Variational microcanonical estimator

Klée Pollock,[1] Peter P. Orth,[1,2,3] and Thomas Iadecola[1,2,*]
[1]*Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA*
[2]*Ames National Laboratory, Ames, Iowa 50011, USA*
[3]*Department of Physics, Saarland University, 66123 Saarbrücken, Germany*

(Received 1 February 2023; accepted 21 August 2023; published 28 September 2023)

We propose a variational quantum algorithm for estimating microcanonical expectation values in models obeying the eigenstate thermalization hypothesis. Using a relaxed criterion for convergence of the variational optimization loop, the algorithm generates weakly entangled superpositions of eigenstates at a given target energy density. An ensemble of these variational states is then used to estimate microcanonical averages of local operators, with an error whose dominant contribution decreases initially as a power law in the size of the ensemble and is ultimately limited by a small bias. We apply the algorithm to the one-dimensional mixed-field Ising model, where it converges for ansatz circuits of depth roughly linear in system size. The most accurate thermal estimates are produced for intermediate energy densities. In our error analysis, we find connections with recent works investigating the underpinnings of the eigenstate thermalization hypothesis. In particular, the failure of energy-basis matrix elements of local operators to behave as *independent* random variables is a potential source of error that the algorithm can overcome by averaging over an ensemble of variational states.

DOI: 10.1103/PhysRevResearch.5.033224

I. INTRODUCTION

Calculating the ground state and thermal equilibrium properties of large and complex quantum systems remains a central task in contemporary quantum physics. While for integrable systems analytical techniques can often solve the problem, in generic nonintegrable systems such methods do not apply. In the last two decades, however, efficient numerical methods have been developed to calculate ground-state and thermal properties in settings where the target state is only modestly entangled. Tensor network (TN) methods exploit the locality of physical Hamiltonians, in particular their area-law entangled ground states [1], to find efficient representations of the wavefunction via truncated matrix product states on classical hardware [2]. Additionally, these efficient representations can be extended to Gibbs states at finite temperature via matrix product operators [3]. Examples of algorithms based on TNs include the minimally entangled typical thermal state (METTS) algorithm [4] for estimating canonical averages, and an algorithm for estimating microcanonical averages using time-evolving block decimation (TEBD) [5]. In higher than one spatial dimension, however, the TN contraction step becomes hard [6], so that classical algorithms may not be sufficient for the simulation of even weakly entangled quantum systems.

It has long been believed that quantum computers are the natural platform to simulate quantum systems [7], but to exploit their full power it is likely that deep quantum circuits and error correction will be required. Currently, we have noisy intermediate-scale quantum (NISQ) devices that cannot yet implement deep circuits with high fidelity, but which can still demonstrate the potential for quantum computing in cases where low-depth circuits are sufficient [8]. There is thus a significant need to develop algorithms that can take advantage of these NISQ devices.

Originating with the variational quantum eigensolver (VQE) [9], one class of algorithms that can potentially achieve this goal in some cases are the hybrid quantum-classical variational quantum algorithms (VQAs) [10–12], which employ a digital quantum computer aided by a classical optimizer. Although generic VQAs suffer from the well-known barren plateau problem [13–15] which suggests unscalability in full generality, there is evidence that VQAs can calculate the ground state of certain Hamiltonians using only polynomial quantum resources, e.g., by using the Hamiltonian variational ansatz for the transverse field Ising model [16]. Recent works have also considered using VQAs to prepare Gibbs states using cost functions such as the relative entropy or relative free energy between the current state and target state [17]; strategies to overcome the costly evaluation of the entropic term have also been proposed [18,19]. Other finite-temperature VQAs prepare thermofield-double (TFD) states, which require doubling the number of qubits in the physical system being simulated—for example, the algorithm of Ref. [20] can prepare the TFD state of free fermions efficiently at any inverse temperature. Alternative quantum algorithms for preparing thermal states include a quantum version of the minimally entangled typical thermal states algorithm (QMETTS algorithm) that involves imaginary time evolution on quantum hardware [21], and an algorithm based on random quantum circuits [22].

In this work, we task a VQA with calculating microcanonical averages of local observables in a one-dimensional (1D)

*iadecola@iastate.edu

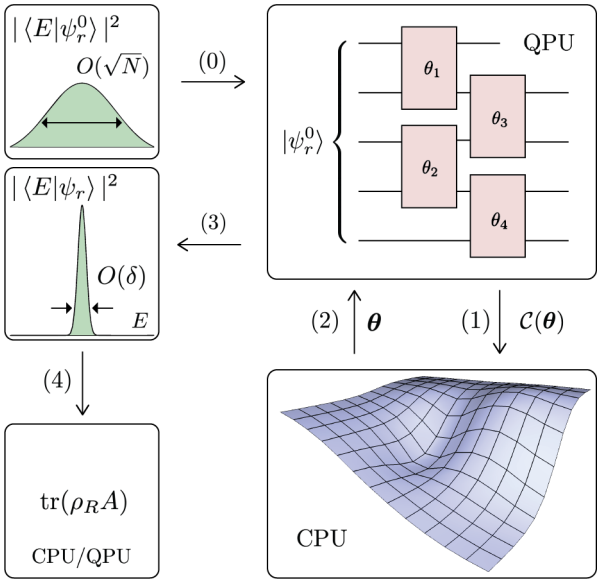

FIG. 1. The VME algorithm. In step (0), the QPU is initialized in a random product state $|\psi_r^0\rangle$ $(r = 1, \ldots, R)$. The VQA repeats steps (1) and (2) that optimize the cost function $C(\boldsymbol{\theta})$ in Eq. (1.1) to "squeeze" the state onto a microcanonical window of size δ as shown in step (3). Steps (0–3) are repeated to produce a pseudorandom ensemble of states $|\psi_r\rangle$ which for large N and R can be used to approximate microcanonical averages of local operators A as in step (4), where $\rho_R = \frac{1}{R}\sum_r |\psi_r\rangle\langle\psi_r|$.

nonintegrable spin model. Our work is partially inspired by analog quantum simulation [23] and classical tensor network [24] algorithms for estimating microcanonical properties. The algorithm takes advantage of the eigenstate thermalization hypothesis (ETH), in particular the "diagonal" ETH which states that in a nonintegrable model the energy-basis diagonal matrix elements $\langle E|A|E\rangle$ of an observable A approach a smooth function $A(E)$ in the thermodynamic limit [25,26].

The algorithm, which we call the variational microcanonical estimator (VME), works as follows (see Fig. 1). We initialize the QPU in a random product state [step (0)] $|\psi_r^0\rangle$, whose energy variance is typically extensive in N (the number of sites) [24]. Given a target energy λ and microcanonical window size δ, a classical optimizer is then tasked with minimizing the cost function

$$\mathcal{C}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,(H-\lambda)^2\,|\psi(\boldsymbol{\theta})\rangle \tag{1.1}$$

[steps (1) and (2)] originally proposed in Ref. [9]. However, instead of trying to reach a local or global minimum, we stop the optimization as soon as $\mathrm{Var}(H) = \langle(H - \langle H\rangle)^2\rangle \leqslant \delta^2$. This produces states whose energy support is roughly limited to the microcanonical window of interest [step (3)], and the resulting variational states $|\psi_r\rangle$ are then used to compute the expectation of a local observable A by averaging $\langle\psi_r|A|\psi_r\rangle$ over R variational states [step (4)]. The ensemble average in step (4) enables a parametric reduction in the error and is essential to the algorithm's performance.

We benchmark the VME algorithm on a nonintegrable Ising chain by comparing its estimates for local observables to corresponding Gaussian microcanonical ensemble predictions obtained from exact diagonalization (ED). Using numerical evidence in combination with the phenomenology of ETH, we conjecture that for local operators A and target energies λ in the bulk of the spectrum, the absolute error in the VME algorithm scales as

$$\epsilon_R \simeq |c| + O(R^{-1/2}) + O(\delta/N) + O(\mathcal{D}^{-1/2}(\lambda)). \tag{1.2}$$

Here, $\mathcal{D}(\lambda)$ is the density of states at the target energy λ, δ is the microcanonical window width, N is the system size, and $c \ll 1$ is a small empirical constant whose magnitude depends on A and other problem parameters. The last two terms in this formula are predicted by ETH and the first two terms we give a phenomenological argument for that we substantiate with numerical evidence.

We then generalize the problem to the reduced state of small subsystems of the chain and find numerically that when choosing $R = O(N^2)$ and for certain λ, the VME appears to approach the corresponding microcanonical state in the thermodynamic limit. The states prepared by the VME are consistent with area-law entanglement for a fixed N, and require roughly linearly deep quantum circuits to prepare. We find that every random initial product state is able to converge, which we attribute to the fact that the algorithm does not seek global minima of the cost function. An additional distinction from other current VQAs for preparing mixed states is that we prepare pure states one at a time, thus avoiding storage of a large ensemble of pure quantum states in a quantum memory. The smallness of the trace distance when choosing $R = O(N^2)$ implies that the microcanonical ensemble, which involves at least one (via ETH) highly entangled (i.e., volume-law) eigenstate is approximately indistinguishable *by local operators* from a polynomially large ensemble of weakly entangled variational states.

The paper is organized as follows. In Sec. II, we introduce (i) the statement of ETH and (ii) a class of states which might be called microcanonical superposition states, which our converged variational states fall under. We then review related works attempting to use these states to estimate thermal averages and the relationship of this problem to ETH. In Sec. III we discuss how averaging over an ensemble of these microcanonical superposition states could significantly improve how well they can estimate microcanonical averages, and then we detail the VME algorithm which can produce these states. Finally in Sec. IV we present the numerical results for the form of the variational ensemble, the error in the algorithm for various local operators, the observable independent trace distance, and finally the quantum resources like circuit depth and entanglement.

II. MOTIVATION

A. Eigenstate thermalization hypothesis

Here we review relevant aspects of the ETH and some recent works which attempt to exploit it to estimate thermal averages. We assume a nonintegrable (i.e., chaotic) Hamiltonian H which has a nondegenerate energy spectrum so that its eigenstates $|E\rangle$ are uniquely labeled by their energies E. Furthermore, we will assume that all operators and states of interest are real in the energy basis for simplicity. The variant of ETH we consider was formulated in Ref. [27] and proposes

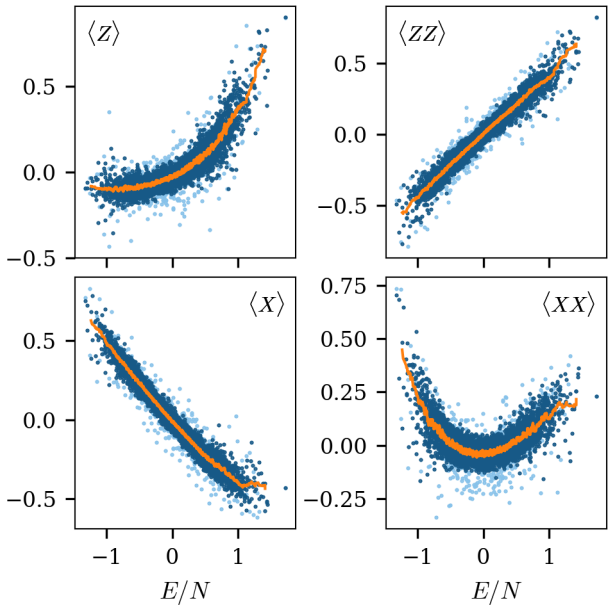

FIG. 2. The energy-basis diagonal matrix elements $\langle E|A|E\rangle$ of various local observables A acting in the middle of the chain, plotted against energy density in the nonintegrable 1D mixed-field Ising model, Eq. (4.1), with parameters $J = 1$, $h_x = -1.05$, and $h_z = -0.5$. Lighter blue colored points are for system size $N = 9$ and darker blue points are for $N = 13$. Orange curves are coarse-grained versions of the $N = 13$ scatter plots which define the "smooth" function $A(E)$ in the thermodynamic limit.

that in a quantum chaotic system, the energy-basis matrix elements of observables have the form

$$\langle E|\,A\,|E'\rangle = \delta_{EE'}A(\bar{E}) + \mathcal{D}^{-1/2}(\bar{E})f(\bar{E},\omega)R_{EE'}, \qquad (2.1)$$

where $\bar{E} = (E + E')/2$, $\omega = E - E'$, $\mathcal{D}(\bar{E})$ is the density of states at energy $\bar{E}$, $A(\bar{E})$ and $f(\bar{E},\omega)$ approach smooth functions in the thermodynamic limit, and $R_{EE'}$ are order-one fluctuations. Examples of such functions $A(\bar{E})$ are shown in Fig. 2 which demonstrates this for local spin operators in the 1D mixed-field Ising model (defined in Sec. IV).

The ansatz (2.1) captures several features of such matrix elements that have been observed in numerical studies. First, because the density of states is exponentially large in system size, the off-diagonal matrix elements are exponentially small. Second, the smooth function $A(E)$ is related to the statistical mechanical prediction for $\langle A\rangle$ at average energy E; this function will play a central role in our algorithm. Finally, the function $f(\bar{E},\omega)$ controls the approach to thermal equilibrium and is related to other spectral properties of the observable [28]; this function figures less prominently in our analysis.

To see how $A(E)$ is related to a thermal average, consider for example a broadened microcanonical ensemble $\rho_{\lambda,\delta}$ centered on energy $E = \lambda$ and of width $O(\delta)$ which we will define more precisely at beginning of Sec. IV. Under certain assumptions about the density of states of the model and away from $\lambda = 0$ (which corresponds to infinite temperature), and assuming the ETH ansatz (2.1), we have in the thermodynamic limit that (see Appendix C for details)

$$A(\lambda) = \langle A\rangle_{\rm mc} + O(\delta^2/N) + O(\mathcal{D}^{-1/2}(\lambda)), \qquad (2.2)$$

where $\langle A\rangle_{\rm mc} = \mathrm{tr}\,\rho_{\lambda,\delta}A$. The ETH thus suggests that, if one could prepare even a single eigenstate $|\lambda\rangle$ of the Hamiltonian with energy λ, then one could accurately estimate thermal averages in sufficiently large systems. However, for a nonintegrable Hamiltonian, a generic excited eigenstate is volume-law entangled, and thus cannot efficiently be prepared by classical algorithms or by VQE-type algorithms [29]. Thus, this feature of ETH does not appear practically useful, expect perhaps in the case of an error corrected quantum computer.

B. Microcanonical superpositions

An alternative approach to using exact eigenstates for computing thermal averages is using pure states of the form

$$|\psi\rangle = \sum_E c_E\,|E\rangle\,, \qquad (2.3)$$

where either c_E are exactly zero outside the energy window defined by $|E - \lambda| \leqslant \delta$, or the states satisfy the weaker condition that $\langle\psi|(H - \lambda)^2|\psi\rangle = O(\delta^2)$. We refer to states of this type as "microcanonical superposition states" and they have been studied in the context of thermal pure quantum (TPQ) states [30], the foundations of quantum statistical mechanics [31,32], algorithms for analog quantum simulators [23], and tensor network algorithms [24].

The practical reason for considering these states is that they appear to be significantly less entangled than exact eigenstates. In fact, there exist matrix product state-based numerical constructions of them such that the maximum entanglement entropy across any cut scales as $k/\delta + \log_2\sqrt{N}$ for some constant k [24] and N being the system size. Thus, by choosing $\delta = O(1/\log_2 N)$, such states can have only $O(\log_2 N)$ entanglement, whereas a single excited eigenstate of a nonintegrable system is expected to have $O(N)$ entanglement. In this work, by choosing $\delta = O(N^{-1/2})$ (for the values of N studied in this paper $N^{-1/2} \approx 1/\log_2 N$) we find a VQA can generate these states using roughly linear circuit depth and which have area-law entanglement for fixed N. In Ref. [24] and in our findings it is clear that generically a smaller δ requires more computational effort.

It is known that if the coefficients c_E are generic, and δ is subextensive in N; then when a state of the form (2.3) is evolved under H, it approaches a state in which small subsystems are approximately thermal [27,33]. Given this fact, one may wonder if a relation like (2.2) holds with $A(\lambda)$ replaced by $\langle\psi|A|\psi\rangle$, just as it did for $|\lambda\rangle$. A key issue, however, is that although the off-diagonal elements of a generic operator are exponentially small, the quantity

$$\langle\psi|A|\psi\rangle = \sum_E c_E^2\,\langle E|A|E\rangle + \sum_{E\neq E'} c_E c_{E'}\,\langle E'|A|E\rangle \qquad (2.4)$$

involves summing exponentially many off-diagonal matrix elements, so long as δ itself is not exponentially small [23]. The reason that the long-time evolved state is locally thermal is that the off-diagonal terms become dephased just enough to counteract the exponentially large sum [33]. As a result, the off-diagonal contribution scales as $O(\mathcal{D}^{-1/2}(\lambda))$, like the off-diagonal matrix elements themselves. Without the additional pseudorandom phases $e^{i(E-E')t}$ appearing in the time-evolved

expectation value, for arbitrary δ there is no *a priori* reason to expect $\langle\psi|A|\psi\rangle$ to closely approximate $\langle A\rangle_{\rm mc}$.

