Supplemental Material to 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 Ave., Boston, MA 02215, USA*
[2] *Theory of quantum and complex systems, Universiteit Antwerpen, B-2610 Antwerpen, Belgium*

I. GLASSY BEHAVIOUR OF DIFFERENT MACHINE LEARNING AND OPTIMAL CONTROL ALGORITHMS WITH LOCAL AND NONLOCAL FLIP UPDATES

A. Stochastic Descent

To benchmark the results obtained using Reinforcement Learning (RL), we use a greedy stochastic descent (SD) algorithm to sample the infidelity landscape minima containing the driving protocols. We restrict our SD algorithm to exploring bang-bang protocols, for which $h_x(t) \in \{\pm 4\}$. The algorithm starts from a random protocol configuration and proposes local field updates at a time t chosen uniformly in the interval $[0, T]$. The updates consist in changing the applied field $h_x(t) \to h'_x(t)$ only if this increases the fidelity. Ideally, the protocol is updated until all possible local field updates can only decrease the fidelity. Practically, for some protocol durations, ensuring that a true local minima with respect to 1-flip is reached can be computational expensive. Therefore, we restrict the number of fidelity evaluations to be at most $20 \times T/\delta t$. In this regard, the obtained protocol is a local minimum with respect to local (1-flip) field updates. The stochastic descent is repeated multiple times with different initial random protocols. The set of protocols $\{h^\alpha | \alpha = 1, \ldots, N_{\text{real}}\}$ obtained with stochastic descent is used to calculate the glass-like order parameter $q(T)$ (see main text). A Python implementation of the algorithm is available on Github.

B. CRAB

Chopped RAndom Basis (CRAB) is a state-of-the-art optimal control algorithm designed to tackle many-body quantum systems [1, 2]. The idea behind CRAB is to decompose the unknown driving protocol into a complete basis (Fourier, Laguerre, etc.), and impose a cut-off on the number of 'harmonics' kept for the optimisation. The algorithm then uses an optimiser to find the values for the expansion coefficients, which optimise the cost function of the problem.

Following Ref. [2], we make a Fourier-basis ansatz for the driving protocol.

$$h_{\text{CRAB}}(t) \;=\; h_0(t)\left(1 + \frac{1}{\lambda(t)}\sum_{i=1}^{N_c} A_i \cos\omega_i t + B_i \sin\omega_i t\right), \tag{1}$$

where the Fourier coefficients $\{A_i, B_i, \omega_i\}$ which parametrise the protocol are found using an implementation of the Nelder-Mead optimization method in the SciPy python library. The number of harmonics kept in the optimisation is given by N_c. The CRAB algorithm uses two auxiliary functions, defined by the user: the first function $h_0(t)$ is a trial initial guess ansatz for the protocol, while the second function, $\lambda(t)$ imposes the boundary conditions $\lambda \to \infty$ for $t \to 0$ and $t \to T$ to the Fourier expansion term.

The cost function which we optimise in the state manipulation problem

$$\mathcal{C}[h(t)] = \mathcal{F}(\{A_i, B_i, \omega_i\}) + \frac{1}{16T}\int_0^T \mathrm{d}t [h(t)]^2 \tag{2}$$

contains the fidelity $\mathcal{F}(\{A_i, B_i, \omega_i\})$ at the end of the protocol, and an additional penalty coming from the L^2 norm of the protocol to keep the optimal protocols bounded. The last constraint is required for a better and honest comparison with the RL, SD, and GRAPE algorithms.

*Electronic address: mbukov@bu.edu
†Electronic address: agrday@bu.edu

Applying CRAB to the state manipulation problem from the main text systematically, we choose $\lambda(t) = 1/\sin(\pi t/T)^2$, and $h_0(t) = -2 + 4t/T$. We also consider $N_c = 10, 20$ to study how much an effect increasing the number of degrees of freedom will have on the optimal protocols found. For each value of N_c we start the optimization algorithm with 10 random initial configurations. We define the optimal protocol as the protocol with the best fidelity out of that group of 10. The random initial frequencies ω_i are chosen the same way as outlined in Ref. [2] while the amplitudes A_i and B_i are chosen uniformly between -10 and 10.

C. GRAPE

GRadient Ascend Pulse Engineering, is a numeric derivative-based optimal control method, first introduced in the context of NMR spectroscopy [3]. As suggested by its name, the method performs gradient optimization. Instead of restricting the protocols to bang-bang type, the method works with quasi-continuous protocols. Protocol magnitudes can take on any value within the allowed manifold but, unlike CRAB, are piecewise constant in time.

