Manipulation of Spin Dynamics by Deep Reinforcement Learning Agent

Jun-Jie Chen\(^1\),\(\ast\) and Ming Xue\(^1\)

\(^1\)State Key Laboratory of Low Dimensional Quantum Physics, Department of Physics, Tsinghua University, Beijing 100084, China

We implement the reinforcement learning agent in spin-1 atomic system to prepare twin-Fock state from given initial state. Proximal policy gradient (PPO) algorithm is used to deal with continuous space of control field and the final optimized protocol is given by a stochastic policy. In both mean-field system and two-body quantum system, RL agent finds the optimal policies. In many-body quantum system, it also gives polices that outperform purely greedy policy and optimized adiabatic passage. These polices given by RL agent have good physical interpretability in phase space and may help us to understand the quantum dynamics. With thorough exploration of state space, RL policy is also robust to noises and have good generalization capability. In fact, RL could be highly versatile in quantum optimal control problems.

I. INTRODUCTION

Understanding the dynamics in quantum system is an important topic in physics. It is conductive to the optimal control of target state preparation in quantum information and precision measurement. The challenge is to find an optimal or sub-optimal protocol of external control field that can evolve the initial state to target state both quickly and accurately. Various methods based on quantum optimal control theory [1–7] or adiabatic shortcut [8–12] are applied in different system to achieve better performance even reach the quantum speed limit. Generally, in a large class of linear quantum system, it is shown that the transition probability landscape has no local sub-optimal when the system is fully controllable [13], which means that perfect control can be find through traditional convex optimization such as gradient decent. However, the prerequisite of full controllability can be easily violated due to limitation of control field and discretion of time, etc., and thus the original landscape crashes. It is also hard to implement such algorithm for large system due to computational complexity. On the other hand, all these methods are based on pure theories instead of experiences data, which sets a gap between simulation and true experiments.

Reinforcement learning (RL) is a class of optimization algorithm that can learn an (sub)optimal policy from interaction with environment. By constantly observe the state of environment, take action and get feedback reward from it, RL agent collects the experiences data and use them to update its policy such that some long-term cumulative rewards are maximized. Compared to traditional optimal control theory, RL has two distinct advantages. First, RL can be implemented in a model-free way, i.e., the agent needs no prior human knowledge of the given system, which makes RL an universal learning framework for many dynamical systems. Second, RL can handle any given object function once we can design a proper reward in the problem and it provides great flexibility to achieve various optimization goals.

In recent years, along with the development of deep learning, RL has achieved great progress in many areas including video and board games, natural language processing, electronic trading, etc [14–16]. In physics, RL is gradually being widely used in quantum state preparation [17, 18], quantum computation and error correction [19–23], quantum phase transitions [24] and quantum robotics [25], etc. For various tasks, RL shows a certain degree of advantages over traditional theories and algorithms. Most works use value-based methods, such as Q-learning, and vanilla policy gradients to deal with dynamical systems with discrete action space, that is the value of control field is discretized. In fact, such discretization could change the landscape and makes the agent unable to learn true policy. The final results may also lack interpretability, even being nonphysical.

In this work, we consider to deploy a RL agent on spin-1 atomic system which learns to generate twin-Fock state evolved from given initial state by controlling external magnetic field. This system is widely studied both theoretically and experimentally [26–31]. In these works, spin squeezing is usually generated by using collective Rabi oscillation or adiabatic passage and high squeezing ratio can be realized. Here we use proximal policy optimization algorithm (PPO) to learn a better control protocol by numerical simulation. PPO belongs to actor-critic type RL algorithm proposed by OpenAI in [32] and has been applied in many challenging problems, such as complicated real-time strategy game and robotics. PPO can easily handle both continuous state and action space, which is also more realistic in physical systems. This work presents a general scheme converting a physical dynamical system to a standard RL task with proper state features, action representation and reward function. In section II, we give a brief introduction to RL from scratch and specify the problems we consider. Section III and IV shows the learning policies of mean-field and quantum dynamical systems with different particle number. Section V summarized the results and identify some problems for further improvement.