On the other hand, in discussions of ETH the off-diagonal fluctuations $R_{EE'}$ are usually stated to behave as random variables [28]. If we take them to be actual *independent* random variables, then the total off-diagonal contribution scales as a random walk and will therefore remain typically of the order $O(\mathcal{D}^{-1/2}(\lambda))$ as shown in Appendix B. However, the validity of such an independence assumption on $R_{EE'}$ is known to depend on the energy scale δ. Sometimes this scale is quoted as $\delta = O(N^{-2})$ which is the scale of $|\omega|$ below which $|f(0,\omega)|$ reaches a plateau, so that the ETH ansatz (at infinite temperature) becomes structureless and reduces to the random-matrix prediction [28]. More recently, however, Refs. [34] and [35] found numerical and analytical evidence that "true" random matrix behavior with effectively independent matrix elements emerges only on the parametrically smaller scale, $\delta = O(N^{-3})$.

Regardless of whether the matrix elements can be treated as independent random variables, it is in general an open question how δ must scale in order for $\langle\psi|A|\psi\rangle$ to converge to the thermal value in the thermodynamic limit. Reference [24] argued that $\delta = O(1/\log_2 N)$ is sufficient for a slow convergence but Ref. [23] argued that $O(N^{-1})$ is needed. Finally, in Ref. [36] it is proposed that in a quantum chaotic system, for a fixed operator A of interest, *every* state of the form (2.3) is thermal with a worst case error x obeying the relation $\delta(x) = \text{poly}(x)$. However, for δ much larger than the random matrix theory scale $O(N^{-2})$ defined above [28], the behavior (and in particular the N scaling) of this polynomial was not completely settled in that work.

In summary, some source of randomness is needed to make the off-diagonal contribution small. In the theory of canonical typicality [31,32], it is the state coefficients; under time evolution it is the effectively random phases; if the window δ is small enough, it is the matrix elements themselves that are effectively random. In this work the source of randomness arises from averaging over an ensemble of variational states $|\psi_r\rangle$ that are prepared by starting from random product states $|\psi_r^0\rangle$. Since we will ultimately use shallow quantum circuits to prepare these variational states, we do not expect them to be typical states on the target microcanonical subspace, nor do we expect them to be typical states in the sense of TPQ states, as in both cases it is likely that deep circuits would be needed to approximate Haar-random states [37,38]. On the other hand, we will see that the states are "random enough" for a certain dephasing mechanism to significantly reduce the off-diagonal contribution in the ensemble-averaged version of Eq. (2.4). Thus we will henceforth refer to the states as pseudorandom, reserving "random" for Haar-random states.

III. VARIATIONAL MICROCANONICAL ESTIMATOR

A. Ensemble of microcanonical superpositions and error analysis

As discussed in Sec. II, we do not necessarily expect a microcanonical superposition state of the form in Eq. (2.3) to closely approximate thermal values when the microcanonical window size δ is too large. Here we adapt known results from the theory of ETH to our problem, and then propose a heuristic mechanism which allows us to capture thermal expectation values using a large pseudorandom ensemble of microcanonical superposition states with "large" (but still subextensive) δ.

For a fixed target energy λ and microcanonical window δ, consider an ensemble of states $\{|\psi_r\rangle\}_{r=1}^R$, each of the form (2.3), and let $\rho_R = \frac{1}{R}\sum_r |\psi_r\rangle\langle\psi_r|$ be their equal weight mixture. We discuss how to prepare these states using a variational algorithm in Sec. III B. In this section, we discuss the error between the actual microcanonical expectation value $\text{tr}(\rho_{\rm mc}A)$ and its average in the ensemble ρ_R. Considering a local operator A, this error can be expressed as

$$\epsilon_R = |\text{tr}(\rho_R A) - \text{tr}(\rho_{\rm mc}A)|. \tag{3.1}$$

Here, we have introduced a microcanonical density matrix $\rho_{\rm mc}$. The microcanonical ensemble could be defined via the standard sharp microcanonical window, or via a smoothed Gaussian version thereof; we will ultimately compare our variational estimates to the latter in Sec. IV. The results of this section do not depend on the exact form, but for now let us take $\rho_{\rm mc} = P_W/n$ with P_W a projector onto the microcanonical window W defined by $|E-\lambda| \leqslant \delta$ and $n = \text{tr}\, P_W$ the number of states in the window. For our purposes, the error (3.1) is best understood as a sum of two distinct types and we therefore decompose it via the triangle inequality as

$$\epsilon_R \leqslant \epsilon_R^{\rm diag} + \epsilon_R^{\rm off}, \tag{3.2a}$$

where

$$\epsilon_R^{\rm diag} = \left|\text{tr}(\rho_{\rm mc}A) - \langle A\rangle_R^{\rm diag}\right|, \tag{3.2b}$$

$$\epsilon_R^{\rm off} = \left|\text{tr}(\rho_R A) - \langle A\rangle_R^{\rm diag}\right|, \tag{3.2c}$$

$$\langle A\rangle_R^{\rm diag} = \sum_E \langle E|A|E\rangle\,\langle E|\rho_R|E\rangle\,. \tag{3.2d}$$

The "diagonal error" $\epsilon_R^{\rm diag}$ captures the difference between the expectation value of A in the microcanonical ensemble and the diagonal ensemble [33] associated with ρ_R. It depends only on diagonal energy-basis matrix elements of A and ρ_R. The "off-diagonal error" $\epsilon_R^{\rm off}$ captures error due to the fact that ρ_R is not diagonal in the energy basis. Plugging Eq. (3.2d) into Eq. (3.2c) yields

$$\epsilon_R^{\rm off} = \sum_{E\neq E'} \langle E|A|E'\rangle\,\langle E'|\rho_R|E\rangle\,, \tag{3.3}$$

which involves only off-diagonal matrix elements of A and ρ_R.

We now consider the dependence of the error ϵ_R on the ensemble size R, window size δ, and system size N, assuming the ETH matrix element ansatz (2.1).

1. Diagonal error $\epsilon_R^{\rm diag}$

We begin with the diagonal contribution $\epsilon_R^{\rm diag}$. Plugging in the ETH ansatz to the expression for the diagonal error gives

$$\epsilon_R^{\rm diag} = \Bigg|\sum_E \langle E|\rho_{\rm mc} - \rho_R|E\rangle\, A(E) + \sum_E \langle E|\rho_{\rm mc} - \rho_R|E\rangle\, \mathcal{D}^{-1/2}(E) f(E,0) R_{EE}\Bigg|. \tag{3.4}$$

The second line is $O(\mathcal{D}^{-1/2}(\lambda))$ [39] since both density matrices have unit trace. The density of states is evaluated at λ since this is a typical energy in W. Via the triangle inequality,

$$\epsilon_R^{\text{diag}} \leqslant |\text{tr}[(\rho_{\text{mc}} - \rho_R)A(H)]| + O(\mathcal{D}^{-1/2}(\lambda)), \tag{3.5}$$

where we have made the first term more compact by writing $A(H) = \sum_E A(E)\,|E\rangle\,\langle E|$. Now we expand the smooth ETH function $A(E)$ near the target energy λ. Repeated uses of the triangle inequality yield

$$\begin{aligned}|\text{tr}[(\rho_{\text{mc}} - \rho_R)A(H)]| &\leqslant |A(\lambda)\text{tr}[\rho_{\text{mc}} - \rho_R]| \\ &+ |(\text{d}A/\text{d}E)(\lambda)\text{tr}[(\rho_{\text{mc}} - \rho_R)(H - \lambda)]| + \cdots,\end{aligned} \tag{3.6}$$

where the ellipsis signifies higher-order derivatives. Both density matrices have unit trace, so the first term vanishes. But then using the fact that the kth derivative of $A(E)$ with respect to E is proportional to N^{-k}, which follows from $A(E) = a(E/N)$ with $a(x)$ becoming N independent in the thermodynamic limit, we see that

$$\epsilon_R^{\text{diag}} \leqslant \frac{\chi_R}{N} + O(N^{-2}) + O(\mathcal{D}^{-1/2}(\lambda)), \tag{3.7}$$

where

$$\chi_R = \left|a'\left(\frac{\lambda}{N}\right)\text{tr}[(\rho_R - \rho_{\text{mc}})(H - \lambda)]\right| \tag{3.8}$$

and $a'(x) = \text{d}a/\text{d}x$. If both ρ_R and ρ_{mc} have support only on the microcanonical window W, then $\chi_R \leqslant 2\delta|a'(\lambda/N)|$ for any R, where we have used that $|\text{tr}[\rho_R - \rho_{\text{mc}}]| \leqslant 2$. If instead they have support on a larger energy interval which is still $O(\delta)$, then χ_R is still $O(\delta)$ and we can still make the rough estimate

$$\epsilon_R^{\text{diag}} \leqslant O(\delta/N) + O(\mathcal{D}^{-1/2}(\lambda)). \tag{3.9}$$

In a very large system, we would not concern ourselves with the difference between Eq. (3.9) and the more accurate Eq. (3.7). However, at the relatively small systems we consider in this work, we can expect that for large R, χ_R and therefore ϵ_R^{diag} will be smaller than $2\delta|a'(\lambda/N)|$ if we compare the variational estimate $\langle A\rangle_R^{\text{diag}}$ to the expectation value of A in a microcanonical ensemble that is similar to the diagonal variational ensemble. In Sec. IV we numerically confirm this for a Gaussian microcanonical ensemble. It is interesting to note that any subextensive choice of δ will lead to vanishing diagonal error in the thermodynamic limit; this is simply a manifestation of statistical-mechanical ensemble equivalence from the perspective of ETH.

2. Off-diagonal error ϵ_R^{off}

We now turn to the off-diagonal error ϵ_R^{off}. First we observe that if the states $|\psi_r\rangle$ have exactly zero energy weight outside the microcanonical window W, the off-diagonal error (3.2c) can be expressed as $\epsilon_R^{\text{off}} = |\text{tr}\rho_R\tilde{A}|$, where

$$\langle E|\tilde{A}|E'\rangle = \begin{cases}\langle E|A|E'\rangle & \text{if } E, E' \in W \text{ and } E \neq E' \\ 0 & \text{otherwise.}\end{cases} \tag{3.10}$$

If the variational states have some nonzero weight outside W, we can expect that the error is still approximately expressible this way, by slightly expanding W. In Sec. IV, we suitably modify the expression $\epsilon_R^{\text{off}} = |\text{tr}\rho_R\tilde{A}|$ in this case. Since we are interested in understanding how averaging over an R-state ensemble can reduce this error, we write the average explicitly as

$$\epsilon_R^{\text{off}} = \left|\frac{1}{R}\sum_r \langle\psi_r|\tilde{A}|\psi_r\rangle\right|. \tag{3.11}$$

Following Refs. [35,36], let $x_r = \langle\psi_r|\tilde{A}|\psi_r\rangle$. For each $r \in \{1, 2, \ldots, R\}$ we have that

$$\lambda_{\min}(\tilde{A}) \leqslant x_r \leqslant \lambda_{\max}(\tilde{A}), \tag{3.12}$$

where the limits are the minimum and maximum eigenvalues of the (purely off-diagonal) operator $\tilde{A}$. We find numerically in Appendix E (Fig. 13) that the maximum (minimum) eigenvalues of various operators are always above (below) zero, respectively, which is expected since $\tilde{A}$ is traceless by construction. If the ensemble of microcanonical superposition states is "random enough" we can expect that for each r, x_r fluctuates between these limits according to some distribution. Should the algorithm perform ideally, the states $|\psi_r\rangle$ would sample $\tilde{A}$ in an unbiased way; i.e., in the limit of infinite samples, $\epsilon_R^{\text{off}} \to \text{tr}(\tilde{A})/n = 0$. This would happen for example if $|\psi_r\rangle$ were drawn from the Haar measure on the microcanonical subspace.

However, let us allow for biased sampling by modeling x_r as independent and identically distributed (i.i.d.) random variables with covariance $\mathbb{E}(x_r x_s) - \mathbb{E}(x_r)\mathbb{E}(x_s) = \sigma^2\delta_{rs}$ and mean $\mathbb{E}(x_r) = c$ which would ideally be zero. With x_r modeled this way, a simple statistical measure of the size of the off-diagonal error would be its mean-square value which has the functional form

$$\mathbb{E}\left[\left(\epsilon_R^{\text{off}}\right)^2\right] = \frac{\sigma^2}{R} + c^2 \equiv y_{c,\sigma}(R). \tag{3.13}$$

For large R, a good approximation to the actual variance σ^2 should be given by the finite-size estimate

$$\sigma_R^2 = \frac{1}{R}\sum_r x_r^2 - \left(\frac{1}{R}\sum_r x_r\right)^2, \tag{3.14}$$

which is bounded from above by the largest square singular value of $\tilde{A}$ which is in turn shown in Appendix E to scale down weakly with N. Under such assumptions, the off-diagonal error in the VME algorithm should thus behave typically as

$$\epsilon_R^{\text{off}} = |c| + O(R^{-1/2}), \tag{3.15}$$

where we have assumed that $\sigma \approx \sigma_R = O(1)$. In Sec. IV A we demonstrate that x_r are indeed well modeled by this phenomenological description with some small nonzero $|c|$ always present. This implies that the variational states indeed are not sampling the microcanonical Hilbert space perfectly uniformly, which is consistent with them also not being close to Haar-random states.

3. Summary

In summary, the basic idea behind our algorithm is as follows. To prepare an approximation to ρ_{mc}, instead of preparing an ensemble of one or more eigenstates $|E\rangle$, which each individually require exponential quantum resources to generate, we will prepare a polynomially large number R of states $|\psi_r\rangle$ which each hopefully require only polynomial

quantum resources to generate, such that $\rho_R \approx \rho_{\rm mc}$ *as measured by local observables*.

The error in the approximation can be understood in terms of two pieces. The first is the diagonal error, which is ultimately about statistical-mechanical ensemble equivalence as it manifests for an isolated quantum system via the diagonal part of the ETH ansatz (2.1). The leading contribution to this type of error thus scales as $O(\delta/N)$, and thus any subextensive window width δ will in principle work. The more prohibitive error is the off-diagonal error, which for a single variational state the ETH alone cannot guarantee to be small at the scale of δ and N practically accessible to the VME algorithm discussed in Sec. III B. To remedy this, we propose to insert randomness by averaging over variational states which have been prepared by initializing the VQA with random product states. We have so far focused on a particular observable A and discussed the error in the context of its matrix elements. The claim that $\rho_R \approx \rho_{\rm mc}$ *as measured by local observables* can be made more precise by introducing the trace distance between certain reduced density matrices, which we examine numerically in Sec. IV.

B. VME algorithm

Algorithm 1 Prepare $|\psi_r(\boldsymbol{\theta}^*)\rangle$.

Data: Hamiltonian H, target energy λ, tolerance δ and random product state $|\psi_r^0\rangle$
Result: converged variational state $|\psi_r\rangle$

```
p = 1;
θ = 0;
while p < ∞ do
    ϵ = 10;
    while ϵ ≥ 10^{-3} do
        while ||∇C||_∞ > ϵ do
            prepare |ψ_r(θ)⟩ = U_p(θ)|ψ_r^0⟩;
            measure C(θ) and {∂_j C(θ)}_j ;
            update θ according to BFGS optimizer;
        if Var(H) ≤ δ^2 then
            p* ← p;
            θ* ← θ;
            converged;
            set |ψ_r⟩ = U(θ*)|ψ_r^0⟩;
        else
            ϵ ← ϵ/2;
    p ← p + 1;
```

We now describe the VME algorithm for preparing microcanonical superposition states $|\psi_r\rangle$ discussed in the previous section. At the beginning we fix a target energy λ, and microcanonical window size

$$\delta = \frac{\Delta E}{N} N^{\alpha}, \tag{3.16}$$

where ΔE is the full energy bandwidth of the Hamiltonian H. In the mixed-field Ising model we later consider, we find $\frac{\Delta E}{N} \approx 3$ independent of N. We focus our numerical studies mainly on the case $\alpha = -1/2$. We initialize the QPU (see Fig. 1) in a random product state

$$|\psi_r^0\rangle = |\varphi_r^1\rangle |\varphi_r^2\rangle \cdots |\varphi_r^N\rangle, \tag{3.17}$$

with $|\varphi_r^j\rangle = \cos(\varphi_r^j)|0\rangle + \sin(\varphi_r^j)|1\rangle$ and φ_r^j drawn from the uniform distribution on $[0, \pi)$, which we can expect to have extensive energy variance [24,33]. We then minimize the "folded-spectrum" cost function [9]

$$\mathcal{C}(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta})|(H-\lambda)^2|\psi(\boldsymbol{\theta})\rangle, \tag{3.18}$$

until $\mathrm{Var}(H) = \langle (H - \langle H\rangle)^2\rangle \leqslant \delta^2$, obtaining the converged variational state $|\psi_r\rangle$. Note that $\mathcal{C}(\boldsymbol{\theta})$ penalizes both large energy variance and deviation of the average energy from the target energy, since

$$\mathcal{C}(\boldsymbol{\theta}) = \mathrm{Var}(H) + \langle H - \lambda\rangle^2, \tag{3.19}$$

but that the convergence criterion only concerns $\mathrm{Var}(H)$. We find in practice that $\langle H-\lambda\rangle^2$ is comparatively small when N is large, so it is also possible to think of $\mathcal{C}(\boldsymbol{\theta}) \lesssim \delta^2$ as the convergence criterion.