In the present case, one can efficiently compute the gradient of the fidelity as follows. Consider the fidelity for some trial protocol $h_x(t)$,

$$F_h(T) = |\langle\psi_*| U(T,0) |\psi_i\rangle|^2, \tag{3}$$

where $U(T,0)$ denotes the time evolution operator from 0 to T. Let us further decompose the Hamiltonian as $H = H_0 + h_x(t)X$, where H_0 is the part over which we have no control, and X denotes the operator we control. The functional derivative of the fidelity with respect to the protocol thus becomes

$$\frac{\delta F_h(T)}{\delta h_x(t)} = i\langle\psi_*| U(T,0) |\psi_i\rangle \langle\psi_i| U(0,t)XU(t,T) |\psi_*\rangle - i\langle\psi_*| U(T,t)XU(t,0) |\psi_i\rangle \langle\psi_i| U(0,T) |\psi_*\rangle \tag{4}$$

Although this expression appears hard to evaluate, it takes on a very simple form

$$\frac{\delta F_h(T)}{\delta h_x(t)} = 2\mathrm{Im}\left[\langle\phi(t)| X |\psi_i(t)\rangle\right], \tag{5}$$

where $|\psi_i(t)\rangle = U(t,0)|\psi_i\rangle$ denotes the initial state propagated to time t and $\langle\phi(t)| = \langle\psi_i(T)|\psi_*\rangle\langle\psi_*| U(T,t)$ denotes the (scaled) final state propagated back in time to time t. Notice that this procedure requires us to exactly know this time evolution operator (i.e. a model of physical system to be controlled). Hence, by propagating both the initial state forward and the target state backward in time one gets access to the full gradient of the control landscape.

To find a local maximum one can simply now gradient ascend the fidelity. A basic algorithm thus goes as follows:

(i) Pick a random initial magnetic field $h_x^0(t)$.

(ii) Compute first $|\psi_i(t)\rangle$ and then $\langle\phi(t)|$ for the current setting of the magnetic field $h_x^N(t)$.

(iii) Update the control field $h_x^{N+1}(t) = h_x^N(t) + \epsilon_N \mathrm{Im}\left[\langle\phi(t)| X |\psi_i(t)\rangle\right]$. Note that this step can be upgraded to a second-order Newton method to improve the performance of the algorithm, see Refs. [4–6].

(iv) Repeat (ii) and (iii) until the desired tolerance is reached.

Here ϵ_N is the step size in each iteration. Choosing a proper step size can be a difficult task. In principle the fidelity should go up after each iteration but if the step size it too large, the algorithm can overshoot the maximum, resulting in a worse fidelity. To avoid this, one should adapt the step size during the algorithm. Numerically we have observed that, in order to avoid overshooting saddles/maxima, $\epsilon_N \propto 1/\sqrt{N}$.

D. Comparison between the RL, SD, CRAB and GRAPE

While a detailed comparison between the ML algorithms and other optimal control algorithms is an interesting topic, it is beyond the scope of the present paper. Below, we only show that both the RL algorithm is capable of finding optimal bang-bang protocols for the quantum control problem from the main text, and that its performance rivals that of SD and the state-of-the-art algorithms for many-body quantum problems CRAB and GRAPE.

The result for the single qubit $L = 1$ are shown in Fig. 2 (left). The most important points can be summarised as follows:

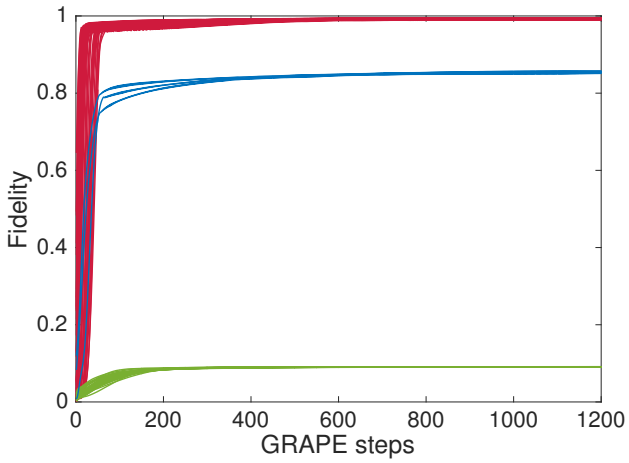


FIG. 1: Fidelity traces of GRAPE for $T = 3.2$ and $L = 6$ as a function of the number of gradient ascend steps for 10^2 random initial conditions. This figure should be compared to Fig. **??** in the main text. GRAPE clearly gets attracted by the same three attractors as SD but has much smaller intra-attractor fluctuations, presumably due to the non-locality of the updates and the continuous values of the control field $h_x \in [-4, 4]$ used in GRAPE. This shows that the glass control phase is present for both local and nonlocal update algorithms.

