Reinforcement Learning in Different Phases of Quantum Control

Marin Bukov,[1,*] Alexandre G. R. Day,[1,†] Dries Sels,[1,2] Phillip Weinberg,[1] Anatoli Polkovnikov,[1] and Pankaj Mehta[1]

[1]*Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA*

[2]*Theory of quantum and complex systems, Universiteit Antwerpen, B-2610 Antwerpen, Belgium*

(Received 12 January 2018; revised manuscript received 1 August 2018; published 27 September 2018)

The ability to prepare a physical system in a desired quantum state is central to many areas of physics such as nuclear magnetic resonance, cold atoms, and quantum computing. Yet, preparing states quickly and with high fidelity remains a formidable challenge. In this work, we implement cutting-edge reinforcement learning (RL) techniques and show that their performance is comparable to optimal control methods in the task of finding short, high-fidelity driving protocol from an initial to a target state in nonintegrable many-body quantum systems of interacting qubits. RL methods learn about the underlying physical system solely through a single scalar reward (the fidelity of the resulting state) calculated from numerical simulations of the physical system. We further show that quantum-state manipulation viewed as an optimization problem exhibits a spin-glass-like phase transition in the space of protocols as a function of the protocol duration. Our RL-aided approach helps identify variational protocols with nearly optimal fidelity, even in the glassy phase, where optimal state manipulation is exponentially hard. This study highlights the potential usefulness of RL for applications in out-of-equilibrium quantum physics.

DOI: 10.1103/PhysRevX.8.031086

Subject Areas: Condensed Matter Physics, Quantum Physics, Statistical Physics

I. INTRODUCTION

Reliable quantum-state manipulation is essential for many areas of physics ranging from nuclear-magnetic-resonance experiments [1] and cold atomic systems [2,3] to trapped ions [4–6], quantum optics [7], superconducting qubits [8], nitrogen-vacancy centers [9], and quantum computing [10]. However, finding optimal control sequences in such experimental platforms presents a formidable challenge due to our limited theoretical understanding of nonequilibrium quantum systems and the intrinsic complexity of simulating large quantum many-body systems.

For long protocol durations, adiabatic evolution can be used to robustly reach target quantum states, provided the change in the Hamiltonian is slow compared to the minimum energy gap. Unfortunately, this assumption is often violated in real-life applications. Typical experiments often have stringent constraints on control parameters, such as a maximum magnetic field strength or a maximal switching frequency. Moreover, decoherence phenomena impose insurmountable time constraints beyond which quantum information is lost irreversibly. For this reason, many experimentally relevant systems are, in practice, uncontrollable; i.e., there are no finite-duration protocols, which prepare the desired state with unit fidelity. In fact, in Anderson and many-body localized or periodically driven systems, which are naturally away from equilibrium, the adiabatic limit does not even exist [11,12]. This has motivated numerous approaches to quantum-state control [13–35]. Despite all advances, at present, surprisingly little is known about how to successfully load a nonintegrable interacting quantum system into a desired target state, especially in short times, or even when this is feasible in the first place [30,36–38].

In this paper, we adopt a radically different approach to this problem based on machine learning (ML) [39–45]. ML has recently been applied successfully to several problems in equilibrium condensed matter physics [46,47], turbulent dynamics [48,49], and experimental design [50,51], and here we demonstrate that reinforcement learning (RL) provides deep insights into nonequilibrium quantum dynamics [52–57]. Specifically, we use a modified version of the Watkins Q-learning algorithm [39] to teach a computer agent to find driving protocols which prepare a quantum system in a target state $|\psi_*\rangle$ starting from an initial state $|\psi_i\rangle$ by controlling a time-dependent field. A far-reaching consequence of our study is the existence of phase transitions in the quantum-control landscape of the generic many-body quantum-control problem. The glassy

*mbukov@bu.edu
†agrday@bu.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

nature of the prevalent phase implies that the optimal protocol is exponentially difficult to find. However, as we demonstrate, the optimal solution is unstable to local perturbations. Instead, we discover classes of RL-motivated stable suboptimal protocols [58], the performance of which rival that of the optimal solution. Analyzing these suboptimal protocols, we construct a variational theory, which demonstrates that the behavior of physical degrees of freedom (d.o.f.) (which scale exponentially with the system size L for ergodic models) in a nonintegrable many-body quantum-spin chain can be effectively described by only a few variables within the variational theory. We benchmark the RL results using stochastic descent (SD) and compare them to optimal control methods such as chopped random basis (CRAB) [30] and (for simplicity) first-order gradient ascent pulse engineering (GRAPE) [59] (without its quasi-Newton extensions [15,60,61]); see discussion in the Supplemental Material [62].

In stark contrast to most approaches to quantum optimal control, RL is a *model-free* feedback-based method which could allow for the discovery of controls even when accurate models of the system are unknown or the parameters in the model are uncertain. A potential advantage of RL over traditional derivative-based optimal control approaches is the fine balance between exploitation of already obtained knowledge and exploration in uncharted parts of the control landscape. Below the quantum speed limit [63], exploration becomes vital and offers an alternative to the prevalent paradigm of multistarting local gradient optimizers [64]. Unlike these methods, the RL agent progressively learns to build a model of the optimization landscape in such a way that the protocols it finds are stable to sampling noise. In this regard, RL-based approaches may be particularly well suited to work with experimental data and do not require explicit knowledge of local gradients of the control landscape [59,62]. This may offer a considerable advantage in controlling realistic systems where constructing a reliable effective model is infeasible, for example, due to disorder or dislocations.

To manipulate the quantum system, our computer agent constructs piecewise-constant protocols of duration T by choosing a drive protocol strength $h_x(t)$ at each time $t = j\delta t$, $j = \{0, 1, \ldots, T/\delta t\}$, with δt the time-step size. In order to make the agent learn, it is given a reward for every protocol it constructs—the fidelity $F_h(T) = |\langle\psi_*|\psi(T)\rangle|^2$ for being in the target state after time T following the protocol $h_x(t)$ under unitary Schrödinger evolution. The goal of the agent is to maximize the reward in a series of attempts. Deprived of any knowledge about the underlying physical model, the agent collects information about already tried protocols based on which it constructs new, improved protocols through a sophisticated biased sampling algorithm. In realistic applications, one does not have access to infinite control fields; for this reason, we restrict to fields $h_x(t) \in [-4, 4]$; see Fig. 1(b). For reasons relating to the simplicity and efficiency of the numerical simulations, throughout this work, we further restrict the RL algorithm to the family of bang-bang protocols [65]. An additional advantage of focusing on bang-bang protocols is that this allows us to interpret the control phase transitions we find using the language of statistical mechanics [67].

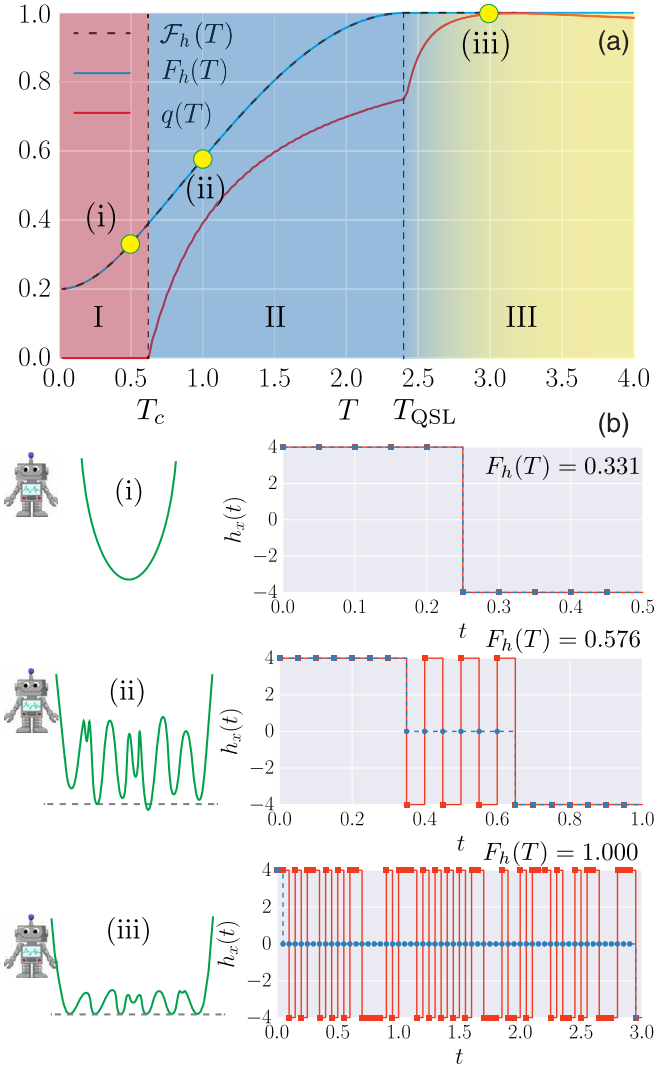

FIG. 1. (a) Phase diagram of the quantum-state manipulation problem for the qubit in Eq. (3) vs protocol duration T, as determined by the order parameter $q(T)$ (red) and the maximum possible achievable fidelity $F_h(T)$ (blue) compared to the variational fidelity $\mathcal{F}_h(T)$ (black, dashed). Increasing the total protocol time T, we go from an overconstrained phase I, through a glassy phase II, to a controllable phase III. (b) Left: The infidelity landscape is shown schematically (green). Right: The optimal bang-bang protocol found by the RL agent at the points (i)–(iii) (red) and the variational protocol [62] (blue, dashed).

II. REINFORCEMENT LEARNING

RL is a subfield of ML in which a computer agent learns to perform and master a specific task by exerting a series of actions in order to maximize a reward function as a result of

interaction with its environment. Here, we use a modified version of Watkins online, off-policy Q-learning algorithm with linear function approximation and eligibility traces [39] to teach our RL agent to find protocols of optimal fidelity. Let us we briefly summarize the details of the procedure. For a detailed description of the standard Q-learning algorithm, we refer the reader to Ref. [39].

