Reinforcement Learning for Digital Quantum Simulation

Adrien Bolens and Markus Heyl
Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Straße 38, 01187 Dresden, Germany

(Received 3 July 2020; revised 9 May 2021; accepted 22 July 2021; published 9 September 2021)

Digital quantum simulation on quantum computers provides the potential to simulate the unitary evolution of any many-body Hamiltonian with bounded spectrum by discretizing the time evolution operator through a sequence of elementary quantum gates. A fundamental challenge in this context originates from experimental imperfections, which critically limits the number of attainable gates within a reasonable accuracy and therefore the achievable system sizes and simulation times. In this work, we introduce a reinforcement learning algorithm to systematically build optimized quantum circuits for digital quantum simulation upon imposing a strong constraint on the number of quantum gates. With this we consistently obtain quantum circuits that reproduce physical observables with as little as three entangling gates for long times and large system sizes up to 16 qubits. As concrete examples we apply our formalism to a long-range Ising chain and the lattice Schwinger model. Our method demonstrates that digital quantum simulation on noisy intermediate scale quantum devices can be pushed to much larger scale within the current experimental technology by a suitable engineering of quantum circuits using reinforcement learning.

DOI: 10.1103/PhysRevLett.127.110502

Introduction.—Digital quantum simulation (DQS) has emerged as one of the most promising applications of quantum computers. Unlike analog simulators, which directly mimic the Hamiltonian of interest, digital simulators reproduce a target time-evolution operator with a sequence of elementary quantum gates. In principle, the unitary time evolution of any spin-type Hamiltonian can be encoded in a quantum computer with arbitrary precision [1]. The experimental implementation of DQS has seen remarkable progress in recent years leading to the simulation of theoretical condensed matter models [2–7], lattice gauge theories [8], and quantum chemistry problems [9–11]. A common and natural approach to factorize time evolution operators into elementary quantum gates is to use Suzuki-Trotter formulas [12,13]. While the theoretical Trotter error can be well controlled [14–16], high accuracy Trotterization requires a large number of quantum gates. This leads to a critical problem because each of these individual gates suffers from experimental imperfections, in particular those which entangle qubits. A key challenge of DQSs is therefore to identify factorizations of time evolution operators utilizing a minimal number of quantum gates in order to exploit currently available hardware resources optimally. How many quantum gates are actually necessary to reproduce the targeted quantum dynamics is, however, an outstanding question.

In this work we introduce a method based on reinforcement learning (RL) to systematically build DQSs constrained to a fixed low number of entangling gates. We apply our method to two models chosen because of their relevance for DQS: the long-range Ising (LRI) model, unsolvable analytically but inheriting a natural Trotter decomposition, and the lattice Schwinger model, a key model to be simulated digitally [8,17]. The real-time dynamics of the lattice Schwinger model is hard to compute theoretically and to quantum simulate digitally as the Hamiltonian involves both short- and long-range couplings on a competing level. Moreover there is no existing quantum computing device which could achieve a natural implementation. As a crucial step in our RL algorithm towards feasible large-scale DQS we propose to optimize the quantum circuits not with respect to the conventionally used global many-body wave function, but rather based on a local reward with the goal to reproduce expectation values of local observables and correlation functions. Remarkably, we find that the dynamics of strongly correlated systems can be digitally realized using just a handful of gates making large system sizes and long-time simulations feasible on current day devices. Specifically, for the lattice Schwinger model, we build quantum circuits using only three entangling gates that correctly reproduce the dynamics of local observables and correlation functions for up to 16 qubits and for large times, reducing the number of entangling gates by one order of magnitude in comparison

to a recent pioneering DQS experiment for 4 qubits [8]. With our RL algorithm we are able to systematically build DQSs with a drastically reduced number of quantum gates for large quantum many-body systems pushing the design of quantum circuits beyond what has been achieved previously utilizing RL methods [18–21] or in the field of quantum control [8,17].

Although quantum supremacy is not possible within the current procedure, we emphasize that current experiments on digitally simulated real-time dynamics using Trotterization are limited still to small system sizes and simulation times. Our work provides a route towards larger-scale DQS in previously inaccessible regimes with currently available hardware resources. Further, our work contributes to the important open question on how many quantum gates are actually necessary to reproduce the targeted quantum dynamics. We show that a substantial reduction is possible which highlights that there can be algorithms with orders of magnitude smaller depth than those used today in the context of Trotterization.