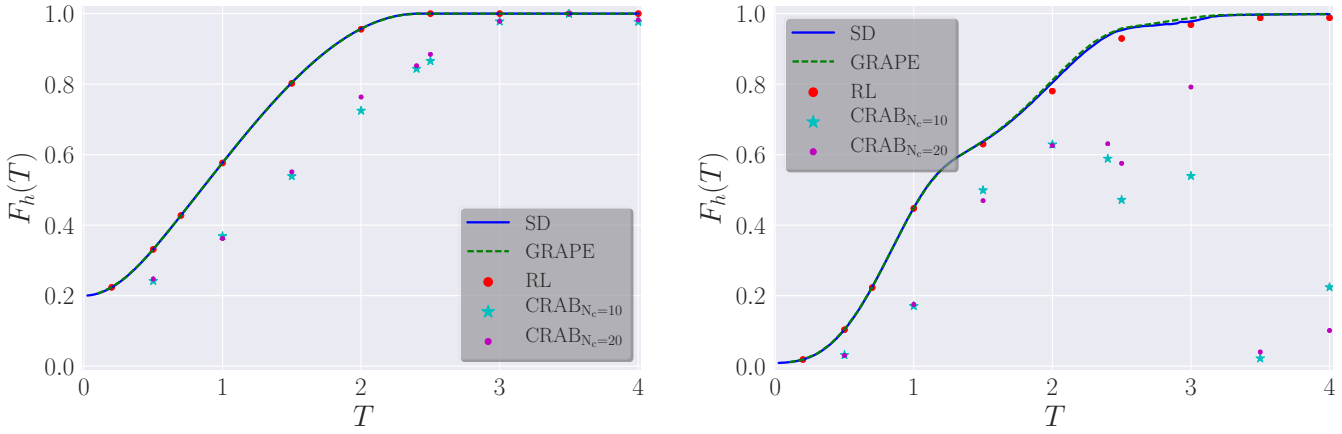


FIG. 2: Comparison between the best fidelities obtained using SD (solid blue line), RL (big red dots), GRAPE (dashed green line) and CRAB (cyan star and small magenta dot) for $L = 1$ (left) and $L = 6$ (right). Here N_c denotes the cap in the number of harmonics kept in the CRAB simulation.

- RL, GS and GRAPE all find the optimal protocol.
- Below the quantum speed limit, T_{QSL}, CRAB finds good, but clearly suboptimal protocols. The plots also show the glassiness represents a generic feature of the constrained optimization problem and not the method used to perform the optimization. Increasing the cutoff N_c, and with it the number of effective degrees of freedom, does not lead to a sizeable improvement in CRAB. One explanation is the following: for the single qubit, we know that the variational protocol, which contains at most two bangs (see Fig. **??**b), is a global minimum of the optimisation landscape. Such a protocol can easily be approximated using up to $N_c = 20$ harmonics. However, allowing more degrees of freedom comes at a huge cost due to the glassiness of the problem: there exist many quasi-degenerate local minima for the algorithm to get stuck in.

The comparison for the Many Coupled Qubits system for $L = 6$ are shown in Fig. 2 (right). Larger values of L do not introduce any change in the behaviour, as we argue in a subsequent section below. In the many-body case, the variational protocol, which contains four bangs, is shown *not* to be the global minimum of the infidelity landscape. Instead, the true global minimum contains many more bangs which, however, only marginally improve the fidelity.

