Deep Reinforcement Learning for Preparation of Thermal and Prethermal Quantum States

Shotaro Z. Baba, Nobuyuki Yoshioka, Yuto Ashida, and Takahiro Sagawa

Phys. Rev. Applied **19**, 014068 — Published 26 January 2023

DOI: 10.1103/PhysRevApplied.19.014068

Deep reinforcement learning for preparation of thermal and prethermal quantum states

Shotaro Z. Baba,[1,*] Nobuyuki Yoshioka,[1,2,3,†] Yuto Ashida,[4] and Takahiro Sagawa[1,5]

[1] *Department of Applied Physics, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*
[2] *Theoretical Quantum Physics Laboratory, RIKEN Cluster for Pioneering Research (CPR), Wako-shi, Saitama 351-0198, Japan*
[3] *JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan*
[4] *Department of Physics, The University of Tokyo, Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan*
[5] *Quantum-Phase Electronics Center (QPEC), The University of Tokyo, Tokyo 113-8656, Japan*

We propose a method based on deep reinforcement learning that efficiently prepares a quantum many-body pure state in thermal or prethermal equilibrium. The main physical intuition underlying the method is that the information on the equilibrium states can be efficiently encoded/extracted by focusing only on a few local observables, relying on the typicality of equilibrium states. Instead of resorting to the expensive preparation protocol that adopts global features such as the quantum state fidelity, we show that the equilibrium states can be efficiently prepared only by learning the expectation values of local observables. We demonstrate our method by preparing two illustrative examples: Gibbs ensembles in non-integrable systems and generalized Gibbs ensembles in integrable systems. Pure states prepared solely from local observables are numerically shown to successfully encode the macroscopic properties of the equilibrium states. Furthermore, we find that the preparation errors, with respect to the system size, decay exponentially for Gibbs ensembles and polynomially for generalized Gibbs ensembles, which are in agreement with the finite-size fluctuation within thermodynamic ensembles. Our method paves a path to studying thermodynamic and statistical properties of quantum many-body systems in quantum hardware.

I. INTRODUCTION

Preparation of a desired quantum many-body state is an essential task that plays a significant role in quantum computing [1, 2], quantum metrology [3], and quantum communication [4]. Specifically, the thermal state is one of the most important targets in quantum state preparation tasks [5–11] from both theoretical and experimental viewpoints. A common strategy is to employ numerical methods such as CRAB [12, 13], GRAPE [14], and Krotov [15]. However, it must be noted that all these methods suffer from the exponential growth of computational cost, and furthermore require detailed knowledge about nonequilibrium properties of the system. Therefore, it is desirable to construct a preparation protocol that employs only a little knowledge during the learning task.

A surging technology to extract the essential feature in quantum systems with prohibitively large exploration space is machine learning, which has exemplified its capacity in a wide range of physics [16–37]. Successful applications include representations of quantum many-body states with neural networks (NNs) [18–30, 38, 39], quantum state tomography [31–33], and phase classification [34, 35, 40], to name a few. In particular, a branch of machine learning called reinforcement learning (RL) [41] has been recognized as a powerful tool to perform quantum state preparation [42–50]. The RL is designed to discover an efficient policy that maximizes a given reward through trial-and-error learning on the behavior of the environment. Several previous studies utilize the algorithm that adopts the deep RL framework; the quantitative evaluation of the action, or the reward, determined by the algorithm makes full use of the capability of NNs to approximate the high-dimensional nonlinear functions [45–52]. While a bulk of previous works choose fidelity as the reward, its computation for quantum many-body systems requires exponentially large resources in either numerics or experiment, and thus fidelity is not practical to scale up.

In this work, we propose a deep-RL-based method that only relies on local measurements to prepare thermal and prethermal pure states described by Gibbs and generalized Gibbs ensembles (GGEs) [53–56]. The underlying physical intuition is that, we may take advantage of the typicality of equilibrium states [7, 57–59] to prepare them using solely local observables, and not rely on global features such as fidelity. We numerically find that, although the deep RL agent is only informed of the local information on the thermodynamic ensembles, the accuracy of the prepared state improves exponentially with the system size for Gibbs ensembles, whereas the improvement is polynomial for GGEs.

The remainder of the paper is organized as follows. In Sec. II, we present an overview of the framework of the deep reinforcement learning. The application of reinforcement learning to quantum state preparation is described in Sec. III, which includes the core proposal of our work, i.e., the local preparation of thermal and prethermal pure quantum states leveraging the typicality of equilibrium states. After the framework is presented, we give the numerical demonstration for preparation of equilibrium states described by Gibbs ensembles and GGEs in Sec. IV and V, respectively. Finally, we give the conclusion and discussion in Sec. VI.

II. REINFORCEMENT LEARNING

The general framework of the reinforcement learning (RL) can be concisely expressed as a procedure to train an agent how to interact with the environment through optimization of cumulative reward [see Fig. 1]. The strengths of deep reinforcement learning, which are why we chose to employ it, are the high degree of freedom in reward design that makes the algorithm independent of the actual model and its ability to handle a huge search space of total actions, which amounts to $15^{200} \approx 10^{235}$ in our demonstration in integrable systems in Sec. V. The corresponding quantities (game tree complexity) of chess, shogi (Japanese chess), and Go, which are canonical environments for high-performance planning, are $\approx 10^{123}$, 10^{226} and 10^{360}, respectively [60–62]. The search space size of our problem is larger than that of shogi, and deep reinforcement learning is reasonable choice to achieve higher performance [63].

The goal of the RL is to discover the best policy π that outputs the sequence of actions $\mathcal{A} = \{a_t\}_t$ based on observations of the environment $\mathcal{O} = \{o_t\}_t$, such that the feedback realizes the most desired behavior quantified by the rewards $\mathcal{R} = \{r_t\}_t$. Typically, all the events are discretized so that each value can be well-defined at each time step t.

A practical strategy widely used in the community of machine learning is Q-learning. Here, one aims to find the best approximation of the optimal action-value function as [41]

$$Q^*(o,a) = \max_\pi \mathbb{E}_\pi \left[r_t + \sum_{n=1}^{\infty} \gamma^n r_{t+n} \middle| o_t = o,\ a_t = a, \pi \right], \tag{1}$$

which is the maximum sum of rewards r_t discounted by γ $(0 < \gamma < 1)$ in a stochastic policy that chooses action according to some probability distribution as $\pi(a|o) = \Pr(a|o)$. A powerful flavor of Q-learning uses the deep NNs to represent the action-value function, and hence referred to as the deep RL algorithms [64, 65]. The extraordinary representative power of NNs have been found to achieve successful applications of the deep RL algorithms in numerous fields that are not necessarily limited to computer science but also natural science, materials sciences, and so on.

In this paper, we focus on a non-distributed implementation [66] of a deep RL algorithm called R2D2 [67]. R2D2 is a type of deep Q-learning algorithm [68], and assumes the agent can obtain partial information about the state of the environment. As we show the architecture in Appendix A, the NN used in R2D2 includes a Long Short-Term Memory (LSTM) layer, and therefore the action-value function Q computed by the NN at step t depends not only on the instant observation o_t, but also on the previous observations $\{o_{t'}\}_{t' \le t}$ [69]. This feature enables the NN to handle time-series inputs and develop the capability in a partially observed environment.

III. REINFORCEMENT LEARNING FOR QUANTUM STATE PREPARATION

A. Global state preparation

Next, we review the general protocol to prepare a desired isolated quantum many-body state using the deep RL framework. Concretely, we aim to prepare a target quantum state $|\psi_{\text{target}}\rangle$ from an easily prepared quantum state $|\psi_0\rangle$, assuming that a set of unitary $\{U_i\}_i$ is available at any time step. By finding the best sequence of unitaries, we try to approximate the target state as

$$|\psi_{\text{target}}\rangle \approx \prod_t U_{i_t} |\psi_0\rangle . \tag{2}$$

It is straightforward to see that such a problem setup is in a great connection with the RL; we identify the quantum many-body system with the environment and the available set of unitaries $\{U_i\}_i$ with the action candidates $\{a_t\}$ at each time step.

Regarding the observation o_t, many works have proposed to use the results for the measurements on the target system [45–47]. Meanwhile, when the both initial and target states are fixed during the whole training, we can expect that the action history $\{a_{t'}\}_{t' \le t}$ contains enough information to find out the desired protocol [43, 44].

As for the reward r_t, numerous existing works have considered global features such as the fidelity $F(\rho_t, \rho_{\text{target}}) = \left(\text{Tr}\sqrt{\sqrt{\rho_t}\rho_{\text{target}}\sqrt{\rho_t}}\right)^2$ [42, 43, 48, 50], where ρ_t is the density operator of the controlled system at time step t. We hereafter refer to such protocols as *global* state preparation protocols. These methods have successfully prepared ground states of quantum many-body spin systems [42], metastable states of the quantum Kapitza oscillator [43], and highest excited states of multi-level quantum systems [50].

Having bridged between the notions in quantum control and the RL, we can train the NN-based agent to find the best control on the quantum system via searching for approximation of the optimal action-value function $Q^*(o,a)$ (1). Note that the evolution of the quantum state may either be experimentally implemented or numerically simulated, as long as the reward function for the deep RL agent can be readily obtained. Once the training is completed, we determine the preparation protocol as $\{a_t^*\}_t$ by choosing the actions so that $a_t^* = \arg\max_a Q(o_t, a)$ at each time step t, which shall maximize the reward.

B. Local state preparation

Now we are ready to describe the preparation protocol for thermal and prethermal pure quantum states that solely relies on local observables, instead of querying for costly global features such as the fidelity. In the following, as opposed to the *global* state preparation, we refer to the following scheme as *local* state preparation.

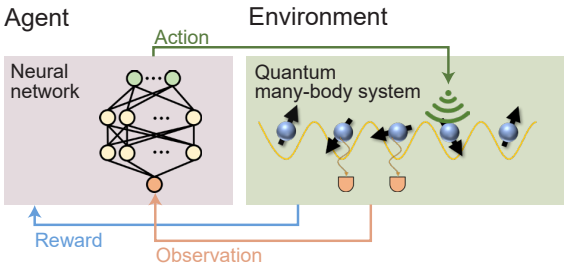

FIG. 1. A graphical illustration of the deep RL framework to prepare thermal and prethermal pure quantum states for thermodynamic ensembles. The agent represented by the deep NN is trained to discover a policy that maximizes the cumulative reward, which is computed based on the local observables of the quantum many-body system (environment). At each time step, the NN takes the result of observation as input and outputs the action-value function, from which the action (or the control operation) on the environment is determined. In this work, the action is chosen from a set of unitaries that are presumably available in the preparation protocol.

The central idea of the local state preparation is to leverage the typicality of pure quantum states [7, 57–59]. The typicality refers to the fact that the overwhelming majority of pure quantum states with the same local conserved quantities are indistinguishable by local observables. Such states are also deemed to be *macroscopically indistinguishable*. It is natural to expect that, by utilizing the typicality, we can prepare a state that encodes the macroscopic property of the equilibrium states by solely controlling local observables. Specifically, we perform local preparation by making the expectation values of local observables close to the equilibrium state, and then letting the system relax to equilibrium through a unitary evolution without any control.

We may take various forms of reward to learn the local observables of the target ensemble. In this work, we formulate the reward function as the inverse of the deviations of the expectation values of local observables from that of the target states:

$$r_t := \frac{1}{|M_t - M_{\text{target}}| + \varepsilon}, \tag{3}$$

where $M_t = (\langle O_1\rangle_t, \langle O_2\rangle_t, \ldots)$ and $M_{\text{target}} = (\langle O_1\rangle_{\text{target}}, \langle O_2\rangle_{\text{target}}, \ldots)$ are vectors consisting of expectation values $\langle O\rangle_{t(\text{target})} := \text{Tr}[\rho_{t(\text{target})}O]$. The small constant ε is also introduced to prevent divergence of the reward function.

As a concrete target for the demonstration of our local preparation protocol, we choose Gibbs ensembles and GGEs as illustrative examples. The evolution of the quantum many-body state is numerically simulated in an exact manner, while in principle we may also employ approximate methods that rely on, e.g., a variational representation such as tensor networks or neural networks. One may alternatively implement the proposed protocol directly on experimental device as well. In the following, we proceed to describe the detailed properties of the thermodynamic ensembles and the expected preparation efficiency under the presence of typicality.