Digital quantum simulation.—Let $\mathcal{H} = \sum_l \mathcal{H}_l$ be such that $U_l(t) = \exp(-i\mathcal{H}_l t)$ can be realized on the chosen quantum computing platform. The targeted dynamics can then be approximately factorized using the Suzuki-Trotter formula: $e^{-i\mathcal{H}\tau} \approx (\prod_l e^{-i\mathcal{H}_l\tau/n})^n$. This Trotterization comes with an error that is rigorously upper bounded as $O(N\tau^2/n)$ [14] with N the number of qubits, whereas the error on local observables can be even much smaller [15]. The central problem is that higher Trotterization accuracy requires larger n. This, however, increases the number of required quantum gates and therefore amplifies the imperfections due to faulty gate operations. In this work we aim to generate optimized quantum circuits for the factorization of time-evolution operators with a minimal number of quantum gates. We focus on trapped ion quantum computing platforms with the following set of universal quantum gates:

$$U_j^x(\theta) = e^{-i\theta\sigma_j^x}, \qquad U_j^z(\theta) = e^{-i\theta\sigma_j^z},$$
$$U^{xx}(\theta) = e^{-i\theta\sum_{j<k}\frac{\sigma_j^x\sigma_k^x}{|k-j|^\alpha}}, \tag{1}$$

where σ_j^x, σ_j^y, and σ_j^z are the Pauli matrices at site j. The exponent α can be theoretically tuned within the range $0 \le \alpha < 3$, but the optimal performance is typically reached either for $\alpha = 0$ or $\alpha \approx 1$. For the following we will focus for concreteness on either $\alpha = 3$ or $\alpha = 1$ while emphasizing that our approach can be straightforwardly applied also to other α or other quantum computing architectures such as superconducting qubits with different sets of universal quantum gates.

The central goal of our work is to find circuits with a small number of quantum gates for the task of reproducing the dynamics of a given Hamiltonian. We translate this task into an optimization problem as follows. Let $|\psi_0\rangle$ denote the initial state and let us fix the resources in terms of quantum gates as in Eq. (1). Then we construct a sequence of gates:

$$|\psi_{\mathrm{DQS}}\rangle = U_n \cdots U_2 U_1 |\psi_0\rangle, \tag{2}$$

$$U_t = U^{xx}(\theta_t^{xx}) \prod_j [U_j^z(\theta_t^{z,j}) U_j^x(\theta_t^{x,j})], \tag{3}$$

as depicted schematically in Fig. 1(a). The main goal now is to choose the underlying variational parameters $\theta = (\theta_t^{xx}, \theta_t^{z,1}, \theta_t^{x,1}, \ldots, \theta_t^{z,N}, \theta_t^{x,N})$ such that $|\psi_{\mathrm{DQS}}\rangle$ is as close as possible to $|\psi_{\mathrm{target}}\rangle = e^{-i\mathcal{H}\tau}|\psi_0\rangle$. From now on the number of entangling gates will be fixed to $n = 3$. As we will show, remarkably, these small quantum circuits will be sufficient to reproduce the dynamics of local observables, see Fig. 1(b).

Method.—We use RL to solve this difficult optimization problem. In RL a software agent learns by interacting with an environment and adapting its behavior accordingly. The agent generates sequences of actions in the environment and learns to perform a given task by maximizing a cumulative reward function. RL has seen a recent surge of applications in the field of quantum control for few-body problems [18,19,21–26] as it suits well optimization problems consisting of successive actions on a state with high dimensionality. Here, we are interested in the dynamics of quantum many-body problems which is a far more challenging problem.