The fidelity optimization problem is defined as an episodic, undiscounted reinforcement learning task. Each episode takes a fixed number of steps $N_T = T/\delta t$, where T is the total protocol duration, and δt the physical (protocol) time step. We define the state $\mathcal{S}$, action $\mathcal{A}$, and reward $\mathcal{R}$ spaces, respectively, as

$$\mathcal{S} = \{s = [t, h_x(t)]\}, \quad \mathcal{A} = \{a = \delta h_x\}, \quad \mathcal{R} = \{r \in [0, 1]\}.$$

The state space $\mathcal{S}$ consists of all tuples $[t, h_x(t)]$ of time t and the corresponding magnetic field $h_x(t)$. Notice that with this choice, no information about the physical quantum state whatsoever is encoded in the RL state, and, hence, the RL algorithm is model-free. Thus, the RL agent will be able to learn circumventing the difficulties associated with the theoretical notions in quantum physics. Including time t to the state is not common in Q-learning, but it is required here in order for the agent to be able to estimate how far away it is from the episode's end and adjust its actions accordingly. Even though there is only one control field, the space of available protocols grows exponentially with the inverse step size δt^{-1}.

The action space $\mathcal{A}$ consists of all jumps δh_x in the protocol $h_x(t)$. Thus, protocols are constructed as piecewise-constant functions. We restrict the available actions of the RL agent in every state s such that at all times the field $h_x(t)$ is in the interval $[-4, 4]$. We verify that RL also works for quasicontinuous protocols with many different steps δh_x [62]. The bang-bang protocols discussed in the next section and the quasicontinuous protocols used in the Supplemental Material [62] are examples of the family of protocol functions we allow in the simulation.

Last but not least, the reward space $\mathcal{R}$ is the space of all real numbers in the interval [0, 1]. The rewards for the agent are given only at the end of each episode according to

$$r(t) = \begin{cases} 0, & \text{if } t < T, \\ F_h(T) = |\langle \psi_* | \psi(T) \rangle|^2, & \text{if } t = T. \end{cases} \quad (1)$$

This definition reflects the fact that we are not interested in which quantum state the physical system is in during the evolution; all that matters for our purpose is to maximize the final fidelity.

An essential part of setting up the RL problem is to define the environment with which the agent interacts in order to learn. We choose the environment to consist of the Schrödinger initial value problem together with the target state

$$\text{Environment} = \{ i\partial_t |\psi(t)\rangle = H(t)|\psi(t)\rangle, \; |\psi(0)\rangle = |\psi_i\rangle, |\psi_*\rangle \},$$

where $H[h_x(t)]$ is the Hamiltonian (see Sec. III) whose time dependence is defined through the magnetic field $h_x(t)$ which the agent is constructing during the episode via online Q-learning updates for specific single-particle and many-body examples.

Let us now briefly illustrate the protocol construction algorithm: For instance, if we start in the initial RL state $s_0 = (t = 0, h_x = -4)$ and take the action $a = \delta h_x = 8$, we go to the next RL state $s_1 = (\delta t, +4)$. As a result of the interaction with the environment, the initial quantum state is evolved forward in time for one time step (from time $t_0 = 0$ to time $t_1 = \delta t$) with the constant Hamiltonian $H[h_x = 4]$: $|\psi(\delta t)\rangle = e^{-iH[h_x=4]\delta t}|\psi_i\rangle$. After each step, we compute the local reward according to Eq. (1) and update the Q function, even though the instantaneous reward at that step might be zero [the update will still be nontrivial in the later episodes, since information is propagated backwards from the end of the episode; see Eq. (2)]. This procedure is repeated until the end of the episode is reached at $t = T$. In general, one can imagine this partially observable Markov decision process as a state-action-reward chain

$$s_0 \to a_0 \to r_0 \to s_1 \to a_1 \to r_1 \to s_2 \to, \ldots, \to s_{N_T}.$$

The above paragraph explains how to choose actions according to some fixed policy $\pi(a|s)$—the probability of taking the action a from the state s. Some RL algorithms such as policy gradient directly optimize the policy. Instead, Watkins Q-learning offers an alternative which allows us to circumvent this. The central object in Q-learning is the $Q(s, a)$ function which is given by the expected total return $R = \sum_{i=0}^{N_T} r_i$ at the end of each episode, starting from a fixed state s, taking the fixed action a, and acting optimally afterwards. Clearly, if we have the optimal Q function Q_*, then the optimal policy is the deterministic policy $\pi_*(a|s) = 1$, if $a = \text{argmax}_{a'} Q(s, a')$, and $\pi_*(a|s) = 0$ for all other actions.

Hence, in Q-learning one looks directly for the optimal Q function. It satisfies the Bellman optimality equation, the solution of which cannot be obtained in a closed form for complex many-body systems [68]. The underlying reason for this can be traced back to the nonintegrability of the dynamical many-body system, as a result of which the solution of the Schrödinger equation cannot be written down as a closed-form expression even for a fixed protocol, and the situation is much more complicated when one starts optimizing over a family of protocols. The usual way of solving the Bellman equation numerically is temporal difference learning, which results in the following Q-learning update rule [39]

$$Q(s_i, a_i) \leftarrow Q(s_i, a_i) + \alpha[r_i + \max_a Q(s_{i+1}, a) - Q(s_i, a_i)], \tag{2}$$

where the learning rate $\alpha \in (0, 1)$. Whenever $\alpha \approx 1$, the convergence of the update rule (2) can be slowed down or even precluded in cases where the Bellman error $\delta_t = r_i + \max_a Q(s_{i+1}, a) - Q(s_i, a_i)$ becomes significant. On the contrary, $\alpha \approx 0$ corresponds to very slow learning. Thus, the optimal value for the learning rate lies in between, and it is determined empirically for the problem under consideration.

To allow for the efficient implementation of piecewise-constant drives, i.e., bang-bang protocols with a large number of bang modes (cf. Ref. [62]), we employ a linear function approximation to the Q function using equally spaced tilings along the entire range of $h_x(t) \in [-4, 4]$ [39]. The variational parameters of the linear approximator are found iteratively using gradient descent. This setup allows the RL agent to generalize, i.e., gain information about the fidelity of not yet encountered protocols.

We iterate the algorithm for 2×10^4 episodes. The exploration-exploitation dilemma [39] requires a fair amount of exploration in order to ensure that the agent visits large parts of the RL state space which prevents it from getting stuck in a local maximum of reward space from the beginning. Too much exploration and the agent will not be able to learn. On the other hand, no exploration whatsoever guarantees that the agent will repeat deterministically a given policy, though it will be unclear whether there exists a better yet unseen one. In the longer run, we cannot preclude the agent from ending up in a local maximum. In such cases, we run the algorithm multiple times starting from a random initial condition and postselect the outcome. Hence, the RL solution is almost optimal in the sense that its fidelity is close to the true global optimal fidelity. Unfortunately, the true optimal fidelity for nonintegrable many-body systems is unknown, and it is a definitive feature of glassy landscapes (see Sec. V) that the true optimal is exponentially hard and, therefore, also impractical to find [67].

We also verify that RL does not depend on the initial condition chosen, provided the change is small. For instance, if one chooses different initial and target states which are both paramagnetic, then RL works with marginal drops in fidelity, which depend parametrically on the deviation from the initial and target states. If, however, the target is, e.g., paramagnetic, and we choose an antiferromagnetic initial state (i.e., the initial and target states are chosen in two different phases of matter), then we observe a drop in the found fidelity.

Because of the extremely large state space, we employ a replay schedule to ensure that our RL algorithm can learn from the high-fidelity protocols it encounters. Our replay algorithm alternates between two different ways of training the RL agent, which we call training stages: an "exploratory" training stage where the RL agent exploits the current Q function to explore, and a "replay" training stage where we replay the best encountered protocol. This form of replay, to the best of our knowledge, has not been used previously. In the exploratory training stage, which lasts 40 episodes, the agent takes actions according to a softmax probability distribution based on the instantaneous values of the Q function. In other words, at each time step, the RL agent looks up the instantaneous values $Q(s, :)$ corresponding to all available actions and computes a probability for each action: $P(a) \sim \exp[\beta_{\mathrm{RL}} Q(s, a)]$. This exploration scheme results in random flips in the bangs of the protocol sequence, which is essentially a variation on the instantaneous RL best solution. Figure 2 shows that some of these variations lead to drastic reduction in fidelity, which we relate to the glassy character of the correlated control phase; see Sec. V.

The amount of exploration is set by β_{RL} with $\beta_{\mathrm{RL}} = 0$ corresponding to random actions and $\beta_{\mathrm{RL}} = \infty$ corresponding to always taking greedy actions with respect to the

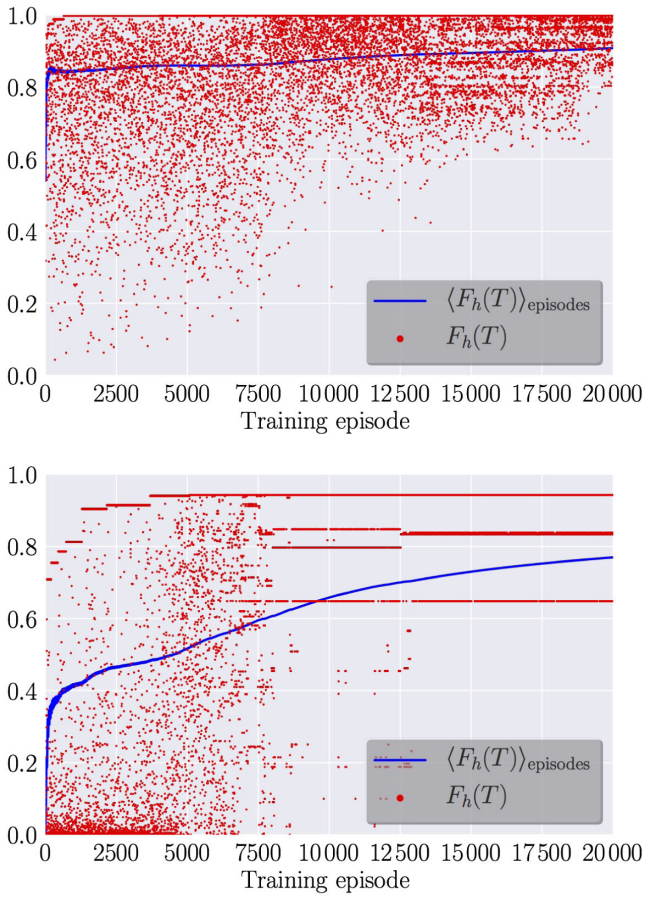

FIG. 2. Learning curves of the RL agent for the problems from Sec. III for $L = 1$ at $T = 2.4$ (upper panel) (see Video 7 of Supplemental Material [62]) and $L = 10$ at $T = 3.0$ (lower panel) (see Video 8 of Supplemental Material [62]). The red dots show the instantaneous reward (i.e., fidelity) at every episode, while the blue line shows the cumulative episode average. The ramp-up of the RL temperature β_{RL} gradually suppresses exploration over time which leads to a smoothly increasing average fidelity. The time step is $\delta t = 0.05$.

current estimate of the Q function. Here, we use an external "learning" temperature scale, the inverse of which, β_{RL}, is linearly ramped down as the number of episodes progresses. In the replay training stage, which is also 40 episodes long, we replay the best-encountered protocol up to the given episode. Through this procedure, when the next exploratory training stage begins again, the agent is biased to do variations on top of the best-encountered protocol, effectively improving it, until it reaches a reasonably good fidelity.