The optimization is repeated R times for different initial random product states to generate the variational ensemble. Notice that this cost function $\mathcal{C}(\boldsymbol{\theta})$ is zero if and only if $|\psi(\boldsymbol{\theta})\rangle = |\lambda\rangle$, the eigenstate with energy λ [40]. Unlike previous explorations [12,29,41] with this cost function, we do not seek local or global minima, since we minimize the cost function only until $\mathrm{Var}(H) \leqslant \delta^2$, which is not a constraint on the gradient, but on the value of the cost function itself. Furthermore, the unique global minimum is a state completely different from the one we target.

For simplicity we restrict our variational states to be real in the computational basis (CB). Because we are interested in the minimal circuit depth needed to prepare the variational states, we employ a periodic structure ansatz (PSA) [14] circuit for which the number of "layers" will be adaptively chosen by the algorithm. The PSA with p layers is defined as

$$U_p(\boldsymbol{\theta}) = \prod_{l=1}^{p} V(\boldsymbol{\theta}_l), \tag{3.20}$$

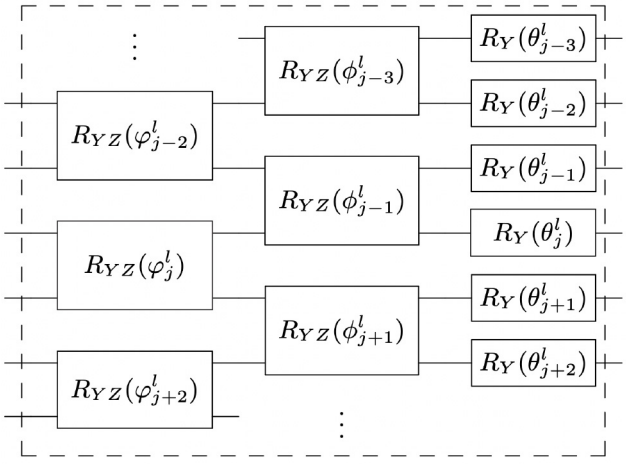

FIG. 3. Layer l of the ansatz circuit, Eq. (3.21), near qubit j.

where each layer is the unitary (see Fig. 3)

$$V(\boldsymbol{\theta}_l) = \prod_{j=1}^{N} e^{i\theta_j^l Y_j} \prod_{\substack{j=1 \\ \text{even}}}^{N} e^{i\phi_j^l Y_j Z_{j+1}} \prod_{\substack{j=1 \\ \text{odd}}}^{N} e^{i\varphi_j^l Y_j Z_{j+1}} \tag{3.21}$$

and $\boldsymbol{\theta}$ stands for the $2Np$ real parameters $\{\theta_j^l, \phi_j^l, \varphi_j^l\}_{jl}$ and Y_j, Z_j are Pauli operators acting on qubit j. The "brick-wall" form of this ansatz breaks down for odd N, so in this case we add an additional gate to entangle the ends of the chain. That is, for odd N, we make the replacement

$$\prod_{\substack{j=1 \\ \text{even}}}^{N} e^{i\phi_j^l Y_j Z_{j+1}} \to e^{i\phi_1^l Y_1 Z_N} \prod_{\substack{j=1 \\ \text{even}}}^{N} e^{i\phi_j^l Y_j Z_{j+1}}. \tag{3.22}$$

The operators appearing in the single-layer unitary $V(\boldsymbol{\theta}_l)$ are chosen based on the findings of Ref. [42]. There it is argued that the pool of $2N-2$ operators $\mathcal{P} = \{iY_jZ_{j+1}\}_{j=1}^{N-1} \cup \{iY_j\}_{j=1}^{N-1}$ is "complete" in the sense that for any state $|\psi\rangle$, the set of states $\{A_k|\psi\rangle\}_k$ form a complete basis, where A_k are nested commutators of operators in $\mathcal{P}$, i.e., elements of the dynamical Lie algebra [43] of $\mathcal{P}$. We have added the extra gates Y_N and (for odd N) Y_1Z_N to the pool, but clearly $\mathcal{P} \cup \{Y_1Y_N, Y_1\}$ is still complete in the above sense.

Since our convergence criterion is based on the value of the cost function and the native convergence criterion of a gradient-based optimizer is based on the size of the gradient, we "wrap" the optimizer in a simple loop (see Algorithm 1) where we repeatedly interrupt the optimizer to check if the convergence criterion is satisfied, which we accomplish by having it only minimize until the gradient norm falls below a relatively large value ϵ which starts at 10^1 and can only be decreased down to 10^{-3}. If the algorithm then cannot achieve convergence using a p-layer ansatz by decreasing ϵ to 10^{-3}, it adds another layer $p \to p+1$ and repeats the procedure.

For the classical optimizer we employ the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimizer which is gradient based. We therefore take advantage of the "parameter shift rule" [44] for computing analytic gradients of the cost function. If the generators are Pauli strings (hence squaring to I), then

$$\frac{\partial}{\partial\theta_j}\mathcal{C}(\boldsymbol{\theta}) = \mathcal{C}\left(\boldsymbol{\theta} + \frac{\pi}{4}\boldsymbol{e}_j\right) - \mathcal{C}\left(\boldsymbol{\theta} - \frac{\pi}{4}\boldsymbol{e}_j\right) \tag{3.23}$$

with $\boldsymbol{e}_j$ a unit vector in the jth direction. Thus, during the optimization both the cost function and its derivatives can be measured using the QPU.

IV. NUMERICAL RESULTS

We test the VME algorithm on the 1D mixed-field Ising model (MFIM) with Hamiltonian

$$H = \sum_{j=1}^{N} (JZ_jZ_{j+1} + h_{x,j}X_j + h_zZ_j) \tag{4.1}$$

and periodic boundary conditions so that site $N+1$ refers to site 1. To ensure there are no accidental degeneracies, we consider weakly nonuniform transverse fields $h_{x,j} = h_x + r_j$, where $r_j \in [-0.01, 0.01]$ are drawn randomly from the uniform distribution. We use a single fixed configuration of the transverse fields for our numerics. We fix parameters $J = 1$, $h_z = 0.5$, and $h_x = -1.05$ such that the system is strongly nonintegrable [45]. In this section we discuss the performance of the VME with respect to this particular model.

TABLE I. The $N = 13$ broadened microcanonical best-fit parameters (μ, σ) at various target energy densities λ/N.

λ/N	μ/N	σ/δ
−0.750	−0.761	0.858
−0.500	−0.511	0.831
−0.250	−0.253	0.821
0.000	0.001	0.832

A. Diagonal variational ensemble

Here we characterize the nature of the converged variational states in the energy basis. To do so, we first define a "broadened" microcanonical ensemble as was done in Ref. [5]. This ensemble is of the form

$$\rho_{\lambda,\delta} = \mathcal{D}_\delta^{-1}(\lambda) G_\delta(H - \lambda), \tag{4.2}$$

where $G_\delta(x) = (2\pi\delta^2)^{-1/2} e^{-x^2/2\delta^2}$ is a normalized Gaussian function, and $\mathcal{D}_\delta(\lambda) = \text{tr}\, G_\delta(H-\lambda)$ is the "broadened" density of states evaluated at energy λ. Here, δ corresponds to the convergence criterion for the VME algorithm, i.e., Eq. (3.16) with $\alpha = -1/2$. We will from now on treat $\mathcal{D}_\delta(\lambda)$ as a good approximation to the density of states in the thermodynamic limit (see Appendix A for further justification).

We claim that the variational algorithm 1 generates diagonal energy-basis matrix elements $\rho_R(E) = \langle E|\rho_R|E\rangle$ approximating a broadened microcanonical ensemble. Figure 4(a) shows the variational-ensemble diagonal energy-basis matrix elements to which we fit the curve $\rho_{\mu,\sigma}(E) = \mathcal{D}^{-1}(\mu) G_\sigma(E - \mu)$ with fitting parameters μ and σ (shown in solid black). Up to fluctuations from eigenstate to eigenstate, we can see that the variational diagonal ensemble is well described by the Gaussian best fit. More precisely, the coarse-grained version, $\overline{\rho_R}(E)$, of the variational-ensemble diagonal elements—where the fluctuations are eliminated—agrees quite well with the Gaussian best fit. The coarse-grained curve is computed as follows: For each E in the spectrum of H, $\overline{\rho_R}(E)$ is defined as the average of $\langle E'|\rho_R|E'\rangle$ over the K eigenenergies E' nearest to E. We set the "resolution" $K = 64$, except for E near the edges of the spectrum where $1 \leqslant K < 64$. At the largest system size of $N = 13$ and for an $R = 288$-state ensemble, we list the best-fit parameters μ and σ in Table I.

Note that away from $\lambda = 0$, the variational ensembles converge with μ slightly differently than the target λ. This is because the variational ensemble is generated by minimizing the first two moments of the operator $(H - \lambda)$, but the density of states determined by the underlying model is nonuniform. In fact if we assume the Gaussian density of states discussed in Appendix A, we have $\text{tr}(\rho_{\mu,\sigma}H) \approx \mu - O(\sigma^2 \frac{\mu}{N})$, where we have assumed that σ decreases with N so that higher-order terms can be neglected in the thermodynamic limit. We can see that when $\mu < 0$, the $\rho_{\mu,\sigma}$ ensemble actually has average

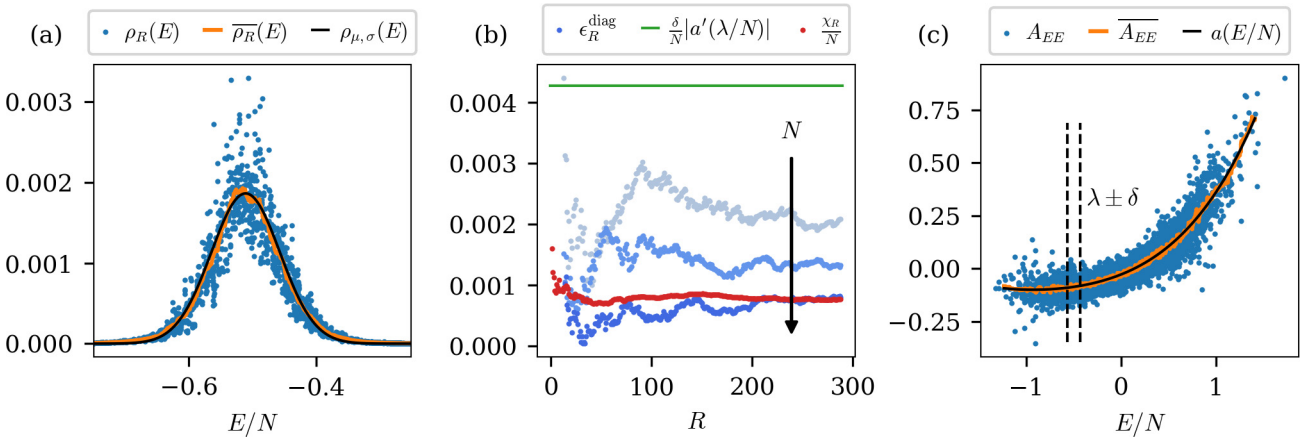

FIG. 4. Analysis of diagonal error in the VME, with operator $A = Z_{\lfloor N/2 \rfloor}$ taken as an example. (a) The variational ensemble diagonal matrix elements $\rho_R(E) = \langle E|\rho_R|E\rangle$ (blue) for $N = 13$, $\lambda/N = -0.5$, $\alpha = -1/2$ [see Eq. (3.16)], and $R = 288$ states in the ensemble. In black, the best Gaussian fit curve $\rho_{\mu,\sigma}(E)$ [see Eq. (4.2)], and in orange a coarse-grained version of the blue scatter points for comparison. (b) The diagonal error [see Eq. (3.2b)] with $\rho_{\text{mc}} = \rho_{\lambda,\delta}$ versus $R \geqslant 12$ for $N = 11, 12, 13$ where increasing N corresponds to darker blue data points. For comparison, we include in green the ensemble-independent estimate $\delta|A'(\lambda)| = \delta|a'(\lambda/N)|/N$ at $N = 13$, as well as the more accurate estimate χ_R [Eq. (3.8)] at $N = 13$. (c) The diagonal matrix elements $A_{EE} = \langle E|A|E\rangle$ in blue, their coarse graining in orange, and the best fourth-order polynomial fit in black which defines $a(E/N)$. This smooth function a is used to compute χ_R. Vertical dashed lines show the scale of the microcanonical window as compared to the whole spectrum.

energy larger than μ. In Appendix D we confirm this statement more quantitatively. As a consequence of this analysis, we conclude that the deviations in μ from λ are a finite-size effect due to the nonconstant density of states, and not due to the fluctuations around the average curve $\overline{\rho_R}(E)$.

In the last row of Table I, we see that all the σ are at least about 15% smaller than δ. This is due to a simplification we have made in the preceding analysis. Note the form of $\overline{\rho_R}(E)$ in Fig. 4(a) at the edges of the window; the orange curve has more weight away from the window than the Gaussian best fit (in black). A more accurate characterization of the ensemble might be, for example, a sum of two Gaussian curves. We nonetheless opt to consider the ensemble as roughly a single Gaussian peak for simplicity. We discuss the quantitative consequences of this simplification in Appendix C and explain why the σ/δ shown in Table I are not closer to unity.

Up to fluctuations around the coarse-grained behavior $\overline{\rho_R}(E)$ and the slight oversimplification of the single-peak Gaussian fit, we have thus established the form of the diagonal part of the variational ensemble. In the following sections we will study the error in the VME estimate of various observables. Clearly, we will need to choose an appropriate microcanonical ensemble to compare to. This ensemble is arguably $\rho_{\mu,\sigma}$ because it closely approximates the (coarse-grained) diagonal ensemble $\overline{\rho_R}(E)$. One could also imagine comparing to the diagonal ensemble itself and focusing solely on the off-diagonal error as was done in Ref. [46]. However, imagining the VME as a practical algorithm for computing broadened microcanonical ensemble averages, one could not know ahead of time what μ and σ were. Furthermore, we have shown that the distinction between λ/N and μ/N is a finite-size effect that is already small at $N = 13$. We will henceforth compare our numerical results to the ensemble $\rho_{\lambda,\delta}$, where λ and δ are precisely the parameters that were initially chosen before running the algorithm; i.e., we set $\rho_{\text{mc}} = \rho_{\lambda,\delta}$ in Eq. (3.2b).

B. Diagonal error

We now briefly discuss the diagonal error with respect to the broadened microcanonical estimates. Figure 4(b) shows, for $N = 11, 12, 13$, the diagonal error versus $R \geqslant 12$ for the operator $A = Z_{\lfloor N/2 \rfloor}$ at $\lambda/N = -0.5$. For comparison we plot a simple estimate of the error $(\delta/N)|a'(\lambda/N)|$ at $N = 13$, as well as the more accurate estimate χ_R defined via Eq. (3.7) which we calculate numerically using the $N = 13$ variational ensemble. We calculate $a'(\lambda/N)$ using a fourth-order polynomial fit to the (coarse-grained) graph of $\langle E|A|E\rangle$ versus E/N, shown as the solid black line in Fig. 4(c). The coarse-grained curve $\overline{A_{EE}}$ is computed in the same way as was $\overline{\rho_R}(E)$ in Sec. IV A, but here we use a resolution of $K = 32$.

For this particular operator and energy density, it is clear that the diagonal error decays with N and is an order-one fraction of the rough estimate $O(\delta/N)$. Furthermore, we can see that for large R its behavior is well captured by the expected estimate χ_R.

