

Fix input state $\rho_{d'}$ and define $\tilde{y}_{d'} = \text{Tr} \{U \rho_{d'} U^\dagger O_C\}$. The cost function is

$$\begin{aligned} \mathcal{C} &= \mathbb{E}_{O_C} \tilde{y}_{d'}^2 \\ &= \mathbb{E}_{O_C} \text{Tr} \{U \rho_{d'} U^\dagger O_C\}^2 \\ &= d_D^2 C_{d'}(U, U). \end{aligned} \quad (\text{L.1})$$

The third line follows from eq. (G.11). We will use Levy's lemma to construct a concentration inequality for \mathcal{C} . First, we compute its average over all possible input states:

$$\begin{aligned}
 \mathcal{C}_{\text{av}} &= \int_{\text{Haar}} dU_A \mathcal{C} \\
 &= d_D^2 \int_{\text{Haar}} dU_A C_{d'}(U, U) \\
 &= \frac{d_D^2}{d_A(d_A + 1)d_B^2} \left[1 + d_A \overline{\text{OTOC}}(U) \right] \\
 &= \frac{d_A^2}{(d_A + 1)d_C^2} \left[\frac{1}{d_A} + \overline{\text{OTOC}}(U) \right] \\
 &= G \left[\frac{1}{d_A} + \overline{\text{OTOC}}(U) \right].
 \end{aligned} \tag{L.2}$$

The third line follows from eq. (G.21).

M Proof of Proposition 4

By applying Levy's lemma to \mathcal{C} , we can show that with probability at least $1 - \epsilon$,

$$|\mathcal{C} - \mathcal{C}_{\text{av}}| \leq \eta_{\mathcal{C}} f(\epsilon), \tag{M.1}$$

where $\eta_{\mathcal{C}}$ is the Lipschitz constant and $f(\epsilon)$ is given by eq. (3.14). As in appendix J, we compute the Lipschitz constant:

$$\begin{aligned}
 \left\| \frac{d}{d\mathbf{w}_d} \mathcal{C} \right\|_2 &= \left\| \frac{d}{d\mathbf{w}_d} \mathbb{E}_{\mathcal{O}_C} \tilde{y}_{d'}^2 \right\|_2 \\
 &= \left\| 2 \mathbb{E}_{\mathcal{O}_C} \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 \\
 &\leq 2 \mathbb{E}_{\mathcal{O}_C} |\tilde{y}_{d'}| \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 \\
 &\leq 2 \mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 \\
 &\leq \frac{4\sqrt{d_A}d_D}{d_B} \sqrt{\overline{\text{OTOC}}(U)} \\
 &= \frac{4\sqrt{d_A}d_A}{d_C} \sqrt{\overline{\text{OTOC}}(U)}.
 \end{aligned} \tag{M.2}$$

The fifth line follows from ineq. (J.19). This produces the Lipschitz constant:

$$\eta_{\mathcal{C}} = \frac{4\sqrt{d_A}d_A}{d_C} \sqrt{\overline{\text{OTOC}}(U)}. \tag{M.3}$$

An application of Levy's lemma to $\partial_{\theta_i} \mathcal{C}$ shows that with probability at least $1 - \epsilon$,

$$|\partial_{\theta_i} \mathcal{C} - \partial_{\theta_i} \mathcal{C}_{\text{av}}| \leq \eta_{\mathcal{C},g} f(\epsilon), \tag{M.4}$$

where $\eta_{\mathcal{C},g}$ is the Lipschitz constant. We compute the Lipschitz constant:

$$\begin{aligned}
\left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \mathcal{C} \right\|_2 &= \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \mathbb{E}_{\mathcal{O}_C} \tilde{y}_{d'}^2 \right\|_2 \\
&= \left\| \frac{d}{d\mathbf{w}_d} \left(2 \mathbb{E}_{\mathcal{O}_C} \tilde{y}_{d'} \partial_{\theta_i} \tilde{y}_{d'} \right) \right\|_2 \\
&= 2 \left\| \mathbb{E}_{\mathcal{O}_C} \left(\partial_{\theta_i} \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} + \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right) \right\|_2 \\
&\leq 2 \mathbb{E}_{\mathcal{O}_C} \left\| \partial_{\theta_i} \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} + \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right\|_2 \\
&\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(\left\| \partial_{\theta_i} \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 + \left\| \tilde{y}_{d'} \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right\|_2 \right) \\
&= 2 \mathbb{E}_{\mathcal{O}_C} \left(|\partial_{\theta_i} \tilde{y}_{d'}| \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 + |\tilde{y}_{d'}| \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right\|_2 \right) \\
&\leq 2 \mathbb{E}_{\mathcal{O}_C} \left(|\partial_{\theta_i} \tilde{y}_{d'}| \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 + \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right\|_2 \right) \\
&\leq 2 \left(2 \|V_l\|_\infty \mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} \tilde{y}_{d'} \right\|_2 + \mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right\|_2 \right) \\
&\leq 2 \left(2 \|V_l\|_\infty \left(\frac{2\sqrt{d_A} d_D}{d_B} \sqrt{\text{OTOC}(U)} \right) + \mathbb{E}_{\mathcal{O}_C} \left\| \frac{d}{d\mathbf{w}_d} \partial_{\theta_i} \tilde{y}_{d'} \right\|_2 \right) \\
&\leq 2 \left(4 \|V_l\|_\infty \frac{\sqrt{d_A} d_D}{d_B} \sqrt{\text{OTOC}(U)} + 4 \|V_l\|_\infty \right) \\
&= 8 \|V_l\|_\infty \left(\frac{\sqrt{d_A} d_A}{d_C} \sqrt{\text{OTOC}(U)} + 1 \right).
\end{aligned} \tag{M.5}$$

Lines eight, nine, and ten follow from ineq. (K.13), ineq. (J.19), and ineq. (K.12), respectively. The Lipschitz constant is therefore

$$\eta_{\mathcal{C},g} = 8 \|V_l\|_\infty \left(\frac{\sqrt{d_A} d_A}{d_C} \sqrt{\text{OTOC}(U)} + 1 \right). \tag{M.6}$$

Open Access. This article is distributed under the terms of the Creative Commons Attribution License ([CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

References

- [1] M. Schuld, I. Sinayskiy and F. Petruccione, *Simulating a perceptron on a quantum computer*, *Phys. Lett. A* **379** (2015) 660.
- [2] K.H. Wan, O. Dahlsten, H. Kristjánsson, R. Gardner and M.S. Kim, *Quantum generalisation of feedforward neural networks*, *npj Quantum Inf.* **3** (2017) 36.
- [3] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe and S. Lloyd, *Quantum machine learning*, *Nature* **549** (2017) 195.
- [4] E. Farhi and H. Neven, *Classification with quantum neural networks on near term processors*, [arXiv:1802.06002](https://arxiv.org/abs/1802.06002).