Two learning curves of the RL agent are shown in Fig. 2. Notice the occurrence of suboptimal protocols even during later episodes due to the stochasticity of the exploration schedule. During every episode, the agent takes the best action (with respect to its current knowledge or experience) with a finite probability or else a random action is chosen. This prevents the algorithm from immediately getting stuck in a high-infidelity (i.e., a bad) minimum. To guarantee convergence of the RL algorithm, the exploration probability is reduced as the number of episodes progresses (cf. discussion above). Convergence of the RL algorithm becomes manifest in Fig. 2, where after many episodes the deviations from the good protocols decrease. In the end, the agent learns the best-encountered protocol as a result of using the replay schedule which speeds up learning (as can be seen by the bad shots becoming rarer with increasing the number
of episodes). We show only these learned protocols in Fig. 1(b) and Fig. 3 of the Supplemental Material [62].

III. PHASES OF QUANTUM CONTROL

A. Single-qubit manipulation

To benchmark the application of RL to physics problems, consider first a two-level system described by

$$H[h_x(t)] = -S^z - h_x(t)S^x, \tag{3}$$

where S^α are the spin-1/2 operators. This Hamiltonian comprises both integrable many-body and noninteracting translational invariant systems, such as the transverse-field Ising model and graphene and topological insulators. The initial $|\psi_i\rangle$ and target $|\psi_*\rangle$ states are chosen as the ground states of Eq. (3) at $h_x = -2$ and $h_x = 2$, respectively. We verify that the applicability of RL does not depend on this specific choice. Although there exists an analytical solution to solve for the optimal protocol in this case [63], it does not generalize to nonintegrable many-body systems. Thus, studying this problem using RL serves a twofold purpose: (i) We benchmark the protocols obtained by the RL agent demonstrating that even though RL is a completely model-free algorithm, it still finds the physically meaningful solutions by constructing a minimalistic effective model on the fly. The learning process is shown in Video 7 of Supplemental Material [62]. (ii) We reveal an important novel perspective on the complexity of quantum-state manipulation which, as we show below, generalizes to many-particle systems. While experimental setups studying single-qubit physics can readily apply multiple control fields (e.g., also control fields in the y direction), in order to test RL on a nontrivial problem with a known solution, we restrict the discussion to a single control parameter.

For fixed total protocol duration T, the infidelity $h_x(t) \mapsto I_h(T) = 1 - F_h(T)$ represents a "potential landscape," the global minimum of which corresponds to the optimal driving protocol. For bang-bang protocols, the problem of finding the optimal protocol becomes equivalent to finding the ground-state configuration of a classical Ising model with complicated interactions [67]. We map out the landscape of local infidelity minima $\{h_x^\alpha(t)\}_{\alpha=1}^{N_{\mathrm{real}}}$ using SD starting from random bang-bang protocol configurations [62]. To study the correlations between the infidelity minima as a function of the total protocol duration T, we define the correlator $q(T)$ closely related to the Edwards-Anderson order parameter for the existence of spin-glass order [69,70] as

$$q(T) = \frac{1}{16N_T}\sum_{j=1}^{N_T}\overline{\{h_x(j\delta t) - \overline{h_x}(j\delta t)\}^2}, \tag{4}$$

where $\overline{h_x}(t) = N_{\mathrm{real}}^{-1}\sum_{\alpha=1}^{N_{\mathrm{real}}} h_x^\alpha(t)$ is the sample-averaged protocol. If the minima $\{h_x^\alpha(t)\}_{\alpha=1}^{N_{\mathrm{real}}}$ are all uncorrelated, then $\overline{h_x}(t) \equiv 0$ and, thus, $q(T) = 1$. On the other hand, if the infidelity landscape contains only one minimum, then $\overline{h_x}(t) \equiv h_x(t)$ and $q(T) = 0$. The behavior of $q(T)$ and the maximum fidelity $F_h(T)$ found using SD together with a qualitative description of the corresponding infidelity landscapes are shown in Fig. 1.

The control problem for the constrained qubit exhibits three distinct control phases as a function of the protocol duration T. If T is greater than the quantum speed limit $T_{\mathrm{QSL}} \approx 2.4$, one can construct infinitely many protocols which prepare the target state with unit fidelity, and the problem is in the *controllable* phase III; cf. Fig. 1. The red line in Fig. 1(b) (iii) shows an optimal protocol of unit fidelity found by the agent whose Bloch sphere representation can be seen in Video 3 of Supplemental Material [62]. In this phase, there is a proliferation of *exactly degenerate*, uncorrelated global infidelity minima corresponding to protocols of unit fidelity, and the optimization task is easy.

At $T = T_{\mathrm{QSL}}$, the order parameter $q(T)$ exhibits a nonanalyticity, and the system undergoes a continuous phase transition to a correlated phase II. For times smaller than T_{QSL} but greater than T_c, the degenerate minima of the infidelity landscape recede to form a correlated landscape with many *nondegenerate* local minima, as reflected by the finite value of the order parameter $0 < q(T) < 1$. As a consequence of this correlated phase, there no longer exists a protocol to prepare the target state with unit fidelity, since it is physically impossible to reach the target state while obeying all constraints. The infidelity minimization problem is nonconvex, and determining the best achievable (i.e., optimal) fidelity

(aka the global minimum) becomes difficult. Figure 1(b) (ii) shows the best bang-bang protocol found by our computer agent [see Video 2 of Supplemental Material [62] and Ref. [62] for protocols with quasicontinuous actions]. This protocol has a remarkable feature: Without any prior knowledge about the intermediate quantum state or its Bloch sphere representation, the model-free RL agent discovers that it is advantageous to first bring the state to the equator—which is a geodesic—and then effectively turns off the control field $h_x(t)$ to enable the fastest possible precession about the z axis [71]. After staying on the equator for as long as optimal, the agent rotates as fast as it can to bring the state as close as possible to the target, thus, optimizing the final fidelity for the available protocol duration.

Decreasing the total protocol duration T further, we find a second critical time $T_c \approx 0.6$. For $T < T_c$, $q(T) \equiv 0$, and the problem has a unique solution, suggesting that the infidelity landscape is convex. This *overconstrained phase* is labeled I in the phase diagram [Fig. 1(a)]. For $T < T_c$, there exists a unique optimal protocol, even though the achievable fidelity can be quite limited; see Fig. 1(b) and Video 1 of Supplemental Material [62]. Since the state precession speed towards the equator depends on the maximum possible allowed field strength h_x, it follows that $T_c \to 0$ for $|h_x| \to \infty$.

1. Relation to counterdiabatic and fast-forward driving

Promising analytical approaches to state manipulation have recently been proposed, known as shortcuts to adiabaticity [21,22,27,29,33,72–76]. They include ideas such as (i) fast-forward (FF) driving, which comprises a protocol that excites the system during the evolution at the expense of gaining speed before taking away all excitations and reaching the target state with unit probability, and (ii) counterdiabatic (CD) driving, which ensures transitionless dynamics by turning on additional control fields. In general, any FF protocol is related to a corresponding CD protocol. While for complex many-body systems it is not possible to construct the mapping between FF and CD, in general, the simplicity of the single-qubit setup (3) allows us to use CD driving to find a FF protocol [34]. For an unbounded control field $h_x(t)$, the FF protocol at the quantum speed limit has three parts which can be understood intuitively on the Bloch sphere: (i) an instantaneous δ-function kick to bring the state to the equator, (ii) an intermediate stage where the control field is off, $h_x(t) \equiv 0$, which allows the state to precess along the equator, and (iii) a complementary δ kick to bring the state from the equator straight to the target [34]. Whenever the control field is bounded $|h_x| \leq 4$, these δ kicks are broadened and take extra time, thus, increasing T_{QSL}. If the RL algorithm finds a unit-fidelity protocol, it is by definition a FF one. Comparing FF driving to the protocol found by our RL agent [cf. Fig. 1(b); see, also, paragraphs above], we find indeed a remarkable similarity between the RL and FF protocols.

B. Many-coupled qubits

The above results raise the natural question of how much more difficult state manipulation is in more complex quantum models. To this end, consider a closed chain of L-coupled qubits, which can be experimentally realized with superconducting qubits [8], cold atoms [77], and trapped ions [6]:

$$H[h_x(t)] = -\sum_{j=1}^{L} [S^z_{j+1} S^z_j + g S^z_j + h_x(t) S^x_j]. \qquad (5)$$

We set $g = 1$ to avoid the antiferromagnet-to-paramagnet phase transition and choose the paramagnetic ground states of Eq. (5) at fields $h_x = -2$ and $h_x = 2$ for the initial and target states, respectively. We verify that the conclusions we draw below do not depend on the choice of initial and target states, provided they both belong to the paramagnetic phase. The details of the control field $h_x(t)$ are the same as in the single-qubit case, and we use the *many-body* fidelity both as the reward and the measure of performance. In this paper, we focus on $L > 2$. The two-qubit optimization problem is shown to exhibit an additional symmetry-broken correlated phase; see Ref. [78].