In Appendix F we present further numerical results for the four local operators Z, ZZ, X, and XX acting on the central one or two sites of the chain at the energy densities $\lambda/N = -0.75$, -0.5, -0.25, and 0.0. At $\lambda/N = -0.5$, all operators have the property that ϵ_R^{diag} decays with N for large R, and the $N = 13$ values are consistent with the estimate χ_R/N. At other energy densities the scaling with N is not always so well established, but the error for large R is always smaller than $(\delta/N)|a'(\lambda/N)|$. The value of χ_R also appears to generally be on the correct scale of ϵ_R^{diag} except for the operator XX, for which χ_R/N undershoots the value of the diagonal error for the higher energy densities. These various

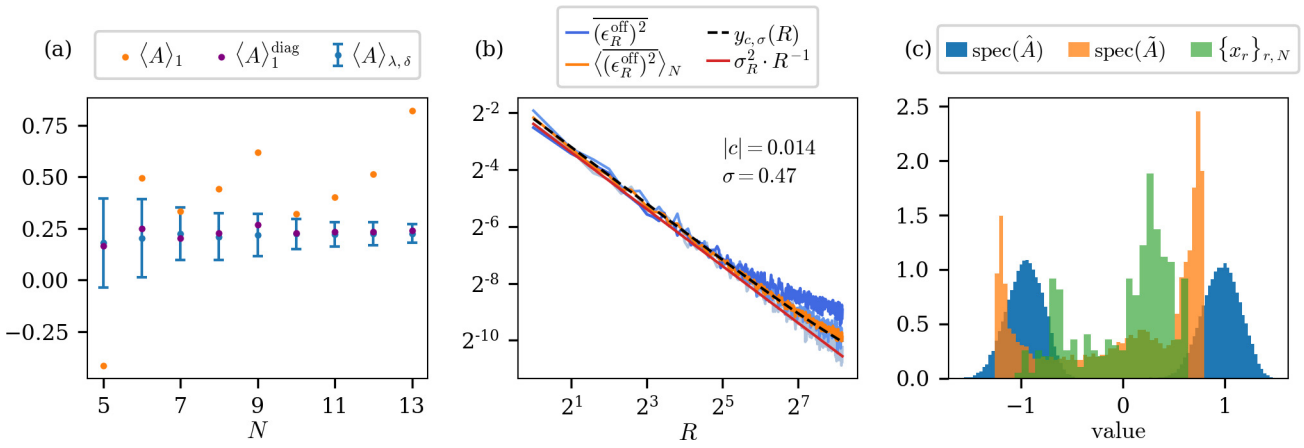

FIG. 5. Off-diagonal error in the VME algorithm, with $A = X_{\lfloor N/2\rfloor}$ at $\lambda/N = -0.5$ taken as an example. (a) The broadened microcanonical ensemble expectation value $\langle A\rangle_{\lambda,\delta} = \mathrm{tr}(\rho_{\lambda,\delta}A)$ and error bars indicating the associated broadened microcanonical standard deviations (obtained from ED) are plotted in blue as a function of N. We compare this to the estimate in a single variational state, $\langle A\rangle_1 = \mathrm{tr}(\rho_1 A)$, along with the corresponding *diagonal ensemble* estimation $\langle A\rangle_1^{\mathrm{diag}}$, i.e., Eq. (3.2d) with $R = 1$. Note that $\langle A\rangle_1^{\mathrm{diag}}$ can only be computed with ED, where off-diagonal contributions can be discarded by hand. (b) The finite-sample mean-square off-diagonal error $\overline{(\epsilon_R^{\mathrm{off}})^2}$ with increasing R for $N = 11, 12, 13$, where increasing N corresponds to a darker blue curve as in Fig. 4. The orange curve, $\langle\overline{(\epsilon_R^{\mathrm{off}})^2}\rangle_N$, is the N-averaged value of $\overline{(\epsilon_R^{\mathrm{off}})^2}$ as discussed in the text. The dashed black line is a two-parameter best-fit to $\langle\overline{(\epsilon_R^{\mathrm{off}})^2}\rangle_N$, and for comparison we plot the "theoretical" error σ_R^2/R in red. (c) Probability density functions of the $N = 13$ eigenvalues spec($\hat{A}$) and spec($\tilde{A}$) with $\hat{A}$ and $\tilde{A}$ as defined in the text. Panel (c) also shows a probability density function of the collection $\{x_r\}_{r,N}$ where N runs over a few system sizes as discussed in the text.

deviations are likely due to ETH not yet strongly setting in at such small system sizes. In particular the $N = 13$ value of $\mathcal{D}^{-1/2}(\lambda)$ is never smaller than 0.04 for the energy densities we consider, so that the ETH fluctuations, i.e., the third term in Eq. (3.7), could be comparable to $\delta/N \approx 0.06$ at $N = 13$ depending on the relative size of $A(E)$ and the ETH function $f(E, 0)$.

In any case, the diagonal error is always quite small across all operators and energy densities, when compared, for example, against the scale of the microcanonical fluctuations themselves. For example, see Fig. 5(a) where in purple we can see that even for a *single* variational state, the diagonal ensemble estimate is highly accurate. We observe this across every considered operator and energy density, as shown in Appendix G.

C. Off-diagonal error

We now turn to discuss the numerical details of the off-diagonal error and how ensemble averaging reduces it. First, Fig. 5(a) illustrates the main problem with using a single variational state with a large δ. We take as an example the operator $A = X_{\lfloor N/2\rfloor}$ at the energy density $\lambda/N = -0.5$ and compare a single variational state estimate $\langle A\rangle_1 = \langle\psi_1|A|\psi_1\rangle$ to the smooth microcanonical average $\langle A\rangle_{\lambda,\delta} = \mathrm{tr}(\rho_{\lambda,\delta}A)$. The estimate is poor even for the largest system size. In Fig. 5(b) we examine how this error is reduced by averaging. It is clear that averaging always reduces the off-diagonal error; however, since we anticipate it to behave as a random variable, we need to measure a statistical quantity to make the analysis precise. Given the finite data set $\{x_r\}_{r=1}^{288}$ of samples, we calculate a finite-size estimate of the mean-square error in whatever underlying distribution x_r is sampled from as a function of R as follows. For each fixed $R \in \mathcal{R} = \{1, 2, \ldots, 288\}$ we calculate

$$\overline{(\epsilon_R^{\mathrm{off}})^2} = \frac{1}{S}\sum_{k=1}^{S}\left|\frac{1}{R}\sum_{r\in P_k(\mathcal{R})|_R} x_r\right|^2, \tag{4.3}$$

where $P_k(\mathcal{R})|_R$ are the first R elements of a random permutation of the ordered index set $\mathcal{R}$. Choosing $S = 100$, we plot $\overline{(\epsilon_R^{\mathrm{off}})^2}$ versus R for $N = 11, 12, 13$ on a log-log scale in anticipation of observing an R^{-1} scaling of the off-diagonal error. Across various operators and energy densities (shown in Appendix G), we observe an initial R^{-1} scaling with prefactor approximately independent of N. However, the large R value can either be larger or smaller depending on N in a nonsystematic way. To remove this dependence we focus our analysis on a system-size averaged version

$$\left\langle\overline{\left(\epsilon_R^{\mathrm{off}}\right)^2}\right\rangle_N = \frac{1}{5}\sum_{N=9}^{13}\overline{\left(\epsilon_R^{\mathrm{off}}\right)^2}, \tag{4.4}$$

which is shown in orange. The data are well described by a two-parameter best-fit function of the form $y_{c,\sigma}(R) = \sigma^2/R + c^2$ which corresponds to x_r being effectively i.i.d. random variables as discussed in Sec. III. The values of $|c|$ and σ are shown in Fig. 5(b). We can see that for the operator X at $\lambda/N = -0.5$ shown here in the main text, the value of $|c| = 0.014$ is quite small. In other cases, in particular for XX, it can be slightly larger: $|c| = 0.071$. Data for all operators and energy densities are shown in Appendix G. We also find that the best-fit parameter σ is on the same scale as the finite-sample estimate σ_R as demonstrated by the collapse of the red curve σ_R^2/R and the best-fit line in the small-R regime. Here σ_R is calculated for $\tilde{A}$ having nonzero energy-basis matrix elements only on an energy window of half width 3δ

[see Eq. (3.10)]. This approximation of computing σ_R based only on the matrix elements of ρ_R and A in energy eigenstates near λ is numerically justified in Appendix E. These results imply that for small R, the statistics of the off-diagonal error are controlled by $\tilde{A}$.

Reference [34] inferred correlations between energy-basis matrix elements of local operators A by the form of the eigenvalue statistics of certain submatrices of A. To help understand the nature of the off-diagonal error, in Fig. 5(c) we also examine the eigenvalue distribution of the operator $\tilde{A}$ defined on an energy window of half width 3δ and centered on energy λ, i.e., as in Eq. (3.10) but with W slightly expanded to accommodate the tails of the roughly Gaussian variational states. See Appendix E for a graphical representation of how this energy window is defined. For comparison we also show in Fig. 5(c) the eigenvalue distribution of the operator $\hat{A}$, which is simply the operator A with its energy-basis diagonal elements deleted. There are a number of interesting qualitative properties displayed by these eigenvalue distributions that are relevant to the off-diagonal error in the VME.

First, we observe that the eigenvalue distribution of $\tilde{A}$ does not appear to qualitatively change shape as N is varied, except for a slight reduction in the total width for increasing N, as demonstrated in Appendix E. Since it is $\tilde{A}$ which roughly determines the off-diagonal error, the qualitative lack of N dependence agrees with the fact that the off-diagonal error does not depend on N in a systematic way at the system sizes we examine.

Interestingly, we observe a correlation between the single-state variational estimates in Fig. 5(a) as N is varied and the eigenvalue distribution of $\tilde{A}$. When $\langle A\rangle_1$ overestimates (underestimates) the microcanonical value across many system sizes, the eigenvalue distribution is biased to the right (left) of zero. For further evidence that this correlation is not an artifact of this energy density or the choice of operator, see Appendix G. To demonstrate this further, we there also plot a histogram of the off-diagonal error present in many individual variational state samples, including samples across a window of system sizes: specifically we show a normalized histogram of the values in the set

$$\{x_r\}_{r,N} = \bigcup_{N=9}^{13} \{x_r^N\}_{r=1}^{288}, \tag{4.5}$$

where x_r^N is the off-diagonal error in variational state r when the system size is N. Using the statistics across multiple system sizes is justified here since the off-diagonal error varies erratically with the system size. On a qualitative level, this latter histogram confirms that the variational states sample $\mathrm{spec}(\tilde{A})$ uniformly enough that a bias in $\mathrm{spec}(\tilde{A})$ on the left or right of zero is reflected in the statistics of x_r. This fact is not too surprising since, roughly speaking, $\tilde{A}$ determines the off-diagonal error via

$$x_r \approx \sum_{\tilde{a}} \tilde{a} |\langle \tilde{a}|\psi_r\rangle|^2 \tag{4.6}$$

with $\tilde{a}$ and $|\tilde{a}\rangle$ the eigenvalues and eigenvectors of $\tilde{A}$. However, it is not obvious that $|\langle \tilde{a}|\psi_r\rangle|^2$ is a uniform distribution and, furthermore, this approximation should only be understood statistically since in actuality, it is the properties of $\hat{A}$ which precisely determine the error:

$$x_r = \sum_{\hat{a}} \hat{a} |\langle \hat{a}|\psi_r\rangle|^2 \tag{4.7}$$

and the truncation of $\hat{A}$ to $\tilde{A}$ is only shown in Appendix E to rigorously hold for the quantity σ_R as opposed to individual realizations x_r.

At all energy densities, the eigenvalue distribution of $\widetilde{XX}$ is much flatter than the other three considered operators, and the histogram of x_r is also qualitatively different for XX, two observations which could be related to the fact that the off-diagonal error for $A = XX$ generally saturates sooner and to a larger value than the other operators do. However, we leave an identification of the precise underlying mechanism to future work.

We conclude the discussion of the off-diagonal error with some observations about the eigenvalue statistics of the full-spectrum off-diagonal operator $\hat{A}$. Even though the variational states in principle have support on the entire energy spectrum, we can see that it is the statistics of $\tilde{A}$ and not of $\hat{A}$ that are correlated with the off-diagonal error, further justifying the truncation to a local energy window. Interestingly, we can see that $\hat{A}$ still has a similar spectral form to that of a Pauli string with eigenvalues ± 1, but the otherwise highly degenerate peaks have been smeared out by removing the diagonal energy-basis elements. In Appendix G we show the eigenvalue distributions of $\hat{A}$ for other A, and note that XX looks the most similar to that of a Pauli string; i.e., its peaks have been broadened the least. When the window is reduced to the scale δ, the distribution becomes less similar to that of a Pauli operator, and we can expect that as $\delta \to 0$, the spectrum approaches that of a random matrix, i.e., the semicircle law [34,35,47]. The fact that the eigenvalue distribution is so far from a semicircle law on the scale $\delta = O(N^{-1/2})$ provides further confirmation that $\langle E|A|E'\rangle$ are not effectively independently distributed and thus we cannot rely on randomness of the matrix elements alone to make the off-diagonal error small.

D. Explicit microcanonical estimates and trace distance

Having examined in some detail the scaling of the absolute diagonal and off-diagonal errors, we now take a step back and consider what the overall statistics of the variational estimates look like when compared to the microcanonical averages and their associated microcanonical fluctuations. For example, in Fig. 6 we show for various system sizes the variational estimate $\mathrm{tr}(\rho_R A)$ for $R = 288$ along with the standard error. This is compared against the broadened microcanonical average calculated from ED, with error bars indicating one microcanonical standard deviation $\Delta A_{\lambda,\delta}$, i.e.,

$$(\Delta A_{\lambda,\delta})^2 = \sum_E \langle E|\rho_{\lambda,\delta}|E\rangle \langle E|A|E\rangle^2 - (\mathrm{tr}\rho_{\lambda,\delta} A)^2, \tag{4.8}$$

which is another scale to which the error can be compared.

Running the VME two different times yields two different variational ensembles ρ_R and ρ_R'. The orange error bars measure how much $\mathrm{tr}(\rho_R A)$ and $\mathrm{tr}(\rho_R' A)$ would differ when $R = 288$. We find the energy densities $\lambda/N = -0.75$ and $\lambda/N = -0.5$ generally have more accurate estimates than

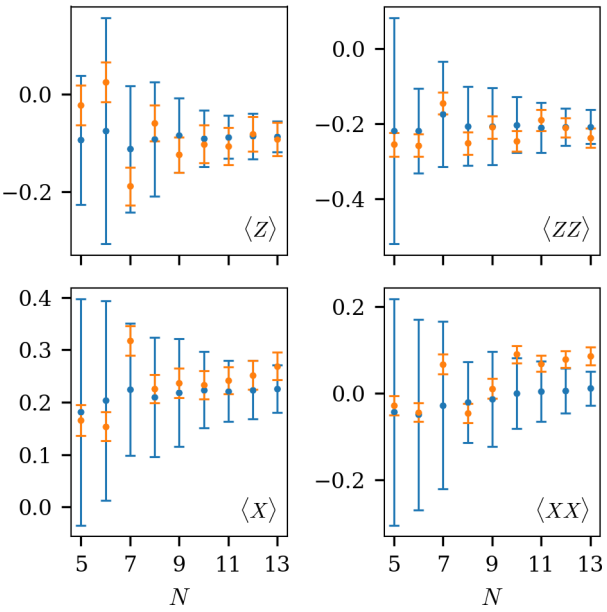

FIG. 6. In blue, smooth microcanonical averages $\langle A\rangle_{\lambda,\delta} = \mathrm{tr}(\rho_{\lambda,\delta}A)$ and their thermal fluctuations, Eq. (4.8). In orange, the corresponding variational estimates for $\alpha = -1/2$ and $R = 288$ with standard error. Averages and their error are plotted as a function of system size N at a fixed energy density of $\lambda/N = -0.5$.

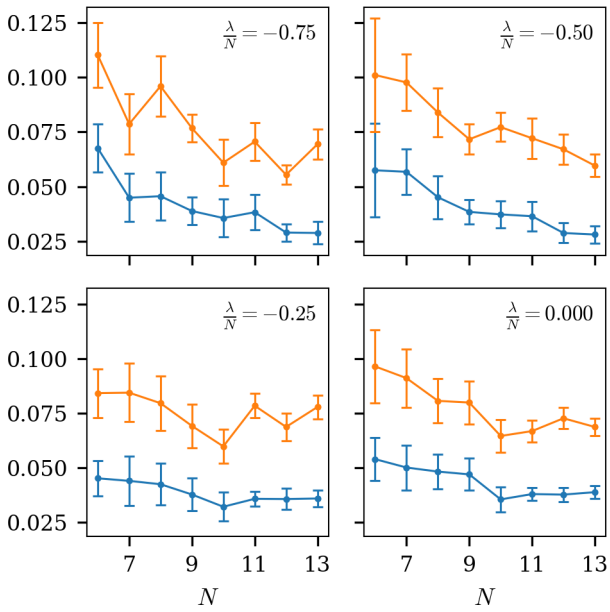

FIG. 7. Spatially averaged trace distance between variational ensembles of size $R(N) = \lfloor 1.5N^2 \rfloor$ and for $\alpha = -1/2$ and the corresponding broadened microcanonical ones. The blue curves are for $|S| = 1$ and the orange ones for $|S| = 2$. Error bars correspond to the standard deviation of this averaged quantity over 20 different ensemble realizations.

$\lambda/N = -0.25$ and $\lambda/N = 0$ (see Appendix H for further results). For a fixed R, the error certainly does not systematically decrease with N. In some cases, it appears to increase with N, for example X at $\lambda/N = -0.5$, shown in Fig. 6, or in some cases for XX as discussed in Appendix H. We do not consider these observations for a fixed R at odds with the off-diagonal error analysis where we stated that over a large range of R the error does not appear to systematically depend on N. We note that the operator XX generally appears to deviate at the larger system sizes more than Z, X, and ZZ across various energy densities, which agrees with the previous observations that XX behaves differently than the other operators.