\(\ast\) chen-jj13@mails.tsinghua.edu.cn
II. SPIN DYNAMICS AS A REINFORCEMENT LEARNING TASK

Reinforcement learning (RL) is an area of machine learning that learns how to act in a large class of dynamical system to maximize given cumulative rewards. Typical RL task is always modeled as a Markov decision process (MDP) as shown in Figs. (1). Such MDP is composed of two major objects, agent and environment. At each time $t$, the agent observes the environment and obtains its state feature $s_t$. Based on the current policy $\pi$ and state $s_t$, the agent takes action $a_t$ and acts on the environment. Then the environment will evolve to state $s_{t+1}$ due to $a_t$ and a reward $r_t$ will be fed back to agent itself. According to $r_t$ and previous experiences, the agent will update its policy $\pi$ by specific RL algorithm. The ultimate object for the agent is to learn an optimal policy $\pi^*$ that satisfies

$$\pi^* = \arg\max_{\pi} J \text{ with } J = \sum_{t=0}^{T_f} \gamma^t \cdot r_t.$$  \hfill (1)

Object function $J$ is a cumulative rewards with discounted factor $\gamma \in [0, 1]$. When $\gamma = 0$, the agent is totally greedy that only tries to maximize instantaneous reward $r_t$ at each time step. Typical $\gamma$ is always chosen to be closed to 1 such that agent can find the true global optimal solution.

FIG. 1. Framework of typical MDP in reinforcement learning.

In RL framework, the policy $\pi$ is a function

$$\pi : \mathcal{S} \rightarrow \mathcal{A},$$  \hfill (2)

that maps the state space onto action space. Common policies are of two types, deterministic and stochastic. The deterministic policy is a function of state as $a = \pi(s)$, i.e., for given state $s$, the policy will give an unique value of action $a$. While the stochastic policy can be represented as $\pi(a|s)$, i.e., for given state $s$ the policy returns a distribution on action space $\mathcal{A}$. In fact, deterministic policy can be viewed as a special stochastic policy with zero variance around $a$. In deep RL, we usually use neural network or other parametric model to approximate the policy function as $\pi_\theta(s)$ and $\pi_\theta(a|s)$ with $\theta$ being trainable parameters. Then the RL task described in (1) can be viewed as a general quadratic optimization problem and $\theta$ can be found by using gradient descent

$$\theta \leftarrow \theta - \alpha \cdot \nabla_\theta J(\theta),$$  \hfill (3)

where $\alpha$ is learning rate. The gradient of $J$ can be further represented in a maximum likelihood form as

$$\nabla_\theta J(\theta) = E_{\tau \sim \pi_\theta} \left[ \sum_{t=0}^{T_f} \nabla_\theta \log \pi_\theta(a_t|s_t) A^\pi_\theta(s_t, a_t) \right],$$  \hfill (4)

where $\tau$ is a trajectory following $\pi_\theta$ and $A^\pi_\theta$ is the advantage function. There are two ways to estimate the advantage function, statistic and parametric ways. Vanilla policy gradient algorithm uses statistic inference from experiences data $(s_t, a_t, r_t, s_{t+1})$ to estimate $A^\pi_\theta$. Actor-critic algorithm uses another neural network (or other parametric model) to approximate $A^\pi_\theta$. Here $\pi_\theta$ is the actor which is responsible for choosing action while $A^\pi_\theta$ is the critic which gives the value of state and action. In this work, we use proximal policy optimization algorithm (PPO) to learn a stochastic Guassian policy $\pi_\varphi(a|s)$. PPO is an advanced actor-critic type algorithm developed in recent years. Compared to traditional methods, PPO has a more robust learning process due to first-order trust region search gradient descent and can handle both discrete and continuous action space. More details can be found in [32].

In this paper, we consider a spin-1 system with its Hamiltonian being

$$H = \int d^3 \sum_{m=-1}^{1} \hat{\psi}_m^\dagger \left[ \frac{\hbar^2}{2m} \nabla^2 + V - pm + q(t)m^2 \right] \hat{\psi}_m + \frac{1}{2} \sum \hat{n}_i^2 + c_2 : \hat{F}_i^2 :,$$  \hfill (5)