- [5] I. Cong, S. Choi and M.D. Lukin, *Quantum convolutional neural networks*, *Nature Phys.* **15** (2019) 1273.
- [6] K. Beer et al., *Training deep quantum neural networks*, *Nature Commun.* **11** (2020) 808.
- [7] Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*, *Nature* **521** (2015) 436.
- [8] J. Tompson, A. Jain, Y. LeCun and C. Bregler, *Joint training of a convolutional network and a graphical model for human pose estimation*, [arXiv:1406.2984](https://arxiv.org/abs/1406.2984).
- [9] A. Krizhevsky, I. Sutskever and G.E. Hinton, *Imagenet classification with deep convolutional neural networks*, *Commun. ACM* **60** (2017) 84.
- [10] G.D. Paparo, V. Dunjko, A. Makmal, M.A. Martin-Delgado and H.J. Briegel, *Quantum speedup for active learning agents*, *Phys. Rev. X* **4** (2014) 031002.
- [11] J. Preskill, *Quantum computing in the nisq era and beyond*, *Quantum* **2** (2018) 79.
- [12] H. Sompolinsky, A. Crisanti and H.J. Sommers, *Chaos in random neural networks*, *Phys. Rev. Lett.* **61** (1988) 259.
- [13] L.P. Wang, E.E. Pichler and J. Ross, *Oscillations and chaos in neural networks: an exactly solvable model*, *Proc. Nat. Acad. Sci.* **87** (1990) 9467.
- [14] A. Potapov and M. Ali, *Robust chaos in neural networks*, *Phys. Lett. A* **277** (2000) 310.
- [15] J. Pathak, B. Hunt, M. Girvan, Z. Lu and E. Ott, *Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach*, *Phys. Rev. Lett.* **120** (2018) 024102.
- [16] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein and S. Ganguli, *Exponential expressivity in deep neural networks through transient chaos*, [arXiv:1606.05340](https://arxiv.org/abs/1606.05340).
- [17] Y.A. Kharkov, V.E. Sotskov, A.A. Karazeev, E.O. Kiktenko and A.K. Fedorov, *Revealing quantum chaos with machine learning*, *Phys. Rev. B* **101** (2020) 064406.
- [18] S. Choudhury, A. Dutta and D. Ray, *Chaos and Complexity from Quantum Neural Network: A study with Diffusion Metric in Machine Learning*, *JHEP* **04** (2021) 138 [[arXiv:2011.07145](https://arxiv.org/abs/2011.07145)] [[INSPIRE](https://inspirehep.net/literature/2011.07145)].
- [19] Y. Wu, J. Yao, P. Zhang and H. Zhai, *Expressivity of quantum neural networks*, [arXiv:2101.04273](https://arxiv.org/abs/2101.04273).
- [20] S.H. Shenker and D. Stanford, *Black holes and the butterfly effect*, *JHEP* **03** (2014) 067 [[arXiv:1306.0622](https://arxiv.org/abs/1306.0622)] [[INSPIRE](https://inspirehep.net/literature/1306.0622)].
- [21] R.J. Lewis-Swan, A. Safavi-Naini, A.M. Kaufman and A.M. Rey, *Dynamics of quantum information*, *Nature Rev. Phys.* **1** (2019) 627.
- [22] R.J. Lewis-Swan, A. Safavi-Naini, J.J. Bollinger and A.M. Rey, *Unifying scrambling, thermalization and entanglement through measurement of fidelity out-of-time-order correlators in the Dicke model*, *Nature Commun.* **10** (2019) 1581 [Erratum *ibid.* **10** (2019) 5007] [[arXiv:1808.07134](https://arxiv.org/abs/1808.07134)] [[INSPIRE](https://inspirehep.net/literature/1808.07134)].
- [23] B. Swingle, G. Bentsen, M. Schleier-Smith and P. Hayden, *Measuring the scrambling of quantum information*, *Phys. Rev. A* **94** (2016) 040302 [[arXiv:1602.06271](https://arxiv.org/abs/1602.06271)] [[INSPIRE](https://inspirehep.net/literature/1602.06271)].
- [24] S. Xu and B. Swingle, *Locality, Quantum Fluctuations, and Scrambling*, *Phys. Rev. X* **9** (2019) 031048 [[arXiv:1805.05376](https://arxiv.org/abs/1805.05376)] [[INSPIRE](https://inspirehep.net/literature/1805.05376)].