1. *Gibbs ensemble for non-integrable systems*

Let us consider a non-integrable system with the energy being the only local conserved quantity such that its equilibrium is described by a Gibbs ensemble. Relatedly, pure states belonging to a given microcanonical shell of the system share their macroscopic properties. This class of typicality is referred to as canonical typicality. One of the most prominent example of canonical typicality can be illustrated under the Haar measure on the space of pure states under some constraint R (e.g., energy) as [58]

$$\langle \|\rho_A - \Omega_A\|_1 \rangle \le \sqrt{\frac{d_A^2}{d_R}}, \tag{4}$$

where $\rho_A = \text{Tr}_{\bar{A}}[|\psi\rangle\langle\psi|]$ is the reduced density operator obtained by tracing out the the complement of subsystem A for $|\psi\rangle\langle\psi|$ with Hilbert space dimension d_A. On the other hand, $\Omega_A = \text{Tr}_{\bar{A}}[\mathbb{1}_R/d_R]$ is the reduced density operator for the projection operator $\mathbb{1}_R$, i.e., the maximally mixed state in the Hilbert subspace under the constraint R whose dimension is d_R. Note that the bracket $\langle\cdot\rangle$ concerns the average regarding the Haar measure on the constrained Hilbert space, and $\|A\|_1 = \text{Tr}\sqrt{A^\dagger A}$.

Equation (4) means that the average distance between a randomly chosen pure state and the maximally mixed state in the constrained Hilbert space decays polynomially with the Hilbert space size as $d_R^{-0.5}$, that is, exponentially with the system size in general quantum systems. The indistinguishability of the pure quantum state from the microcanonical ensemble leads us to expect that, once a state is prepared to be within the target energy shell, the prepared state captures the macroscopic properties with an accuracy that improves exponentially with the system size [70].

We emphasize that the key of local preparation protocol is to encode the prepared state into the target energy shell, which requires more than merely learning the expectation value of the energy. In other words, even if the expectation value is correctly learned, the prepared state may correspond to a superposition of pure states that belong to other energy shells. Such a situation may cause deviation in other physical observables. In this work, we attempt to address this problem by letting the RL agent learn other macroscopic observables in addition to energy. It is in fact highly nontrivial to determine how many observables we need to embed the prepared state into the energy shell. We find that, for the non-integrable transverse-field Ising chain, it suffices to take

only the total magnetization $\sum_l \sigma_l^z/L$ [see Sec. IV B for the numerical demonstration].

As another possibly effective method to assure the prepared state to be encoded in the desired energy shell, we propose to incorporate the variance of observables $\langle O^2\rangle_{\rm target} - \langle O\rangle^2_{\rm target}$ into the reward. For instance, it is obvious that the energy variance is suppressed when the prepared state is in the target energy shell. This is actually not limited to the energy; thanks to the typicality, we may employ any macroscopic observable for this purpose.

2. *Generalized Gibbs ensemble for integrable systems*

As opposed to the non-integrable systems, integrable systems have an extensive number of conserved quantities, which are also called integrals of motion (IOM) in the literature. The equilibrium states in integrable systems are known to be described rather by GGEs [54]:

$$\rho_{\rm GGE} = \frac{\exp\left[-\sum_m \lambda_m \hat{\mathcal{I}}_m\right]}{{\rm Tr}\left[\exp\left[-\sum_m \lambda_m \hat{\mathcal{I}}_m\right]\right]}, \tag{5}$$

where $\{\hat{\mathcal{I}}_m\}$ is the full set of the IOMs, and $\{\lambda_m\}$ is the corresponding set of the Lagrange multipliers which dictates the distribution over expectation values of the IOMs.

To discuss the typicality in the set of pure states with close expectation values of IOM, the authors of Ref. [71] introduced a notion of a statistical ensemble named the generalized microcanonical ensemble (GME). In parallel to the ordinary microcanonical ensemble, the GME is constructed by assigning equal weight to all eigenstates whose IOMs are close to some certain values that identify the ensemble. It has been pointed out Ref. [71] that the standard deviations of local observables within such "a window of IOM" decay polynomially as

$$\sigma_{\rm loc} \propto L^{-0.5}, \tag{6}$$

where L is the system size. Therefore, by following a parallel discussion as in the case for Gibbs ensembles in non-integrable systems, we may also expect that the local preparation protocol works for GGEs in integrable systems as well, with its accuracy improving polynomially with the system size.

Let us remark on another supporting argument based on the truncation of GGEs itself. While Eq. (5) takes all possible IOMs into account, we expect that the macroscopic behavior in terms of local observables can be extracted by considering local conserved quantities. As such, here, we aim to capture the truncated alternative of the statistical ensemble by focusing on the *local* integrals of motion (LIOM). Concretely, we denote the LIOMs that acts at most $n+1$ neighboring sites as $\hat{I}_n^\sigma$ with σ denoting some additional label, and introduce a locality-constrained variant of GGE which is known as the truncated GGE (tGGE) [72]:

$$\rho_{{\rm tGGE},n_{\rm local}} = \frac{\exp\left[-\sum_{n=0}^{n_{\rm local}}\sum_\sigma \lambda_n^\sigma \hat{I}_n^\sigma\right]}{{\rm Tr}\left[\exp\left[-\sum_{n=0}^{n_{\rm local}}\sum_\sigma \lambda_n^\sigma \hat{I}_n^\sigma\right]\right]}, \tag{7}$$

which only includes the LIOMs with $n \le n_{\rm local}$. It is natural to expect that a tGGE gives a good approximation of $\rho_{\rm GGE}$ in terms of local quantities, and furthermore ${\rm Tr}_{\bar{A}}\left[\rho_{{\rm tGGE},n_{\rm local}}\right] \approx {\rm Tr}_{\bar{A}}\left[\rho_{\rm GGE}\right]$. For instance, Ref. [72] has investigated the transverse field Ising chain in the integrable regime, and found that tGGEs approximates the corresponding GGEs when $n_{\rm local}$ is larger than the size of subsystem A.

We remark that the local preparation protocol for the integrable systems aims to construct a tGGE rather than the original GGE. In this sense, we expect that the validity is not assured for observables with higher $n_{\rm local}$, for which the discrepancy between the GGE and tGGE is non-negligible. We discuss this point more in detail in Sec. V B.

IV. APPLICATION TO GIBBS ENSEMBLES

A. Model and setup

As a demonstration for local preparation of thermal pure states described by Gibbs ensembles, we consider the transverse field Ising model on a chain with the periodic boundary condition:

$$\hat{H}_{\rm Ising} = \sum_{l=1}^{L}\left[J\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z + h\hat{\sigma}_l^z + g\hat{\sigma}_l^x\right], \tag{8}$$

where $\hat{\sigma}_l^x$, $\hat{\sigma}_l^y$, $\hat{\sigma}_l^z$ are the Pauli operators at site l, L is the system size, J is the amplitude of the Ising interaction, and h (g) is the strength of the longitudinal (transverse) magnetic field. In the following, the parameters are fixed as $J=1$, $h=0.8090$, $g=0.9045$ so that the system is non-integrable [73]. As a target state, we aim to prepare a thermal pure quantum state corresponding to the Gibbs ensemble with inverse temperature $\beta = 0.2$, where the initial state before any quantum control is taken to be a product state $|\downarrow\downarrow\cdots\downarrow\rangle$. The total preparation time is fixed as $T=24$ with the time step set as $\delta t = 0.1$, which also determines the total time step to be 240.

The set of action candidates $\{a_t\}$ available for the RL agent is given as $\{e^{-iG_k\delta t}\}_{k=1}^6$, where the time evolution generator G_k is chosen from the following six operators:

$$\begin{aligned}&\hat{H}_{\rm Ising},\ \sum_{l=1}^{L} J\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z + h\hat{\sigma}_l^z,\ g\sum_{l=1}^{L}\hat{\sigma}_l^x,\ \sum_{l=1}^{L}\hat{\sigma}_l^y,\\ &\sum_{l=1}^{L}\hat{\sigma}_l^x\hat{\sigma}_{l+1}^y + \hat{\sigma}_l^y\hat{\sigma}_{l+1}^x,\ \sum_{l=1}^{L}\hat{\sigma}_l^y\hat{\sigma}_{l+1}^z + \hat{\sigma}_l^z\hat{\sigma}_{l+1}^y.\end{aligned} \tag{9}$$

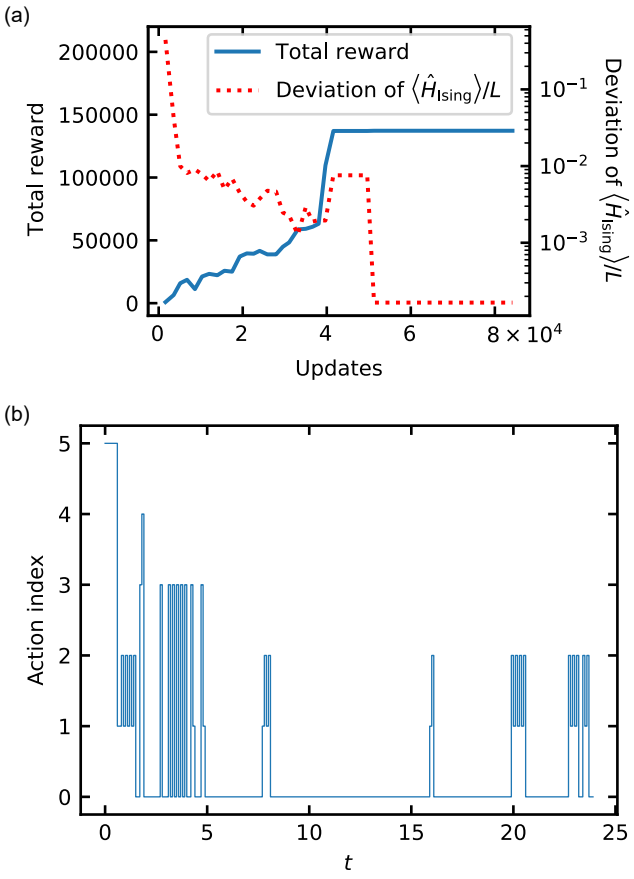

FIG. 2. (a) The learning curves for local preparation of the Gibbs ensemble of transverse field Ising model in the non-integrable regime. The blue and red curves denote the total reward and the deviation of the energy density computed for the prepared states, respectively. The horizontal axis shows the number of parameter updates. (b) The optimal action sequence $\{a_t^*\}$ for local preparation of Gibbs ensemble discovered by the RL agent whose training profile is provided in Fig. 2 (a). The six-fold action indices correspond to 0: $\hat{H}_{\rm Ising}$, 1: $\sum_l J\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z + h\hat{\sigma}_l^z$, 2: $g\sum_l \hat{\sigma}_l^x$, 3: $\sum_l \hat{\sigma}_l^y$, 4: $\sum_l \hat{\sigma}_l^x\hat{\sigma}_{l+1}^y + \hat{\sigma}_l^y\hat{\sigma}_{l+1}^x$, 5: $\sum_l \hat{\sigma}_l^y\hat{\sigma}_{l+1}^z + \hat{\sigma}_l^z\hat{\sigma}_{l+1}^y$, respectively. After quickly manipulating the macroscopic observables to encode the property of the thermal ensemble, the agent mainly chooses to evolve the system under the target Hamiltonian $\hat{H}_{\rm Ising}$. For all plots, the system size is $L = 16$ with the inverse temperature set as $\beta = 0.2$.

These terms are also used in Ref. [44], which discusses how to accelerate the preparation of the ground state using counter-diabatic driving. Regarding the local reward r_t, we include the energy density $L^{-1}\hat{H}_{\rm Ising}$ and the magnetization density $L^{-1}\sum_l \hat{\sigma}_l^z$. Note that these are the sum of local operators acting on neighboring two sites at most.

B. Numerical results

Now we proceed to present numerical results that successfully prepare thermal pure quantum states, which encode both local observables used during the training and also more non-local ones that are excluded from the reward function. Figure 2 (a) shows an example of the learning curve of the RL agent for $L = 16$. We can see that, as the number of training episodes increases, the RL agent learns the better protocols that achieve higher total reward and smaller energy deviation.