To see if converged ensembles tend towards the microcanonical state with increasing N in an observable-independent way, we check if, for a subsystem S of the chain, the state $\rho_R^S = \mathrm{tr}_{\bar{S}}(\rho_R)$ approaches the subsystem broadened microcanonical state $\rho_{\lambda,\delta}^S = \mathrm{tr}_{\bar{S}}(\rho_{\lambda,\delta})$. As a distance measure we consider the trace distance, $T(\rho, \rho') = \frac{1}{2}||\rho - \rho'||_1$, which measures the distinguishability of ρ and ρ' in an operationally meaningful way [48]. More importantly for our purposes, however, it also bounds the difference in the expected value of any observable A as

$$T(\rho, \rho') \geqslant \frac{|\mathrm{tr}(\rho A) - \mathrm{tr}(\rho' A)|}{2\sigma_{\max}(A)}, \tag{4.9}$$

where $\sigma_{\max}(A)$ is the maximum singular value of A. This can be derived by using von Neumann's trace inequality $|\mathrm{tr}(XY)| \leqslant \sum_i \sigma_i(X)\sigma_i(Y)$. In particular for a Pauli string operator, we have $\sigma_{\max}(A) = 1$. We plot $T(\rho_R^S, \rho_{\lambda,\delta}^S)$ versus N, averaged over all contiguous subsystems S of fixed size $|S|$. We observe that the trace distance for a single-site subsystem $|S| = 1$ is always smaller (by about a factor of $1/2$) than for a system of two nearest-neighbor sites $|S| = 2$. In light of the previous analysis where the off-diagonal error did not appear to depend systematically on N, we choose R to scale with N as $R(N) = O(N^2)$, and further average over 20 ensemble realizations in each case. Since we have only 288 samples total, we compute the statistics of 20 random (possibly overlapping) subsets of size R.

The reason one may want to increase R with N is that the trace distance satisfies the inequality

$$T\left(\rho_R^S, \rho_{\lambda,\delta}^S\right) \leqslant \frac{1}{R}\sum_{r=1}^{R} T\left(\rho_r^S, \rho_{\lambda,\delta}^S\right), \tag{4.10}$$

with $\rho_r^S = \mathrm{tr}_{\bar{S}}\,|\psi_r\rangle\langle\psi_r|$, so that a large-$R$ variational ensemble can only do better than single pure states can on average. Since we are interested in understanding the behavior of the algorithm in the thermodynamic limit, and in light of our above results suggesting that the off-diagonal error does not appear to depend strongly on N, we consider the case of $R = O(N^2)$ so that we can observe a continual decrease of the trace distance with N at the lower energy densities. It is possible, however, that for a larger number of samples the spatially averaged trace distance will also eventually saturate to a finite value as it did for $A = (XX)_{\lfloor N/2 \rfloor}$ in particular.

The trace distance results in Fig. 7 are consistent with the estimates for particular local operators; for example, when comparing Fig. 6 at $N = 13$, we checked numerically that the deviation of the average estimate (corresponding to $R = 288$) from the microcanonical one never exceeds twice the trace distance at the corresponding energy density. However, the $N = 13$ value of the trace distances shown correspond to $R = 253$, whereas the observables correspond to $R = 288$. Furthermore, the trace distance is averaged over all contiguous

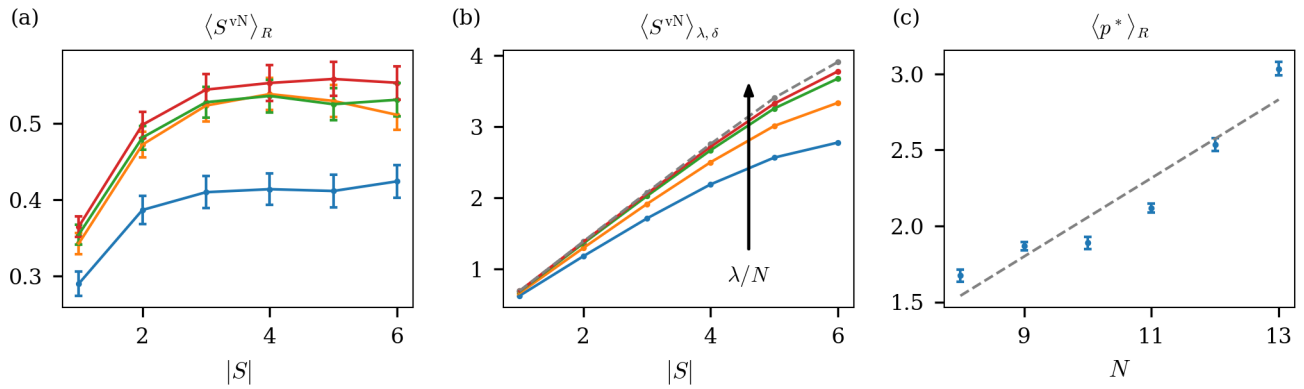

FIG. 8. [(a) and (b)] The entanglement entropy S^{vN} of a contiguous subregion S for different energy densities, where the blue, orange, green, and red curves correspond to $\lambda/N = -0.75, -0.5, -0.25$, and 0, respectively. Panel (a) contains the ensemble average entanglement entropy of converged variational states at $N = 13$ for an $R = 144$-state variational ensemble with error bars indicating the standard error. For comparison, panel (b) shows the average entanglement entropy in the corresponding broadened microcanonical ensembles [see Eq. (4.12)]. The gray dashed curve is the Page entropy for a Haar-random state. (c) Graph of the number of layers in the PSA circuit for $\lambda/N = -0.5$, averaged over $R = 144$ samples with error bars indicating the standard error. The dashed gray line is the best linear fit $p(N) = 0.26N - 0.52$.

subsystems of size $|S|$ so in principle this value no longer exactly upper bounds the observables as in Eq. (4.10), which are obtained for a given site $j = \lfloor N/2 \rfloor$, but in this case the results are nonetheless consistent.

E. Quantum resources

In Fig. 8(a) we plot the ensemble-averaged von Neumann entanglement entropy for a contiguous subsystem S of the chain

$$\langle S^{\mathrm{vN}} \rangle_R = \frac{1}{R} \sum_r S^{\mathrm{vN}}(|\psi_r\rangle) \tag{4.11}$$

versus $|S|$ at $N = 13$. Here, the von Neumann entanglement entropy of a pure state $|\psi\rangle$ is $S^{\mathrm{vN}}(|\psi\rangle) = -\mathrm{tr}(\rho^S \ln \rho^S)$ with $\rho^S = \mathrm{tr}_{\bar{S}} |\psi\rangle \langle\psi|$. We find that on average the variational states appear to be area-law entangled because the values saturate with increasing $|S|$. Regardless of the scaling with $|S|$, we note that the entanglement of the variational states is much smaller than the average entanglement of eigenstates within the broadened microcanonical window. In particular, in Fig. 8(b) we compute the broadened microcanonical average of the von Neumann entropy,

$$\langle S^{\mathrm{vN}} \rangle_{\lambda,\delta} = \sum_E \mathcal{D}^{-1}(\lambda) G_\delta(E - \lambda) S^{\mathrm{vN}}(|E\rangle). \tag{4.12}$$

Comparing Figs. 8(a) and 8(b), we observe that the variational states have an order of magnitude less entanglement than the eigenstates which they are superpositions of. Figure 8(b) also contains the Page entropy, i.e., the average entanglement of states drawn Haar-randomly from the full Hilbert space [49]. We find that the $\lambda = 0$ microcanonical average of the entanglement entropy is quite close to the Page value, providing an additional check that the MFIM is highly nonintegrable for the chosen parameters.

The low entanglement of the variational states can be attributed to the fact that they are generated by low-depth circuits, which makes them atypical. In Fig. 8(c) we plot the average number of layers yielding convergence, which is a proxy for the circuit depth needed to prepare the converged variational state (the depth of a p-layer ansatz circuit is $3p$; see Fig. 3). We show the behavior for $\lambda/N = -0.5$, but we find a linear scaling in N for the other energy densities as well (with slightly different slopes) such that they require a similar circuit depth. Since the total number of variational parameters is $2Np^*$, the total number of gates scales roughly quadratically in N. The actual number of variational parameters at $N = 13$ and $\lambda/N = -0.5$ is only about 78.

We also briefly consider how the entanglement and the number of layers in the variational ansatz depend on the tolerance δ. In particular we consider exponents $\alpha = -0.5, -0.75, -1$ [see Eq. (3.16)]; the addition of tolerances stricter than $\alpha = -0.5$ used elsewhere in this work limits the numerically accessible system sizes to around $N = 8$, similar to Ref. [29]. Figure 9 shows that both entanglement and circuit depth increase with $|\alpha|$. For example, for $|\alpha| = 1$, already at $N = 8$ we need $p^* \approx 5$ layers for convergence, making this scaling prohibitive for the classical optimizer. It appears that $p^*(N)$ grows much faster with N at $|\alpha| = 1$ than it does for $|\alpha| \leqslant 0.75$. The fact that larger entanglement and higher circuit depths would be needed to prepare variational states

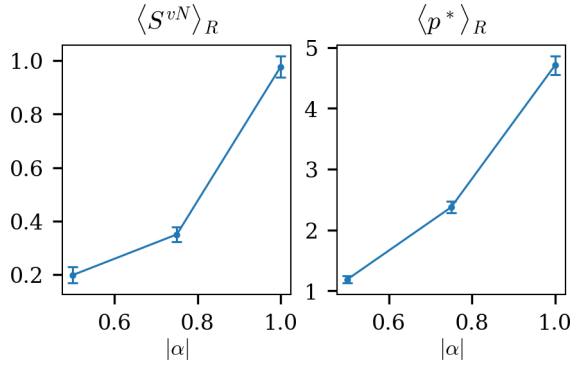

FIG. 9. Effect of the tolerance parameter on the half-chain ($|S| = 4$) entanglement entropy and circuit depth of the VME, for $N = 8$ at target energy density $\lambda/N = -0.75$. The results are averaged over 48 variational states with error bars being the standard error.

with smaller energy variance is consistent with other studies (e.g., Ref. [24]).

V. CONCLUSION AND OUTLOOK

In this work we propose a VQA for estimating Gaussian microcanonical averages of local operators at intermediate energy density. Given the target average energy λ and a microcanonical width of $O(N^{-1/2})$, the variational algorithm evolves random product states into weakly entangled states whose diagonal ensembles are approximately Gaussian microcanonical on average. We have systematically examined what we call the diagonal and off-diagonal contributions to the error in this estimation, and found that the latter is the dominant source of error. The mean-square off-diagonal error is on the one hand parametrically reduced by ensemble averaging as R^{-1} for small R, but on the other hand for large R saturates to a small value $|c|$ whose size depends principally on the observable under consideration. We have left the identification of the mechanism behind this bias and methods to remove it for future work.

We have also examined the performance of the algorithm in an observable-independent way by computing the trace distance between the subsystem variational ensemble and the subsystem microcanonical one, which appears to continually decrease with N when we take $R = O(N^2)$ and $\lambda/N = -0.75, -0.5$ (though we cannot rule out saturation), whereas for $\lambda/N = -0.25, 0$ there is not a consistent decay. We find this result interesting because for other finite-temperature VQA methods, intermediate temperatures (as opposed to infinite temperature) are more difficult to simulate [17,19]. Since the number of variational parameters appears to scale roughly as $O(N^2)$, this suggests that a classical optimizer could handle the optimization at larger system sizes. The main bottleneck in the classical simulation of our proposed VQA is the repeated evaluation of the cost function; it would be useful to implement a more sophisticated classical simulation of the QPU, for example by tensor network methods so that larger systems could be reached. This would help to decide if the trend in the trace distance continues for larger N.

Before concluding, let us briefly discuss the complexity of the VME algorithm. Neglecting the nonzero bias constant $|c|$ on the basis that it is small, we note that the computational time complexity of VME is $O(MRGS)$ where M is the number of times the cost function is requested during the optimization, R is the number of states in the variational ensemble, G is the number of gates in the variational circuit, and S is an upper bound on the number of shots needed to estimate the cost function during each evaluation.

Let $\Omega = (H - \lambda)^2$. In the beginning of the optimization, when the state $|\psi(\boldsymbol{\theta})\rangle$ is basically a product state, we have

$$\mathrm{Var}(\Omega) \sim O(N^2), \tag{5.1}$$

implying that $O(N^2)$ shots are needed to get a system-size-independent statistical error. This can then be further reduced to the desired tolerance by an $O(1)$-in-system-size multiplicative factor. Towards the end of the optimization, we will end up with a state having

$$\mathrm{Var}(\Omega) \sim O(N^{-2}) \tag{5.2}$$

(neglecting the small non-Gaussian tails of $\rho_R(E)$ that were discussed in Sec. IV B). Thus, the most "shot costly" part of the optimization is in the beginning. To simply upper bound the resources, we thus take $S = O(N^2)$.

Now, we have observed empirically that $G = O(N^2)$, and that the off-diagonal error scales as $R^{-1/2}$ (again neglecting $|c|$). Therefore, since the diagonal error is $O(N^{-1})$, we may take $R = O(N^2)$ and conclude that the time complexity of VME is $O(MN^6)$ in achieving a statistical error of $O(N^{-1})$. Whether or not the number of calls, M, is exponentially large in N is still an open question for all variational quantum algorithms.

In Sec. IV A we saw that the diagonal error vanishes as $O(\delta/N)$ with increasing system size, so that even product states [whose typical energy width is $O(\sqrt{N})$] with target energy λ would suffice for a vanishing diagonal error, provided the ETH holds. An ensemble of R random product states would also yield an off-diagonal error proportional to $R^{-1/2}$, but it is unclear if one can variationally squeeze those states onto a microcanonical window without violating the unbiased condition $\mathbb{E}[x_r] = 0$.

As far as variational algorithms are concerned, the VME could be considered as an example of a broader class of VQAs where the convergence criterion is based on the value of the cost function rather than its gradient. Furthermore, the smallness of the cost function at convergence is only $O[1/\mathrm{poly}(N)]$ while all initial random product states were able to converge, so it would be interesting to study if the well-known barren-plateau problem [13,14] is less significant in this setting.

ACKNOWLEDGMENTS

The authors acknowledge valuable discussions with Anatoly Dymarsky, Andreas Elben, Ronak Tali, Milan Kornjača, Yihua Qiang, Ana-Marija Nedić, Niladri Gomes, Yong-Xin Yao, and Laura James. This material is based upon work supported by the National Science Foundation under Grant No. DMR-2038010.

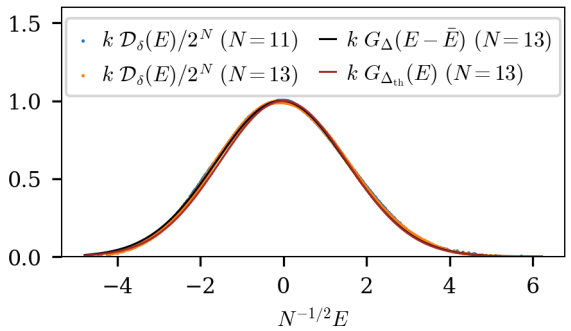

FIG. 10. The broadened density of states of the MFIM plotted against $N^{-1/2}E$ so that the form of the curves is N independent. At $N = 13$, a Gaussian best fit yields parameters $\gamma = \Delta^2/N = 2.47$ and $\bar{E}/\Delta = -0.03$ at $N = 13$. The theoretical value calculated directly from H with $\gamma_{\mathrm{th}} = \Delta_{\mathrm{th}}^2/N = 2.35$ is also plotted and seen to agree well. The prefactor k depends on the curve and is chosen so that the maximum value of each curve is 1 [the factor is $(2\pi\Delta^2)^{1/2}$ when the width is Δ].

APPENDIX A: DENSITY OF STATES

Let $G_y(x) = (2\pi y^2)^{-1/2} e^{-x^2/2y^2}$ be a normalized Gaussian window function with zero mean and standard deviation y. We maintain the convention that $f(Q) = \sum_q f(q)\,|q\rangle\,\langle q|$ for some Hermitian operator Q with eigenvalues q. We find numerically that the broadened density of states $\mathcal{D}_\delta(\lambda) = \mathrm{tr}\, G_\delta(H-\lambda)$, with broadening parameter $\delta = (\Delta E/N)N^{-1/2}$, is smooth (away from the tails of the spectrum) and can be approximated by a Gaussian of the form $2^N G_\Delta(E) = 2^N(2\pi\Delta^2)^{-1/2} e^{-E^2/2\Delta^2}$ with $\Delta^2 = \gamma N$ and γ becoming N independent in the thermodynamic limit. The density of states and a Gaussian best-fit curve are shown in Fig. 10. Following the method discussed in the Appendix of Ref. [50], we can check if this approximation is reasonable by estimating the parameter γ from the Hamiltonian directly. Note that in terms of the supposed form of the density of states and in the thermodynamic limit the following equality should hold:

$$\mathrm{tr}(H^2) = \int_{-\infty}^{\infty} \mathrm{d}E\, \mathcal{D}(E) E^2 = 2^N \gamma N. \tag{A1}$$

With periodic boundary conditions it can be checked for the Hamiltonian (4.1) that $\mathrm{tr}(H^2) = 2^N[(1+h_z^2)N + \sum_j h_{xj}^2]$, but since the random fields h_{xj} have been chosen to all be within 1% of the central value h_x, we can safely approximate $\mathrm{tr}(H^2) \approx 2^N(1+h_x^2+h_z^2)N$, yielding the theoretical estimate $\gamma_{\mathrm{th}} = 1 + h_x^2 + h_z^2 = 2.35$, which is within about 5% of the best-fit value of γ. We also checked that the density of states is roughly unaffected when it is computed using the broadening parameter σ (with which the variational states converge) instead of the broadening parameter δ.