where $p$ is linear Zeeman and $q(t)$ is time-dependent quadratic Zeeman, $\hat{n}_i = \sum_{m} \hat{n}_m$ ($m = \pm 1, 0$) is total density operator and $\hat{F}_i = \sum_{m,n} (\hat{F}_i)_{mn} \hat{\psi}_m^\dagger \hat{\psi}_n$ ($i = x, y, z$) is the spin-1 operator. The Hamiltonian conserves total magnetic moment $F_z$ and particle number. In the following, we always consider the dynamics in $F_z = 0$ subspace and ferromagnetic interaction with $c_2 < 0$. The ultimate goal is to find (a) an (sub-)optimal protocol $q(t)$ that evolves the quantum state from given initial state $|\psi_i\rangle$ to target state $|\psi_f\rangle$. Here we choose the target state to be the twin-Fock state $|\psi_f\rangle = |N/2, 0, N/2\rangle$ which is a squeezed state in spin space and very valuable for quantum precise measurements. The initial state $|\psi_i\rangle$ can be arbitrarily given.

Now we are ready to convert our problem of quantum state preparation in spin-1 system to a standard RL task. To this end, we will first specify the definition of $\mathcal{S}, \mathcal{A}, \mathcal{R}$ for our system.

- State space $\mathcal{S}$: The most straight-forward representation of $s_t$ for a quantum state is its wave function $|\psi\rangle$ because it contains all the information we need. However, $|\psi\rangle$ is not always the best choice, especially for a many-body quantum system since its dimension increase (exponentially) as particle number increasing. An alternative is to use representative physical observables to describe state $s_t$. 

Compared to wave function, there are two advantages. First, the dimension of state features is unchanged which makes it possible to generalize the policy to different particle number. Second, the output policy has more interpretability when state feature has clear physical meaning. In this work, we use the second option.

- **Action space**: Action is clear in this problem that \( a_t = q(t) \), the second-order quadratic Zeeman term at time \( t \). Experimentally, \( q \) can be tuned by external magnetic field or by microwave dressing.

- **Reward space **: Reward \( r_t \) selection depends on the optimization target itself. Here we want to achieve maximum fidelity at final time \( T \), that is 

\[
J = |\langle \psi(T) | \psi_f \rangle|^2. \tag{6}
\]

However, feedback might be too sparse if we only give reward at final time \( T \) and the training process will be very hard, even fail under current setup. In fact, we can decompose object function (6) into a summation as

\[
J = \sum_{i=0}^{n} \left( |\langle \psi(t_i) | \psi_f \rangle|^2 - |\langle \psi(t_{i-1}) | \psi_f \rangle|^2 \right). \tag{7}
\]

Here, total evolution time \( T \) is discretized into consecutive period ended at \( t_i \) \( i = 0, \ldots, n \). At each time step \( t_i \), a reward

\[
r_i = |\langle \psi(t_i) | \psi_f \rangle|^2 - |\langle \psi(t_{i-1}) | \psi_f \rangle|^2, \tag{8}
\]

which is the instantaneous change of fidelity, is fed back to agent. Dense rewards scheme makes the training more quickly and stable. Though \( r_i \) seems to be greedy on fidelity increasing, RL algorithm always try to maximize \( J \), i.e., the summation of \( r_i \). It won’t restrict us on greedy policy and RL agent still learns to find global optimal. We can further modified the reward as

\[
r_i = -\ln \left( \frac{1 - |\langle \psi(t_i) | \psi_f \rangle|^2}{1 - |\langle \psi(t_{i-1}) | \psi_f \rangle|^2} \right), \tag{9}
\]

which ensures the agent can still learn well even when the state has evolved to near target.

In the following sections, we show the performance of RL agent on various environments, including mean-field and quantum dynamics. Some interpretabilities are also extracted from the final policies.

### III. MEAN-FIELD DYNAMICS

Under the mean-field approximation, we have \( \hat{\psi} = (\xi_1, \xi_0, \xi_{-1})^T \) with field parameterization \( \xi_1 = \sqrt{(1 - \rho_0)/2} \cdot e^{i \chi_+}, \xi_0 = \sqrt{\rho_0} \) and \( \xi_{-1} = \sqrt{(1 - \rho_0)/2} \cdot e^{i \chi_-} \). Here \( \rho_0 \) is the population density on \( m_F = 0 \) state and \( \chi_{\pm} \) is the phase of \( m_F = \pm 1 \) components. Then the spin dynamics in \( F_z = 0 \) subspace is governed by

\[
\dot{\rho}_0 = \frac{2c_2}{\hbar} \rho_0(1 - \rho_0) \sin \theta_s, \tag{10}
\]