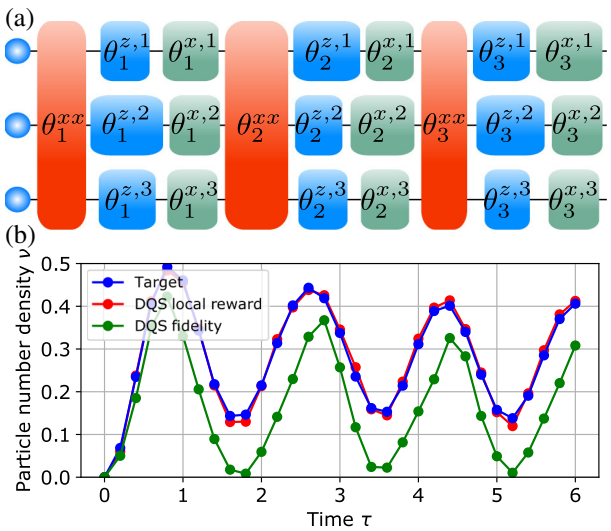

FIG. 1. (a) Quantum circuit used for the DQSs, here for a three-site system. The boxes represent different quantum gates. (b) Particle number density ν in the 10-site lattice Schwinger model starting from the bare vacuum for the parameters used in a recent DQS experiment [8]: $J = w = 2m$ in Eq. (7). We show the DQS results using the fidelity (green) and the local reward [Eq. (4)] (red), and the exact time evolution (blue).

In this work, inspired by the method used in Ref. [19], we use a modified version of a deep Q-network algorithm [27], a variant of the original Watkins off-policy Q-learning algorithm using artificial neural networks as function approximators [28,29]. While we now summarize the central aspects of the algorithm, further details can be found in Refs. [29,30].

The optimization problem is defined as an episodic RL problem: each episode is divided into a finite number of steps $t = 1, \ldots, n$, corresponding to the steps of the DQS. At $t = 0$, the quantum wave function is in a given initial state $|\psi_0\rangle$. Then, at each step t the agent chooses an action $a_t = (\theta_t^{xx}, \theta_t^{z,1}, \theta_t^{x,1}, \ldots, \theta_t^{z,N}, \theta_t^{x,N})$ defining the unitary U_t in Eq. (3). After each action the agent receives a reward r_t. At the end of the episode the reward $r_n \equiv R$ characterizes how close the final state $|\psi_{\mathrm{DQS}}\rangle$ is to the target state $|\psi_{\mathrm{target}}\rangle$. For intermediate steps, the reward is set to 0 as we do not constraint the specific evolution of the quantum wave function between the initial and target state. In deep Q-learning, a neural network is trained to predict the value $Q(s, a)$ of choosing an action a given a state s of the environment, following the update rule $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$ for the learning rate α, while the action $a_t = \mathrm{argmax}_a Q(s_t, a) +$ Gaussian noise [29,30]. Here, s describes the state of the quantum wave function. Importantly, the actions a take continuous values in our case, which is not standard for Q-learning. We have modified our algorithm accordingly so that the $\mathrm{argmax}_a Q(s, a)$ operation is done by maximizing the output of the neural network with respect to part of its input [30].

Reward.—A central quantity in the problem is the reward quantifying how close $|\psi_{\mathrm{DQS}}\rangle$ is to $|\psi_{\mathrm{target}}\rangle$. First, we define a global reward as the square of the fidelity $R_{\mathrm{global}} = |\langle\psi_{\mathrm{DQS}}|\psi_{\mathrm{target}}\rangle|^2$, which is commonly used to compare the two states. With a limited number of entangling gates, we find, however, that it is challenging to obtain high fidelities for large system sizes or times. As a consequence, we now introduce an alternative reward, which takes into account that in quantum simulation we are not so much interested in the global many-body wave function but rather in reproducing local observables and correlation functions. Let $\rho = |\psi_{\mathrm{target}}\rangle\langle\psi_{\mathrm{target}}|$ and $\sigma = |\psi_{\mathrm{DQS}}\rangle\langle\psi_{\mathrm{DQS}}|$. We then define a local reward

$$R_{\mathrm{local}} = 1 - \frac{2}{N(N-1)} \sum_{j<k} \sqrt{D(\rho^{jk}||\sigma^{jk})} \tag{4}$$

measuring the closeness of reduced density matrices ρ^{jk} and σ^{jk} of the subsystem made of sites j and k for ρ and σ, respectively. Here $D(\rho||\sigma) = \mathrm{Tr}\rho(\log\rho - \log\sigma)$ is the relative entropy and N denotes the number of qubits. A reward of $R_{\mathrm{local}} = 1$ means that all expectation values and correlation functions are reproduced exactly. In practice, we further cap negative values to zero such that $R_{\mathrm{local}} \in [0, 1]$. It is a crucial observation that a high local reward $R_{\mathrm{local}} = 1 - \epsilon$ can be directly translated into a high accuracy for local observables and correlations functions. For a two-body operator $O = \{2/[N(N-1)]\} \sum_{j<k} O^{jk}$ we have