Figure 3 shows the phase diagram of the coupled-qubits model. First, notice that while the overconstrained-to-glassy critical point T_c survives, the quantum-speed-limit critical point T_{QSL} is (if existent at all) outside the short protocol time range of interest. Thus, the glassy phase extends over to long and probably infinite protocol durations, which offers an alternative explanation for the difficulty of preparing many-body states with high fidelity. The glassy properties of this phase are analyzed extensively in Ref. [67]. Second,

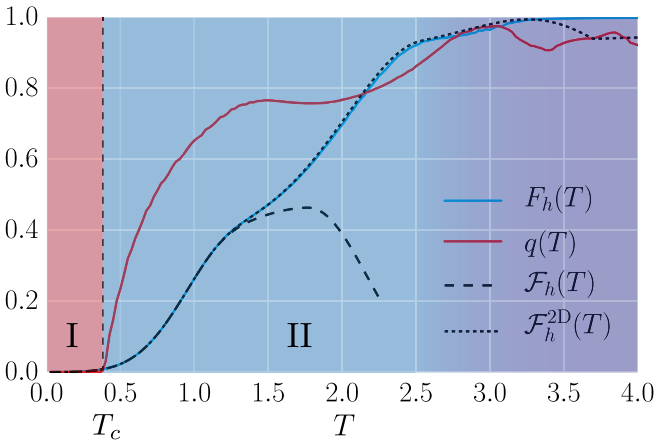

FIG. 3. Phase diagram of the many-body quantum-state manipulation problem. The order parameter (red) shows a kink at the critical time $T_c \approx 0.4$ when a phase transition occurs from an overconstrained phase (I) to a glassy phase (II). The best fidelity $F_h(T)$ (blue) obtained using SD is compared to the variational fidelity $\mathcal{F}_h(T)$ (dashed) and the 2D-variational fidelity $\mathcal{F}_h^{2D}(T)$ (dotted) [62].

observe that even though unit fidelity is no longer achievable, there exist nearly optimal protocols with extremely high many-body fidelity [79] at short protocol durations. This fact is striking because the Hilbert space of our system grows *exponentially* with L, and we use only one control field to manipulate exponentially many degrees of freedom in a short time. Nonetheless, it has been demonstrated that two states very close to each other in or of equal fidelity can possess sufficiently different physical properties or be very far in terms of physical resources [78,80–82]. Hence, one should be cautious when using the fidelity as a measure for preparing many-body states, and exploring other possible reward functions for training RL agents is an interesting avenue for future research.

Another remarkable characteristic of the optimal solution is that for the system sizes $L \geq 6$ both $q(T)$ and $-L^{-1}\log F_h(T)$ converge to their thermodynamic limit values with no visible finite-size corrections [62]. This is likely related to the Lieb-Robinson bound for information propagation which suggests that information should spread over approximately $JT = 4$ sites for the longest protocol durations considered.

IV. VARIATIONAL THEORY FOR NEARLY OPTIMAL PROTOCOLS

An additional feature of the optimal bang-bang solution found by the agent is that the entanglement entropy of the half-system generated during the evolution always remains small, satisfying an area law [62]. This implies that the system likely follows the ground state of some local yet *a priori* unknown effective Hamiltonian [83]. This emergent behavior motivates us to use the best protocols found by ML to construct simple variational protocols consisting of just a few bangs. Let us now demonstrate how to construct variational theories by giving specific examples which, to our surprise, capture the essence of the phase diagram of quantum control both qualitatively and quantitatively.

A. Single qubit

By carefully studying the optimal driving protocols the RL agent finds in the case of the single qubit, we find a few important features. Focusing for the moment on bang-bang protocols, in the overconstrained and correlated phases (cf. Fig. 1(b) and Videos 1–3 of Supplemental Material [62]), we recognize an interesting pattern: For $T < T_c$, as we explain in Sec. III, there is only one minimum in the infidelity landscape, which dictates a particularly simple form for the bang-bang protocol—a single jump at half the total protocol duration $T/2$. On the other hand, for $T_c \leq T \leq T_{\mathrm{QSL}}$, there appears a sequence of multiple bangs around $T/2$, which grows with increasing the protocol duration T. By looking at the Bloch sphere representation (see Videos 1–3 of Supplemental Material [62]), we identify this protocol structure as an attempt to turn off the h_x field once the state is rotated to the equator. This trick allows for the instantaneous state to be moved in the direction of the target state in the shortest possible distance (i.e., along a geodesic).

Hence, it is suggestive to try out a three-pulse protocol as an ansatz for the optimal solution [see Fig. 4(a)]: The first (positive) pulse of duration $\tau^{(1)}/2$ brings the state to the equator. Then, the h_x field is turned off for a time $\tilde{\tau}^{(1)} = T - \tau^{(1)}$, after which a negative pulse directs the state off the equator towards the target state. Since the initial value problem is time-reversal symmetric for our choice of initial and target states, the duration of the third pulse must be the same as that of the first one. We thus arrive at a variational protocol parametrized by $\tau^{(1)}$; see Fig. 4(a).

The optimal fidelity is thus approximated by the variational fidelity $\mathcal{F}_h(\tau^{(1)}, T-\tau^{(1)})$ for the trial protocol [Fig. 4(a)] and can be evaluated analytically in a straightforward manner:

$$\mathcal{F}_h(\tau^{(1)}, T-\tau^{(1)}) = |\langle\psi_*|e^{-i\frac{\tau^{(1)}}{2}H[-h_{\max}]}e^{-i(T-\tau^{(1)})H[0]}e^{-i\frac{\tau^{(1)}}{2}H[h_{\max}]}|\psi_i\rangle|^2,$$
$$H[h_x] = -S^z - h_x S^x. \tag{6}$$

However, since the exact expression is rather cumbersome, we choose not to show it explicitly. Optimizing the variational fidelity at a fixed protocol duration T, we solve the corresponding transcendental equation to find the extremal value $\tau^{(1)}_{\mathrm{best}}$ and the corresponding optimal variational fidelity $\mathcal{F}_h(T)$ shown in Figs. 4(b) and 4(c). For times $T \leq T_c$, we find $\tau^{(1)} = T$ which corresponds to $\tilde{\tau}^{(1)} = 0$, i.e., a single bang in the optimal protocol. The overconstrained-to-correlated phase transition at T_c is marked by a nonanalyticity at $\tau^{(1)}_{\mathrm{best}}(T_c) = T_c \approx 0.618$. This is precisely the minimal time the agent can take to bring the state to the equator of the Bloch sphere, and it depends on the value of the maximum magnetic field allowed (here, $h_{\max} = 4$). Figure 4(d) shows that in the overconstrained phase, the fidelity is optimized at the boundary of the variational domain, although $\mathcal{F}_h(\tau^{(1)}, T-\tau^{(1)})$ is a highly nonlinear function of $\tau^{(1)}$ and T.

For $T_c \leq T \leq T_{\mathrm{QSL}}$, the time $\tau^{(1)}$ is kept fixed (the equator being the only geodesic for a rotation along the $\hat{z}$ axis of the Bloch sphere), while the second pulse time $\tilde{\tau}^{(1)}$ grows linearly until the minimum time $T_{\mathrm{QSL}} \approx 2.415$ is eventually reached. The minimum time is characterized by a bifurcation in our effective variational theory, as the corresponding variational infidelity landscape develops

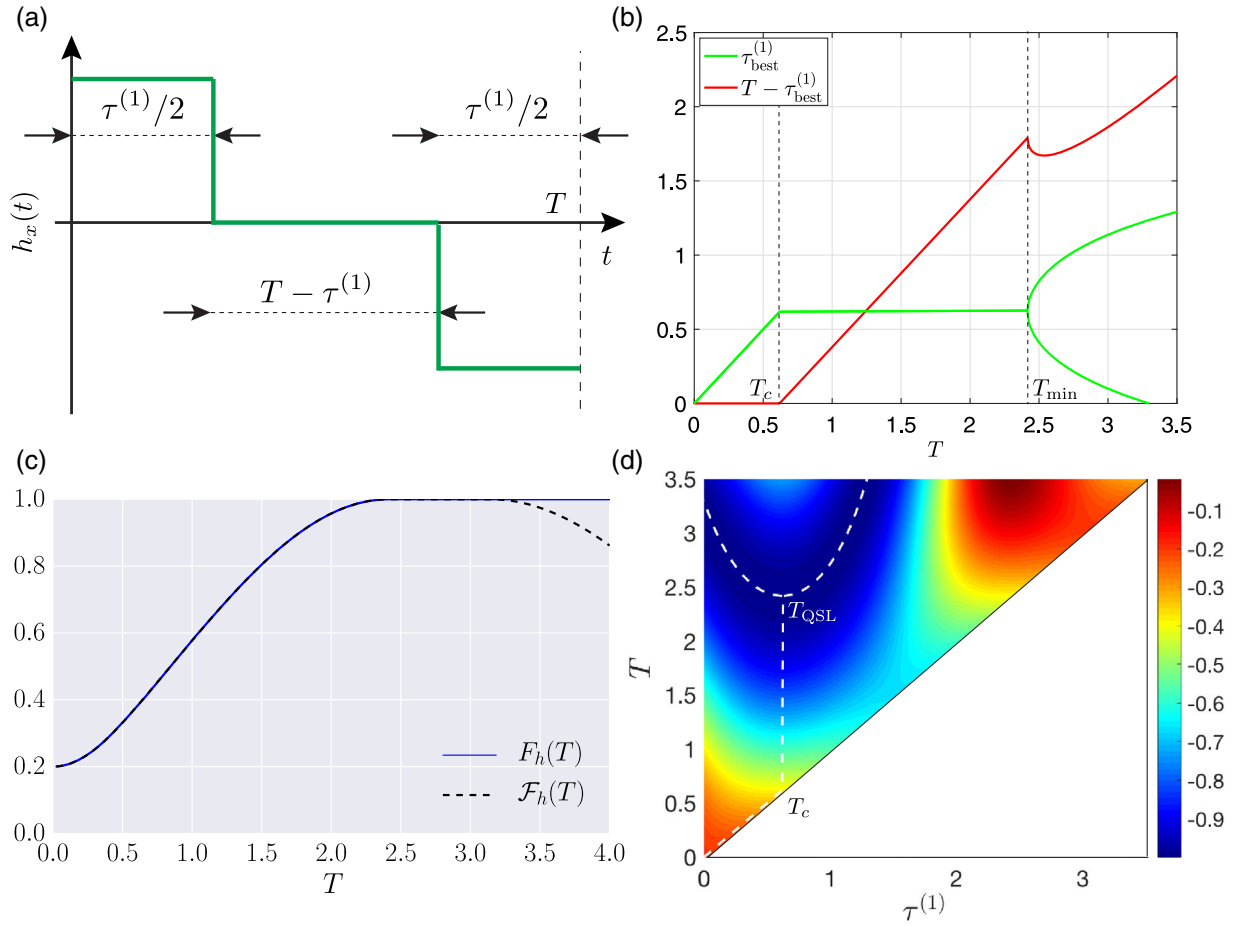

FIG. 4. (a) Three-pulse variational protocol which allows us to capture the optimal protocol found by the computer in the overconstrained and the glassy phases of the single-qubit problem. (b) $\tau^{(1)}_{\rm best}$ (green) with the nonanalytic points of the curve marked by dashed vertical lines corresponding to $T_c \approx 0.618$ and $T_{\rm QSL} \approx 2.415$. (c) Best fidelity obtained using SD (solid blue) and the variational ansatz (dashed black). (d) The variational infidelity landscape with the minimum for each T slice designated by the dashed line which shows the robustness of the variational ansatz against small perturbations.

two minima; see Figs. 4(b) and 4(d). Past that protocol duration, our simplified ansatz is no longer valid, and the system is in the controllable phase. Furthermore, a sophisticated analytical argument based on optimal control theory can give exact expressions for T_c and $T_{\rm QSL}$ [63], in precise agreement with the values we obtain. The Bloch sphere representation of the variational protocols in Fig. 1(b) (dashed blue lines) for the single qubit are shown in Videos 4–6 of Supplemental Material [62].