Here, we briefly discuss the control protocol found by the RL agent. Figure 2 (b) shows the control protocol whose learning curve is shown in Fig. 2 (a) to realize a thermal pure state described by the Gibbs ensemble. It can be seen that the optimal protocol consists of two stages: first is the sequence of nontrivial actions that change macroscopic observables, and the second is the relaxation of the system under free evolution using $\hat{H}_{\rm Ising}$, which leads the system to the typical state of $\hat{H}_{\rm Ising}$. While there is a small number of actions other than the free relaxation at the late stage of the control protocol, we argue that they do not contribute significantly to the result even if we substitute them with $\hat{H}_{\rm Ising}$ because they consist almost entirely of alternating use of action 1 and action 2. This can be considered effectively equivalent to $\hat{H}_{\rm Ising}$ in the sense of the Trotter decomposition.

Figures 3 (a)–(d) display the dynamics of local observables obtained by the preparation protocol learned by the RL agent. Figures 3 (a) and (b) show the observables $\hat{H}_{\rm Ising}$ and $\sum_l \hat{\sigma}_l^z$, respectively, which are used for the reward function. Both of them converge to the corresponding values of the Gibbs ensemble represented by the red horizontal lines. It is more intriguing to see the convergent behaviors in Figs. 3 (c) and (d), which show the dynamics of local observables that are not included in the reward function, namely the two-point correlator $\sum_l \hat{\sigma}_l^z\hat{\sigma}_{l+1}^z$ and three-point correlator $\sum_l \hat{\sigma}_l^z\hat{\sigma}_{l+1}^z\hat{\sigma}_{l+2}^z$, respectively.

It is natural to wonder how accurately the general observables, or the reduced density operator of local systems, are captured by the local preparation protocol. For this purpose, we evaluate the distance between reduced density operators of the prepared pure state and the target Gibbs ensemble as

$$D(\rho, \rho') := \frac{\|\rho - \rho'\|_{\rm F}}{\sqrt{\|\rho\|_{\rm F}^2 - \|\rho'\|_{\rm F}^2}}, \tag{10}$$

where $\|A\|_{\rm F} := \sqrt{{\rm Tr}\,[A^\dagger A]}$ is the Frobenius norm [74].

We can see from Fig. 4 (a) that the time average of the distance function (10), which we denote as $\bar{D}$, is suppressed for the smaller subsystem size LA. What is more interesting is the scaling with respect to the total system size; we observe nontrivial suppression of error along with the increase of the system size L.

We further zoom into the scaling of the error suppression, by performing finite-size scaling on the averaged

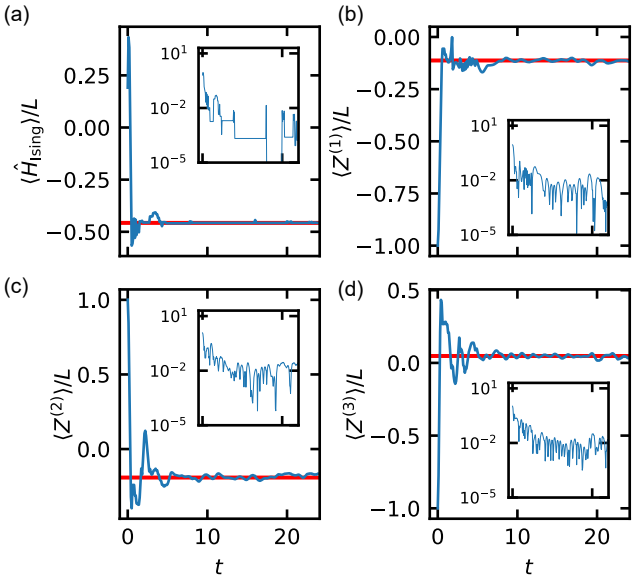

FIG. 3. (a)–(d) The dynamics of local observables generated by the state preparation protocol learned by the deep RL agent, with the insets showing the absolute error from the corresponding expectation values of the target Gibbs state (red line). Each panel displays (a) energy density $\hat{H}_{\rm Ising}/L$, (b) $Z^{(1)}/L = L^{-1}\sum_l \hat{\sigma}^z_l$, (c) $Z^{(2)}/L = L^{-1}\sum_l \hat{\sigma}^z_l\hat{\sigma}^z_{l+1}$, and (d) $Z^{(3)}/L = L^{-1}\sum_l \hat{\sigma}^z_l\hat{\sigma}^z_{l+1}\hat{\sigma}^z_{l+2}$. Note that local observables in (a) and (b) are both included in the reward for the deep RL agent. We observe convergence not only for physical quantities that are included in the RL reward as in (a) and (b), but also local observables that are absent in the reward as in (c) and (d). For all plots, the system size is $L = 16$ with the inverse temperature set as $\beta = 0.2$.

distance function as shown in Fig. 4 (b). Here, we observe that the scaling of the suppression is given as

$$\bar{D} = \mathcal{O}(d^{-b}), \tag{11}$$

where d is the dimension of the corresponding energy shell. To be concrete, d is obtained by counting the number of eigenstates included within the energy window $[\langle \hat{H}_{\rm Ising}\rangle_\beta - \varepsilon_{\rm e} L,\ \langle \hat{H}_{\rm Ising}\rangle_\beta]$, where $\langle \hat{H}_{\rm Ising}\rangle_\beta$ is the energy expectation value of the target Gibbs ensemble at inverse temperature β, and the shell width is fixed as $\varepsilon_{\rm e} = 0.5$ [see Appendix C for further discussion on the choice of $\varepsilon_{\rm e}$].

The scaling of the distance (11) implies that $\bar{D}$ decays exponentially with the system size L. We argue that this is compatible with the scaling of canonical typicality (4). This feature supports an expectation that our local preparation protocol for the Gibbs ensembles becomes exponentially more precise as the system size increases.

As a technical remark, we mention that the dimension shown in the figure corresponds to the size of the symmetry-resolved Hilbert space, namely the parity and the momentum.

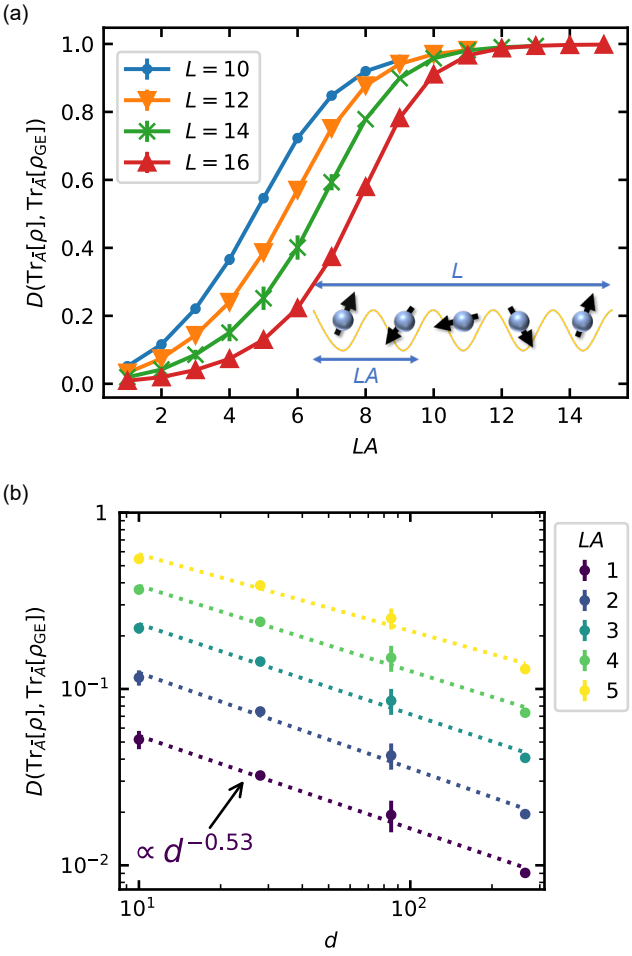

FIG. 4. Distance of reduced density operator between the target Gibbs state and the prepared state, which are averaged over time and random training instances. (a) Error of distance function $\bar{D}$ at various subsystem sizes LA. The blue circles, orange downward triangles, green crosses, and red upward triangles denote the data for $L = 10, 12, 14, 16$, respectively. (b) Decay of error in the reduced density operator along with the corresponding energy shell dimension d, i.e., the number of the eigenstates in the energy shell. The dotted lines are guides to the eye showing the fits with $D = ad^{-b}$, where powers b are summarized in Table III of Appendix. B. The error bars correspond to the standard deviations of different protocols learned independently with random seeds. In both plots, the system is controlled by the deep RL agent at $0 \le t \le 24$, and then undergoes free evolution $\exp[-i\hat{H}_{\rm Ising}t]$ in $24 < t \le 48$. Note that the distance $\bar{D}$ concerns average over the free evolution period and also the random training instance.

V. APPLICATION TO GENERALIZED GIBBS ENSEMBLE

A. Model and setup

We next describe an even more intriguing system that is integrable so that the thermodynamic ensembles is de-

scribed by GGEs. Concretely, we consider the XX model on a periodic chain with a longitudinal field:

$$\hat{H}_{\mathrm{XX}} = -\sum_{l=1}^{L}\left[\frac{J}{2}\left(\hat{\sigma}_l^x\hat{\sigma}_{l+1}^x + \hat{\sigma}_l^y\hat{\sigma}_{l+1}^y\right) + h\hat{\sigma}_l^z\right], \tag{12}$$

where J is the amplitude of interaction, h is the strength of the magnetic field along the z-axis, and L is the system size. In the following, the parameters are fixed as $J = 1$ and $h = 2$.

The integrability of the XX model can be verified straightforwardly by mapping into a non-interacting fermionic system via the Jordan-Wigner transformation:

$$\begin{aligned}\hat{H}_{\mathrm{tb}} &= -\sum_{l=1}^{L}[J(\hat{a}_l^\dagger\hat{a}_{l+1} + \hat{a}_{l+1}^\dagger\hat{a}_l) + 2h(\hat{a}_l^\dagger\hat{a}_l - \frac{1}{2})], &(13)\\ &= -\sum_k\left[2\left(J\cos k + h\right)\hat{n}_k - h\right], &(14)\end{aligned}$$

where $\hat{a}_l^\dagger$ and $\hat{a}_l$ are fermionic creation and annihilation operators that are related to the Pauli operators as $\hat{a}_l^{(\dagger)} = \prod_{l'<l}\hat{\sigma}_{l'}^z\hat{\sigma}_l^{+(-)}$ where $\hat{\sigma}_l^\pm = (\hat{\sigma}_l^x \pm i\hat{\sigma}_l^y)/2$. In the second row (14), we moved into the Fourier space by introducing the mode occupation operator $\hat{n}_k = \hat{a}_k^\dagger\hat{a}_k$ for $\hat{a}_k = L^{-1/2}\sum_l \hat{a}_l e^{-ikl}$. We can construct LIOMs as many as the system size L by taking the linear combination as

$$\begin{aligned}\hat{I}_n^+ &= -2J\sum_k \cos(nk)\hat{a}_k^\dagger\hat{a}_k = -J\sum_l\left(\hat{a}_l^\dagger\hat{a}_{l+n} + \hat{a}_{l+n}^\dagger\hat{a}_l\right),\\ \hat{I}_n^- &= -2J\sum_k \sin(nk)\hat{a}_k^\dagger\hat{a}_k = iJ\sum_l\left(\hat{a}_l^\dagger\hat{a}_{l+n} - \hat{a}_{l+n}^\dagger\hat{a}_l\right).\end{aligned}$$

For the convenience in the later discussion, we mention the rightmost sides to remark that $\hat{I}_n^\pm$ can be explicitly expressed as sums over hopping terms between the n-th nearest neighboring sites in the fermionic picture.

As the target thermodynamic ensemble, we aim for the GGE constructed from the LIOMs as

$$\rho_{\mathrm{GGE}} = \frac{\exp\left[-\sum_n\sum_{\sigma=\pm}\lambda_n^\sigma\hat{I}_n^\sigma\right]}{\mathrm{Tr}\left[\exp\left[-\sum_n\sum_{\sigma=\pm}\lambda_n^\sigma\hat{I}_n^\sigma\right]\right]}, \tag{15}$$

where $\{\lambda_n^\sigma\}$ are the Lagrange multipliers. More precisely, we set the initial state of the control system to be the ground state of $\hat{H}_{\mathrm{tb}}$ in the space of the total particle number N_{f}, and aim to prepare a prethermal pure state corresponding to a given set of Lagrange multipliers $\{\lambda_n^\sigma\}$ [see Appendix D for details regarding the parameter choice of $\{\lambda_n^\sigma\}$]. The total preparation time is fixed as $T = 40$ and the time step as $\delta t = 0.2$, thus the total number of time steps is 200.