\[
\dot{\theta}_s = -\frac{2q}{\hbar} + \frac{2c_2}{\hbar} (1 - 2\rho_0)[1 + \cos \theta_s], \tag{11}
\]

where \( \theta_s = \chi_+ + \chi_- \). The solution of the coupled equations is a classical non-rigid pendulum, which is a typical RL benchmark task. Our target for this task is to evolve initial state to \( \rho_0 = 0 \) state. In this system, the state is \( s_t = (\rho_0(t), \theta_s(t)) \) and the action is \( q \) itself. Since the phase space of mean-field dynamics is 2-dimensional, state tuple \((\rho_0, \theta_s)\) is sufficient to represent the whole system. That is, this system is a pure MDP.

Now we deploy PPO agent on this mean-field system. We choose a very small neural network to approximate the actor and critic. The hidden layer contains two fully-connected layers (or MLP) with 32 and 16 neurons on layer and the activation function is tanh. The output connected layers (or MLP) with 32 and 16 neurons on the actor and critic. The hidden layer contains two fully-connected layers (or MLP) with 32 and 16 neurons on layer and the activation function is tanh. The output connected layers (or MLP) with 32 and 16 neurons on the actor and critic. The hidden layer contains two fully-connected layers (or MLP) with 32 and 16 neurons on layer and the activation function is tanh.

| TABLE I. Mean-field system | Hyperparameters | Value |
|-----------------------------|----------------|-------|
| hidden size | \([32, 16]\) |       |
| activation | tanh |       |
| discounted factor \( \gamma \) | 0.999 |       |
| actor-network learning rate | 3E-4 |       |
| critic-network learning rate | 1E-3 |       |
| steps/episode | 100 |       |
| time/step\((dt)\) | 0.05 |       |
| target KL-divergence | 0.01 |       |
Dynamics is almost frozen while magnetic phase \( \theta \) then moves along this geodesic path to the target \( \rho \). Trajectories cumulate rapidly, which corresponds to the horizontal evolution of the trajectories on phase space guided by generalized policy \( \pi \).

In Figs.(3a), we can see that all trajectories have the same characteristics. They always move horizontally first until they reach \( \theta_s = \pi/2 \) and then move along this geodesic path to the target \( \rho \). Figs.(3b) plots the time series of \( q(t) \). At the beginning of evolution, \( q \) is much larger than \( c_2 \) such that the spin dynamics is almost frozen while magnetic phase \( \theta_s \) is accumulating rapidly, which corresponds to the horizontal trajectory in Figs.(3a). Then \( q \) decays to small value comparable with \( c_2 \) that activates spin dynamics and the population on \( m_F = \pm 1 \) states increases. In fact, it is obvious to show that such policy is optimal. From RHS of (10), the decay rate of \( \rho_0 \) is maximized only when \( \theta = \pi/2 \). To keep \( \theta \) constant, we have to set RHS of (11) to be zero all the time. We have

\[
q(t) = c_2(1 - \rho_0(t)),
\]

which is quite similar to the protocol shown in Figs.(3b) when \( t > 0.5 \). In fact, if \( q \) is unlimited, we can set \( q = \pm \infty \) at the initial time and tune the magnetic phase \( \theta_s \) to \( \pi/2 \) with no time cost. To sum up, in this simple mean-field spin dynamics system, the RL agent is able to learn an optimal policy that can evolve any initial state to target state.

IV. QUANTUM DYNAMICS

The quantum dynamics of spin-1 system can be revealed from single mode approximation (SMA). Under SMA, we assume that all particles share the same spatial mode. This allows the field operator to be approximated as \( \hat{\psi}_m = \hat{a}_m \cdot \phi(\vec{r}) \) \( (m = 0, \pm 1) \) where \( \phi \) is the wave function of spatial mode and \( \hat{a}_m \) is the annihilation operator of spin state \( m \). The Hamiltonian under SMA is

\[
H = \frac{c_2}{2N} \left[ 2(N_0 - 1)(\hat{N}_1 + \hat{N}_{-1}) + 2(\hat{a}_1^\dagger \hat{a}_{-1}^\dagger \hat{a}_0 \hat{a}_0 + \text{h.c.}) \right] - q\hat{N}_0,
\]

where \( N \) is total particle number and \( c_2 \) is the coupling strength of spin-exchange interaction. Still we assume \( c_2 = -1.0 \) and total magnetic moment \( F_z = 0 \) in the following discussion.