APPENDIX B: OFF-DIAGONAL CONTRIBUTION ASSUMING INDEPENDENT AND IDENTICALLY DISTRIBUTED RANDOM VARIABLES

In Sec. II, we claimed that should the order-one fluctuations $R_{EE'}$ be actual independent and identically distributed random variables, then the off-diagonal contribution to Eq. (2.4) would be typically $O(\mathcal{D}^{-1/2}(\lambda))$. To see this, assume for $E > E'$ that $R_{EE'}$ are samples from an underlying distribution satisfying $\mathbb{E}[R_{EE'}] = 0$, $\mathbb{E}[R_{EE'}^2] = 1$, and $\mathbb{E}[R_{EE'}R_{E''E'''}] = 0$ for $E \neq E''$, $E' \neq E'''$, and $E'' > E'''$. We restrict the energies to the upper triangular part of the R matrix since $R_{EE'} = R_{E'E}$ for energy-basis real observables. Let

$$\begin{aligned} x_1 &= 2 \sum_{E>E'\in W} c_E c_{E'}\, \langle E|A|E'\rangle \\ &= 2 \sum_{E>E'\in W} c_E c_{E'} \mathcal{D}^{-1/2}(\bar{E}) f(\bar{E},\omega) R_{EE'}, \end{aligned} \tag{B1}$$

where in the second equality we have inserted the ETH matrix element ansatz. The above is notation consistent with Sec. III A, where x_1 is the "off-diagonal error" for a single microcanonical superposition state supported only on the microcanonical window W. Clearly we have $\mathbb{E}[x_1] = 0$ and

$$\begin{aligned} \mathbb{E}[x_1^2] = 4 \sum_{\substack{E>E'\in W \\ E''>E'''\in W}} & c_E c_{E'} c_{E''} c_{E'''} \mathcal{D}^{-1/2}(\bar{E}) \mathcal{D}^{-1/2}(\bar{E}'') \\ & \times f(\bar{E},\omega) f(\bar{E}'',\omega'') \mathbb{E}[R_{EE'}R_{E''E'''}], \end{aligned} \tag{B2}$$

where $\bar{E}' = (E''+E''')/2$ and $\omega'' = E''-E'''$. The independence assumption collapses the quadruple sum, giving

$$\mathbb{E}\big[x_1^2\big] = 4 \sum_{E>E'\in W} c_E^2 c_{E'}^2 \mathcal{D}^{-1}(\bar{E}) f^2(\bar{E},\omega). \tag{B3}$$

Normalization of the state implies $c_E^2 = O(1/n)$, where n is the number of eigenenergies in W, and the double sum runs over $n(n-1) = O(n^2)$ energies. Altogether we have $\mathbb{E}[x_1^2] = O(\mathcal{D}^{-1}(\lambda))$ since λ is a typical energy in the window, and where we have neglected N dependence of $f(\bar{E},\omega)$ (see Ref. [39]). In the thermodynamic limit then, the central limit theorem implies that $x_1 \approx O(\mathcal{D}^{-1/2}(\lambda))$.

APPENDIX C: EXPECTATION VALUE OF $(H-\lambda)$ IN BROADENED MICROCANONICAL ENSEMBLE

Here we justify Eq. (2.2), which expresses the relation between the smooth function $A(E)$ appearing in the ETH matrix element ansatz and the microcanonical expectation value $\langle A\rangle_{\lambda,\delta}$, by considering the expectation value of $(H-\lambda)$ in the smooth microcanonical ensemble. We take the density of states to be the Gaussian function described in Appendix A. Under such an assumption, the first few terms of the Taylor series for the density of states $\mathcal{D}(E) = 2^N(2\pi\Delta^2)^{-1/2} e^{-E^2/2\Delta^2}$ near energy λ read

$$\begin{aligned} \mathcal{D}(E)/\mathcal{D}(\lambda) = 1 - \frac{\lambda}{\Delta^2}(E-\lambda) + \left[\left(\frac{\lambda}{\Delta^2}\right)^2 - \frac{1}{\Delta^2}\right]\frac{(E-\lambda)^2}{2} \\ + \frac{1}{\Delta^4}\left(3\lambda - \frac{\lambda^3}{\Delta^2}\right)\frac{(E-\lambda)^3}{6} + \cdots. \end{aligned} \tag{C1}$$

Here, $\Delta^2 = \gamma N$ with $\gamma \approx 1 + h_x^2 + h_z^2$, and we have not assumed that $E-\lambda$ is small in any sense yet, only that the density of states admits a Taylor expansion in the thermodynamic

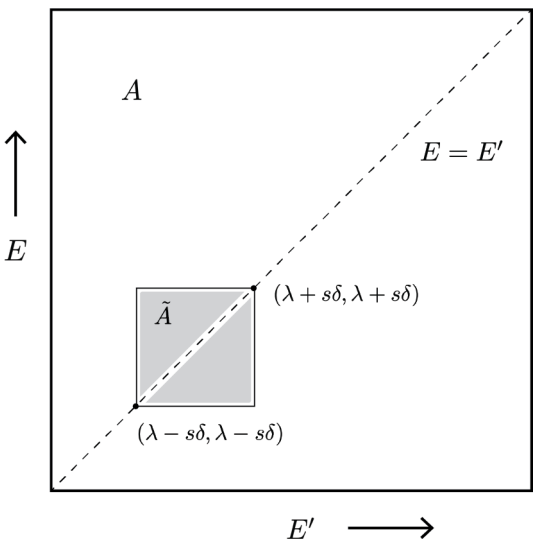

FIG. 11. In shaded gray, the off-diagonal-only submatrix $\tilde{A}$ of A relevant to off-diagonal error for microcanonical superpositions whose weight is mostly within s standard deviations δ of the central energy λ. Matrix elements are shown on the energy scale rather than the eigenvalue index scale.

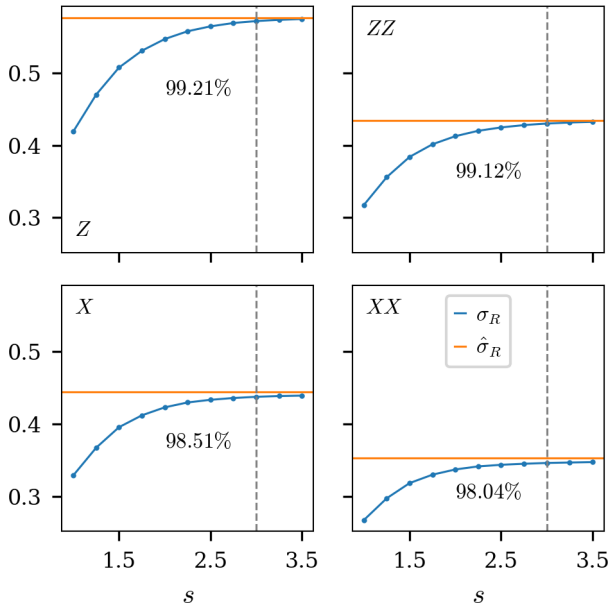

FIG. 12. For various operators A, the effect of expanding the window to include s standard deviations around λ for $N = 13$, $R = 288$, and when $\lambda/N = -0.5$. We see that at $s = 3$, we have captured basically all of $\hat{\sigma}_R$ as shown by the percentages. Other energy densities are unremarkable except that at $\lambda/N = -0.75$ only about 97% of XX is captured.

limit. If we ignore any error incurred in replacing sums by integrals in the thermodynamic limit, where the energy bandwidth approaches infinity and the level spacing zero, the first moment of $(H - \lambda)$ in the broadened microcanonical ensemble $\rho_{\lambda,\delta} = \sum_E \mathcal{D}^{-1}(\lambda) G_\delta(E - \lambda) |E\rangle \langle E|$ reads

$$\mathrm{tr}[(H - \lambda)\rho_{\lambda,\delta}] = -\frac{\delta^2 \lambda}{\Delta^2} + \frac{3\delta^4}{6\Delta^4}\left(3\lambda - \frac{\lambda^3}{\Delta^2}\right) + \cdots . \quad \text{(C2)}$$

We can then use these formulas to estimate the expectation value of an ETH-obeying operator in this broadened microcanonical ensemble. In doing so, an important consequence of the ETH is that the smooth function $A(E)$ should be expressible as a function of energy density in the thermodynamic limit (see Fig. 2 in the main text). In this paper we consider only Pauli-string-type observables. In this case, note that $A(E) = a(E/N)$ is $O(1)$ because $A(E)$ is defined via (a best-fit curve to) the averaging procedure

$$\frac{1}{K}\sum_{E'} \langle E'|A|E'\rangle \quad \text{(C3)}$$

over K eigenstates near $|E\rangle$, and $|\langle E'|A|E'\rangle| \leqslant 1$ for Pauli strings. Now since $A(E) = a(E/N)$, it follows that $a(x) = O(1)$ and

$$\left.\frac{\mathrm{d}A}{\mathrm{d}E}\right|_E = \frac{1}{N}\left.\frac{\mathrm{d}a}{\mathrm{d}x}\right|_{E/N} = O\left(\frac{1}{N}\right), \quad \text{(C4)}$$

and similarly for higher derivatives. Now consider a broadened microcanonical ensemble at energy λ, i.e., $\rho_{\lambda,\delta} = \mathcal{D}^{-1}(\lambda) G_\delta(H - \lambda)$. Then, the expectation value of the *smooth ETH function* in this ensemble is obtained by going to the continuum and combining Eqs. (C1) and (C2). The result

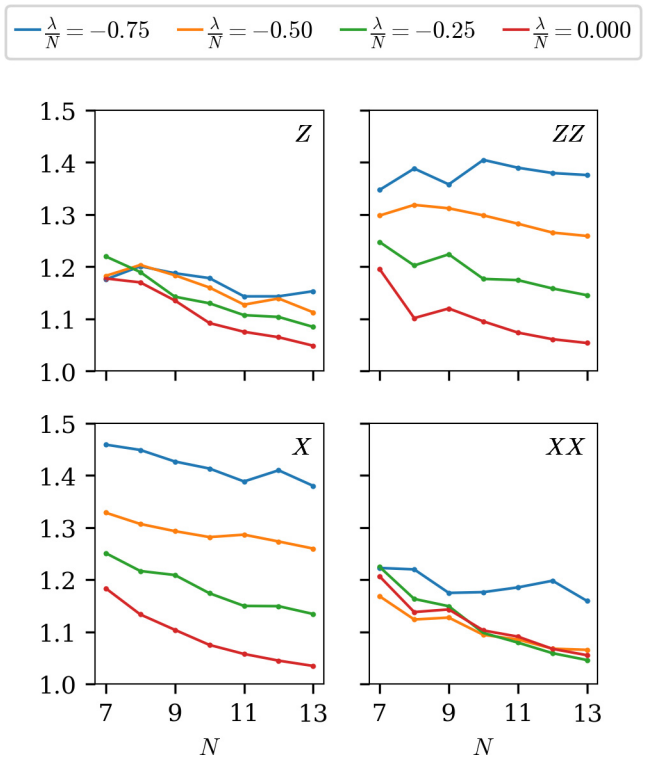

FIG. 13. At various energy densities λ/N, the maximum singular value of $\tilde{A}$ for various A, with $\tilde{A}$ the 3δ large submatrix of A and $\delta = O(N^{-1/2})$, i.e., with $s = 3$ as in Fig. 11.

reads

$$\int \mathrm{d}E \frac{\mathcal{D}(E)}{\mathcal{D}(\lambda)} G_\delta(E - \lambda) A(E) = A(\lambda) + O(\delta^2/N). \quad \text{(C5)}$$

From this and the ETH ansatz follows Eq. (2.2).

APPENDIX D: ADDITIONAL DISCUSSION OF DIAGONAL ENSEMBLE

In this Appendix we explain the deviations in μ and σ from λ and δ, respectively, when fitting the converged variational

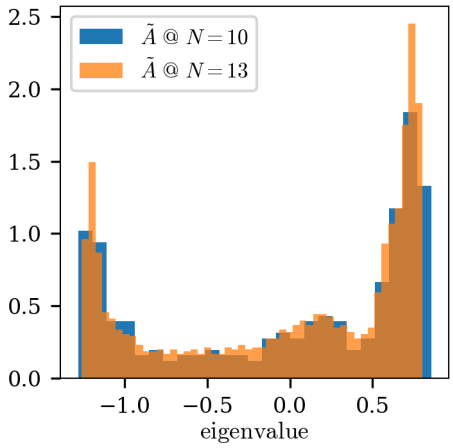

FIG. 14. For the operator $A = X$ at energy density $\lambda/N = -0.5$, probability density functions of the eigenvalues of $\tilde{A}$ as in the orange histogram of Fig. 5(c) in the main text, but here shown for $N = 10$ in blue and for $N = 13$ in transparent orange.

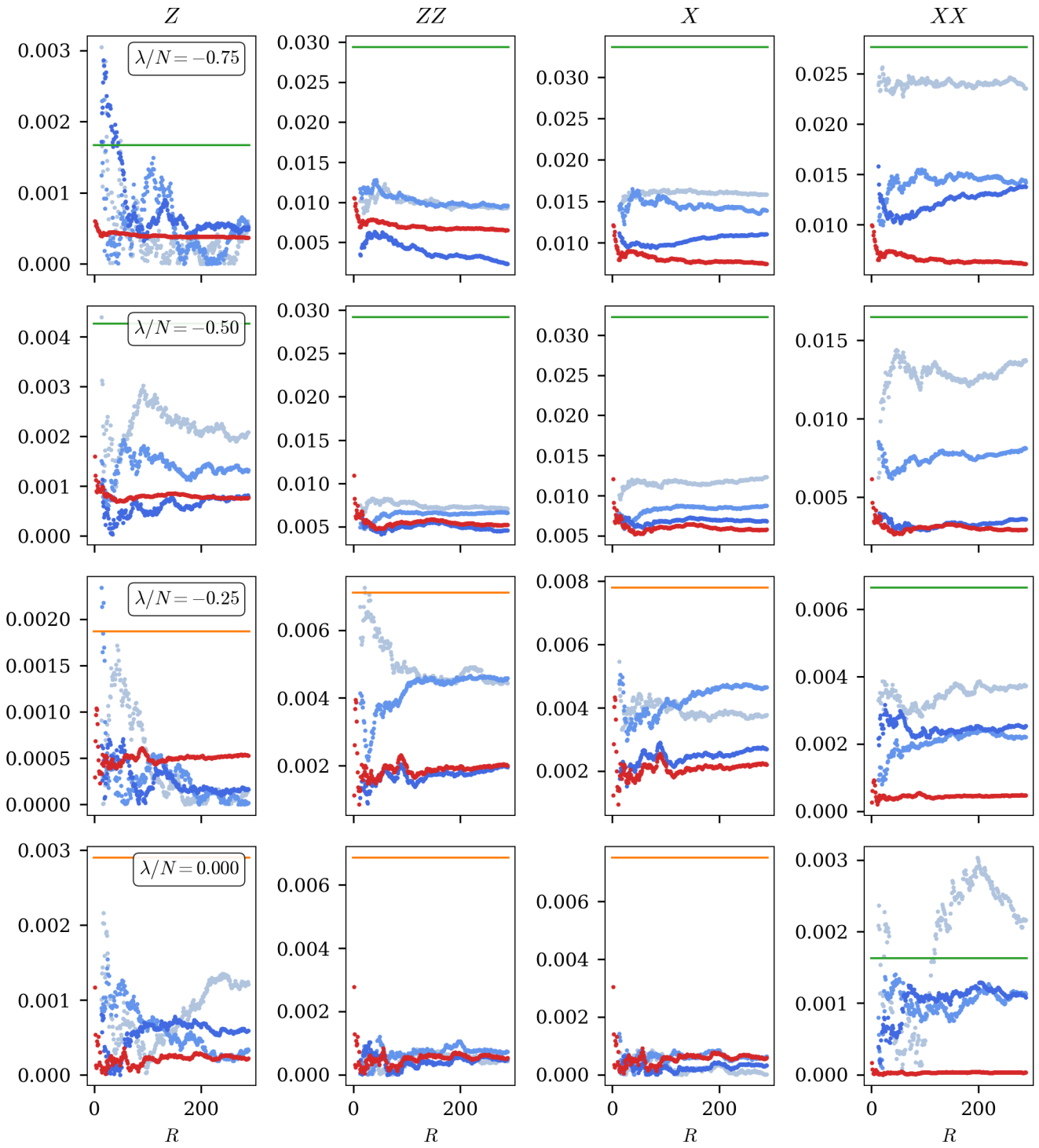

FIG. 15. Diagonal error in the VME. Curves are interpreted identically to those in Fig. 4(b) in the main text, except to bring the estimate $(\delta/N)|a'(\lambda/N)|$ down to scale we plot instead in some cases $(\delta/4N)|a'(\lambda/N)|$. The latter are shown in orange instead of green to indicate the use of a constant scale factor.

ensembles ρ_R to the best-fit curves $\rho_{\mu,\sigma}$. As was described in Sec. IV A of the main text, μ will undershoot λ because the density of states is nonuniform. We can confirm this more precisely as follows. Using the best-fit parameters to the variational ensemble (i.e., the values in Table I), we treat the spectrum as continuous and compute the numerical integral $\mathrm{tr}[\rho_{\mu,\sigma}(H-\lambda)] \approx -0.042$ at $\lambda/N = -0.5$. Doing the same for the ensemble whose central energy is λ, we find that the value of $\mathrm{tr}[\rho_{\lambda,\delta}(H-\lambda)] \approx 0.142$. In the latter calculation, we emphasize that this value is roughly the same whether using δ or σ for the width of the Gaussian. Thus, the ensemble with central energy μ actually minimizes the operator $(H-\lambda)$ much better than the ensemble with central energy λ. Thus the deviation in μ from λ is a finite-size effect due to a nonuniform density of states.