To summarize, for the single-qubit example, the variational fidelity $\mathcal{F}_h(T)$ agrees nearly perfectly with the optimal fidelity $F_h(T)$ obtained using SD and optimal control; cf. Fig. 1(a). We further demonstrate that our variational theory fully captures the physics of the two critical points T_c and $T_{\rm QSL}$ [62]. Interestingly, the variational solution for the single-qubit problem coincides with the global minimum of the infidelity landscape all the way up to the quantum speed limit [62].

B. Many-coupled qubits

Let us also discuss the variational theory for the many-body system. Consider first the same one-parameter variational ansatz from Sec. IV A; see Fig. 5(a). Since the variational family is one dimensional, we refer to this ansatz as the 1D-variational theory. The dashed black line in Fig. 5(c) shows the corresponding 1D-variational fidelity. We see that once again this ansatz captures correctly the critical point T_c separating the overconstrained and the glassy phases. Nevertheless, a comparison with the optimal fidelity [see Fig. 5(c)] reveals that this variational ansatz breaks down in the glassy phase, although it rapidly converges to the optimal fidelity with decreasing T. Looking at Fig. 5(b), we note that the value $\tau^{(1)}_{\rm best}$, which maximizes the variational fidelity, exhibits a few kinks. However, only the kink at $T = T_c$ captures a physical transition of the original control problem, while the others appear as artifacts of the simplified variational theory, as can be seen by the regions of agreement between the optimal and variational fidelities.

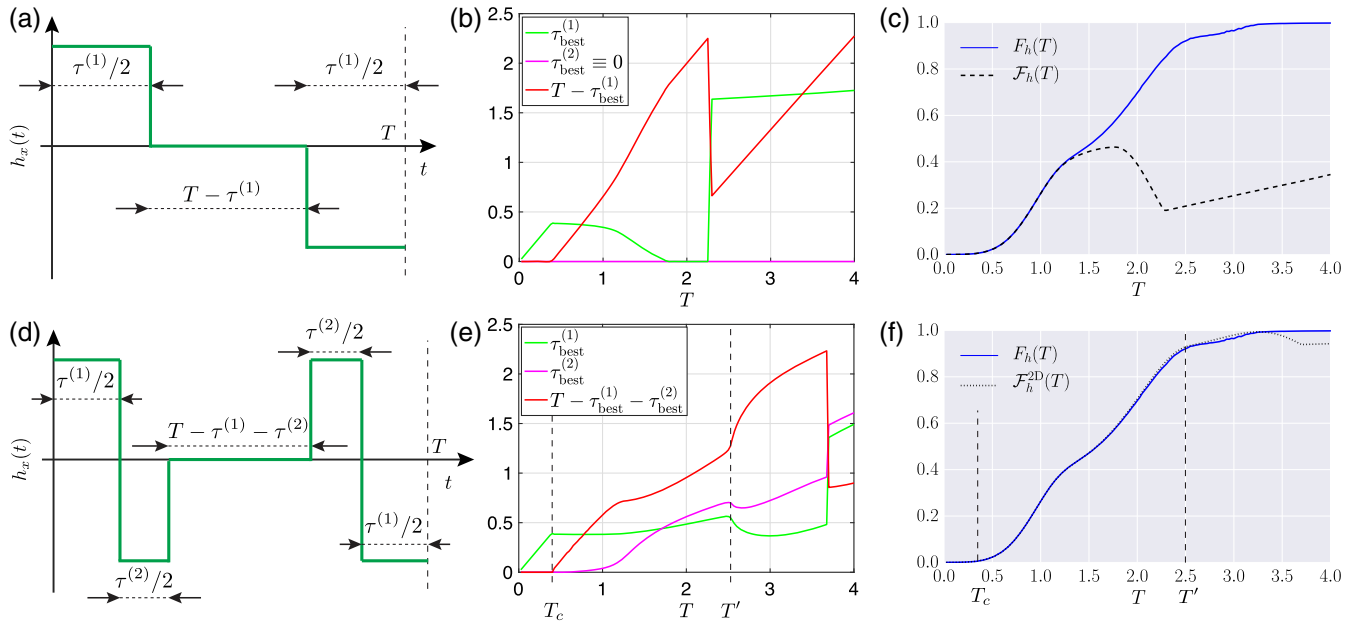

FIG. 5. (a) Three-pulse variational protocol which allows us to capture the optimal protocol found by the computer in the overconstrained phase but fails the glassy phase of the nonintegrable many-body problem. This ansatz captures the nonanalytic point at $T_c \approx 0.4$ but fails in the glassy phase. (b) The pulse durations $\tau^{(1)}_{\rm best}$ (green) and $\tau^{(2)}_{\rm best}$ (magenta) for highest-fidelity variational protocol of length T of the type shown in (a). The fidelity of the variational protocols exhibits a physical nonanalyticity at $T_c \approx 0.4$ and unphysical kinks outside the validity of the ansatz. (c) 1D maximal variational fidelity (dashed back) compared to the best numerical protocol (solid blue). (d) Five-pulse variational protocol which allows us to capture the optimal protocol found by the computer in the overconstrained phase and parts of the glassy phase of the nonintegrable many-body problem. (e) The pulse durations $\tau^{(1)}_{\rm best}$ (green) and $\tau^{(2)}_{\rm best}$ (magenta) for the best variational protocol of length T of the type shown in (d). These variational protocols exhibit physical nonanalyticities at $T_c \approx 0.4$ and $T' \approx 2.5$ (vertical dashed lines) (f) 2D maximal variational fidelity (dashed-dotted back) compared to the best numerical protocol (solid blue).

Inspired by the structure of the protocols found by our RL agent once again (see Video 8 of Supplemental Material [62]), we now extend the qubit variational protocol, as shown in Fig. 5(d). In particular, we add two more pulses to the protocol, retaining its symmetry structure: $h_x(t) = -h_x(T-t)$, whose length is parametrized by a second independent variational parameter $\tau^{(2)}/2$. Thus, the pulse length where the field is set to vanish is now given by $\tilde{\tau} = T - \tau^{(1)} - \tau^{(2)}$. These pulses are reminiscent of spin-echo protocols and appear to be important for entangling and disentangling the state during the evolution. Notice that this extended variational ansatz includes by definition the simpler ansatz from the single-qubit problem discussed above by setting $\tau^{(2)} = 0$.

Let us now turn on the second variational parameter $\tau^{(2)}$ and consider the full two-dimensional variational problem:

$$\mathcal{F}^{\rm 2D}_h(\tau^{(1)},\tau^{(2)},T-\tau^{(1)}-\tau^{(2)}) = |\langle\psi_*|e^{-i\frac{\tau^{(1)}}{2}H[-h_{\rm max}]}e^{-i\frac{\tau^{(2)}}{2}H[h_{\rm max}]}e^{-i(T-\tau^{(1)})H[0]}e^{-i\frac{\tau^{(2)}}{2}H[-h_{\rm max}]}e^{-i\frac{\tau^{(1)}}{2}H[h_{\rm max}]}|\psi_i\rangle|^2,$$
$$H[h_x] = -\sum_{j=1}^{L}(S^z_{j+1}S^z_j + gS^z_j + h_xS^x_j). \tag{7}$$

For the maximum-fidelity variational protocol, we show the best variational fidelity $\mathcal{F}^{\rm 2D}_h$ [Fig. 5(f)] and the corresponding values of $\tau^{(1)}_{\rm best}$ and $\tau^{(2)}_{\rm best}$ [Fig. 5(e)]. There are two important points here: (i) Fig. 5(f) shows that the 2D-variational fidelity seemingly reproduces the optimal fidelity on a much longer scale compared to the 1D-variational ansatz, i.e., for all protocol durations $T \lesssim 3.3$. (ii) The 2D-variational ansatz reduces to the 1D one in the overconstrained phase $T \leq T_c$. In particular, both pulse lengths $\tau^{(1)}_{\rm best}$ and $\tau^{(2)}_{\rm best}$ exhibit a nonanalyticity at $T = T_c$ but also at $T' \approx 2.5$. Interestingly, the 2D-variational ansatz captures the optimal fidelity on both sides of T', which suggests that there

is likely yet another transition within the glassy phase, hence, the different shading in the many-body phase diagram (Fig. 3). Similar to the 1D-variational problem, here we also find artifact transitions (nonanalytic behavior in $\tau^{(i)}_{\max}$ outside of the validity of the variational approximation).

In summary, in the many-body case, the same one-parameter variational ansatz describes only the behavior in the overconstrained phase [cf. Fig. 3 (dashed line)] up to and including the critical point T_c but fails for $T > T_c$. Nevertheless, a slightly modified two-parameter variational ansatz motivated again by the solutions found by the ML agent (see Video 8 of Supplemental Material [62]) appears to be fully sufficient to capture the essential features of the optimal protocol much deeper into the glassy phase, as shown by the $\mathcal{F}^{\mathrm{2D}}_h(T)$ curve in Fig. 3. This many-body variational theory features an additional pulse reminiscent of spin echo, which appears to control and suppress the generation of entanglement entropy during the drive [62]. Indeed, while the two-parameter ansatz is strictly better than the single-parameter protocol for all $T > T_c$, the difference between the two grows slowly as a function of time. It is only at a later time, $T \approx 1.3$, that the effect of the second pulse really kicks in, and we observe the largest entanglement in the system for the optimal protocol.

Using RL, we identify nearly optimal control protocols [58] which can be parametrized by a few d.o.f. Such simple protocols have been proven to exist in weakly entangled one-dimensional spin chains [38]. However, the proof of the existence does not imply that these d.o.f. are easy to identify. Initially, the RL agent is completely ignorant about the problem and explores many different protocols while it tries to learn the relevant features. In contrast, optimal control methods, such as CRAB [30], usually have a much more rigid framework, where the d.o.f. of the method are fixed from the beginning. This restriction can limit the performance of those methods below the quantum speed limit [62,64].