In quantum system, we can directly use wave function as the representation of state \( s_t \) which makes the spin dynamics a pure MDP. For particle number \( N \), the Hilbert space has dimension \( N/2 + 1 \) and \( s_t \) feature size is \( N + 2 \) (module and phase). However, using wave function as features will lose some interpretability and generalization ability as discussed before. Alternatively, we use physical observables \( \rho_0 \) and \( \theta_s \) as state features, in which

\[
\rho_0 = \langle \hat{N}_0 \rangle / N, \quad \theta_s = \text{args } \langle \hat{a}_1^\dagger \hat{a}_{-1}^\dagger \hat{a}_0 \hat{a}_0 \rangle.
\]

These two observables are almost identical to those used in mean-field dynamics. In fact, all representative observables in this spin-1 quantum system, including total angular momentum \( \langle \hat{L}_i \rangle \), \( \langle \Delta \hat{L}_i \rangle \) \( (i = x, y, z) \) even squeezing ratio \( \xi \), can be inferred from \( \rho_0 \) and \( \theta_s \). On the other hand, \( \rho_0 \) and \( \theta_s \) lose some information of the true quantum state which makes it a partially observable MDP (POMDP) and may have negative impacts on policy performance. The hyper-parameters used for our simulations are listed in Table-II and no sophisticated tuning is made either. The framework is quite similar to that used in mean-field system.
TABLE II. Quantum system

| Hyperparameters          | Value |
|-------------------------|-------|
| hidden size             | 64, 32|
| activation              | tanh  |
| discounted factor $\gamma$ | 0.999 |
| actor-network learning rate | 3E-4  |
| critic-network learning rate | 1E-3  |
| steps/episode           | 200   |
| time/step(dt)           | 0.1   |
| target KL-divergence    | 0.01  |

A. Two-body problem $N = 2$

First, we consider the simplest case with $N = 2$. The wave function can be represented as a pseudo-spin-1/2 on Bloch sphere

$$|\psi\rangle = \cos \frac{\theta}{2} |1\rangle + \sin \frac{\theta}{2} e^{i\phi} |2\rangle,$$

where $|1\rangle = |0, 2, 0\rangle$, $|2\rangle = |1, 0, 1\rangle$ and it is easy to show that $\rho_0 = \cos^2 \theta/2$ and $\theta_s = -\phi$. The equation of motion is

$$\dot{\theta} = -\sqrt{2}c_2 \sin \phi,$$

$$\phi = -c_2 q + \sqrt{\frac{5}{2}} \cos \phi \left( \tan \frac{\theta}{2} - \cot \frac{\theta}{2} \right).$$

The optimal policy that evolves state $|1\rangle$ to $|2\rangle$ is obtained when $\phi = \pi/2$ ($\theta_s = -\pi/2$) and $\dot{\phi} = 0$, which implies

$$q(t) = c_2/4.$$

It is just a simple Rabi-oscillation between initial state $|0, 2, 0\rangle$ and target state $|1, 0, 1\rangle$. In Figs. (4), we show the policies learned by RL agent with fixed ($\pi_s$) or random ($\pi_g$) initial state. Here we also observe that the phase space (Hilbert space) is explored more thoroughly with randomized initial state. Figs.(5) shows the evolution results of three typical initial states $|\psi\rangle = |1, 0, 1\rangle + 0.1e^{i\phi} |2\rangle$ and $0.9|1\rangle + 0.1e^{-i\pi/2}|2\rangle$ based on policy $\pi_g$. The trajectories learned by the agent have same characteristics as those of mean-field system and magnetic field $q$ is closed to theoretical optimal value $c_2/4$ when $\theta_s$ is adjusted to near $-\pi/2$.

We can further obtain the quantum speed limit $T_{QSL}$ of this simple system

$$T_{QSL} = \min \int_0^{\pi} \frac{1}{\sqrt{2} |c_2|} d\theta,$$

which is almost achieved by our policy shown in Figs. (5-c). We also notice that equation(20) is identical to the Bhattacharyya bound $T_{QSL} = \Delta E_0^{-1} \arccos (\psi_i | \psi_f)$ where $|\psi_i\rangle = |1\rangle$, $|\psi_f\rangle = |2\rangle$ and $\Delta E_0$ is the energy variance of initial state $|\psi_i\rangle$. In our case, $\Delta E_0 = \sqrt{2} |c_2|/2$. This is obvious because the dimension of Hilbert space is only two when $N = 2$. To sum up, RL agent is able to learn the optimal policy in simple two-body system.