In parallel with the case for Gibbs ensembles, we allow the deep RL agent to choose a_t as an appropriate unitary from a set $\{e^{-iG_k\delta t}\}_{k=1}^{15}$, where the time evolution generator G_k is chosen from the following fifteen operators:

$$\begin{aligned}&\hat{H}_{\mathrm{XX}},\ \sum_{l=1}^{L}\hat{a}_l^\dagger\hat{a}_l\cos\frac{m\pi}{L}l,\ \sum_{l=1}^{L}\hat{a}_l^\dagger\hat{a}_l\sin\frac{m\pi}{L}l,\\ &\sum_{l=1}^{L}\left(\hat{a}_l^\dagger\hat{a}_{l+j} + \hat{a}_{l+j}^\dagger\hat{a}_l\right)\cos\frac{n\pi}{L}l, \qquad (16)\\ &\sum_{l=1}^{L}\left(\hat{a}_l^\dagger\hat{a}_{l+j} + \hat{a}_{l+j}^\dagger\hat{a}_l\right)\sin\frac{n\pi}{L}l,\end{aligned}$$

where $m \in \{2, 4, L\}$, and $(j, n) \in \{(1,2),(2,2),(1,L),(2,L)\}$. Note that these operators are chosen so that they are not diagonal in the position or momentum basis.

Regarding the local reward r_t, we employ normalized LIOMs $\{\hat{I}_n^+/L\}_{1\le n\le n_{\mathrm{local}}}$, where we fix $n_{\mathrm{local}} = 4$ in the following. All $\hat{I}_n^-$ terms are excluded since they are constantly zero not only for the initial and target states, but also for any intermediate states ρ_t evolved with the above unitaries.

B. Numerical results

We now present the numerical results obtained by running the deep RL algorithm to prepare prethermal pure quantum states that capture the characteristics of the GGE. As we show the learning curve in Fig. 5(a), the deep RL agent successfully learns to improve the local preparation protocol. This can be more quantitatively understood from Fig. 5(b), which evaluates the absolute difference in the expectation values of LIOMs between the target GGE and the prepared state.

Let us further investigate the dynamics of the LIOMs generated by the preparation protocol discovered by the RL agent. As shown in Fig. 6, the behavior of LIOMs seem to be qualitatively different depending on n in the sense that, the only LIOMs with separation $n \le n_{\mathrm{local}} = 4$, which we added to the reward, seem to converge to the expectation values of the GGE.

Similar behavior can also be observed for non-conserved quantities such as the correlation function

$$\Gamma_n \coloneqq \frac{1}{L}\sum_l\left(\hat{\sigma}_l^+\hat{\sigma}_{l+n}^- + \hat{\sigma}_{l+n}^+\hat{\sigma}_l^-\right), \tag{17}$$

which can be seen as a representative of local operators acting on neighboring $n+1$ sites. As we can see in Fig. 7, we see a notable convergence into the GGE values for $n \le n_{\mathrm{local}}$, while the errors seem to remain for $n > n_{\mathrm{local}}$.

Furthermore, we focus on the scaling of the error suppression. As we can see from the finite-size scaling of the averaged distance function $\bar{D}(\rho_t, \rho_{\mathrm{GGE}})$ in Fig. 8, we observe that the distance of the reduced density operators is suppressed as $\bar{D} = \mathcal{O}(L^{-b})$. These behaviors are

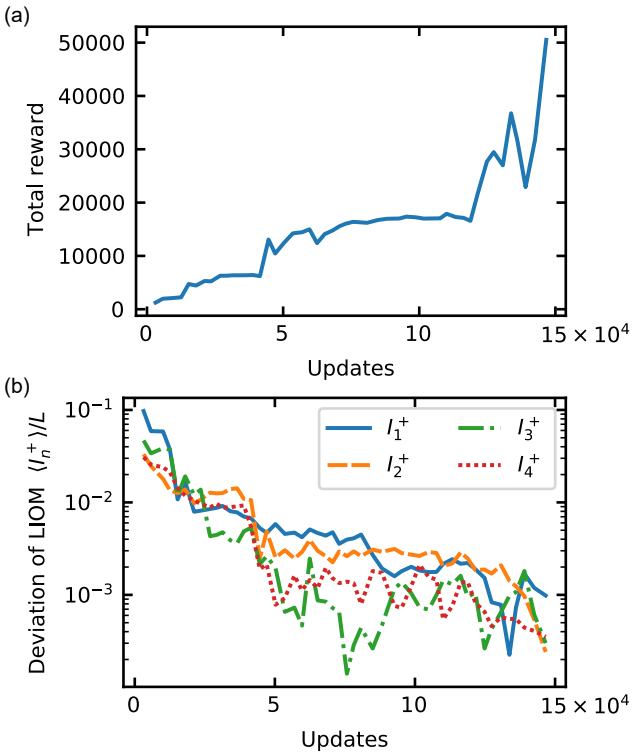

FIG. 5. The learning curves for local preparation of the GGE. These figures show the total rewards and the deviations of the LIOM densities at the end of the episodes in evaluating the training progress. The horizontal axis shows the number of parameter updates. (a) The total rewards. (b) The deviations of the LIOM densities. The corresponding model is the transverse field Ising model where the parameter is set so that the model is integrable. We added the LIOMs $\hat{I}_n^+$ with separation $n = 1, 2, 3, 4$ to the reward of the RL. The system size $L = 120$. The lines show the median of the corresponding values in evaluation with random seeds.

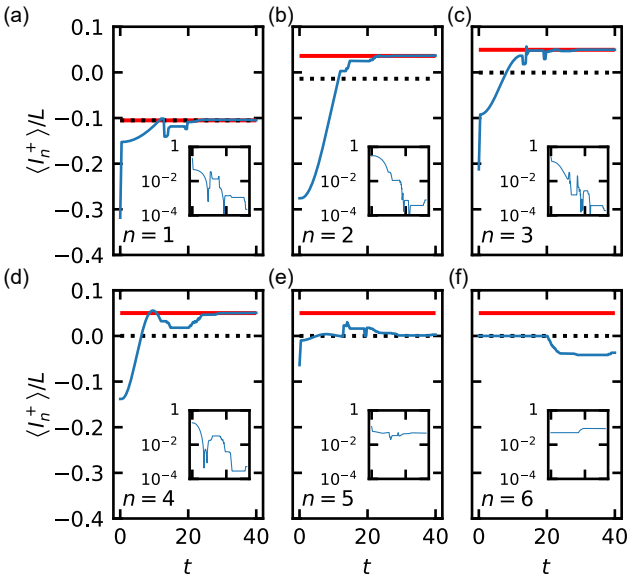

FIG. 6. The expectation values of the LIOMs in the protocol found by RL for the GGE (blue). Each figure shows the LIOM with different separations. The red horizontal lines show the corresponding expectation values of the target GGE. The dotted black lines represent the relevant Gibbs ensemble, which shares the same expectation values of total particle number and energy. The insets show the absolute deviations in the expectation values $\langle \hat{I}_n^+ \rangle / L$ between the states in the protocol and the target GGE. We added the LIOMs $\hat{I}_n^+$ with separation $n = 1, 2, 3, 4$ to the reward of the RL. Only LIOMs with separation $n \leq 4$ converge to the expectation values of the GGE.

compatible with the scaling of the fluctuation of local observables in the GME (6).

Here we conjecture that, while the errors for $LA \lesssim n_{\rm local}+1$ shall be suppressed polynomially even for larger system size, the errors for $LA \gtrsim n_{\rm local}+1$ may saturate at finite values. This is because the current local preparation scheme learns the LIOMs $\hat{I}_n^\sigma$ with $n \leq n_{\rm local}$ to encode the prepared state into the "LIOM shell" only for such conserved quantities. This means that, the prepared state fully encodes the macroscopic properties of the tGGE, but not those of the GGE. We numerically find that, the distance of the tGGE and GGE remains to be finite even if the total system is in the thermodynamic limit [See Appendix E], which is in agreement with Ref. [72] which investigated integrable region of the transverse-field Ising chain. This supports our conjecture that the distance between the GGE and the local-prepared state shall not be suppressed in the thermodynamic limit. Meanwhile, it is possible that the error from the tGGE itself is suppressed polynomially.

VI. CONCLUSION

In this work, we have proposed a deep-RL-based quantum state preparation framework for thermodynamic ensembles that solely relies on a few local observables but not on global features such as the fidelity. The core idea is to leverage the typicality of pure states in quantum many-body systems; the macroscopic properties can be encoded simply via learning a few local observables and undergoing free evolution. We have provided numerical demonstrations that have successfully trained the deep RL agent to learn the macroscopic properties of the Gibbs ensembles (Fig. 2) and the GGEs (Fig. 5). We find that the accuracy of the prepared state has improved exponentially with the system size for the former (Fig. 4) and polynomially for the latter (Fig. 8), which is consistent with the argument of the typicality within a given shell of local conserved quantities.

We envision four future directions of our work. First, the application to interacting integrable models is an important issue for local preparation. This issue is related to the previous work on which conserved quantities should be considered to predict local properties of the steady state in interacting integrable systems that can be solved with the Bethe ansatz [e.g., Ref. [75]]. Based

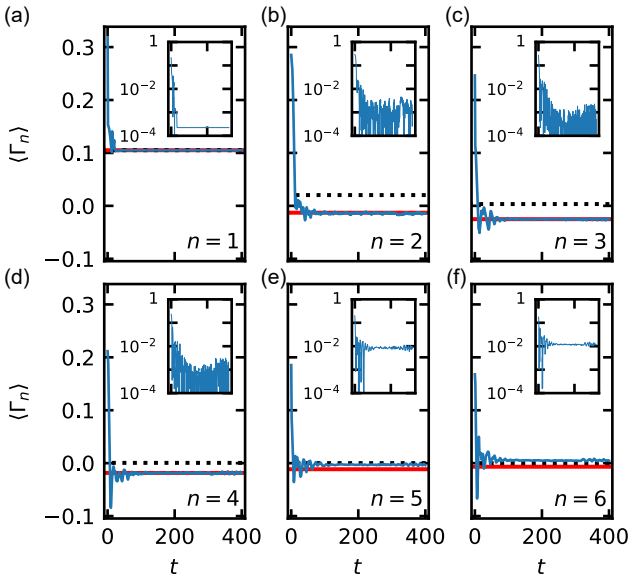

FIG. 7. The relaxation processes of the expectation values of correlation functions (17), which are non-conserved quantities of $\hat{H}_{\rm XX}$ (12) (blue). Each figure shows the correlation function with different separation n, which means the correlation function corresponds to the two spins separated by $n-1$ sites. The red horizontal lines show the corresponding expectation values of the target GGE, whereas the dotted black lines represent the relevant Gibbs ensemble which shares the same expectation values of total particle number and energy. Note that the system is controlled with the learned protocol in $0 \leq t \leq 40$, and subsequently undergoes free evolution with $\hat{H}_{\rm XX}$. The only correlation functions with separation $n \leq 4$ fluctuate around the expectation values of the GGE. Note that the correlation function with $n = 1$, where the Jordan-Wigner string does not appear, is proportional to the LIOM with $n = 1$, i.e., it is the conserved quantity of $\hat{H}_{\rm XX}$.

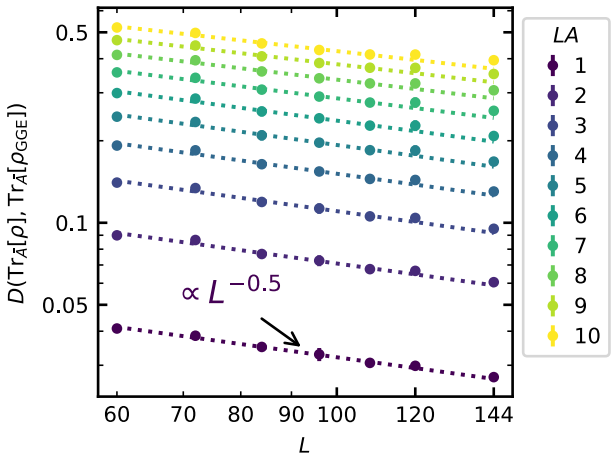

FIG. 8. Finite-size scaling of the distance function of the reduced density operators between the prepared states and the target GGE. The dotted lines are guides to eye that show power-law fitting $\bar{D} = \mathcal{O}(L^{-b})$, where powers b are summarized in Table IV of Appendix. B. The error bars correspond to the standard deviations of different protocols learned independently with random seeds. The system is controlled by the deep RL agent at $0 \leq t \leq 40$, and then undergoes free evolution by $\hat{H}_{\rm XX}$. Note that the distance $\bar{D}$ concerns average over temporal, i.e., the free evolution period, and the random training instance. The number of random seeds is 5 at most.

on the prior work, we can expect to be able to perform local preparation if we also include *quasi-local* conserved quantities to the reward. Note that, in such systems the finite size effect on the fluctuation of local observables is severe in system size tractable by exact diagonalization. It is an open problem how to simulate interacting integrable systems efficiently in a scalable way so that local preparation strategy can be pursued.