- [25] Z.-W. Liu, S. Lloyd, E.Y. Zhu and H. Zhu, *Entanglement, quantum randomness, and complexity beyond scrambling*, *JHEP* **07** (2018) 041 [[arXiv:1703.08104](#)] [[INSPIRE](#)].
- [26] L. Leone, S.F.E. Oliviero, Y. Zhou and A. Hamma, *Quantum chaos is quantum*, *Quantum* **5** (2021) 453.
- [27] Z. Holmes, A. Arrasmith, B. Yan, P.J. Coles, A. Albrecht and A.T. Sornborger, *Barren Plateaus Preclude Learning Scramblers*, *Phys. Rev. Lett.* **126** (2021) 190501 [[arXiv:2009.14808](#)] [[INSPIRE](#)].
- [28] H. Shen, P. Zhang, Y.-Z. You and H. Zhai, *Information Scrambling in Quantum Neural Networks*, *Phys. Rev. Lett.* **124** (2020) 200504 [[arXiv:1909.11887](#)] [[INSPIRE](#)].
- [29] Y. Wu, P. Zhang and H. Zhai, *Scrambling ability of quantum neural networks architectures*, [arXiv:2011.07698](#).
- [30] S. Sachdev and J. Ye, *Gapless spin fluid ground state in a random, quantum Heisenberg magnet*, *Phys. Rev. Lett.* **70** (1993) 3339 [[cond-mat/9212030](#)] [[INSPIRE](#)].
- [31] A. Kitaev, *A simple model of quantum holography (part 1)*, talk at KITP, April 7, 2015, <http://online.kitp.ucsb.edu/online/entangled15/kitaev/>.
- [32] A. Kitaev, *A simple model of quantum holography (part 2)*, talk at KITP, May 27, 2015, <http://online.kitp.ucsb.edu/online/entangled15/kitaev2/>.
- [33] Y. Sekino and L. Susskind, *Fast Scramblers*, *JHEP* **10** (2008) 065 [[arXiv:0808.2096](#)] [[INSPIRE](#)].
- [34] J. Maldacena, S.H. Shenker and D. Stanford, *A bound on chaos*, *JHEP* **08** (2016) 106 [[arXiv:1503.01409](#)] [[INSPIRE](#)].
- [35] P. Hayden and J. Preskill, *Black holes as mirrors: Quantum information in random subsystems*, *JHEP* **09** (2007) 120 [[arXiv:0708.4025](#)] [[INSPIRE](#)].
- [36] B. Yoshida and A. Kitaev, *Efficient decoding for the hayden-preskill protocol*, [arXiv:1710.03363](#).
- [37] D.A. Roberts, D. Stanford and L. Susskind, *Localized shocks*, *JHEP* **03** (2015) 051 [[arXiv:1409.8180](#)] [[INSPIRE](#)].
- [38] B. Swingle and D. Chowdhury, *Slow scrambling in disordered quantum systems*, *Phys. Rev. B* **95** (2017) 060201 [[arXiv:1608.03280](#)] [[INSPIRE](#)].
- [39] D. Chowdhury and B. Swingle, *Onset of many-body chaos in the $O(N)$ model*, *Phys. Rev. D* **96** (2017) 065005 [[arXiv:1703.02545](#)] [[INSPIRE](#)].
- [40] A.W. Harrow, L. Kong, Z.-W. Liu, S. Mehraban and P.W. Shor, *Separation of out-of-time-ordered correlation and entanglement*, *PRX Quantum* **2** (2021) 020339.
- [41] M. Gärttner, J.G. Bohnet, A. Safavi-Naini, M.L. Wall, J.J. Bollinger and A.M. Rey, *Measuring out-of-time-order correlations and multiple quantum spectra in a trapped-ion quantum magnet*, *Nature Phys.* **13** (2017) 781.
- [42] J. Li et al., *Measuring Out-of-Time-Order Correlators on a Nuclear Magnetic Resonance Quantum Simulator*, *Phys. Rev. X* **7** (2017) 031011 [[arXiv:1609.01246](#)] [[INSPIRE](#)].
- [43] K.A. Landsman et al., *Verified Quantum Information Scrambling*, *Nature* **567** (2019) 61 [[arXiv:1806.02807](#)] [[INSPIRE](#)].