B. Many-body problem $N = 10$

In the following we consider the many-body quantum dynamics with particle number $N = 10$. Now the system has two major differences from two-body and mean-field dynamics. First, the system becomes a POMDP when we use $s_t = (\rho_0, \theta_s)$. But since $\rho_0$ and $\theta_s$ catch most of the important features of this system, we may still learn a good policy. Second, when the dimension of Hilbert space gets larger, there will be a longer frozen time period at the early stage of evolution in which the fidelity on target state almost remains zero. This frozen time makes the
reward being naturally sparse and training process could be hard. However, we still get applicable policies under such setup at least for many body system with tens of particle.

\[ q(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu(t))^2}{2\sigma^2}}, \]  

where \( \mu(t) \) is the exact value given by the policy and \( \sigma \) indicates the strength of noise. Figs.(8)(a1,b1) show the evolution of fidelity under noise (21) with \( \sigma = 0.1|c_2| \). The width of the shaded band represents the standard deviation of fidelity over 100 sampled trajectories. It is shown that \( \pi_g \) is much more stable than \( \pi_s \) even the protocol seems to be oscillating. We also apply the policy learned from \( N = 10 \) to other particle number case to observe its generalization capability. As shown in Figs.(8)(a2,b2), \( \pi_g \) also has better performance than \( \pi_s \).

For \( N < 10 \), both \( \pi_s \) and \( \pi_g \) are applicable because the dimension of Hilbert space is smaller than that of \( N = 10 \). A policy learned in higher dimension should be available in its subspace. For \( N > 10 \), only \( \pi_s \) works because the state space is explored more thoroughly and true physical rules are learned under its setup. On the other hand, we notice that policy can be generalized to different particle number only when we use physical observables as state representation because the dimension of state input is fixed.

V. CONCLUSION

In conclusion, we use deep reinforcement learning to optimize twin-Fock state preparation in spin-1 atomic system. It not only gives better control protocols than some traditional methods but also presents a general scheme to convert a physical dynamical system into a standard RL task. We have noticed that state representation is important to the physical interpretability of final policy and reward signal should not be sparse. One central problem of RL algorithm is exploration and exploit. Only when the environment is effectively explored and experiences data are collected, agent could learn optimal policy without falling into local minimums (maxima). In our problem, we expect that using randomized or multiple initial state in each training episode could help to ease such problems and makes the policies being more robust and easily to be generalized to different environments.

In section-IV we also point out two major problems in many body system. For the first problem caused by POMDP, it could be partially solved by using recurrent RL [33] or using multiple steps trajectory as state input [14]. The second problem of sparse reward is even more essential and might be promoted by using hierarchal reinforcement learning [34, 35] framework.
FIG. 7. Results of initial state $|0, N, 0\rangle$ (a1, b1) evolution trajectory on phase space guided by policy $\pi_s$ or $\pi_g$. (a2, b2) magnetic field $q(t)$ given by $\pi_s$ or $\pi_g$. (a3) time evolution of fidelity on target state $|N/2, 0, N/2\rangle$ following $\pi_s$ or $\pi_g$. Here the adiabatic passage refers to linear ramp in time period $[0, t]$ with $q$ changing from $q_i$ to $q_f$. $q_i$ and $q_f$ is optimized such that this linear ramp shows best performance. In greedy policy, $q$ is chosen at each time step to maximize instantaneous reward, that is the one-step fidelity increment.