$$|\langle O\rangle_{\mathrm{target}} - \langle O\rangle_{\mathrm{DQS}}| \leq \sqrt{2} \max_{j,k} \|O^{jk}\|_\infty \cdot \epsilon, \tag{5}$$

where $\|\cdots\|_\infty$ denotes the operator norm. This can be derived using Hölder's inequality for Schatten norms and Pinsker inequality [31,32]: $|\mathrm{Tr}[(\rho^{jk} - \sigma^{jk})O^{jk}]| \leq \|\rho^{jk} - \sigma^{jk}\|_1 \|O^{jk}\|_\infty \leq \sqrt{2D(\rho^{jk}||\sigma^{jk})}\|O^{jk}\|_\infty$. Similarly, for a single-body operator $O = (1/N)\sum_j O^j$ we have $|\langle O\rangle_{\mathrm{target}} - \langle O\rangle_{\mathrm{DQS}}| \leq \sqrt{2} \max_j \|O^j\|_\infty \epsilon$.

Results.—As a first proof of concept, we apply our method to the LRI model

$$\mathcal{H}_{\mathrm{LRI}} = J \sum_{j<k} \frac{1}{|k-j|^\alpha} \sigma_j^x \sigma_k^x + m_x \sum_j \sigma_j^x + m_z \sum_j \sigma_j^z. \tag{6}$$

For this system we can directly compare our approach to a conventional Trotterization procedure, as there exists a straightforward decomposition of the Hamiltonian into the universal set of quantum gates in Eq. (1) upon choosing $\theta_n^{xx} = J\tau/n$, $\theta_n^{z,j} = m_z\tau/n$, and $\theta_n^{x,j} = m_x\tau/n$. For concreteness, we will consider $J = 1$, $m_x = m_z = 2$, and $\alpha = 3$ starting from a fully polarized state $|\psi_0\rangle = |\uparrow \ldots \uparrow\rangle$. Let us emphasize, however, that we obtain similar results also for other choices of system parameters.

The learning of the agent is witnessed by the evolution of the reward as a function of episodes shown in Fig 2(a). Starting from the Trotterized circuit, the agent progressively improves the circuit until convergence. The mean value of the maximum rewards throughout each independent run is shown in Fig. 2(b) upon varying the system size N. As opposed to the Trotter fidelity, which decays exponentially, the DQS rewards remain at large values. The obtained fidelity decays with the system size N but only linearly at the considered N, and the local reward is remarkably unaffected. Now if we fix the system size and increase τ, the Trotterization also fails eventually. In Fig. 2(c) we show this together with the DQS results for both types of rewards. To give a more physical perspective to our results, we also compare the values of physical observables resulting from DQS, Trotterization, and from the actual dynamics (using exact diagonalization). Figures 2(d), 2(e), and 2(f) show the magnetization $\langle\psi|(1/2N)\sum_i \sigma_j^z|\psi\rangle$, the energy $\langle\psi|\mathcal{H}|\psi\rangle$, and the Loschmidt echo $|\langle\psi_0|\psi\rangle|^2$. For the 10-qubit system, after $\tau = 1$ it is clear that the system enters a regime where the Trotterization with $n = 3$ fails. At the same time, there is a drop in performance of our algorithm, but the reward converges to a finite value as τ increases. Importantly, when