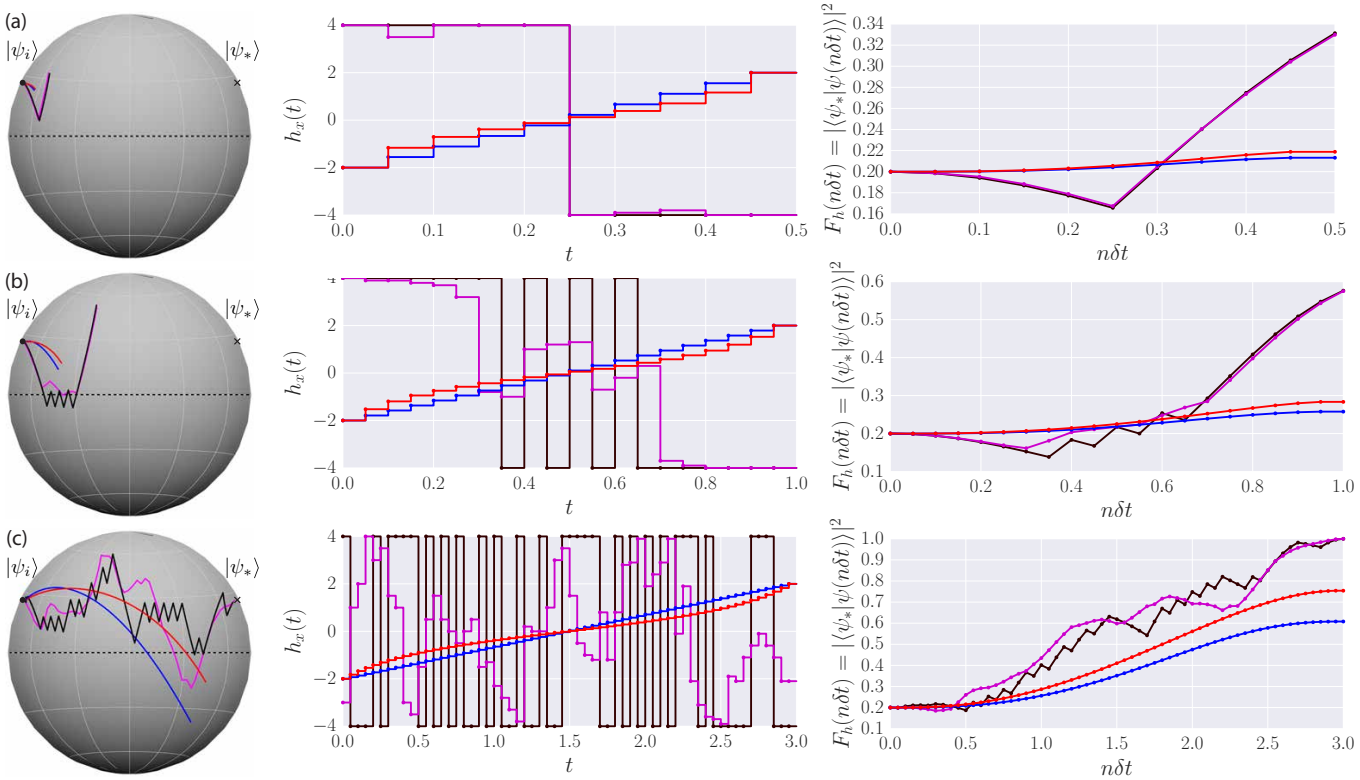


FIG. 3: Comparison between the bang-bang (black) and quasi-continuous (magenta) protocols found by the RL agent, and the Landau-Zener (blue) and geodesic (red) protocols computed from analytical theory in the overconstrained phase for $T = 0.5$ (a), the glassy phase for $T = 1.0$ (b), and the controllable phase for $T = 3.0$ (c). The left column shows the representation of the corresponding protocol on the Bloch sphere, the middle one – the protocols themselves, and the right column – the instantaneous fidelity in the target state $\psi_*\rangle$.

- All algorithms give reasonable fidelities, see Fig. 2 (right)
- Even though GRAPE seems to display better performance out of the four methods, one should not forget that this algorithm, which uses global flips, requires knowledge of all fidelity gradients – valuable information which is not easily accessible through experimental measurements. One has to keep in mind though, that this comparison is not completely honest, since GRAPE allows for the control field h_x to take any value in the interval $[-4, 4]$, which offers a further advantage over the bang-bang based RL. On the other hand, the model-free RL rivals the performance of GRAPE at all protocol durations, and outperforms CRABS, even for the single qubit below the quantum speed limit, where the problem enters the glassy phase. The reason for the seemingly slowly decreasing performance of RL with T is that, since the total protocol duration is held fixed, the number of bangs increases exponentially with $T \sim N_T$, and hence the state space which has to be explored by the agent also grows exponentially. A detailed study of the scaling is postponed to future studies.
- All algorithms suffer from the glassiness in the optimisation landscape. This is not surprising, since the glass phase is an intrinsic property of the infidelity landscape, as defined by the optimisation problem, and does not depend on which algorithm is used to look for the optimal solution. Fig. 1 shows the fidelity traces as a function of running time for GRAPE. Comparing this to the corresponding results for SD, see Fig. **??**, we see a strikingly similar behaviour, even though GRAPE uses nonlocal flip updates in contrast to SD. This means that, in the glassy phase, GRAPE also gets stuck in suboptimal attractors, similar to SD and RL. Thus, as an important consequence, the glassy phase affects both local and nonlocal-update algorithms.

II. PERFORMANCE OF THE DIFFERENT DRIVING PROTOCOLS FOR THE QUBIT

It is interesting to compare the bang-bang and quasi-continuous driving protocols found by the agent to a simple linear protocol, which we refer to as Landau-Zener (LZ), and the geodesic protocol, which optimizes local fidelity

close to the adiabatic limit essentially slowing down near the minimum gap [7]. We find that the RL agent offers significantly better solutions in the overconstrained and glassy phases, where the optimal fidelity is always smaller than unity. The Hamiltonian of the qubit together with the initial and target states read:

$$H(t) = -S^z - h_x(t)S^x, \qquad |\psi_i\rangle \sim (-1/2 - \sqrt{5}/2, 1)^{\mathsf{T}}, \qquad |\psi_*\rangle \sim (1/2 + \sqrt{5}/2, 1)^{\mathsf{T}}, \tag{6}$$

where $|\psi_i\rangle$ and $|\psi_*\rangle$ are the ground state of $H(t)$ for $h_i = -2$ and $h_* = +2$ respectively. Note that for bang-bang protocols, the initial and target states are not eigenstates of the control Hamiltonian since $h_x(t)$ takes on the values ± 4.