One might wonder how the nearly optimal protocols found using RL and SD correlate with the best variational protocols. For the problem under consideration, averaging parts of the set of bang-bang protocols, which contain randomly generated local minima of the infidelity landscape $\{h^{\alpha}_x(t)\}^{N_{\mathrm{real}}}_{\alpha=1}$ (see insets in Fig. 7) results in protocols which resemble the continuous ones we find using GRAPE. The variational solutions are indeed close to these averaged solutions, although they are not exactly the same, since the variational protocols are constrained to take on three discrete values (positive, zero, and negative), while the averaged protocols can take on any values in the interval $[-4, 4]$. The RL agent cannot find these variational solutions because we limit the actions space to having h_x take the minimum or maximum allowable value, and there is no way to take an action where $h_x = 0$.

We also show how by carefully studying the driving protocols found by the RL agent, one can obtain ideas for effective theories which capture the essence of the underlying physics. This approach is similar to using an effective ϕ^4 theory to describe the physics of the Ising phase transition. The key difference is that the present problem is out of equilibrium, where no general theory of statistical mechanics exists so far. We hope that an underlying pattern between such effective theories can be revealed with time, which might help shape the guiding principles of a theory of statistical physics away from equilibrium.

V. GLASSY BEHAVIOR

It is quite surprising that the dynamics of a nonintegrable many-body quantum system associated with the optimal protocol is so efficiently captured by such a simple two-parameter variational protocol, even in the regimes where there is no obvious small parameter and where spin-spin interactions play a significant role. Upon closer comparison of the variational and the optimal fidelities, one can find regions in the glassy phase where the simple variational protocol outperforms the numerical "best" fidelity; cf. Fig. 3.

To better understand this behavior, we choose a grid of $N_T = 28$ equally spaced time steps and compute all 2^{28} bang-bang protocols and their fidelities. The corresponding density of states (DOS) in fidelity space is shown in Fig. 6 for two choices of T in the overconstrained and glassy phase. This approach allows us to unambiguously determine the ground state of the infidelity landscape (i.e., the optimal protocol). Starting from this ground state, we then construct all excitations generated by local in time flips of the bangs of the optimal protocol. The fidelity of the "one-flip" excitations is shown using red circles in Fig. 6. Notice how in the glassy phase these 28 excitations have relatively low fidelities compared to the ground state and are surrounded by approximately 10^6 other states. This result has profound consequences: As we are "cooling" down in the glassy phase, searching for the optimal protocol and coming from a state high up in the infidelity landscape, if

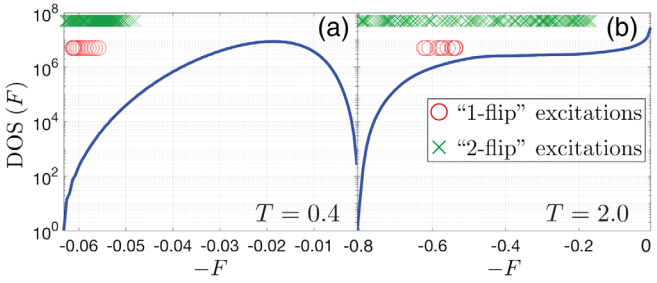

FIG. 6. Density of states (protocols) in the overconstrained phase at $T = 0.4$ (a) and the glassy phase at $T = 2.0$ (b) as a function of the fidelity F. The red circles and the green crosses show the fidelity of the "one-spin" flip and "two-spin" flip excitation protocols above the absolute ground state (i.e., the optimal protocol). The system size is $L = 6$ and each protocol has $N_T = 28$ bangs.

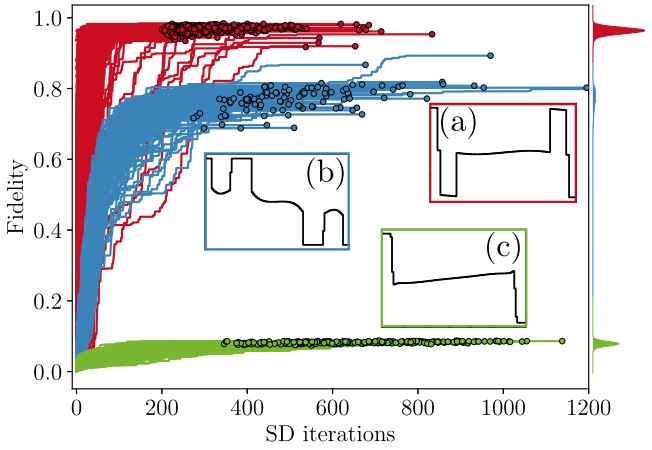

FIG. 7. Fidelity traces of SD for $T = 3.2$, $L = 6$, and $N_T = 200$ as a function of the number of iterations of the algorithm for 10^3 random initial conditions. The traces are characterized by three main attractors marked by the different colors. The termination of each SD run is indicated by a colored circle. The relative population of the different attractors is shown as a density profile on the right-hand side. Insets (a)–(c): Averaged profile of the protocols obtained for the red, blue, and green attractor, respectively.

we miss one of the 28 elementary excitations, it becomes virtually impossible to reach the global ground state, and the situation becomes much worse if we increase the number of steps N_T. On the contrary, in the overconstrained phase, the smaller value of the DOS at the one-flip excitation (approximately 10^2) makes it easier to reach the ground state.

The green crosses in Fig. 6 show the fidelity of the "two-flip" excitations. By the above argument, a two-flip algorithm will not see the phase as a glass for $T \lesssim 2.5$, yet it does so for $T \gtrsim 2.5$ marked by the different shading in Fig. 3. Correlated with this observation, we find a signature of a transition also in the improved two-parameter variational theory in the glassy phase [see Sec. IV B and kinks at T' in Fig. 5(e)]. In general, we expect the glassy phase to exhibit a series of phase transitions reminiscent of the random k-satisfiability (k-SAT) problems [84,85]. The glassy nature of this correlated phase has been studied in detail in Ref. [67] by mapping this optimal control problem to an effective classical spin-energy function which governs the control phase transitions.

In contrast to the single-qubit system, there are also multiple attractors present in the glassy phase of the many-body system (see Fig. 7). Each attractor has a typical representative protocol (Fig. 7 insets). Even though intra-attractor protocols share the same averaged profile, they can nevertheless have a small mutual overlap comparable to the overlap of interattractor protocols. This indicates that in order to move in between protocols within an attractor, highly nonlocal moves are necessary. For this reason, GRAPE [59], an algorithm which performs global updates on the protocol by computing exact gradients in the control landscape, also performs very well on our optimization problem. Similar to SD, in the glassy phase GRAPE cannot escape local minima in the infidelity landscape, and, therefore, the same three attractors are found with comparable relative populations to SD, but intra-attractor fluctuations are significantly suppressed due to GRAPE's nonlocal character.

VI. OUTLOOK AND DISCUSSION

In this work, we demonstrate the usefulness of Q-learning to manipulate single-particle and many-body quantum systems. Q-learning is only one of many reinforcement learning algorithms, including the SARSA, policy gradient, and actor critic methods, just to name a few. In the Supplemental Material, we show that Q-learning's performance is comparable to many of the leading optimal control algorithms [62]. It is interesting and desirable to compare different RL algorithms among themselves on physical quantum systems. An exciting future direction is to investigate which advantages deep learning offers in the context of quantum control, and there exist recent studies exploring deep RL in a physics [53,54,56].

Looking forward to controlling nonintegrable many-body systems, an important question arises as to how the computational cost of Q-learning scales with the system size L. As we explain in Sec. II, the Q-learning algorithm can be decomposed into a learning part and an "interaction with the environment" part where all physics or dynamics happens. The learning part does not know about the state of the quantum system; it keeps track of only the value of the magnetic field at a given time $[t, h_x(t)]$. As a result of this choice, for a single global drive, the learning part of the algorithm is independent of the system size L since it depends only on a single scalar reward: the fidelity of the final state. The RL algorithm is instead computationally limited by the size of the action and state spaces. As currently implemented, this means that the RL algorithm is limited to finding short protocols (since the state space scales exponentially with the number of bangs). However, it may be possible to circumvent this bottleneck by using deep RL which uses neural networks to represent the Q function.

One place where the system size implicitly enters the computational costs of the RL protocol is through the number of episodes needed to train the RL algorithm. At every time step, one solves Schrödinger's equation to simulate the dynamics. The solver's scaling with L depends on how the time evolution is implemented: In spin-1/2 systems, for exact diagonalization (used here), the computational cost scales exponentially 2^{2L}, while a more sophisticated Krylov method alleviates this somewhat to $L^2 2^L$ [86], and matrix product states scale only as L^2 (in the worst case) [87]. Therefore, combining RL with existing approximate techniques to evolve quantum states can lead

to a significant reduction of CPU time, provided applying these techniques is justified by the underlying physics.

The present work demonstrates the suitability of RL for manipulating or controlling quantum systems. Yet, it does not explore how one can improve the Q-learning algorithm and adjust it to the specific needs of quantum control. Let us briefly list a few possible directions that the interested reader may want to keep in mind: (i) Alternative definitions of the RL state space (see Ref. [39]) may prove advantageous, depending on the need of the problem setup, since the RL state space defines the agent's knowledge about the physical system. For instance, if the RL agent is to be coupled to an experiment, one cannot use the wave function for this purpose, whereas wave functions may be accessible in numerical simulations. We find that the choice of RL state space influences the learning capabilities of the RL agent. (ii) Another way to increase performance is to add more controls. This increases only the possibility to reach a higher fidelity, but it comes at a cost of a potential slowdown due to a higher computational demand to explore the increased RL state space. (iii) In addition, choosing a suitable family of protocols and how to parametrize it may also lead to increased performance in RL. We use bang-bang protocols because of their computational simplicity, yet the needs of a given problem may justify another choice: The experimental realization of bang-bang protocols is limited by the resolution with which a pulse can be stabilized, which is set by the experimental apparatus. Alternatively, the RL setup can be formulated to control the size of some generalized Fourier coefficients, an idea underlying the CRAB algorithm. (iv) On the algorithmic side, one can also optimize the exploration and replay schedules which control the learning efficiency with increasing the number of training episodes and influence the RL agent's learning speed.

Reinforcement learning algorithms are versatile enough and can be suitably combined with existing ones. For instance, applying RL to complex problems with glassy landscapes is likely to benefit from a pretraining stage. Such a beneficial behavior has already been observed in the context of deep RL [41,42]. For the purpose of pretraining, in certain cases it may be advantageous to combine RL with existing derivative-based optimal control methods, such as GRAPE and CRAB, or even exhaustive search, so that one starts the optimization from a reasonable "educated guess." In recent years, it was shown that derivative-based and feedback-loop control methods can be efficiently combined to boost performance [88]. Vice versa, RL's exploration schedule defined on a suitable abstract RL-state space may prove a useful addition to improve on already existing algorithms.