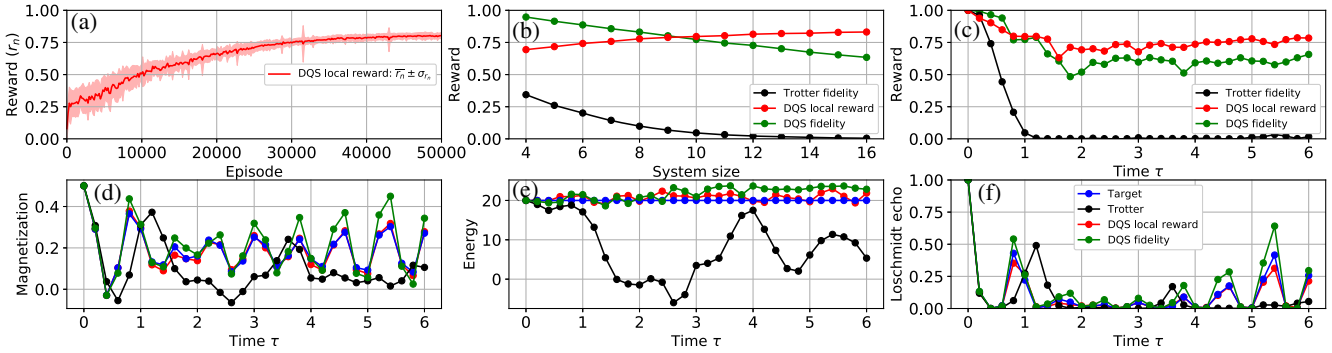

FIG. 2. Results for the DQS of the long-range Ising model with three entangling gates. (a) Evolution of the local reward during training for 100 independent runs for a 16-qubit system (average and standard deviation). (b) Rewards of the DQS as a function of the system size (for $\tau = 1.0$), and the corresponding Trotter fidelity. (c) Rewards of the DQS as a function of time (for a 10-qubit system), and the corresponding Trotter fidelity. (d) Magnetization obtained as function of time optimizing for the fidelity (green) and the local reward (red), the corresponding results using Trotterization (black), and the exact time evolution (blue) (for a 10-site system). (e) Same as (d) for the energy. (f) Same as (d) for the Loschmidt echo.

translated in terms of physical observables the resulting quantum circuits are much more successful than Trotterization. All quantities are well reproduced by the DQS, especially when the local reward is used. This indicates that our algorithm can systematically find a circuit bringing the initial state to an arbitrary target state using only three entangling gates.

Having demonstrated that our RL based method with local reward exhibits a remarkable performance for the LRI model, we now aim to go one step ahead by studying a system where no natural decomposition into a Trotter sequence exists. For that purpose we focus on the lattice Schwinger model:

$$\mathcal{H}_S = w\sum_j [\sigma_j^+ \sigma_{j+1}^- + \text{H.c.}] + \frac{m}{2}\sum_j (-1)^n \sigma_n^z + \frac{J}{2}\sum_{j=1}^{N-1}\left[\sum_{m=1}^{j}[\sigma_m^z + (-1)^m]\right]^2, \quad (7)$$

which is represented here in the Kogut-Susskind Hamiltonian formulation [33,34], as it has been recently realized experimentally using DQS based on Trotterization [8]. Concerning the nonequilibrium protocol we closely follow the experiment [8]. We start from the Néel state and apply $e^{-i\mathcal{H}_S\tau}$ with $w = J = 1$ and $m = 0.5$. Further, we use $\alpha = 1$ for the entangling gates in the DQS in Eq. (1), as this represents one of the optimal working points in systems of trapped ions.

Even more so than with the LRI model, optimizing with the fidelity only results in suboptimal sets of parameters as can be seen in Fig. 3. Both short-range and long-range couplings are present in the lattice Schwinger model, and thus reproducing the dynamics with only three entangling gates is particularly challenging. Nevertheless, we show that better sets of parameters do exist and are obtained when using the local reward. Interestingly, as for the LRI model, the performance of the algorithm with the local reward does not plummet as the system size increases, and