The second important question is the generalization of the local preparation protocol to include, e.g., dissipative terms, measurement and feedback, or postselection. We naturally expect that the powerful explorability of the deep RL framework is not limited to coherent control but could be applied to broader operation sets.

Third, we may consider the application of the local preparation protocol for the task of Hamiltonian learning [76–78] by attempting to encode the macroscopic properties using unitaries that do not explicitly contain the information of Hamiltonian itself.

Finally, it is intriguing to seek how the local preparation protocol is affected by various noises, such as the statistical noise accompanied by sampling over observables. Efficient estimation methods such as the randomized measurement schemes [79] shall be essential to boost the training accuracy of the RL agent.

VII. ACKNOWLEDGMENTS

The authors wish to thank fruitful discussion with Ryusuke Hamazaki and Jiahao Yao. S. B. is supported by Materials education program for the future leaders in research, industry, and Technology (MERIT) of The University of Tokyo. N.Y. wishes to thank JST PRESTO No. JPMJPR2119 and JST Grant Number JPMJPF2221. Y.A. acknowledges support from the Japan Society for the Promotion of Science through Grant Nos. JP19K23424 and JP21K13859. T.S. is supported by JSPS KAKENHI Grant Number JP19H05796, JST CREST Grant Number JPMJCR20C1, Japan, and JST ERATO-FS Grant Number JPMJER2204, Japan. N.Y. and T.S. are also supported by Institute of AI and Beyond of The University of Tokyo. We implemented the exact time evolution in Sec. IV with QuSpin [80, 81] and the RL algorithm with PyTorch [82]. The RL in Sec. V is performed on AI Bridging Cloud Infrastructure (ABCI) of National Institute of Advanced Industrial Science and Technology (AIST).

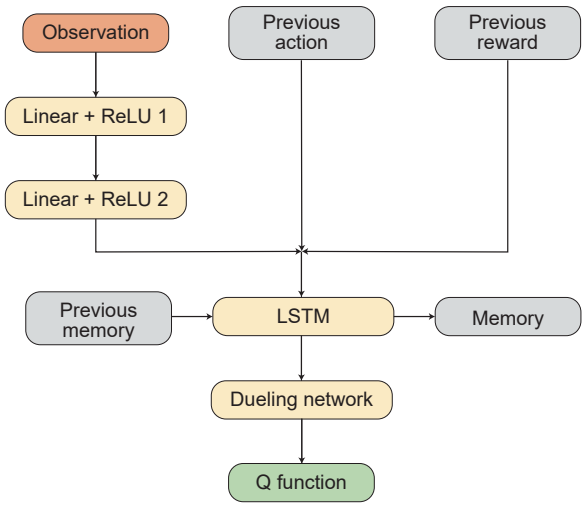

FIG. A1. The abstract deep NN architecture used for the RL in the present work. At each time step t, the deep NN takes as input the observation o_t, the previous action a_{t-1}, and previous reward r_{t-1}. After intermediate computations by fully-connected layers, LSTM, and the dueling network, the deep NN outputs the estimate of the Q function.

	Sec. IV	Sec. V
Linear + ReLU 1	$240 \to 512$	$200 \to 1024$
Linear + ReLU 2	$512 \to 512$	$1024 \to 1024$
LSTM	$(512, 6, 1) \to 512$	$(1024, 15, 1) \to 1024$
Dueling network	$512 \to 6$	$1024 \to 15$

TABLE I. The Input and Output sizes of each layer are shown in the former and latter of the arrow, respectively. The input of the NN consists of the action history, the one-hot representation of the previous action, and the previous reward. The input of Linear + ReLU 1 layer is only the action history. The one-hot representation of the previous action and previous reward is concatenated with the output of Linear + ReLU 2 before the LSTM layer.

Reward discount γ		0.997
Minibatch size		324(Sec. IV)
		380(Sec. V)
Sequence length		40
Optimizer		Adam [83]
Optimizer setting	learning rate	10^{-4}
	ε	10^{-3}
	β	$(0.9, 0.999)$
Replay ratio		1
Gradient norms clip		80

TABLE II. The hyperparameters used in the NNs. The agent performs updates on batches of (minibatch size $\times$ sequence length) observations. Replay ratio means the effective number of times each experienced observation is being replayed for the training. See Ref. [67] and its previous non-LSTM version, Ref. [84], for the details of the hyperparameters. The other parameters follow the ones in Ref. [66].

Appendix A: Layout of the neural network and hyperparameters

Here, we describe the architecture of the deep NN which is used to estimate the action-value function (Q function). Figure A1 shows the overall picture; the input for the LSTM at time step t is the observation o_t, the previous action a_{t-1}, and the previous reward r_{t-1}, whereas the output is Q function whose optimal expression is given in Eq. (1) in the main text. Refer to Table I for the size of the input and output of each layer and Table II for the hyperparameter used in the NN. In the following, we further describe the details of the structure.

The observation o_t is chosen to be the action history $o_t = (a_0, a_1, \ldots, a_{t-1}, -1, \ldots, -1)$, which is fed to the fully-connected layers. Fully-connected layers first perform linear transformation, and then apply a non-linear activation function which is chosen to be the rectified linear unit (ReLU) in the present work. The intermediate output from the second fully-connected layer is concatenated with the previous choice of action a_{t-1} and the reward in the previous time step r_{t-1}, and then fed to the Long Short-Term Memory (LSTM) layer.

The LSTM layer is introduced so that the network can refer to the history of computational results at $t' < t$ to estimate the Q function at time step t. Namely, the input of the LSTM layer is not only the ones mentioned above, but also its "memory" including a *hidden state* (short-term memory) and a *cell state* (long-term memory) [69]. Refer to literature such as Ref. [85] for detailed information. This output memory is fed to the LSTM layer in the next time step t, which enables the deep NN to successfully deal with time-series inputs.

In the subsequent dueling network [86], the input is separated into two branches. One branch evaluates the value of the observation $V(o)$, and the other branch evaluates the advantage of actions regarding the observation $A(o, a) = Q(o, a) - V(o)$. The output Q function of the dueling network is obtained by summing the output of the two branches: $Q(o, a) = V(o) + A(o, a)$. This separation may contribute to better training stability, faster convergence, and better performance.

As a computational resource, we have used a single CPU and four GPUs (Intel Xeon E5-2698 v4, 4×NVIDIA Tesla V100) in Sec. IV and two CPUs and four GPUs (2×Intel Xeon Gold 6148, 4×NVIDIA Tesla V100) in Sec. V, respectively.

Appendix B: Fitting parameters for the finite-size scaling of the preparation accuracy

In Tables III and IV, we summarize the powers b obtained by the fit for the finite-size scaling of the local preparation accuracy in Sec. IV B and Sec. V B, respec-

Subsystem size LA	Power b
1	0.53(4)
2	0.54(4)
3	0.51(4)
4	0.48(4)
5	0.43(4)

TABLE III. The powers obtained by the fit for the result of the finite-size scaling of the distance function $\bar{D}$, given by Eq. (10), between the prepared states and the target Gibbs ensembles for the system size $L = 10, 12, 14, 16$ in Sec. IV B.

subsystem size LA	power b
1	0.50(3)
2	0.50(5)
3	0.46(4)
4	0.43(4)
5	0.41(4)
6	0.41(3)
7	0.41(3)
8	0.41(3)
9	0.41(3)
10	0.41(3)

TABLE IV. The powers obtained by the fit for the result of the finite-size scaling with the system size $L = 60, 72, 84, 96, 108, 120$ for the distance function $\bar{D}$, given by Eq. (10), between the prepared states and the target GGE in Sec. V B.

tively.

Appendix C: Relationship between energy shell width ε_e and scaling of preparation accuracy

Here we discuss the relationship between the energy shell width ε_e and the scaling behavior of the accuracy of prepared state. Recall that the distance function in the Eq. (10) of the main text is given as

$$D(\rho, \rho') = \frac{\|\rho - \rho'\|_F}{\sqrt{\|\rho\|_F^2 - \|\rho'\|_F^2}}, \tag{C1}$$

where $\|A\|_F = \sqrt{\mathrm{Tr}[A^\dagger A]}$ denoting the Frobenius norm. In Fig. C2, we show how the scaling exponent b defined from $\bar{D}(\rho_{\mathrm{Gibbs}}, \rho_t) = O(d^{-b})$ varies according to ε_e. At $\varepsilon_e = 0.1$ for $L = 10$, the corresponding energy shell only includes a single eigenstate. Meanwhile, $\varepsilon_e = 0.75$ corresponds to the extreme case where almost all eigenstates below $\langle \hat{H}_{\mathrm{Ising}} \rangle_\beta$ are included in the energy shell $[\langle \hat{H}_{\mathrm{Ising}} \rangle_\beta - \varepsilon_e L, \langle \hat{H}_{\mathrm{Ising}} \rangle_\beta]$. In Sec. IV B, we have chosen intermediate ε_e so that the power b is stable against the choice of ε_e.

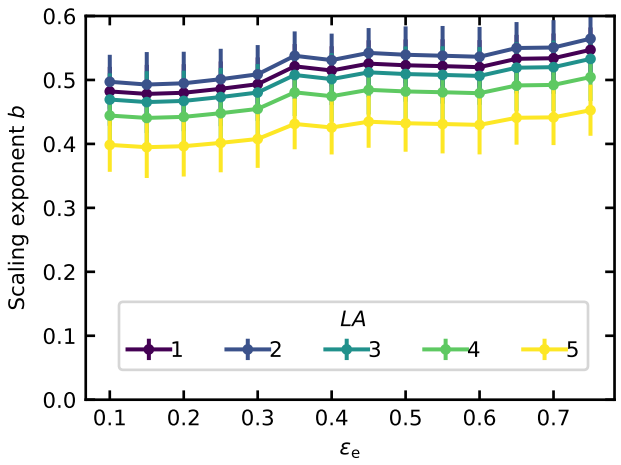

FIG. C2. Relationship between the energy shell width ε_e and the scaling exponent b of the distance $\bar{D}$, concerning the local preparation of the Gibbs ensemble. The parameters of the system are identical to those used in Sec. IV B.

Appendix D: Choosing Lagrange multipliers for the target GGE

The Lagrange multipliers for the GGE in Sec. V is chosen so that the expectation values of local conserved quantities partly reproduce those of the Gibbs ensemble, while some deviate from it. Concretely, we impose the following equality:

$$\mathrm{Tr}\left[\hat{I}_n^+ \rho_{\mathrm{GGE,target}}\right] = \langle \hat{I}_n^+ \rangle_\beta \quad \text{for } n = 0,\ 1, \tag{D1}$$

$$\mathrm{Tr}\left[\hat{I}_n^+ \rho_{\mathrm{GGE,target}}\right] = \langle \hat{I}_n^+ \rangle_\beta + \varepsilon_I L \text{ for } 2 \le n < \frac{L}{2}, \tag{D2}$$

$$\mathrm{Tr}\left[\hat{I}_n^- \rho_{\mathrm{GGE,target}}\right] = \langle \hat{I}_n^- \rangle_\beta , \tag{D3}$$

where $\langle \hat{I}_n^\sigma \rangle_\beta$ is the expectation values of LIOMs of the Gibbs ensemble at the inverse temperature $\beta = 0.4$. Note that ε_I determines the deviation between the target GGE and the Gibbs ensemble. For simplicity, we constantly take $\varepsilon_I = 0.05$ for every L. In addition, we set $\mathrm{Tr}\left[\hat{I}_{L/2}^+ \rho_{\mathrm{GGE,target}}\right] = 0$ because the fermionic tight-binding Hamiltonian obtained from the Jordan-Wigner transformation (Eq. (13) in the main text) is anti-periodic when the fermionic particle number N_f is even.