Using RL, we reveal the existence of control phase transitions and show their universality in the sense that they also affect the behavior of state-of-the-art optimal control methods. The appearance of a glassy phase, which dominates the many-body physics, in the space of protocols of the quantum-state manipulation problem, could have far-reaching consequences for efficiently manipulating systems in condensed matter experiments. Quantum computing relies heavily on our ability to prepare states with high fidelity, yet finding high-efficiency state manipulation routines remains a difficult problem. Highly controllable quantum emulators, such as ultracold atoms and ions, depend almost entirely on the feasibility to reach the correct target state before it can be studied. We demonstrate how a *model-free* RL agent can provide valuable insights into constructing variational theories which capture almost all relevant features of the dynamics generated by the optimal protocol. Unlike the optimal bang-bang protocol, the simpler variational protocol is robust to small perturbations while giving comparable fidelities. This implies the existence of nearly optimal protocols, which do not suffer from the exponential complexity of finding the global minimum of the entire optimization landscape. Finally, in contrast with optimal control methods such as stochastic gradient descent (SGD), GRAPE, and CRAB that assume an exact model of the physical system, the model-free nature of RL suggests that it can be used to design protocols even when our knowledge of the physical systems we wish to control is incomplete or our system is noisy or disordered [89].

The existence of phase transitions in quantum-control problems may have profound consequences beyond physical systems. We suspect that the glassy behavior observed here maybe a generic feature of many control problems, and it will be interesting to see if this is indeed the case. It is our hope that given the close connections between optimal control and RL, the physical interpretation of optimization problems in terms of a glassy phase will help in developing novel efficient algorithms and help spur new ideas in RL and artificial intelligence.

ACKNOWLEDGMENTS

We thank J. Garrahan, M. Heyl, M. Schiró, and D. Schuster for illuminating discussions. M. B., P. W., and A. P. are supported by National Science Foundation Grants No. DMR-1813499, No. ARO W911NF1410540, and No. AFOSR FA9550-16-1-0334. A. D. is supported by a NSERC PGS D. A. D. and P. M. acknowledge support from Simon's Foundation through the MMLS Fellow program. D. S. acknowledges support from the FWO as postdoctoral fellow of the Research Foundation—Flanders and CMTV. We use QuSpin for simulating the dynamics of the qubit systems [90,91]. The authors are pleased to acknowledge that the computational work reported in this paper is performed on the Shared Computing Cluster which is administered by Boston University's Research Computing Services. The authors also acknowledge the Research Computing Services group for providing consulting support which contributed to the results reported within this paper.

[1] L. M. K. Vandersypen and I. L. Chuang, *NMR Techniques for Quantum Control and Computation*, Rev. Mod. Phys. **76**, 1037 (2005).

[2] S. van Frank, M. Bonneau, J. Schmiedmayer, S. Hild, C. Gross, M. Cheneau, I. Bloch, T. Pichler, A. Negretti, T. Calarco *et al.*, *Optimal Control of Complex Atomic Quantum Systems*, Sci. Rep. **6**, 34187 (2016).

[3] P. B. Wigley, P. J. Everitt, A. van den Hengel, J. W. Bastian, M. A. Sooriyabandara, G. D. McDonald, K. S. Hardman, C. D. Quinlivan, P. Manju, C. C. N. Kuhn *et al.*, *Fast Machine-Learning Online Optimization of Ultra-Cold-Atom Experiments*, Sci. Rep. **6**, 25890 (2016).

[4] R. Islam, E. E. Edwards, K. Kim, S. Korenblit, C. Noh, H. Carmichael, G.-D. Lin, L.-M. Duan, C.-C. Joseph Wang, J. K. Freericks, and C. Monroe, *Onset of a Quantum Phase Transition with a Trapped Ion Quantum Simulator*, Nat. Commun. **2**, 377 (2011).

[5] C. Senko, P. Richerme, J. Smith, A. Lee, I. Cohen, A. Retzker, and C. Monroe, *Realization of a Quantum Integer-Spin Chain with Controllable Interactions*, Phys. Rev. X **5**, 021026 (2015).

[6] P. Jurcevic, B. P. Lanyon, P. Hauke, C. Hempel, P. Zoller, R. Blatt, and C. F. Roos, *Quasiparticle Engineering and Entanglement Propagation in a Quantum Many-Body System*, Nature (London) **511**, 202 (2014).

[7] C. Sayrin, I. Dotsenko, X. Zhou, B. Peaudecerf, T. Rybarczyk, S. Gleyzes, P. Rouchon, M. Mirrahimi, H. Amini, M. Brune, J.-M. Raimond, and S. Haroche, *Real-Time Quantum Feedback Prepares and Stabilizes Photon Number States*, Nature (London) **477**, 73 (2011).

[8] R. Barends, A. Shabani, L. Lamata, J. Kelly, A. Mezzacapo, U. L. Heras, R. Babbush, A. G. Fowler, B. Campbell, Y. Chen *et al.*, *Digitized Adiabatic Quantum Computing with a Superconducting Circuit*, Nature (London) **534**, 222 (2016).

[9] B. B. Zhou, A. Baksic, H. Ribeiro, C. G. Yale, F. J. Heremans, P. C. Jerger, A. Auer, G. Burkard, A. A. Clerk, and D. D. Awschalom, *Accelerated Quantum Control Using Superadiabatic Dynamics in a Solid-State Lambda System*, Nat. Phys. **13**, 330 (2017).

[10] M. A. Nielsen and I. Chuang, *Quantum Computation and Quantum Information* (AAPT, 2002), ISBN-13: 978-1107002173.

[11] V. Khemani, R. Nandkishore, and S. L. Sondhi, *Nonlocal Adiabatic Response of a Localized System to Local Manipulations*, Nat. Phys. **11**, 560 (2015).

[12] P. Weinberg, M. Bukov, L. D'Alessio, A. Polkovnikov, S. Vajna, and M. Kolodrubetz, *Minimizing Irreversible Losses in Quantum Systems by Local Counter-Diabatic Driving*, Phys. Rep. **688**, 1 (2017).

[13] A. Baksic, H. Ribeiro, and A. A. Clerk, *Speeding Up Adiabatic Quantum State Transfer by Using Dressed States*, Phys. Rev. Lett. **116**, 230503 (2016).

[14] X. Wang, M. Allegra, K. Jacobs, S. Lloyd, C. Lupo, and M. Mohseni, *Quantum Brachistochrone Curves as Geodesics: Obtaining Accurate Minimum-Time Protocols for the Control of Quantum Systems*, Phys. Rev. Lett. **114**, 170501 (2015).

[15] R. R. Agundez, C. D. Hill, L. C. L. Hollenberg, S. Rogge, and M. Blaauboer, *Superadiabatic Quantum State Transfer in Spin Chains*, Phys. Rev. A **95**, 012317 (2017).

[16] S. Bao, S. Kleer, R. Wang, and and A. Rahmani, *Optimal Control of gmon Qubits Using Pontyagin's Minimum Principle: Preparing a Maximally Entangled State with Singular Bang-Bang Protocols*, Phys. Rev. A **97**, 062343 (2018).

[17] G. M. Rotskoff, G. E. Crooks, and E. Vanden-Eijnden, *Geometric Approach to Optimal Nonequilibrium Control: Minimizing Dissipation in Nanomagnetic Spin Systems*, Phys. Rev. E **95**, 012148 (2017).

[18] N. Leung, M. Abdelhafez, J. Koch, and D. I. Schuster, *Speedup for Quantum Optimal Control from GPU-Based Automatic Differentiation*, Phys. Rev. A **95**, 042318 (2017).

[19] Z.-C. Yang, A. Rahmani, A. Shabani, H. Neven, and C. Chamon, *Optimizing Variational Quantum Algorithms Using Pontryagin's Minimum Principle*, Phys. Rev. X **7**, 021027 (2017).

[20] C. Jarzynski, *Generating Shortcuts to Adiabaticity in Quantum and Classical Dynamics*, Phys. Rev. A **88**, 040101 (2013).

[21] M. Kolodrubetz, D. Sels, P. Mehta,, and A. Polkovnikov, *Geometry and Non-Adiabatic Response in Quantum and Classical Systems*, Phys. Rep. **697**, 1 (2017).

[22] D. Sels and A. Polkovnikov, *Minimizing Irreversible Losses in Quantum Systems by Local Counterdiabatic Driving*, Proc. Natl. Acad. Sci. U.S.A. **114**, E3909 (2017).

[23] S. J. Glaser, T. Schulte-Herbrüggen, M. Sieveking, O. Schedletzky, N. C. Nielsen, O. W. Sørensen, and C. Griesinger, *Unitary Control in Quantum Ensembles: Maximizing Signal Intensity in Coherent Spectroscopy*, Science **280**, 421 (1998).

[24] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, *Whither the Future of Controlling Quantum Phenomena?*, Science **288**, 824 (2000).

[25] N. Khaneja, R. Brockett, and S. J. Glaser, *Time Optimal Control in Spin Systems*, Phys. Rev. A **63**, 032308 (2001).

[26] S. E. Sklarz and D. J. Tannor, *Loading a Bose-Einstein Condensate onto an Optical Lattice: An Application of Optimal Control Theory to the Nonlinear Schrödinger Equation*, Phys. Rev. A **66**, 053619 (2002).

[27] M. Demirplak and S. A. Rice, *Adiabatic Population Transfer with Control Gields*, J. Phys. Chem. A **107**, 9937 (2003).

[28] 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*, J. Magn. Reson. **172**, 296 (2005).

[29] M. V. Berry, *Transitionless Quantum Driving*, J. Phys. A **42**, 365303 (2009).

[30] T. Caneva, T. Calarco, and S. Montangero, *Chopped Random-Basis Quantum Optimization*, Phys. Rev. A **84**, 022326 (2011).

[31] E. Zahedinejad, S. Schirmer, and B. C. Sanders, *Evolutionary Algorithms for Hard Quantum Control*, Phys. Rev. A **90**, 032310 (2014).

[32] E. Zahedinejad, J. Ghosh, and B. C. Sanders, *High-Fidelity Single-Shot Toffoli Gate via Quantum Control*, Phys. Rev. Lett. **114**, 200502 (2015).