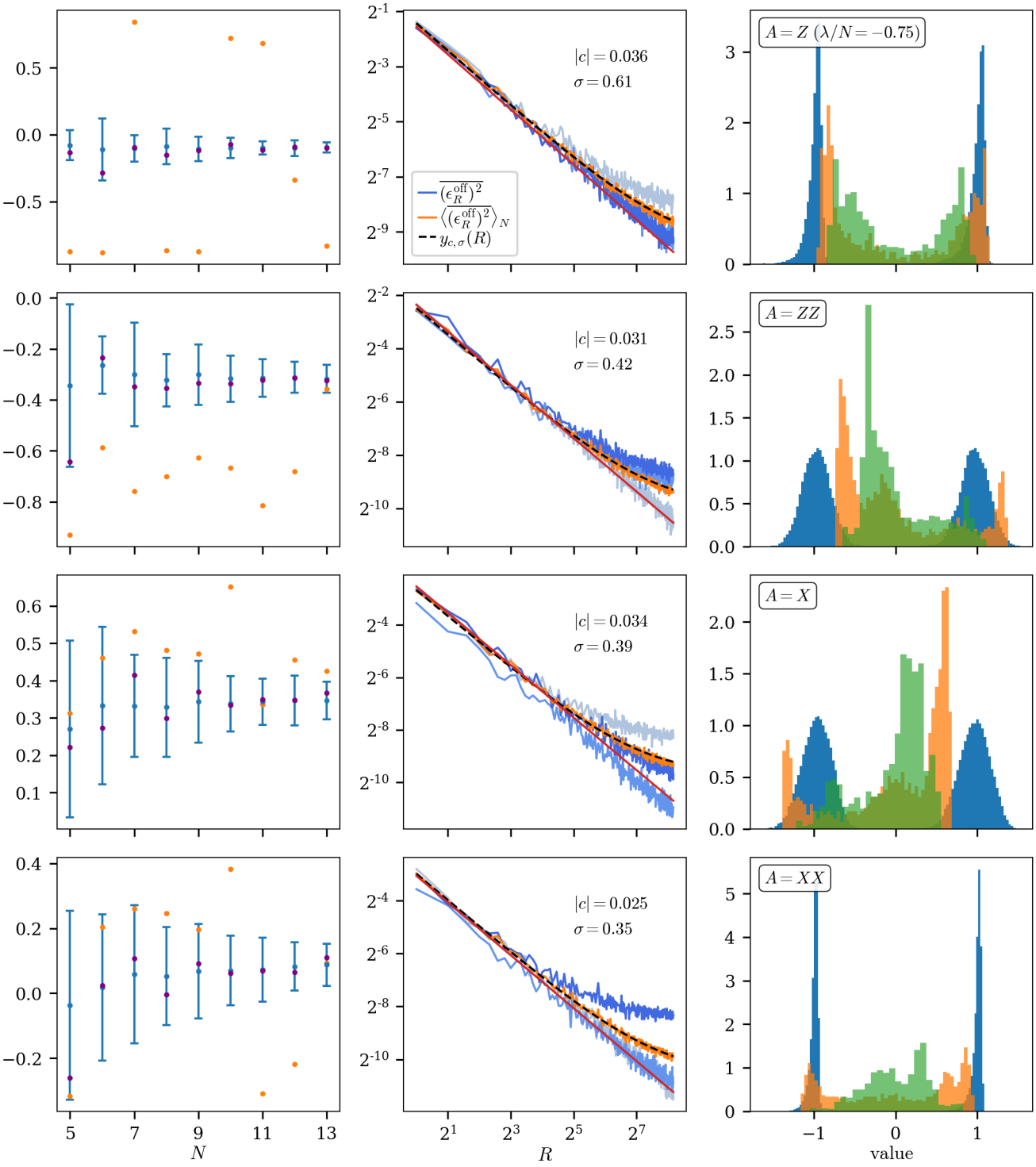

FIG. 16. Off-diagonal error in the VME for observables $A = Z, ZZ, X, XX$ at $\lambda/N = -0.75$. See the caption of Fig. 5 in the main text for further explanation of what is shown in the plots.

We now address the deviations in σ from δ, which we claim to be due to the slight "non-Gaussianity" of $\overline{\rho_R}(E)$, i.e., the excess weight outside the Gaussian window that we see in Fig. 4. Best-fit curves aside, we first check that the fluctuations of $\rho_R(E)$ around $\overline{\rho_R}(E)$ contribute negligibly to the expectation value of $(H-\lambda)^2$. We directly compute $\mathrm{tr}[\overline{\rho_R}(H)(H-\lambda)^2] \approx 0.92\,\delta^2$, and we can see that this value is consistent with the actual ensemble average value of the cost function, $\mathrm{tr}[\rho_R(H-\lambda)^2] \approx 0.90\,\delta^2$. The fact that these values are slightly less than δ^2 can be attributed to the convergence criterion only requiring that the variance (which is approximately the cost) be *at most* δ^2. Now comparing this to the Gaussian model best-fit ensemble,which predicts a cost function value of only $\mathrm{tr}[\rho_{\mu,\sigma}(H-\lambda)^2] \approx 0.69\delta^2$, we see that

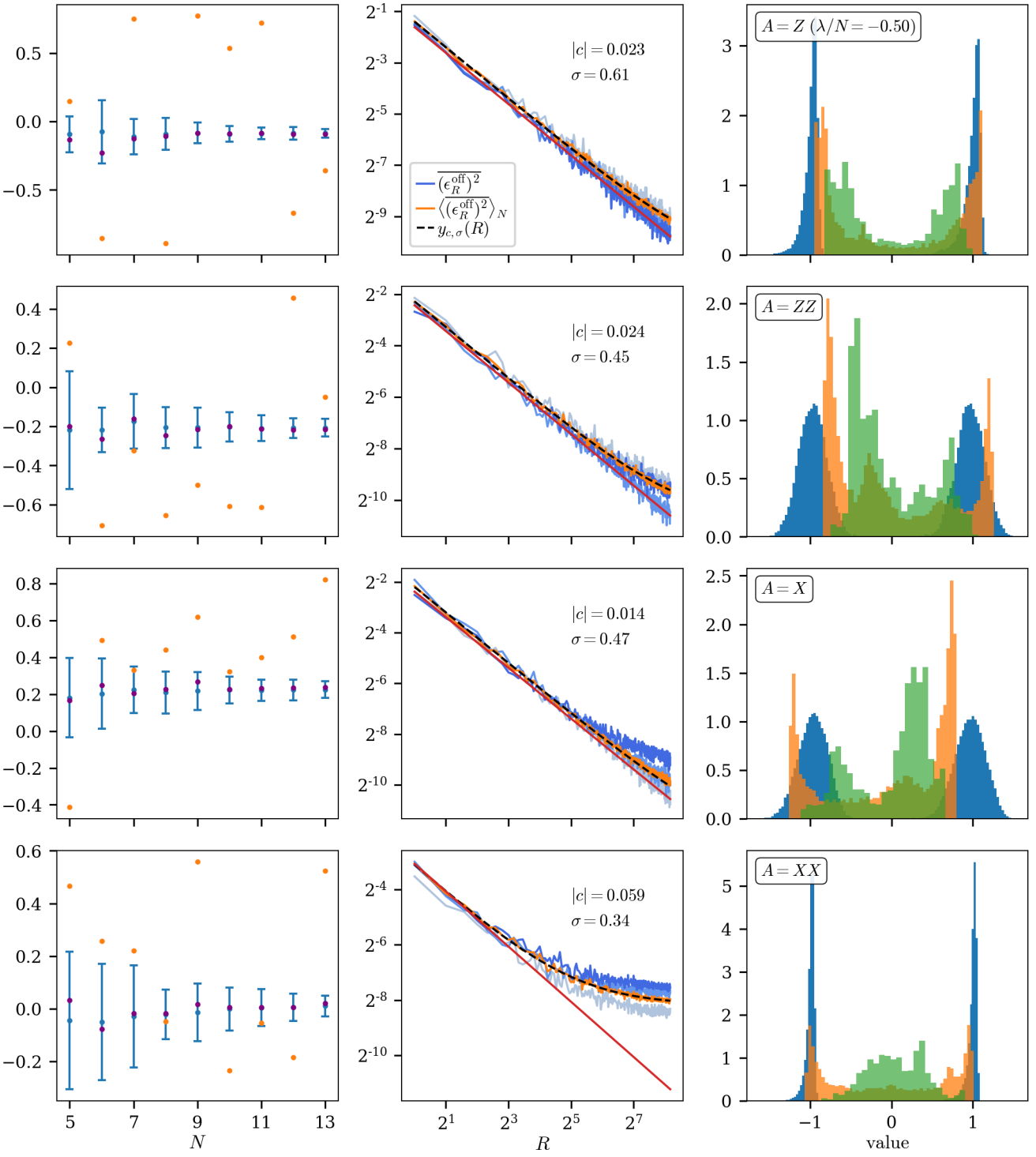

FIG. 17. Off-diagonal error in the VME for observables $A = Z, ZZ, X, XX$ at $\lambda/N = -0.5$. See the caption of Fig. 5 in the main text for further explanation of what is shown in the plots.

it undershoots δ^2 since it is missing the contribution from the excess energy weight.

APPENDIX E: JUSTIFICATION FOR REPLACING $\hat{A}$ BY $\tilde{A}$ AND FURTHER PROPERTIES OF $\tilde{A}$

In this Appendix we justify deleting certain energy-basis matrix elements of A based on the form of the variational states, for the purposes of gaining some intuition about the nature of the off-diagonal error. Let the operator $\hat{A}$ be A with its energy-basis diagonal elements set to zero, i.e.,

$$\langle E|\hat{A}|E'\rangle = \begin{cases} \langle E|A|E'\rangle & \text{if } E \neq E' \\ 0 & \text{if } E = E'. \end{cases} \tag{E1}$$

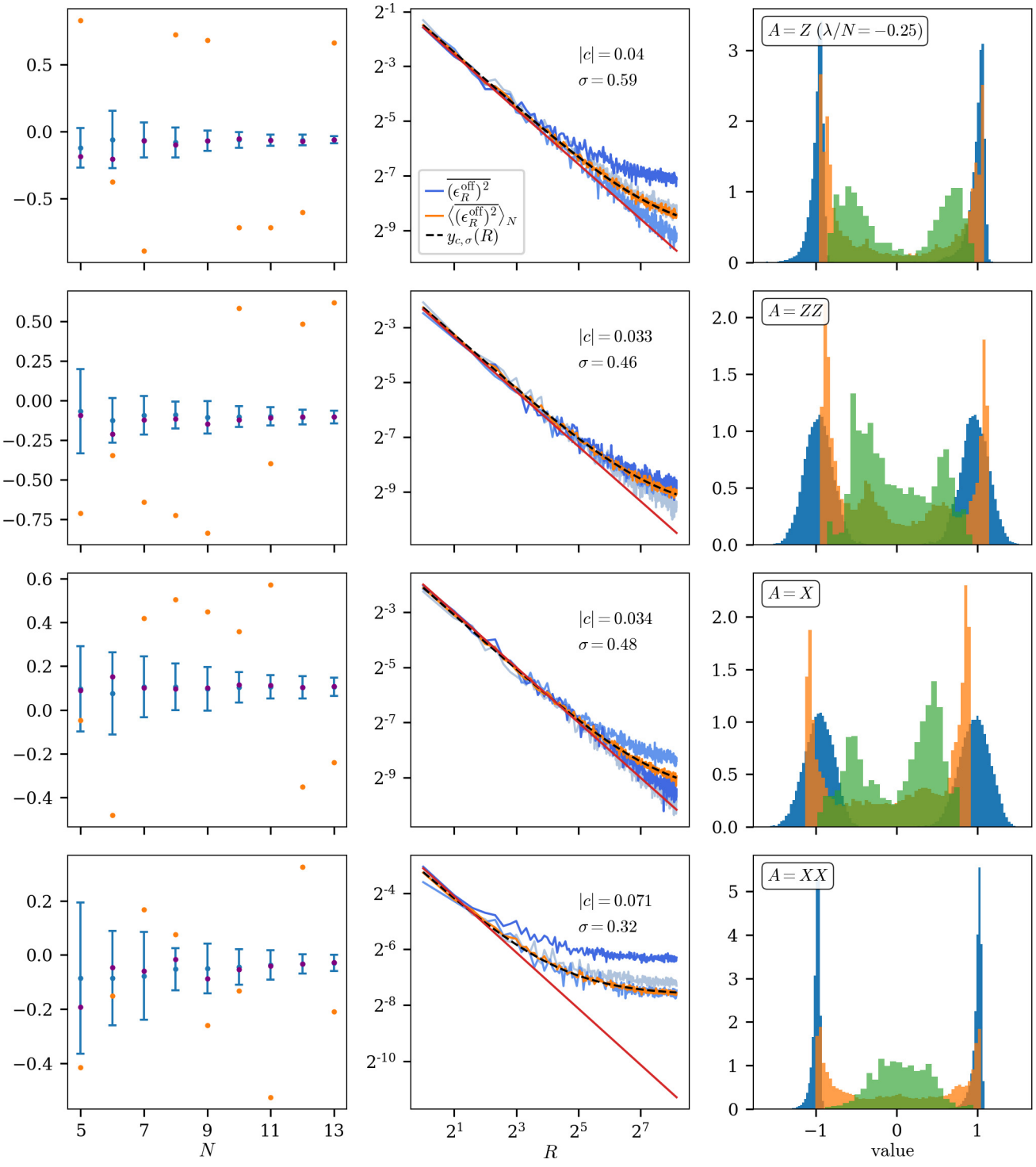

FIG. 18. Off-diagonal error in the VME for observables $A = Z, ZZ, X, XX$ at $\lambda/N = -0.25$. See the caption of Fig. 5 in the main text for further explanation of what is shown in the plots.