We remark that the LIOMs with $n = 0, 1$ correspond to the total particle number and energy, respectively. Thus, this target GGE shares only the expectation values of the total particle number and energy with the Gibbs ensemble. Therefore, in order to prepare the subsystem whose size is larger than two, we need to control additional LIOMs other than the particle number and energy.

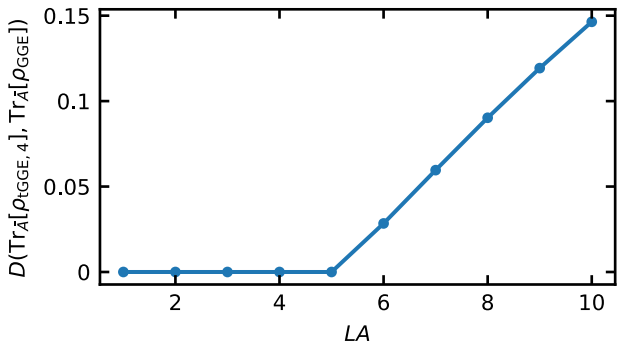

FIG. E3. The distance, given by Eq. (10), between the target GGE in Sec. V B and the corresponding tGGE with $n_{\mathrm{local}} = 4$.

Appendix E: The distance between the GGE and the tGGE

To quantify the difference between the tGGE and the target GGE considered in Sec. V B, we show the distance $D(\rho_{\mathrm{GGE}}, \rho_{\mathrm{tGGE}})$ in Fig. E3 (see Eq. (10) for the definition). We observe that the tGGE and the GGE agree well when $LA \leq n_{\mathrm{local}} + 1$, whereas they deviate when $LA > n_{\mathrm{local}} + 1$. This result is compatible with Ref. [72], which consider the integrable parameter region of transverse-field Ising chain.

Appendix F: System-size dependence of learning progress

In this section, we analyze the system-size dependence of the reinforcement learning progresses. Figures F4 show the learning curves of physical observables for different system sizes. They tell us that the number of updates required to learn the optimal protocols is almost independent of the system size. We suppose that this feature is related to the fact that our method considers only local properties, which are independent of the system size.

Appendix G: Learning for other initial states and unitaries

In this section, we provide the results of the deep RL considering different unitary generators and initial states. The choices of unitary generators and initial state are shown in Table V. The target state is the same as the one in Sec. IV B, and the system size $L = 14$.

Figure G5 (a) shows the learning curves of the RL agent corresponding to choices (i)–(v), respectively. We can see that, as the number of training episodes increases, all RL agents corresponding to the different choices learn the better protocols that achieve smaller energy deviation.

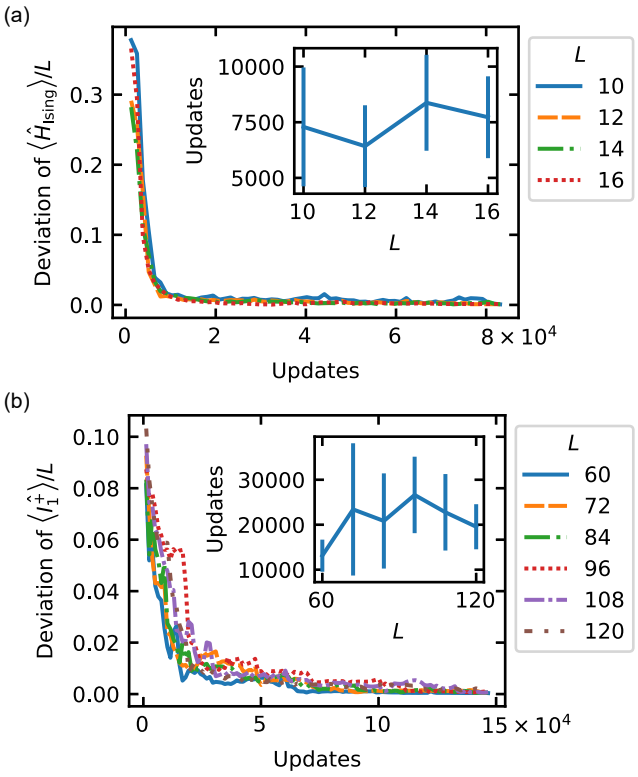

FIG. F4. The system-size dependence of the learning progresses for local preparation of (a) Gibbs ensembles in Sec. IV B and (b) GGE in Sec. V B. These figures show the deviation of (a) the energy density and (b) the LIOM densities, respectively. Each figure represents the average regarding different protocols learned independently to perform the finite-size scaling in Fig. 4 or Fig. 8. The insets show the number of updates required to bring the deviations below a certain threshold of (a) 0.02 and (b) 0.01, respectively.

	Initial states	Generators
(i)	Ground	Same as Sec. IV B
(ii)	Ground	$\hat{H}_{\mathrm{Ising}}$, $\sum_{l=1}^{L} J\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z + h\hat{\sigma}_l^z$, $g\sum_{l=1}^{L}\hat{\sigma}_l^x$
(iii)	Product	$\hat{H}_{\mathrm{Ising}}$, $\sum_{l=1}^{L} J\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z + h\hat{\sigma}_l^z$, $g\sum_{l=1}^{L}\hat{\sigma}_l^x$
(iv)	Product	$\sum_{l=1}^{L} J\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z + h\hat{\sigma}_l^z$, $g\sum_{l=1}^{L}\hat{\sigma}_l^x$
(v)	Product	$\sum_{l=1}^{L}\hat{\sigma}_l^z\hat{\sigma}_{l+1}^z$, $\sum_{l=1}^{L}\hat{\sigma}_l^z$, $\sum_{l=1}^{L}\hat{\sigma}_l^x$

TABLE V. The choices of initial state and unitary generators in Appendix G. *Ground* means the corresponding initial state of the preparation is the ground state. *Product* means the corresponding initial state of the preparation is the product state which is the same as the one used in Sec. IV B.

Figures G5 (b)–(e) display the dynamics of local observables obtained by the preparation protocol learned by the RL agent. All of them converge to the corresponding values of the Gibbs ensemble represented by the red horizontal lines.

In Fig. G6, we show the time average of the distance function (10) of the subsystems between the target Gibbs

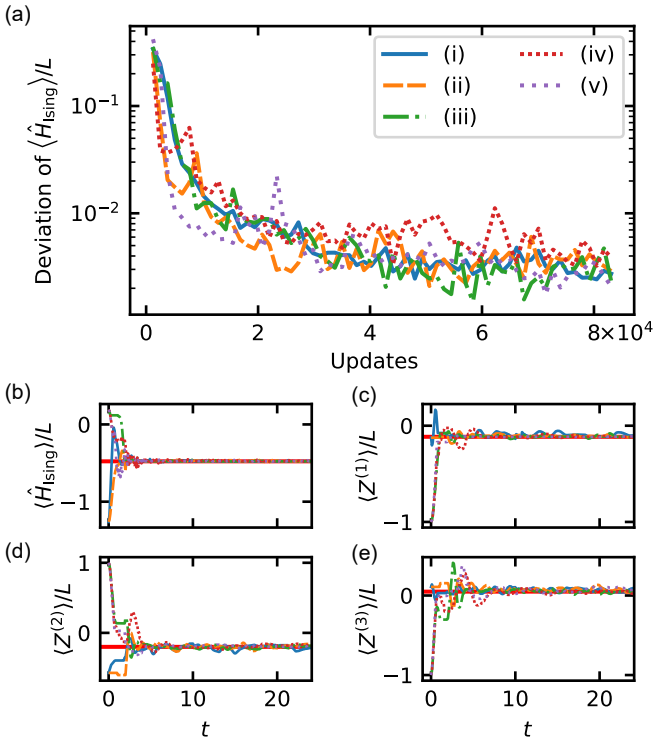

FIG. G5. The result of the deep RL considering different unitary generators and initial states. The choices of unitary generators and initial state are shown in Table V. (a) The learning curves for local preparation of the Gibbs ensemble of transverse field Ising model which is the same as the one used in Sec. IV B. Each curve shows the energy deviation calculated for the prepared states by learning corresponding to each choice (i)–(v). The horizontal axis shows the number of parameter updates. (b)–(e) The dynamics of local observables generated by the state preparation protocol learned by the deep RL. The red horizontal lines show the corresponding expectation values of the target Gibbs ensemble. Each panel displays (b) energy density $\hat{H}_{\rm Ising}/L$, (c) $Z^{(1)}/L = L^{-1}\sum_l \hat{\sigma}^z_l$, (d) $Z^{(2)}/L = L^{-1}\sum_l \hat{\sigma}^z_l \hat{\sigma}^z_{l+1}$, and (e) $Z^{(3)}/L = L^{-1}\sum_l \hat{\sigma}^z_l \hat{\sigma}^z_{l+1}\hat{\sigma}^z_{l+2}$. The system size is $L = 14$, with the inverse temperature set as $\beta = 0.2$.

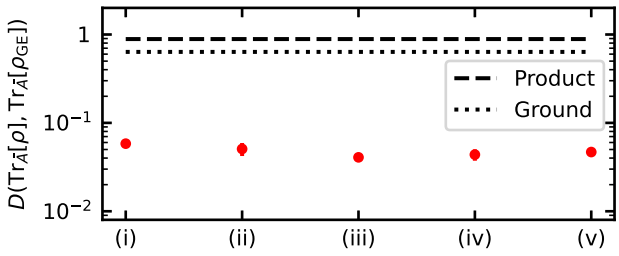

FIG. G6. The distance of the reduced density operators whose size is two from the target Gibbs state used in Sec. IV B. The red circles represent the prepared state considering different unitary generators and initial states. The choices of unitary generators and initial state are shown in Table V. The horizontal black dashed line shows the result for the product state which is the same as the one used in Sec. IV B. The horizontal black dotted line represents the ground state. The system size $L = 14$.

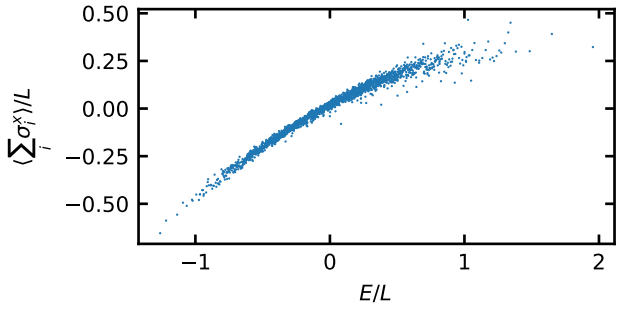

FIG. H7. The eigenenergy density dependence of the eigenstate expectation values of $L^{-1}\sum_{l=0}^{L-1}\sigma^x_l$, that is, we plot $\{\langle E_\alpha|\sum_{l=0}^{L-1}\sigma^x_l|E_\alpha\rangle/L\}_\alpha$, where $\{|E_\alpha\rangle\}_\alpha$ are the eigenstates. The Hamiltonian is $\hat{H}_{\rm Ising}$, whose parameters are the same as the ones used in Sec. IV B. We can observe that eigenstates in an energy shell share a typical value, which is a microcanonical ensemble average. Furthermore, these typical values show the nonlinear dependence on the eigenenergy density. The system size $L = 16$.

state and the prepared state. The values in these results are comparable to those in Sec. IV B, and we can conclude that the prepared states under choices (i)–(v) are typical.

Surprisingly, even for choices (iv) and (v), which cannot use $\hat{H}_{\rm Ising}$, the observables (d)–(e), which are not used for the reward, converge to the target values and the distances between the subsystems get small. We suppose that the unitary time evolutions, which maintain a steady state with respect to the observables added to the reward (the energy and total magnetization), are equivalent to the time evolutions by $\hat{H}_{\rm Ising}$ effectively, which makes the prepared state typical. We point out the connection to the studies [77] where the steady state has embedded Hamiltonian information that can be used to infer the parameters of the Hamiltonian. Of course, this phenomenon may be model-dependent and needs to be verified more precisely.

Appendix H: Eigenstate thermalization hypothesis

In this section, we briefly describe the eigenstate thermalization hypothesis (ETH), a hypothesis about the typical behavior of eigenstates in an energy shell. Furthermore, by looking at the dependence of eigenstate expectation values of the total magnetization $\sum_l \sigma^z_l/L$ on eigenstate energy density, we can infer the cause of the successful local preparation by using the total magnetization as an additional local observable of the reward for RL in Sec. IV B.