[33] M. Theisen, F. Petiziol, S. Carretta, P. Santini, and S. Wimberger, *Superadiabatic Driving of a Three-Level Quantum System*, Phys. Rev. A **96**, 013431 (2017).

[34] M. Bukov, D. Sels, and A. Polkovnikov, *The Geometric Bound of Accessible Quantum State Preparation*, arXiv: 1804.05399.

[35] R.-B. Wu, B. Chu, D. H. Owens, and H. Rabitz, *Data-Driven Gradient Algorithm for High-Precision Q Control*, Phys. Rev. A **97**, 042122 (2018).

[36] P. Doria, T. Calarco, and S. Montangero, *Optimal Control Technique for Many-Body Quantum Dynamics*, Phys. Rev. Lett. **106**, 190501 (2011).

[37] T. Caneva, A. Silva, R. Fazio, S. Lloyd, T. Calarco, and S. Montangero, *Complexity of Controlling Quantum Many-Body Dynamics*, Phys. Rev. A **89**, 042322 (2014).

[38] S. Lloyd and S. Montangero, *Information Theoretical Analysis of Quantum Optimal Control*, Phys. Rev. Lett. **113**, 010502 (2014).

[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2017).

[40] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. (2006), ISBN: 978-0-387-31073-2.

[41] 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 (London) **518**, 529 (2015).

[42] 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 (London) **529**, 484 (2016).

[43] V. Dunjko, J. M. Taylor, and H. J. Briegel, *Quantum-Enhanced Machine Learning*, Phys. Rev. Lett. **117**, 130501 (2016).

[44] 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*, arXiv:1803.08823.

[45] R. S. Judson and H. Rabitz, *Teaching Lasers to Control Molecules*, Phys. Rev. Lett. **68**, 1500 (1992).

[46] G. Carleo and M. Troyer, *Solving the Quantum Many-Body Problem with Artificial Neural Networks*, Science **355**, 602 (2017).

[47] J. Carrasquilla and R. G. Melko, *Machine Learning Phases of Matter*, Nat. Phys. **13**, 431 (2017).

[48] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, *Learning to Soar in Turbulent Environments*, Proc. Natl. Acad. Sci. U.S.A. **113**, E4877 (2016).

[49] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, *Flow Navigation by Smart Microswimmers via Reinforcement Learning*, Phys. Rev. Lett. **118**, 158004 (2017).

[50] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger, *Automated Search for New Quantum Experiments*, Phys. Rev. Lett. **116**, 090405 (2016).

[51] A. A. Melnikov, H. P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel, *Active Learning Machine Learns to Create New Quantum Experiments*, Proc. Natl. Acad. Sci. U.S.A. **115**, 1221 (2018).

[52] C. Chen, D. Dong, H.-X. Li, J. Chu, and T.-J. Tarn, *Fidelity-Based Probabilistic q-Learning for Control of Quantum Systems*, IEEE Trans. Neural Netw. Learn. Syst. **25**, 920 (2014).

[53] M. August and J. M. Hernández-Lobato, *Taking Gradients through Experiments: LSTMs and Memory Proximal Policy Optimization for Black-Box Quantum Control*, arXiv: 1802.04063.

[54] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, *Reinforcement Learning with Neural Networks for Quantum Feedback*, Phys. Rev. X **8**, 031084 (2018).

[55] X.-M. Zhang, Z.-W. Cui, X. Wang, and M.-H. Yung, *Automatic Spin-Chain Learning to Explore the Quantum Speed Limit*, Phys. Rev. A **97**, 052333 (2018).

[56] M. Y. Niu, S. Boixo, V. Smelyanskiy, and H. Neven, *Universal Quantum Control through Deep Reinforcement Learning*, arXiv:1803.01857.

[57] F. Albarran-Arriagada, J. C. Retamal, E. Solano, and L. Lamata, *Measurement-Based Adaptation Protocol with Quantum Reinforcement Learning*, arXiv:1803.05340.

[58] T. R. Gingrich, G. M. Rotskoff, G. E. Crooks, and P. L. Geissler, *Near-Optimal Protocols in Complex Nonequilibrium Transformations*, Proc. Natl. Acad. Sci. U.S.A. **113**, 10263 (2016).

[59] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrueggen, and S. J. Glaser, *Optimal Control of Coupled Spin Dynamics: Design of NMR Pulse Sequences by Gradient Ascent Algorithms*, J. Magn. Reson. **172**, 296 (2005).

[60] S. Machnes, U. Sander, S. J. Glaser, P. de Fouquières, A. Gruslys, S. Schirmer, and T. Schulte-Herbrüggen, *Comparing, Optimizing, and Benchmarking Quantum-Control Algorithms in a Unifying Programming Framework*, Phys. Rev. A **84**, 022305 (2011).

[61] P. De Fouquieres, S. G. Schirmer, S. J. Glaser,, and Ilya Kuprov, *Second Order Gradient Ascent Pulse Engineering*, J. Magn. Reson. **212**, 412 (2011).

[62] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.8.031086 for (i) a comparison between RL, SD, GRAPE and CRAB, (ii) RL with quasi-continuous protocols, and (iii) a critical scaling analysis of the control phase transitions.

[63] G. C. Hegerfeldt, *Driving at the Quantum Speed Limit: Optimal Control of a Two-Level System*, Phys. Rev. Lett. **111**, 260501 (2013).

[64] J. J. W. H. Sørensen, M. K. Pedersen, M. Munch, P. Haikka, J. H. Jensen, T. Planke, M. G. Andreasen, M. Gajdacz, K. Mølmer, A. Lieberoth, and J. F. Sherson, *Exploring the Quantum Speed Limit with Computer Games*, Nature (London) **532**, 210 (2016).

[65] We note that there is no proof that the family of bang-bang protocols contains the optimal protocol for the single-qubit control problem in Sec. III, since the present control problem is of bilinear type [66].

[66] H. Schättler and U. Ledzewicz, *Geometric Optimal Control: Theory, Methods and Examples* (Springer-Verlag, New York, 2012).

[67] A. G. R. Day, M. Bukov, P. Weinberg, P. Mehta, and D. Sels, *The Glassy Phase of Optimal Quantum Control*, arXiv: 1803.10856.

[68] Bellman's equation probably admits a closed solution for the single-qubit example from Sec. III.

[69] T. Castellani and A. Cavagna, *Spin-Glass Theory for Pedestrians*, J. Stat. Mech. (2005) P05012.
[70] L. O. Hedges, R. L. Jack, J. P. Garrahan, and D. Chandler, *Dynamic Order-Disorder in Atomistic Models of Structural Glass Formers*, Science **323**, 1309 (2009).
[71] Notice that the agent does not have control over the z field.
[72] M. Demirplak and S. A. Rice, *Assisted Adiabatic Passage Revisited*, J. Phys. Chem. B **109**, 6838 (2005).
[73] M. Demirplak and S. A. Rice, *On the Consistency, Extremal, and Global Properties of Counterdiabatic Fields*, J. Chem. Phys. **129**, 154111 (2008).
[74] A. del Campo, *Shortcuts to Adiabaticity by Counterdiabatic Driving*, Phys. Rev. Lett. **111**, 100502 (2013).
[75] S. Deffner, C. Jarzynski, and A. del Campo, *Classical and Quantum Shortcuts to Adiabaticity for Scale-Invariant Driving*, Phys. Rev. X **4**, 021013 (2014).
[76] F. Petiziol, B. Dive, F. Mintert, and S. Wimberger, *Fast Adiabatic Evolution by Oscillating Initial Hamiltonians*, arXiv:1807.10227.
[77] J. Simon, W. S. Bakr, R. Ma, M. E. Tai, P. M. Preiss, and M. Greiner, *Quantum Simulation of Antiferromagnetic Spin Chains in an Optical Lattice*, Nature (London) **472**, 307 (2011).
[78] M. Bukov, A. G. R. Day, P. Weinberg, A. Polkovnikov, P. Mehta, and D. Sels, *Broken Symmetry in a Two-Qubit Quantum Control Landscape*, Phys. Rev. A **97**, 052114 (2018).
[79] We expect the many-body overlap between two states to be exponentially small in the system size L.
[80] C. Benedetti, A. P. Shurupov, M. G. A. Paris, G. Brida, and M. Genovese, *Experimental Estimation of Quantum Discord for a Polarization Qubit and the Use of Fidelity to Assess Quantum Correlations*, Phys. Rev. A **87**, 052136 (2013).
[81] M. Bina, A. Mandarino, S. Olivares, and M. G. A. Paris, *Drawbacks of the Use of Fidelity to Assess Quantum Resources*, Phys. Rev. A **89**, 012305 (2014).
[82] A. Mandarino, M. Bina, C. Porto, S. Cialdi, S. Olivares, and M. G. A. Paris, *Assessing the Significance of Fidelity as a Figure of Merit in Quantum State Reconstruction of Discrete and Continuous-Variable Systems*, Phys. Rev. A **93**, 062118 (2016).
[83] L. Vidmar, D. Iyer, and M. Rigol, *Emergent Eigenstate Solution to Quantum Dynamics Far from Equilibrium*, Phys. Rev. X **7**, 021012 (2017).
[84] M. Mézard, G. Parisi, and R. Zecchina, *Analytic and Algorithmic Solution of Random Satisfiability Problems*, Science **297**, 812 (2002).
[85] S. C. Morampudi, B. Hsu, S. L. Sondhi, R. Moessner, and C. R. Laumann, *Clustering in Hilbert Space of a Quantum Optimization Problem*, Phys. Rev. A **96**, 042303 (2017).
[86] Y. Saad, *Analysis of Some Krylov Subspace Approximations to the Matrix Exponential Operator*, SIAM J. Numer. Anal. **29**, 209 (1992).
[87] J. J. García-Ripoll, *Time Evolution of Matrix Product States*, New J. Phys. **8**, 305 (2006).
[88] D. J. Egger and F. K. Wilhelm, *Adaptive Hybrid Optimal Quantum Control for Imprecisely Characterized Systems*, Phys. Rev. Lett. **112**, 240503 (2014).
[89] M. Bukov, *Reinforcement Learning to Autonomously Prepare Floquet-Engineered States: Inverting the Quantum Kapitza Oscillator*, arXiv:1808.08910.
[90] P. Weinberg and M. Bukov, *QuSpin: A PYTHON Package for Dynamics and Exact Diagonalisation of Quantum Many Body Systems Part I: Spin Chains*, SciPost Phys. **2**, 003 (2017).
[91] 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*, arXiv:1804.06782.