At this point we also define $\tilde{A}$ for general s, where s is the number of standard deviations δ around λ that are *not* deleted:

$$\langle E|\tilde{A}|E'\rangle = \begin{cases} \langle E|A|E'\rangle & \text{if } E \neq E',\ |E-\lambda| \leqslant s\delta, \\ & \text{and } |E'-\lambda| \leqslant s\delta \\ 0 & \text{otherwise.} \end{cases} \tag{E2}$$

An equivalent definition is given graphically in Fig. 11. Since the ensemble of variational states $|\psi_r\rangle$ approximates a Gaussian microcanonical ensemble near with average energy near λ *on average* [see Fig. 4(a)], we can anticipate that the value of $\langle\psi_r|\hat{A}|\psi_r\rangle^2$ will also be unaffected *on average* by replacing $\hat{A}$ with $\tilde{A}$ when $\tilde{A}$ has a sufficiently large energy support. Because of the established Gaussian form, we might expect that two

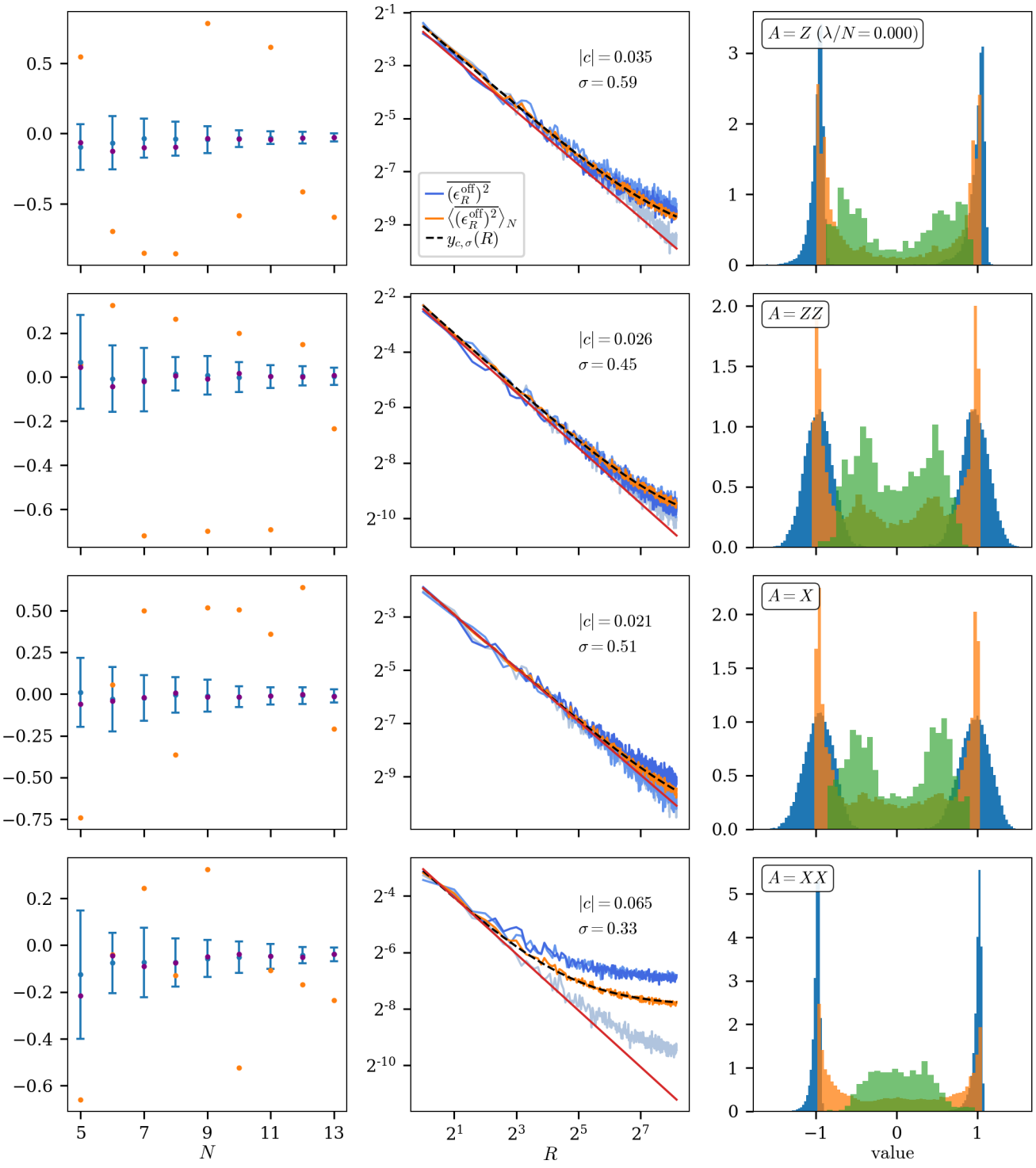

FIG. 19. Off-diagonal error in the VME for observables $A = Z, ZZ, X, XX$ at $\lambda/N = 0.0$. See the caption of Fig. 5 in the main text for further explanation of what is shown in the plots.

standard deviations around λ is always sufficient to capture 95% of the average absolute error. However, there are fluctuations around this behavior, and the coarse-grained variational states have some excess energy weight beyond the Gaussian best-fit curve. Furthermore, the variational ensembles are not exactly centered on λ. Thus, we justify replacing $\hat{A}$ with an appropriately chosen $\tilde{A}$ numerically as follows. In Fig. 12 we compare σ_R when computed on a window of size $2s \times 2s$ (see Fig. 11 for clarification) and $\hat{\sigma}_R$, which is computed from the entire spectrum. We show on the plots the fraction of $\hat{\sigma}_R$ that is captured by σ_R at $s = 3$, i.e., roughly three standard deviations, which we consider to be sufficiently large to capture basically all of $\hat{\sigma}_R$.

For the 3δ truncated operators $\tilde{A}$ we have just discussed, in this Appendix we also consider how the maximum singular value of $\tilde{A}$, i.e., the larger of $|\lambda_{\max}(\tilde{A})|$, $|\lambda_{\min}(\tilde{A})|$, scales

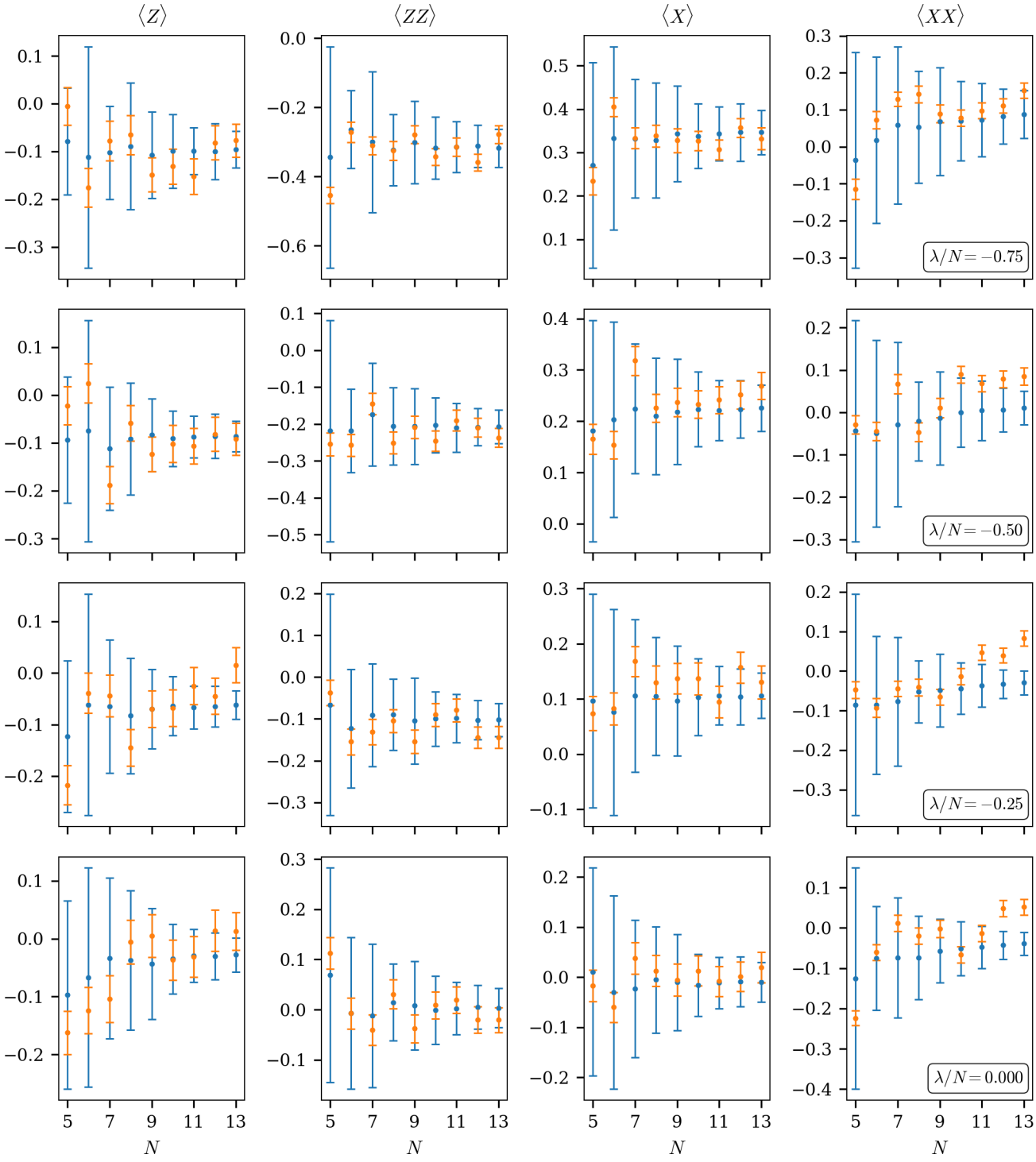

FIG. 20. Ensemble-averaged VME observable expectation values (orange) and broadened microcanonical averages (blue) plotted versus system size for the full range of energy densities and operators considered in this work. See the caption of Fig. 6 in the main text for further explanation of what is shown in the plots.

with N. The results are shown in Fig. 13. These results provide evidence that the numerical prefactor σ_R appearing in the statistical description of the off-diagonal error is $O(1)$ in system size. In this Appendix we also consider how the eigenvalue statistics qualitatively vary with N, with an example shown in Fig. 14 demonstrating that the distribution is qualitatively independent of N, except for the slight decrease of the distribution's width with N as reflected in Fig. 13.

APPENDIX F: ADDITIONAL NUMERICAL DATA FOR DIAGONAL ERROR

Figure 15 shows the diagonal error in the VME estimate for operators Z, ZZ, X, XX acting in the middle of the chain. As stated in the main text, the results for energy density $\lambda/N = -0.5$ appear to agree best with the prediction of ETH that the error should decay as $1/N$ and agree with χ_R/N [see

Eq. (3.8)] for large R. While in general such a clear scaling with N is missing, all cases show that the $N = 13$ diagonal error is never larger than some order-one fraction of the rough estimate $a'(\lambda/N)\delta/N$, with XX at $\lambda/N = 0$ showing the case where the diagonal error comes closest to the estimate. We also note that the ETH prediction χ_R/N for the diagonal error is always on the correct scale of the error for $N = 13$. The operator XX is also an outlier in this sense: χ_R/N significantly underestimates the actual error except for $\lambda/N = -0.5$.

APPENDIX G: ADDITIONAL NUMERICAL DATA FOR OFF-DIAGONAL ERROR

In Sec. IV of the main text, Fig. 5 demonstrates the behavior of the off-diagonal error for the operator X acting on the middle of the chain at the energy density $\lambda/N = -0.5$. The purpose of this Appendix is to establish that the trends observed there hold more generally across different operators and target energy densities. The plots shown also demonstrate that the operator XX systematically differs from the other considered observables. We show the equivalent of Fig. 5 for $\lambda/N = -0.75, -0.5, -0.25, 0$ and $A = Z, ZZ, X, XX$ acting on the central one or two sites of the chain. The results are shown in Figs. 16–19, respectively.

APPENDIX H: ADDITIONAL NUMERICAL DATA FOR EXPECTATION VALUES

Here we show in Fig. 20 the VME estimates for the four local observables $A = Z, ZZ, X, XX$ acting on the central one or two sites of the chain. We show these estimates when targeting four different energy densities $\lambda/N = -0.75, -0.5, -0.25, 0$ and when $R = 288$. As we observed in Sec. IV, the off-diagonal error generally does not systematically depend on N. Here we can see that for fixed R, beyond $N \sim 9$, the accuracy of the VME estimates indeed does not depend systematically on N except again for XX where it appears to increase with N, consistent with the $R = 288, N = 13$ value of the off-diagonal error for XX being largest in the plots in Appendix G. We see that the microcanonical estimates can be quite good, as for Z at $\lambda/N = -0.5$, or quite poor, as for XX at $\lambda/N = 0$. The better results are generally for the operators Z, ZZ, and X and at the lower target energy densities $\lambda/N = -0.75$ and $\lambda/N = -0.5$.

[1] M. B. Hastings, An area law for one-dimensional quantum systems, J. Stat. Mech. (2007) P08024.

[2] G. Vidal, Efficient Classical Simulation of Slightly Entangled Quantum Computations, Phys. Rev. Lett. **91**, 147902 (2003).

[3] T. Kuwahara, Á. M. Alhambra, and A. Anshu, Improved Thermal Area Law and Quasilinear Time Algorithm for Quantum Gibbs States, Phys. Rev. X **11**, 011047 (2021).

[4] E. M. Stoudenmire and S. R. White, Minimally entangled typical thermal state algorithms, New J. Phys. **12**, 055026 (2010).

[5] F. Schrodi, P. Silvi, F. Tschirsich, R. Fazio, and S. Montangero, Density of states of many-body quantum systems from tensor networks, Phys. Rev. B **96**, 094303 (2017).

[6] N. Schuch, M. M. Wolf, F. Verstraete, and J. I. Cirac, Computational Complexity of Projected Entangled Pair States, Phys. Rev. Lett. **98**, 140506 (2007).

[7] R. P. Feynman, Simulating physics with computers, Int. J. Theor. Phys. **21**, 467 (1982).

[8] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[9] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[10] J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, E. Grant, L. Wossnig, I. Rungger, G. H. Booth, and J. Tennyson, The variational quantum eigensolver: A review of methods and best practices, Phys. Rep. **986**, 1 (2022).

[11] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[12] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[13] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[14] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, Diagnosing barren plateaus with tools from quantum optimal control, Quantum **6**, 824 (2022).

[15] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[16] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the Hamiltonian variational ansatz, PRX Quantum **1**, 020319 (2020).

[17] G. Verdon, J. Marks, S. Nanda, S. Leichenauer, and J. Hidary, Quantum Hamiltonian-based models and the variational quantum thermalizer algorithm, arXiv:1910.02071.

[18] J. Foldager, A. Pesah, and L. K. Hansen, Noise-assisted variational quantum thermalization, Sci. Rep. **12**, 3862 (2022).

[19] J. Selisko, M. Amsler, T. Hammerschmidt, R. Drautz, and T. Eckl, Extending the variational quantum eigensolver to finite temperatures, arxiv:2208.07621.

[20] J. Wu and T. H. Hsieh, Variational Thermal Quantum Simulation via Thermofield Double States, Phys. Rev. Lett. **123**, 220502 (2019).

[21] M. Motta, C. Sun, A. T. K. Tan, M. J. O'Rourke, E. Ye, A. J. Minnich, F. G. S. L. Brandão, and G. K.-L. Chan, Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution, Nat. Phys. **16**, 205 (2020).

[22] O. Shtanko and R. Movassagh, Preparing thermal states on noiseless and noisy programmable quantum processors, arXiv:1910.02071.

[23] S. Lu, M. C. Bañuls, and J. I. Cirac, Algorithms for quantum simulation at finite energies, PRX Quantum **2**, 020321 (2021).

[24] M. C. Bañuls, D. A. Huse, and J. I. Cirac, Entanglement and its relation to energy variance for local one-dimensional Hamiltonians, Phys. Rev. B **101**, 144305 (2020).

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

[26] M. Srednicki, Chaos and quantum thermalization, Phys. Rev. E **50**, 888 (1994).

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

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

[29] F. Zhang, N. Gomes, Y. Yao, P. P. Orth, and T. Iadecola, Adaptive variational quantum eigensolvers for highly excited states, Phys. Rev. B **104**, 075159 (2021).

[30] S. Sugiura and A. Shimizu, Thermal Pure Quantum States at Finite Temperature, Phys. Rev. Lett. **108**, 240401 (2012).

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

[32] S. Goldstein, D. A. Huse, J. L. Lebowitz, and R. Tumulka, Thermal Equilibrium of a Macroscopic Quantum System in a Pure State, Phys. Rev. Lett. **115**, 100402 (2015).

[33] M. Rigol, V. Dunjko, and M. Olshanii, Thermalization and its mechanism for generic isolated quantum systems, Nature (London) **452**, 854 (2008).

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

[35] A. Dymarsky, Bound on Eigenstate Thermalization from Transport, Phys. Rev. Lett. **128**, 190601 (2022).

[36] A. Dymarsky and H. Liu, New characteristic of quantum many-body chaotic systems, Phys. Rev. E **99**, 010102(R) (2019).

[37] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[38] J. Haferkamp, Random quantum circuits are approximate unitary t-designs in depth $O(nt^{5+o(1)})$, Quantum **6**, 795 (2022).

[39] Technically, the function $f(\bar{E}, \omega)$ can have some N dependence for $|\omega| < O(N^{-1})$ but is expected to be N independent at larger $|\omega|$; see the Supplemental Material of Ref. [36] and see Ref. [28] where $|f(0, \omega)|$ is argued to increase as $N^{1/2}$ in the former regime. For $\bar{E}$ away from zero, less appears to be known about its N dependence, but in any case we neglect it, assuming that here the exponentially large density of states will suppress any N dependence of $f(E, 0)$ in the thermodynamic limit.

[40] That is, assuming λ is in the spectrum of H, or else the cost function is not zero but minimized for the nearest eigenstate, so that the cost function will still be exponentially small in the state $|\lambda\rangle$.

[41] Y. Shen, X. Zhang, S. Zhang, J.-N. Zhang, M.-H. Yung, and K. Kim, Quantum implementation of the unitary coupled cluster for simulating molecular electronic structure, Phys. Rev. A **95**, 020501(R) (2017).

[42] H. L. Tang, V. O. Shkolnikov, G. S. Barron, H. R. Grimsley, N. J. Mayhall, E. Barnes, and S. E. Economou, Qubit-ADAPT-VQE: An adaptive algorithm for constructing hardware-efficient Ansätze on a quantum processor, PRX Quantum **2**, 020310 (2021).

[43] D. D'Alessandro, *Introduction to Quantum Control and Dynamics* (Chapman and Hall/CRC, New York, 2007).

[44] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[45] M. C. Bañuls, J. I. Cirac, and M. B. Hastings, Strong and Weak Thermalization of Infinite Nonintegrable Quantum Systems, Phys. Rev. Lett. **106**, 050405 (2011).

[46] A. Çakan, J. I. Cirac, and M. C. Bañuls, Approximating the long time average of the density operator: Diagonal ensemble, Phys. Rev. B **103**, 115113 (2021).

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

[48] M. M. Wilde, *Quantum Information Theory* (Cambridge University Press, Cambridge, UK, 2013).

[49] D. N. Page, Average Entropy of a Subsystem, Phys. Rev. Lett. **71**, 1291 (1993).

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