The idea that the energy eigenstates satisfy the canoni-

cal typicality is called the ETH [87–89]. The ETH claims that every energy eigenstate in the energy shell represents thermal equilibrium. More concretely, the energy eigenstates give the same expectation values of macroscopic observables as the relevant microcanonical ensemble for a large system:

$$\langle E_\alpha|\hat{O}|E_\alpha\rangle \approx \langle\hat{O}\rangle_{\mathrm{MC}} \tag{H1}$$

for every energy eigenstate $|E_\alpha\rangle$ in an energy shell, where $\langle\cdot\rangle_{\mathrm{MC}}$ is the corresponding microcanonical ensemble average. Based on the ETH, we can explain the thermalization mechanism of isolated quantum many-body systems [53, 56, 90]. The ETH has been verified numerically for few-body observables in a variety of non-integrable quantum many-body lattice models [73, 91–96].

In addition, a power-law decay with the dimension of the corresponding energy shell is observed for the variance of the energy eigenstate expectation values in the energy shell:

$$\sigma^2_{\mathrm{ETH}} \coloneqq \frac{1}{d}\sum_\alpha \left[\langle E_\alpha|\hat{O}|E_\alpha\rangle - \langle\hat{O}\rangle_{\mathrm{MC}}\right]^2, \tag{H2}$$

where d is the dimension of the energy shell, that is, the variance decays exponentially with the system size [91].

In Fig. H7, we show the eigenenergy density dependence of the eigenstate expectation values of total magnetization $L^{-1}\sum_{l=0}^{L-1}\sigma_l^x$, which is the additional local observable of the reward for RL in Sec. IV B. The Hamiltonian is $\hat{H}_{\mathrm{Ising}}$, whose parameters are the same as the ones used in Sec. IV B. Figure H7 shows that eigenstates in an energy shell have a typical value. This behavior is consistent with the ETH.

As we noted in Sec. III B 1, it is nontrivial to determine how many observables we need to embed the prepared state into a single energy shell. In Fig. H7, we also observe the nonlinear dependence of the typical values on energy densities. In particular, the dependence looks like strictly convex. We conjecture that this strict convexity helps the prepared state consist of eigenstates within a single energy shell for our demonstration in the non-integrable transverse-field Ising model.

Appendix I: Numerical methods

1. Non-integrable systems

For the numerical calculations regarding non-integrable systems, we adopt rigorous standard methods considering the Hilbert space whose dimension d scales exponentially with the system size. To calculate efficiently, the Hilbert space is resolved by the parity and momentum symmetry. Specifically, the calculation is limited to the zero-momentum sector and the parity-symmetric sector. The time evolution is performed by simply calculating $U|\psi\rangle$, which is multiplication of a $d\times d$ matrix by a d-dimensional vector. The construction of the symmetry-resolved basis and the time-evolution are implemented with QuSpin [80, 81].

2. Non-interacting integrable systems

In contrast to the numerical calculation regarding non-integrable systems, the calculations regarding integrable systems are done by exploiting the fact that the XX model can be mapped to a free fermionic system. In this section, we provide the details on the numerical methods used in Sec. V B, which corresponds to the preparations in the XX model.

Specifically, we will first discuss the Slater determinant, which efficiently describes free fermionic states, and then the time evolution of the Slater determinant. Next, we will explain how to calculate the expectation values of the fermionic observables, and finally how to calculate the expectation values of the observables consisting of hard-core bosons (HCBs), which is used to calculate the observables consisting of Pauli operators.

a. *Slater determinant*

The wave function $|\psi_{\mathrm{F}}\rangle$ of free-fermionic systems can be represented by a Slater determinant, namely a product of single-particle states:

$$|\psi_{\mathrm{F}}\rangle = \prod_{m=1}^{N_{\mathrm{f}}}\left(\sum_{l=1}^{L}P_{l,m}\hat{a}_l^\dagger\right)|0\rangle\,, \tag{I1}$$

where P is the $L\times N_{\mathrm{f}}$ matrix of components of $|\psi_{\mathrm{F}}\rangle$ and $|0\rangle$ represents a vacuum.

b. *Time evolution*

The time-evolution of $|\psi_{\mathrm{F}}\rangle$ under the unitary operator $\hat{U}$ generated by a quadratic Hamiltonian $\hat{H}_{\mathrm{q}} \coloneqq \sum_{m,n=1}^{L}H_{m,n}\hat{a}_m^\dagger\hat{a}_n$ with time length δt can be calculated as:

$$\hat{U}|\psi_{\mathrm{F}}\rangle = \prod_{m=1}^{N_{\mathrm{f}}}\left(\sum_{l=1}^{L}(UP)_{l,m}\,\hat{a}_l^\dagger\right)|0\rangle\,. \tag{I2}$$

This calculation is performed by multiplication of an $L\times L$ unitary matrix $U=\exp\left[-iH\delta t\right]$ by an $L\times N_{\mathrm{f}}$ matrix P.

c. *Fermionic observables*

Consider observables which are quadratic in fermions: $\hat{A} \coloneqq \sum_{m,n=1}^{L}A_{mn}\hat{a}_m^\dagger\hat{a}_n$. The expectation values of such

observables are calculated as:

$$\langle\psi_{\rm F}|\hat{A}|\psi_{\rm F}\rangle = \sum_{m,n} A_{mn}\,\langle\psi_{\rm F}|\hat{a}_m^\dagger\hat{a}_n|\psi_{\rm F}\rangle \tag{I3}$$

$$= \sum_m A_{mm} - \sum_{m,n} A_{mn}\,\langle\psi_{\rm F}|\hat{a}_n\hat{a}_m^\dagger|\psi_{\rm F}\rangle \tag{I4}$$

$$= \sum_m A_{mm} - \sum_{m,n} A_{mn} G^{\rm F}_{nm}, \tag{I5}$$

where $G^{\rm F}_{nm} := \langle\psi_{\rm F}|\hat{a}_n\hat{a}_m^\dagger|\psi_{\rm F}\rangle$ is the equal-time Green's function for fermions.

The creation of a particle at site m by acting $\hat{a}_m^\dagger$ on $|\psi_{\rm F}\rangle$ is represented by the addition of one column to P with the m-th element $P_{m(N_{\rm f}+1)} = 1$ and the rest are 0. In what follows, we denote the new component matrix of Slater determinant by $P^{{\rm F}(m)}$, which is $L\times(N_{\rm f}+1)$ matrix and generated by creating a fermion at site m on the Slater determinant represented by P. Because the inner product of two Slater determinants is calculated by the determinant of the product of the component matrices, the equal-time Green's function for fermions is calculated as:

$$G^{\rm F}_{nm} = \det\left[\left(P^{{\rm F}(n)}\right)^\dagger P^{{\rm F}(m)}\right]. \tag{I6}$$

When the columns of P are orthonormal vectors, we can derive $\det\left[\left(P^{{\rm F}(n)}\right)^\dagger P^{{\rm F}(m)}\right] = \delta_{nm} - \sum_{k=1}^{N_{\rm f}} P_{nk}P^*_{mk}$, which results in

$$\langle\psi_{\rm F}|\hat{A}|\psi_{\rm F}\rangle = {\rm Tr}\left[P^\dagger A P\right]. \tag{I7}$$

d. Hard-core bosonic observables

Next, we consider how to compute the expectation values of observables consisting of Pauli operators. In this section, we consider HCBs in order to introduce the creation and annihilation picture of particles. Here, we denote the creation and annihilation operators for a HCB acting on site m by $\hat{b}_m^\dagger$ and $\hat{b}_m$, respectively. HCB operators are introduced as $\hat{b}_m^\dagger = \left(\hat{\sigma}_m^x + i\hat{\sigma}_m^y\right)/2$, $\hat{b}_m = \left(\hat{\sigma}_m^x - i\hat{\sigma}_m^y\right)/2$. The calculating method described here follows the technique used in Refs. [54, 97–99].

Consider observables which is quadratic in HCBs: $\hat{B} := \sum_{m,n=1}^{L} B_{mn}\hat{b}_m^\dagger\hat{b}_n$. The expectation values of such observables are calculated as:

$$\langle\psi_{\rm F}|\hat{B}|\psi_{\rm F}\rangle = \sum_m B_{mm}\left(1 - G^{\rm B}_{mm}\right) + \sum_{m\neq n} B_{mn}G^{\rm B}_{nm}, \tag{I8}$$

where $G^{\rm B}_{nm} := \langle\psi_{\rm F}|\hat{b}_n\hat{b}_m^\dagger|\psi_{\rm F}\rangle$ is the equal-time Green's function for HCBs.

The action of $\hat{b}_m^\dagger = \hat{a}_m^\dagger\prod_{l=1}^{m-1} e^{-i\pi\hat{a}_l^\dagger\hat{a}_l}$ on $|\psi_{\rm F}\rangle$ is represented by a change of sign on the element P_{kl} for $k \le m-1$, and then the addition of one column to P with the m-th element $P_{m(N_{\rm f}+1)} = 1$ and the rest are 0. As a result, the Green's function for HCBs is calculated as:

$$G^{\rm B}_{nm} = \det\left[\left(P^{{\rm B}(n)}\right)^\dagger P^{{\rm B}(m)}\right], \tag{I9}$$

where $P^{{\rm B}(m)}$ is the new component matrix of Slater determinant, which is generated by creating a HCB at site m on the Slater determinant represented by P.

* baba@noneq.t.u-tokyo.ac.jp
† nyoshioka@ap.t.u-tokyo.ac.jp

[1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, 10th ed. (Cambridge University Press, Cambridge ; New York, 2010).
[2] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien, Quantum computers, Nature **464**, 45 (2010).
[3] V. Giovannetti, S. Lloyd, and L. Maccone, Advances in quantum metrology, Nature Photonics **5**, 222 (2011).
[4] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Quantum cryptography, Reviews of Modern Physics **74**, 145 (2002).
[5] B. M. Terhal and D. P. DiVincenzo, The problem of equilibration and the computation of correlation functions on a quantum computer, Physical Review A **61**, 022301 (2000).
[6] D. Poulin and P. Wocjan, Sampling from the Thermal Quantum Gibbs State and Evaluating Partition Functions with a Quantum Computer, Physical Review Letters **103**, 220502 (2009).
[7] S. Sugiura and A. Shimizu, Canonical Thermal Pure Quantum State, Physical Review Letters **111**, 010401 (2013).
[8] F. G. S. L. Brandão and M. J. Kastoryano, Finite correlation length implies efficient preparation of quantum thermal states, Communications in Mathematical Physics **365**, 1 (2019).
[9] J. Wu and T. H. Hsieh, Variational Thermal Quantum Simulation via Thermofield Double States, Physical Review Letters **123**, 220502 (2019).
[10] A. N. Chowdhury, G. H. Low, and N. Wiebe, A Variational Quantum Algorithm for Preparing Quantum Gibbs States, arXiv:2002.00055.
[11] Y. Wang, G. Li, and X. Wang, Variational Quantum Gibbs State Preparation with a Truncated Taylor Series, Physical Review Applied **16**, 054035 (2021).
[12] P. Doria, T. Calarco, and S. Montangero, Optimal Control Technique for Many-Body Quantum Dynamics, Physical Review Letters **106**, 190501 (2011).
[13] T. Caneva, T. Calarco, and S. Montangero, Chopped random-basis quantum optimization, Physical Review A **84**, 022326 (2011).
[14] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-