The RL agent is initiated at the field $h(t = 0) = h_{\min} = -4.0$. The RL protocols are constructed from the following set of jumps, δh_x, allowed at each protocol time step δt:

- *bang-bang* protocol: $\delta h_x \in \{0.0, \pm 8.0\}$ which, together with the initial condition, constrains the field to take the values $h_x(t) \in \{\pm 4.0\}$.
- *quasi-continuous* protocol: $\delta h_x \in \{0.0, \pm 0.1, \pm 0.2, \pm 0.5, \pm 1.0, \pm 2.0, \pm 4.0, \pm 8.0\}$. We restrict the actions available in a state to ensure $h_x(t) \in [-4.0, 4.0]$.

Interestingly, the RL agent figures out that it is always advantageous to first jump to $h_{\max} = +4.0$ before starting the evolution, as a consequence of the positive value of the coefficient in front of S^z.

The analytical adiabatic protocols are required to start and end in the initial and target states, which coincide with the ground states of the Hamiltonians with fields $h_i = -2.0$ and $h_* = 2.0$, respectively. They are defined as follows:

- *Landau-Zener(LZ)* protocol: $h_x(t) = (h_* - h_i)t/T + h_i$
- *geodesic* protocol: $h_x(t) = \tan(at + b)$, where $b = \arctan(h_i)$ and $a = \arctan(h_* - b)/T$.

Figure 3 shows a comparison between these four protocol types for different values of T, corresponding to the three quantum control phases. Due to the instantaneous gap remaining small compared to the total protocol duration, the LZ and geodesic protocols are very similar, irrespective of T. The two protocols significantly differ only at large T, where the geodesic protocol significantly outperforms the linear one. An interesting general feature for the short protocol durations considered is that the fidelities obtained by the LZ and geodesic protocols are clearly worse than the ones found by the RL agent. This points out the far-from-optimal character of these two approaches, which essentially reward staying close to the instantaneous ground state during time evolution. Looking at the fidelity curves in Fig. 3, we note that, before reaching the optimal fidelity at the end of the ramp for the overconstrained and glassy phases, the instantaneous fidelity drops below its initial value at intermediate times. This suggests that the angle between the initial and target states on the Bloch sphere becomes larger in the process of evolution, before it can be reduced again. Such situation is very reminiscent of counter-diabatic or fast forward driving protocols, where the system can significantly deviate from the instantaneous ground state at intermediate times [8–10]. Such problems, where the RL agent learns to sacrifice local rewards in view of obtaining a better total reward in the end are of particular interest in RL [11].