- [44] M.K. Joshi et al., *Quantum information scrambling in a trapped-ion quantum simulator with tunable range interactions*, *Phys. Rev. Lett.* **124** (2020) 240505 [[arXiv:2001.02176](#)] [[INSPIRE](#)].
- [45] R.J. Garcia, Y. Zhou and A. Jaffe, *Quantum scrambling with classical shadows*, *Phys. Rev. Res.* **3** (2021) 033155.
- [46] Z. Webb, *The clifford group forms a unitary 3-design*, *Quant. Inf. Comput.* **16** (2016) 1379.
- [47] H. Zhu, *Multiqubit clifford groups are unitary 3-designs*, *Phys. Rev. A* **96** (2017) 062336.
- [48] B. Yan, L. Cincio and W.H. Zurek, *Information Scrambling and Loschmidt Echo*, *Phys. Rev. Lett.* **124** (2020) 160603 [[arXiv:1903.02651](#)] [[INSPIRE](#)].
- [49] A. Chenu, I.L. Egusquiza, J. Molina-Vilaplana and A. del Campo, *Quantum work statistics, Loschmidt echo and information scrambling*, *Sci. Rep.* **8** (2018) 12634 [[arXiv:1711.01277](#)] [[INSPIRE](#)].
- [50] S.H. Shenker and D. Stanford, *Multiple Shocks*, *JHEP* **12** (2014) 046 [[arXiv:1312.3296](#)] [[INSPIRE](#)].
- [51] J. Iaconis, *Quantum state complexity in computationally tractable quantum circuits*, *PRX Quantum* **2** (2021) 010329.
- [52] D.A. Roberts and B. Yoshida, *Chaos and complexity by design*, *JHEP* **04** (2017) 121 [[arXiv:1610.04903](#)] [[INSPIRE](#)].
- [53] T. Zhou and X. Chen, *Operator dynamics in a Brownian quantum circuit*, *Phys. Rev. E* **99** (2019) 052212 [[arXiv:1805.09307](#)] [[INSPIRE](#)].
- [54] V. Khemani, A. Vishwanath and D.A. Huse, *Operator spreading and the emergence of dissipation in unitary dynamics with conservation laws*, *Phys. Rev. X* **8** (2018) 031057 [[arXiv:1710.09835](#)] [[INSPIRE](#)].
- [55] A. Nahum, S. Vijay and J. Haah, *Operator Spreading in Random Unitary Circuits*, *Phys. Rev. X* **8** (2018) 021014 [[arXiv:1705.08975](#)] [[INSPIRE](#)].
- [56] F.G.S.L. Brandão, A.W. Harrow and M. Horodecki, *Local Random Quantum Circuits are Approximate Polynomial-Designs*, *Commun. Math. Phys.* **346** (2016) 397 [[arXiv:1208.0692](#)] [[INSPIRE](#)].
- [57] A.W. Harrow and R.A. Low, *Random quantum circuits are approximate 2-designs*, *Commun. Math. Phys.* **291** (2009) 257.
- [58] P. Hosur, X.-L. Qi, D.A. Roberts and B. Yoshida, *Chaos in quantum channels*, *JHEP* **02** (2016) 004 [[arXiv:1511.04021](#)] [[INSPIRE](#)].
- [59] J. Duchi, E. Hazan and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, *J. Mach. Learn. Res.* **12** (2011) 2121.
- [60] S.J. Reddi, S. Kale and S. Kumar, *On the convergence of adam and beyond*, [arXiv:1904.09237](#).
- [61] J.R. McClean, S. Boixo, V.N. Smelyanskiy, R. Babbush and H. Neven, *Barren plateaus in quantum neural network training landscapes*, *Nature Commun.* **9** (2018) 4812.
- [62] M. Cerezo, A. Sone, T. Volkoff, L. Cincio and P.J. Coles, *Cost function dependent barren plateaus in shallow parametrized quantum circuits*, *Nature Commun.* **12** (2021) 1791.

- [63] Z. Liu, L.-W. Yu, L.M. Duan and D.-L. Deng, *The presence and absence of barren plateaus in tensor-network based machine learning*, [arXiv:2108.08312](#).
- [64] S. Wang et al., *Noise-induced barren plateaus in variational quantum algorithms*, [arXiv:2007.14384](#).
- [65] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A.T. Sornborger and P.J. Coles, *Absence of Barren Plateaus in Quantum Convolutional Neural Networks*, *Phys. Rev. X* **11** (2021) 041011 [[arXiv:2011.02966](#)] [[INSPIRE](#)].
- [66] B. Yoshida and N.Y. Yao, *Disentangling Scrambling and Decoherence via Quantum Teleportation*, *Phys. Rev. X* **9** (2019) 011006 [[arXiv:1803.10772](#)] [[INSPIRE](#)].