Herbrüggen, and S. J. Glaser, Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms, Journal of Magnetic Resonance **172**, 296 (2005).
[15] V. F. Krotov, *Global methods in optimal control theory*, Monographs and textbooks in pure and applied mathematics No. 195 (M. Dekker, New York, 1996).
[16] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists, Physics Reports **810**, 1 (2019).
[17] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Reviews of Modern Physics **91**, 045002 (2019).
[18] N. Yoshioka, Y. Akagi, and H. Katsura, Transforming generalized Ising models into Boltzmann machines, Physical Review E **99**, 032113 (2019).
[19] Y. Nomura, N. Yoshioka, and F. Nori, Purifying Deep Boltzmann Machines for Thermal Quantum States, Physical Review Letters **127**, 060601 (2021).
[20] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Neural-Network Quantum States, String-Bond States, and Chiral Topological States, Physical Review X **8**, 011006 (2018).
[21] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, Nature Communications **8**, 662 (2017).
[22] S. Lu, X. Gao, and L.-M. Duan, Efficient representation of topologically ordered states with restricted Boltzmann machines, Physical Review B **99**, 155136 (2019).
[23] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, Recurrent neural network wave functions, Physical Review Research **2**, 023358 (2020).
[24] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems, Physical Review Letters **124**, 020503 (2020).
[25] D. Pfau, J. S. Spencer, A. G. d. G. Matthews, and W. M. C. Foulkes, Ab-Initio Solution of the Many-Electron schrödinger Equation with Deep Neural Networks, Physical Review Research **2**, 033429 (2020).
[26] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).
[27] I. E. Lagaris, A. Likas, and D. I. Fotiadis, Artificial neural network methods in quantum mechanics, Computer Physics Communications **104**, 1 (1997).
[28] N. Yoshioka and R. Hamazaki, Constructing neural stationary states for open quantum many-body systems, Physical Review B **99**, 214306 (2019).
[29] F. Vicentini, A. Biella, N. Regnault, and C. Ciuti, Variational Neural-Network Ansatz for Steady States in Open Quantum Systems, Physical Review Letters **122**, 250503 (2019).
[30] M. J. Hartmann and G. Carleo, Neural-Network Approach to Dissipative Quantum Many-Body Dynamics, Physical Review Letters **122**, 250502 (2019).
[31] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, Nature Machine Intelligence **1**, 155 (2019).
[32] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nature Physics **14**, 447 (2018).
[33] G. Torlai and R. G. Melko, Latent Space Purification via Neural Density Operators, Physical Review Letters **120**, 240503 (2018).
[34] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nature Physics **13**, 431 (2017).
[35] N. Yoshioka, Y. Akagi, and H. Katsura, Learning disordered topological phases by statistical recovery of symmetry, Physical Review B **97**, 205110 (2018).
[36] P. A. Erdman and F. Noé, Identifying optimal cycles in quantum thermal machines with reinforcement-learning, npj Quantum Information **8**, 1 (2022).
[37] P. A. Erdman and F. Noé, Driving black-box quantum thermal machines with optimal power/efficiency trade-offs using reinforcement learning, arXiv:2204.04785.
[38] N. Yoshioka, W. Mizukami, and F. Nori, Solving quasiparticle band spectra of real solids using neural-network quantum states, Communications Physics **4**, 106 (2021).
[39] A. Nagy and V. Savona, Variational quantum monte carlo method with a neural-network ansatz for open quantum systems, Physical Review Letter **122**, 250501 (2019).
[40] S. Tibaldi, G. Magnifico, D. Vodola, and E. Ercolessi, Unsupervised and supervised learning of interacting topological phases from single-particle correlation functions, arXiv:2202.09281.
[41] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, 2nd ed., Adaptive computation and machine learning series (The MIT Press, Cambridge, Massachusetts, 2018).
[42] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Reinforcement Learning in Different Phases of Quantum Control, Physical Review X **8**, 031086 (2018).
[43] M. Bukov, Reinforcement learning for autonomous preparation of Floquet-engineered states: Inverting the quantum Kapitza oscillator, Physical Review B **98**, 224305 (2018).
[44] J. Yao, L. Lin, and M. Bukov, Reinforcement Learning for Many-Body Ground-State Preparation Inspired by Counterdiabatic Driving, Physical Review X **11**, 031070 (2021).
[45] V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios, and M. H. Devoret, Model-Free Quantum Control with Reinforcement Learning, Physical Review X **12**, 011059 (2022).
[46] Z. T. Wang, Y. Ashida, and M. Ueda, Deep Reinforcement Learning Control of Quantum Cartpoles, Physical Review Letters **125**, 100401 (2020).
[47] S. Borah, B. Sarma, M. Kewming, G. J. Milburn, and J. Twamley, Measurement-Based Feedback Quantum Control with Deep Reinforcement Learning for a Double-Well Nonlinear Potential, Physical Review Letters **127**, 190403 (2021).
[48] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, Universal quantum control through deep reinforcement learning, npj Quantum Information **5**, 33 (2019).
[49] Y. Ding, Y. Ban, J. D. Martín-Guerrero, E. Solano, J. Casanova, and X. Chen, Breaking adiabatic quantum control with deep learning, Physical Review A **103**, L040401 (2021).
[50] Z. An, H.-J. Song, Q.-K. He, and D. L. Zhou, Quantum optimal control of multilevel dissipative quantum systems with reinforcement learning, Physical Review A

103, 012404 (2021).
[51] P. Sgroi, G. M. Palma, and M. Paternostro, Reinforcement Learning Approach to Nonequilibrium Quantum Thermodynamics, Physical Review Letters **126**, 020601 (2021).
[52] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Reinforcement Learning with Neural Networks for Quantum Feedback, Physical Review X **8**, 031084 (2018).
[53] M. Rigol, V. Dunjko, and M. Olshanii, Thermalization and its mechanism for generic isolated quantum systems, Nature **452**, 854 (2008).
[54] M. Rigol, V. Dunjko, V. Yurovsky, and M. Olshanii, Relaxation in a Completely Integrable Many-Body Quantum System: An *Ab Initio* Study of the Dynamics of the Highly Excited States of 1D Lattice Hard-Core Bosons, Physical Review Letters **98**, 050405 (2007).
[55] L. Vidmar and M. Rigol, Generalized Gibbs ensemble in integrable lattice models, Journal of Statistical Mechanics: Theory and Experiment **2016**, 064007 (2016).
[56] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics, Advances in Physics **65**, 239 (2016).
[57] S. Goldstein, J. L. Lebowitz, R. Tumulka, and N. Zanghì, Canonical Typicality, Physical Review Letters **96**, 050403 (2006).
[58] S. Popescu, A. J. Short, and A. Winter, Entanglement and the foundations of statistical mechanics, Nature Physics **2**, 754 (2006).
[59] S. Sugiura and A. Shimizu, Thermal Pure Quantum States at Finite Temperature, Physical Review Letters **108**, 240401 (2012).
[60] L. V. Allis, *Searching for solutions in games and artificial intelligence*, Ph.D. thesis, Transnational University Limburg, Maastricht, Netherlands (1994).
[61] H. Iida, M. Sakuta, and J. Rollason, Computer shogi, Artificial intelligence **134**, 121 (2002).
[62] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature **529**, 484 (2016).
[63] The authors of Ref. [100] applied deep reinforcement learning to defeat a world champion program of shogi, which is based on a highly optimized alpha-beta search engine with many domain-specific adaptations.
[64] Y. Li, Deep Reinforcement Learning, arXiv:1810.06339.
[65] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, Deep Reinforcement Learning that Matters, arXiv:1709.06560.
[66] A. Stooke and P. Abbeel, rlpyt: A Research Code Base for Deep Reinforcement Learning in PyTorch, arXiv:1909.01500.
[67] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, Recurrent Experience Replay in Distributed Reinforcement Learning, in *International Conference on Learning Representations* (2019).
[68] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning, Nature **518**, 529 (2015).
[69] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation **9**, 1735 (1997).
[70] Gibbs ensembles, which is our target, differs from microcanonical ensembles. Nevertheless, a microcanonical ensemble is locally equivalent to a Gibbs (canonical) ensemble for translation invariant short-ranged Hamiltonians, even if the system is finite [101, 102]. In other words, the microcanonical and the canonical expectation values almost coincide when we look at the subsystem that is not too large.
[71] A. C. Cassidy, C. W. Clark, and M. Rigol, Generalized Thermalization in an Integrable Lattice System, Physical Review Letters **106**, 140405 (2011).
[72] M. Fagotti and F. H. L. Essler, Reduced density matrix after a quantum quench, Physical Review B **87**, 245107 (2013).
[73] H. Kim, T. N. Ikeda, and D. A. Huse, Testing whether all eigenstates obey the eigenstate thermalization hypothesis, Physical Review E **90**, 052105 (2014).
[74] Note that the chosen distance function can be efficiently computed for Slater determinants, which is used to express free-fermionic states obtained from Jordan-Wigner transformation. See Ref. [72] for the detailed property of this distance.
[75] E. Ilievski, J. De Nardis, B. Wouters, J.-S. Caux, F. H. L. Essler, and T. Prosen, Complete Generalized Gibbs Ensemble in an interacting Theory, Physical Review Letters **115**, 157201 (2015).
[76] K. Rudinger and R. Joynt, Compressed sensing for Hamiltonian reconstruction, Physical Review A **92**, 052322 (2015).
[77] E. Bairey, I. Arad, and N. H. Lindner, Learning a Local Hamiltonian from Local Measurements, Physical Review Letters **122**, 020504 (2019).
[78] A. Anshu, S. Arunachalam, T. Kuwahara, and M. Soleimanifar, Sample-efficient learning of interacting quantum systems, Nature Physics **17**, 931 (2021).
[79] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nature Physics **16**, 1050 (2020).
[80] P. Weinberg and M. Bukov, QuSpin: a Python package for dynamics and exact diagonalisation of quantum many body systems part I: spin chains, SciPost Physics **2**, 003 (2017).
[81] P. Weinberg and M. Bukov, QuSpin: a Python package for dynamics and exact diagonalisation of quantum many body systems. Part II: bosons, fermions and higher spins, SciPost Physics **7**, 020 (2019).
[82] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.
[83] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in *International Conference on Learn-*

ing Representations (2015).
[84] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. v. Hasselt, and D. Silver, Distributed Prioritized Experience Replay, in *International Conference on Learning Representations* (2018).
[85] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Physica D: Nonlinear Phenomena **404**, 132306 (2020).
[86] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, Dueling Network Architectures for Deep Reinforcement Learning, in *Proceedings of The 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 48 (PMLR, New York, New York, USA, 2016) pp. 1995–2003.
[87] J. M. Deutsch, Quantum statistical mechanics in a closed system, Physical Review A **43**, 2046 (1991).
[88] M. Srednicki, Chaos and quantum thermalization, Physical Review E **50**, 888 (1994).
[89] J. v. Neumann, Beweis des Ergodensatzes und desH-Theorems in der neuen Mechanik, Zeitschrift für Physik **57**, 30 (1929).
[90] T. Mori, T. N. Ikeda, E. Kaminishi, and M. Ueda, Thermalization and prethermalization in isolated quantum systems: a theoretical overview, Journal of Physics B: Atomic, Molecular and Optical Physics **51**, 112001 (2018).
[91] W. Beugeling, R. Moessner, and M. Haque, Finite-size scaling of eigenstate thermalization, Physical Review E **89**, 042112 (2014).
[92] T. Yoshizawa, E. Iyoda, and T. Sagawa, Numerical Large Deviation Analysis of the Eigenstate Thermalization Hypothesis, Physical Review Letters **120**, 200604 (2018).
[93] R. Steinigeweg, J. Herbrych, and P. Prelovšek, Eigenstate thermalization within isolated spin-chain systems, Physical Review E **87**, 012118 (2013).
[94] S. Sorg, L. Vidmar, L. Pollet, and F. Heidrich-Meisner, Relaxation and thermalization in the one-dimensional Bose-Hubbard model: A case study for the interaction quantum quench from the atomic limit, Physical Review A **90**, 033606 (2014).
[95] A. Khodja, R. Steinigeweg, and J. Gemmer, Relevance of the eigenstate thermalization hypothesis for thermal relaxation, Physical Review E **91**, 012120 (2015).
[96] R. Mondaini, K. R. Fratus, M. Srednicki, and M. Rigol, Eigenstate thermalization in the two-dimensional transverse field Ising model, Physical Review E **93**, 032104 (2016).
[97] M. Rigol and A. Muramatsu, Emergence of quasicondensates of hard-core bosons at finite momentum, Physical Review Letters **93**, 230404 (2004).
[98] M. Rigol and A. Muramatsu, Universal properties of hard-core bosons confined on one-dimensional lattices, Physical Review A **70**, 031603(R) (2004).
[99] M. Rigol and A. Muramatsu, Free expansion of impenetrable bosons on one-dimensional optical lattices, Modern Physics Letters B **19**, 861 (2005).
[100] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science **362**, 1140 (2018).
[101] F. G. S. L. Brandão and M. Cramer, Equivalence of Statistical Mechanical Ensembles for Non-Critical Quantum Systems, arXiv:1502.03263.
[102] H. Tasaki, On the local equivalence between the canonical and the microcanonical distributions for quantum spin systems, Journal of Statistical Physics **172**, 905 (2018).