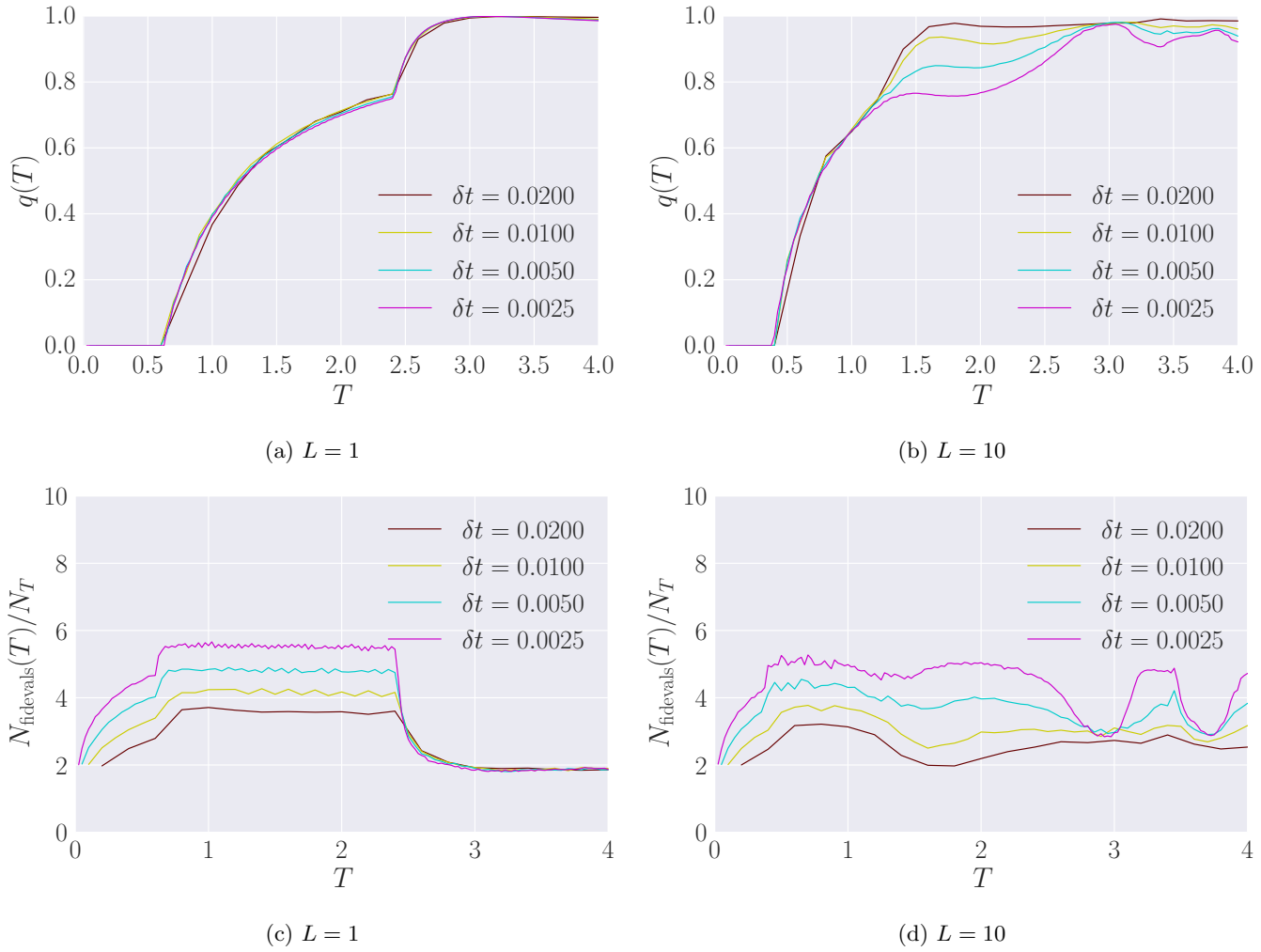
III. CRITICAL SCALING ANALYSIS OF THE CONTROL PHASE TRANSITIONS

In this section, we show convincing evidence for the existence of the phase transitions in the quantum state manipulation problem discussed in the main text. We already argued that there is a phase transition in the optimization problem as a function of the protocol time T. Mathematically the problem is formulated as

$$h_x^{\text{optimal}}(t) = \underset{h_x(t):|h_x(t)|\leq 4}{\arg\min} \{I_h(T)\} = \underset{h_x(t):|h_x(t)|\leq 4}{\arg\max} |\langle\psi_*|\mathcal{T}_t e^{-i\int_0^T \mathrm{d}t H[h_x(t)]}|\psi_i\rangle|^2, \tag{7}$$

i.e. as finding the optimal driving protocol $h_x^{\text{optimal}}(t)$, which transforms the initial state (the ground state of the Hamiltonian H at $h_x = -2$) into the target state (the ground state of H corresponding to $h_x = 2$) maximizing the fidelity in protocol duration T under unitary Schrödinger evolution. We assume that the ramp protocol is bounded, $h_x(t) \in [-4, 4]$, for all times during the ramp. In this section, we restrict the analysis to bang-bang protocols only, for which $h_x(t) \in \{\pm 4\}$. The minimum protocol time step is denoted by δt. There are two different scaling limits in the problem. We define a continuum limit for the problem as $\delta t \to 0$ while keeping the total protocol duration $T = const.$ Additionally, there is the conventional thermodynamic limit, where we send the system size $L \to \infty$.

As we already alluded to in the main text, one can think of this optimization problem as a minimization in the infidelity landscape, determined by the mapping $h_x(t) \mapsto I_h(T) = 1 - F_h(T)$, where each protocol is assigned a point



(a) $L = 1$ (b) $L = 10$

(c) $L = 1$ (d) $L = 10$

FIG. 4: Finite time step-size, δt, scaling of the order parameter $q(T)$ [top] and the number of fidelity evaluations per time step, N_{fidevals}/N_T [bottom], to reach a local minimum of the infidelity landscape.

in fidelity space – the probability of being in the target state after evolution for a fixed protocol duration T. Finding the global minimum of the landscape then corresponds to obtaining the optimal driving protocol for any fixed T.

To obtain the set of local minima $\{h_x^\alpha(t)|\alpha = 1, \ldots, N_{\text{real}}\}$ of the infidelity landscape at a fixed total protocol duration T and protocol step size δt, we apply Stochastic Descent(SD), see above, starting from a random protocol configuration, and introduce random *local* changes to the bang-bang protocol shape until the fidelity can no longer be improved. This method is guaranteed to find a set of representative local infidelity minima with respect to "1-flip" dynamics, mapping out the bottom of the landscape of $I_h(T)$. Keeping track of the mean number of fidelity evaluations N_{fidevals} required for this procedure, we obtain a measure for the average time it takes the SD algorithm to settle in a local minimum. While the order parameter $q(T)$ (see below) was used in the main text as a measure for the static properties of the infidelity landscape, dynamic features are revealed by studying the number of fidelity evaluations N_{fidevals}.