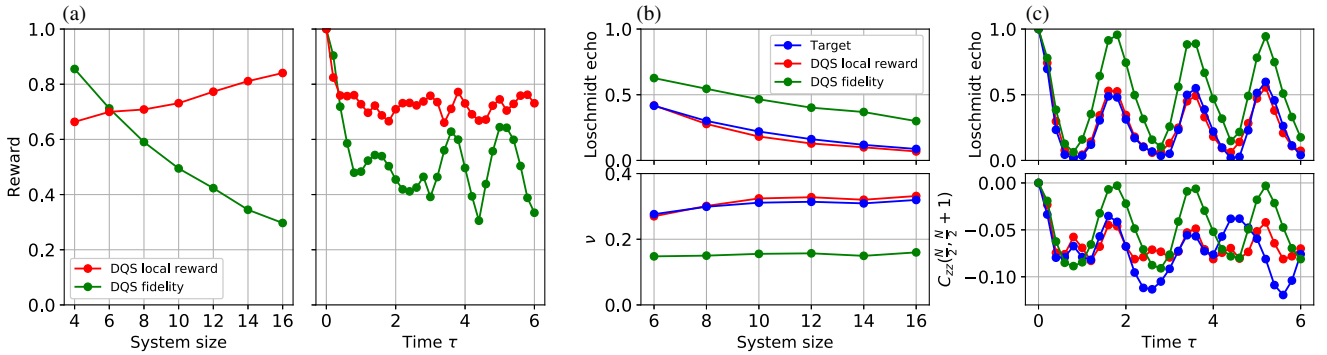

FIG. 3. Results for DQS of the Schwinger model with three entangling gates. (a) Final reward of the DQSs using the fidelity and the local reward Eq. (4) as a function of system size (for $\tau = 4.0$) and time (for a 10-site system). (b) Loschmidt echo and particle number density ν as a function of system size (for $\tau = 4.0$). (c) Loschmidt echo and nearest-neighbor quantum correlations in the middle of the chain of the DQS as a function of time using the fidelity (green) and the local reward Eq. (4) (red), and the exact time evolution (blue) (for a 10-site system).

physical observables are significantly better reproduced with the local reward than with the fidelity, as shown Figs. 3(b) and 3(c) and in Fig. 1(b), where the particle number density $\nu = (1/2N)\sum_{j=1}^{N}\langle(-1)^j\sigma_j^z + 1\rangle$ is shown, which has also been measured in the recent experiment [8].

While ν as a few-body operator is directly covered by the local reward, the Loschmidt echo is a global quantity, but can be nevertheless reproduced remarkably well. To explore further the performance of our RL approach, we compare in Fig. 3(c) the obtained dynamics for a two-body quantum correlation function against the exact solution. There we show results for the connected correlator $C_{zz}(N/2, N/2+1)$ in the middle of the chain where $C_{zz}(j, j+1) = \langle\sigma_j^z\sigma_{j+1}^z\rangle - \langle\sigma_j^z\rangle\langle\sigma_{j+1}^z\rangle$. While the two-body operator seems not as well reproduced as the single-body operator for long times, this is different for $\tau \lesssim 2.0$ when using the local reward. This is remarkable as the overall signal strength of $C_{zz}(N/2, N/2+1)$ is much smaller than what one would expect on the basis of the bound in Eq. (5).

Outlook.—For the considered problems three entangling gates have turned out to be typically sufficient for an accurate DQS of local observables, remarkably. In the future it might be important to increase the number of gates for higher precision, where convergence of our algorithm turns out to become progressively challenging. This might be remedied for instance by either utilizing more advanced neural network structures, e.g., recurrent neural networks or long short-term memories, or by reducing the number of independent variational parameters in the optimization problem using physical insights, in particular, by utilizing symmetries.

The current scheme requires an exact theoretically known reference of the target state, which we obtain using exact diagonalization. The overarching goal of DQS, however, is to address scenarios which are beyond such a theoretical description and therefore without such exact reference available. For current typical DQS scenarios such a regime of quantum supremacy is not yet reached, so that our algorithm represents a central contribution to push DQS significantly beyond what has been achieved up to now in terms of system size and simulation time. In particular, our work represents a key benchmark showing that a tremendous reduction of the circuit depth is possible. For instance, the algorithm used in the recent experiment [8] required roughly $2N$ entangling gates per time step. For $N = 10$ and the 30 time steps shown in Fig. 3 this implies more than 600 gates, which is orders of magnitude more than the 3 gates required with our algorithm. Our work therefore represents a first step in utilizing reinforcement learning for DQS with the key goal to reach an algorithm which doesn't rely anymore on an exactly known referenc.

We acknowledge Peter Zoller, Rick van Bijnen, and Christian Kokail for the fruitful discussions. This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (Grant Agreement No. 853443), and M. H. further acknowledges support by the Deutsche Forschungsgemeinschaft via the Gottfried Wilhelm Leibniz Prize program.