FIG. 8. (a1, b1) stability of policy $\pi_s$ and $\pi_g$ added by Gaussian white noise. The results are extracted from 100 sampled trajectories. The solid blue line is the average value of fidelity and the shaded band represents the standard deviation. (a2, b2) generalization capability of policy $\pi_s$ and $\pi_g$ learned from $N = 10$ without noise.
[1] I. Brouzos, A. I. Streltsov, A. Negretti, R. S. Said, T. Caneva, S. Montangero, and T. Calarco, Phys. Rev. A 92, 062110 (2015).
[2] F. Caruso, S. Montangero, T. Calarco, S. F. Huelga, and M. B. Plenio, Phys. Rev. A 85, 042331 (2012).
[3] C. Brif, R. Chakrabarti, and H. Rabitz, New Journal of Physics 12, 075008 (2010).
[4] V. Beltrani, J. Dominy, T.-S. Ho, and H. Rabitz, The Journal of chemical physics 134, 194106 (2011).
[5] J. Yan, D. Hocker, R. Long, T.-S. Ho, and H. Rabitz, Phys. Rev. A 89, 063408 (2014).
[6] A. Rothman, T.-S. Ho, and H. Rabitz, Phys. Rev. A 73, 053401 (2006).
[7] H. Rabitz, T.-S. Ho, M. Hsieh, R. Kosut, and M. Demiralp, Phys. Rev. A 74, 012721 (2006).
[8] A. Sala, D. L. Nunez, J. Martorell, L. De Sarlo, T. Zibold, F. Gerbier, A. Polis, and B. Julii-Diaz, Phys. Rev. A 94, 043623 (2016).
[9] S. Campbell, G. De Chiara, M. Paternostro, G. M. Palma, and R. Fazio, Phys. Rev. Lett. 114, 177206 (2015).
[10] T. c. v. Opatrný, H. Saberi, E. Brion, and K. Molmer, Phys. Rev. A 93, 023815 (2016).
[11] B.-H. Huang, Y.-H. Kang, Y.-H. Chen, Z.-C. Shi, J. Song, and Y. Xia, Phys. Rev. A 97, 012333 (2018).
[12] D. Sels and A. Polkovnikov, Proceedings of the National Academy of Sciences 114, E3909 (2017).
[13] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, Science 303, 1998 (2004).
[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, Nature 518, 529 (2015).
[15] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot, Nature 529, 484 (2016).
[16] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, and T. Graepel, arXiv:1712.01815 (2017).
[17] M. Bukov, Physical Review B 98, 224305 (2018).
[18] S. Yu, F. Albaran-Arriagada, J. Retamal, Y.-T. Wang, W. Liu, Z.-J. Ke, Y. Meng, Z.-P. Li, J.-S. Tang, and E. Solano, arXiv:1808.09241 (2018).
[19] P. Andreasson, J. Johansson, S. Liljestrand, and M. Granath, arXiv:1811.12338 (2018).
[20] S. Herbert and A. Sengupta, arXiv:1812.11619 (2018).
[21] J. Lin, Z. Y. Lai, and X. Li, arXiv:1812.10797 (2018).
[22] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, arXiv:1802.08451 (2018).
[23] R. Sveke, M. S. Kesselring, E. P. van Nieuwenburg, and J. Eisert, arXiv:1810.07207 (2018).
[24] M. Bukov, A. G. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, arXiv:1705.00565 (2017).
[25] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, Phys. Rev. X 4, 031002 (2014).
[26] M. Kitagawa and M. Ueda, Phys. Rev. A 47, 5138 (1993).
[27] O. E. Mustecaplo glu, M. Zhang, and L. You, Phys. Rev. A 66, 033611 (2002).
[28] B. Luecke, M. Scherer, J. Kruse, L. Pezzé, F. Deuretzbacher, P. Hyllus, O. Topic, J. Peise, W. Ertmer, J. Arlt, L. Santos, A. Smerzi, and C. Klempt, Science 334, 773 (2011).
[29] C. Gross, H. Strobel, E. Nicklas, T. Zibold, N. Bar-Gill, G. Kurizki, and M. Oberthaler, Nature 480, 219 (2011).
[30] E. M. Bookjans, C. D. Hamley, and M. S. Chapman, Phys. Rev. Lett. 107, 210406 (2011).
[31] X.-Y. Luo, Y.-Q. Zou, L.-N. Wu, Q. Liu, M.-F. Han, M. K. Tey, and L. You, Science 355, 620 (2017).
[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, arXiv:1707.06347 (2017).
[33] M. Hausknecht and P. Stone, arXiv:1507.06527 (2015).
[34] T. G. Dietterich, Journal of Artificial Intelligence Research 13, 227 (2000).
[35] A. G. Barto and S. Mahadevan, Discrete event dynamic systems 13, 41 (2003).