As discussed in the main text, the rich phase diagram of the problem can also be studied by looking at the order parameter function q (closely related to the Edwards-Anderson order parameter for detecting glassy order in spin systems [12]):

$$q(T) = \frac{1}{16N_T}\overline{\sum_{n=1}^{N_T}\{h_x(n\delta t) - \overline{h_x}(n\delta t)\}^2}, \qquad \overline{h_x}(t) = \frac{1}{N_{\text{real}}}\sum_{\alpha=1}^{N_{\text{real}}} h_x^\alpha(t). \tag{8}$$

Here, N_T is the total number of protocol time steps of fixed width δt, N_{real} is the total number of random protocol realisations $h_x^\alpha(t)$ probing the minima of the infidelity landscape (see previous paragraph), and the factor $1/16$ serves

to normalise the squared bang-bang drive protocol $h_x^2(t)$ within the range $[-1, 1]$.

A. Single Qubit

For $T > T_{\rm QSL}$, the optimization problem of the single qubit ($L = 1$) is controllable, and there exist infinitely many protocols which can prepare the target state with unit fidelity. In analogy with the random k-SAT problem [13], we call this the *controllable* (or underconstrained) phase of the quantum control problem. Intuitively, this comes about due to the large total protocol durations available which allow one to correct a set of 'wrong moves' at a later time in order to achieve a better fidelity at the end of the ramp. We have seen that both the Reinforcement Learning (RL) and Stochastic Descent (SD) agents readily and quickly learn to exploit this feature for optimal performance. In this phase, which is not present in the thermodynamic limit $L \to \infty$, there exist analytical results to compute driving protocols of unit fidelity based on the geodesic and counter-diabatic approaches [8, 14, 15]. The driving protocols $h_x^\alpha(t)$, corresponding to the minima of the infidelity landscape $I_h(T)$, are completely uncorrelated, resulting in $\overline{h_x}(t) = 0$ and, thus, $q = 1$. As $T \searrow T_{\rm QSL}$, the infidelity minima start becoming correlated, reflected in a drop in the value of $q(T)$. At the time $T = T_{\rm QSL}$, a phase transition occurs to a glassy phase with shallow, quasi-degenerate infidelity minima corresponding to many almost equally optimal protocols. Fig. 4a shows that the order parameter $q(T)$ develops a clear non-analyticity in the continuum limit $\delta t \to 0$, which proves the existence of a phase transition in protocol space. At the same critical time $T_{\rm QSL}$, a drastic rapid increase is detected in the number of fidelity evaluations required to map out the infidelity minima, see Fig. 4c.

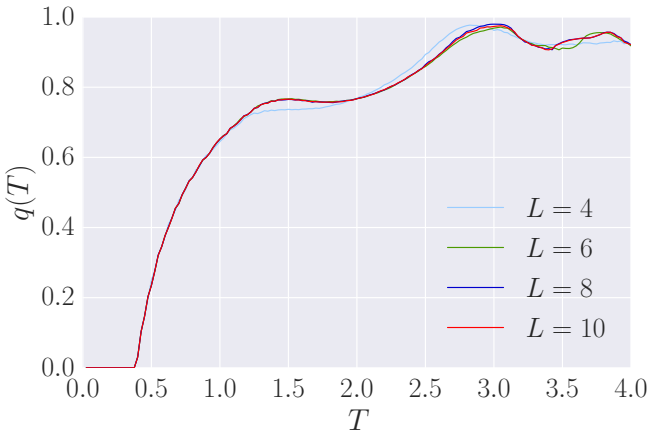
For $T_c < T < T_{\rm QSL}$ the control problem is in the glassy phase. We showed in the main text by examining the DOS of protocols with respect to local flips of the field, that finding the optimal protocol appears as complicated as finding the ground state of a glass. This is reflected in the observed increase of the average number of fidelity evaluations $N_{\rm fidevals}$ with decreasing δt (c.f Fig. 4c), and a decrease in the learning rate of the RL and SD agents. The local minima protocols $\{h_x^\alpha(t)\}$ are strongly correlated, as can be seen from the finite value of the order parameter $q(T)$ in Fig. 5a. More importantly, for any practical purposes, unit fidelity can no longer be obtained under the given dynamical constraints.