[1] R. P. Feynman, Int. J. Theor. Phys. **21**, 467 (1982).
[2] J. T. Barreiro, M. Müller, P. Schindler, D. Nigg, T. Monz, M. Chwalla, M. Hennrich, C. F. Roos, P. Zoller, and R. Blatt, Nature (London) **470**, 486 (2011).
[3] B. P. Lanyon, C. Hempel, D. Nigg, M. Müller, R. Gerritsma, F. Zähringer, P. Schindler, J. T. Barreiro, M. Rambach, G. Kirchmair *et al.*, Science **334**, 57 (2011).
[4] R. Barends, L. Lamata, J. Kelly, L. García-Álvarez, A. Fowler, A. Megrant, E. Jeffrey, T. White, D. Sank, J. Mutus *et al.*, Nat. Commun. **6**, 7654 (2015).
[5] Y. Salathé, M. Mondal, M. Oppliger, J. Heinsoo, P. Kurpiers, A. Potočnik, A. Mezzacapo, U. Las Heras, L. Lamata, E. Solano *et al.*, Phys. Rev. X **5**, 021027 (2015).
[6] N. Langford, R. Sagastizabal, M. Kounalakis, C. Dickel, A. Bruno, F. Luthi, D. Thoen, A. Endo, and L. DiCarlo, Nat. Commun. **8**, 1715 (2017).
[7] K. X. Wei, C. Ramanathan, and P. Cappellaro, Phys. Rev. Lett. **120**, 070501 (2018).
[8] E. A. Martinez, C. A. Muschik, P. Schindler, D. Nigg, A. Erhard, M. Heyl, P. Hauke, M. Dalmonte, T. Monz, P. Zoller *et al.*, Nature (London) **534**, 516 (2016).
[9] P. J. J. O'Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tranter, N. Ding *et al.*, Phys. Rev. X **6**, 031007 (2016).
[10] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Nature (London) **549**, 242 (2017).
[11] C. Hempel, C. Maier, J. Romero, J. McClean, T. Monz, H. Shen, P. Jurcevic, B. P. Lanyon, P. Love, R. Babbush *et al.*, Phys. Rev. X **8**, 031022 (2018).
[12] H. F. Trotter, Proc. Am. Math. Soc. **10**, 545 (1959).
[13] M. Suzuki, Prog. Theor. Phys. **56**, 1454 (1976).
[14] S. Lloyd, Science **273**, 1073 (1996).
[15] M. Heyl, P. Hauke, and P. Zoller, Sci. Adv. **5**, eaau8342 (2019).
[16] L. M. Sieberer, T. Olsacher, A. Elben, M. Heyl, P. Hauke, F. Haake, and P. Zoller, npj Quantum Inf. **5**, 78 (2019).
[17] C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos *et al.*, Nature (London) **569**, 355 (2019).
[18] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Phys. Rev. X **8**, 031084 (2018).
[19] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Phys. Rev. X **8**, 031086 (2018).
[20] M. Bukov, Phys. Rev. B **98**, 224305 (2018).
[21] J. Yao, M. Bukov, and L. Lin, arXiv:2002.01068.
[22] C. Chen, D. Dong, H.-X. Li, J. Chu, and T.-J. Tarn, IEEE Trans. Neural Networks Learn. Systems **25**, 920 (2013).
[23] F. Albarrán-Arriagada, J. C. Retamal, E. Solano, and L. Lamata, Phys. Rev. A **98**, 042315 (2018).

[24] X.-M. Zhang, Z.-W. Cui, X. Wang, and M.-H. Yung, Phys. Rev. A **97**, 052333 (2018).
[25] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, npj Quantum Inf. **5**, 33 (2019).
[26] Y.-H. Zhang, P.-L. Zheng, Y. Zhang, and D.-L. Deng, Phys. Rev. Lett. **125**, 170501 (2020).
[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, Nature (London) **518**, 529 (2015).
[28] C. J. Watkins and P. Dayan, Mach. Learn. **8**, 279 (1992).
[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).
[30] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.110502 for additional information on the optimization algorithm.
[31] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes* (Holden-Day, San Francisco, 1964).
[32] E. A. Carlen and E. H. Lieb, J. Math. Phys. (N.Y.) **55**, 042201 (2014).
[33] J. Schwinger, Phys. Rev. **128**, 2425 (1962).
[34] J. Kogut and L. Susskind, Phys. Rev. D **11**, 395 (1975).