When we reach the second critical point $T = T_c$, another phase transition occurs from the glassy to an overconstrained phase. At $T = T_c$, the order parameter reaches zero, suggesting that the infidelity landscape contains a single minimum. In this phase, i.e. for $T < T_c$, the protocol duration is too short to achieve a good fidelity. Nonetheless, in the continuum limit $\delta t \to 0$, there exists a single optimal protocol, although the corresponding maximum fidelity is far from unity. In this overconstrained phase, the optimization problem becomes convex and easy to solve. This is reflected by the observation both the optimal quasi-continuous and bang-band protocols found by the RL agent are nearly identical, cf. Fig. 3. The dynamic character of the phase transition is revealed by a sudden drop in the number of fidelity evaluations $N_{\rm fidevals}$.

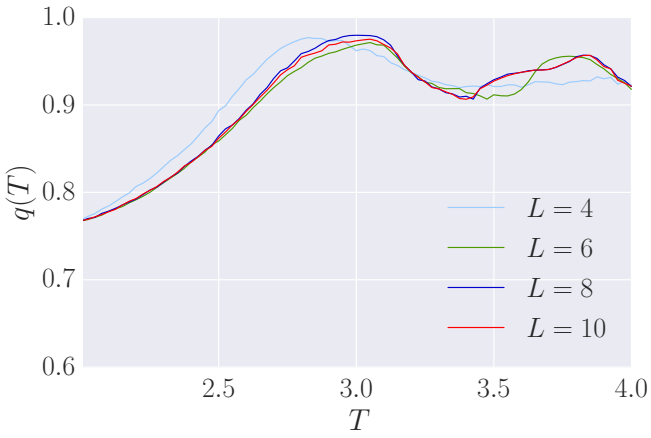
B. Coupled Qubits

One can also ask the question what happens to the quantum control phases in the thermodynamic limit, $L \to \infty$. To this end, we repeat the calculation for a series of chain lengths L. We omit the case $L = 2$, in which the physics of the control problem has a different character, exhibiting spontaneous symmetry breaking in the glasy phase [16]. Due to the non-integrable character of the many-body problem, we are limited to small system sizes. However, Fig. 5 shows convincing data that we capture the behaviour of the system in the limit $L \to \infty$ for the relatively short protocol durations under consideration. Moreover to our surprise the finite-size effects almost entirely disappear for $L \geq 6$ for all range of protocol durations we are considering. It seems that system is able to find an optimal solution, where the information simply does not propagate outside of a very small region and hence the optimal protocol rapidly becomes completely insensitive to the system size.

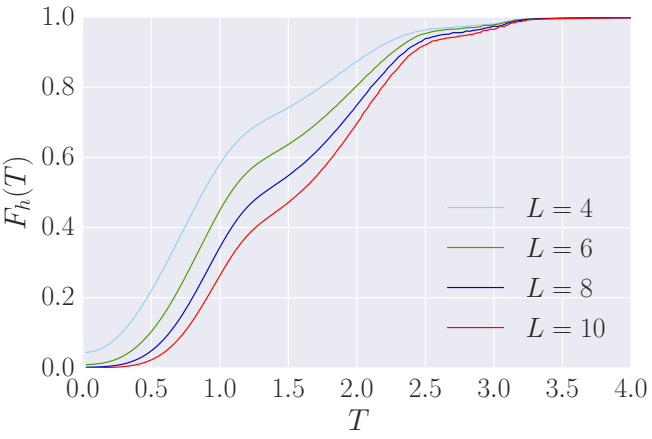
Figure 5c-d shows the system size scaling of the negative logarithmic *many-body* fidelity. While at $L = 4$ we do see remnants of finite-size effects, starting from $L = 6$ the curves are barely changing. A similar rapid system-size convergence is observed also for the order-parameter $q(T)$ (see Fig. 5a-b) and the entanglement entropy of the half chain (Fig. 5e). The protocol step size dependence of the order parameter $q(T)$, the average fidelity evaluations $N_{\rm fidevals}$, and the entanglement entropy $S_{\rm ent}^{L/2}$ of the half-chain are shown in Figs. 4b, 4d and 5f.



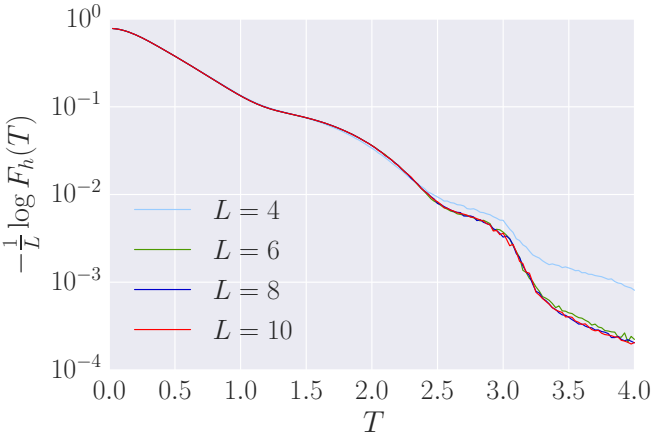
(a) Order parameter $q(T)$ vs. total ramp duration T.



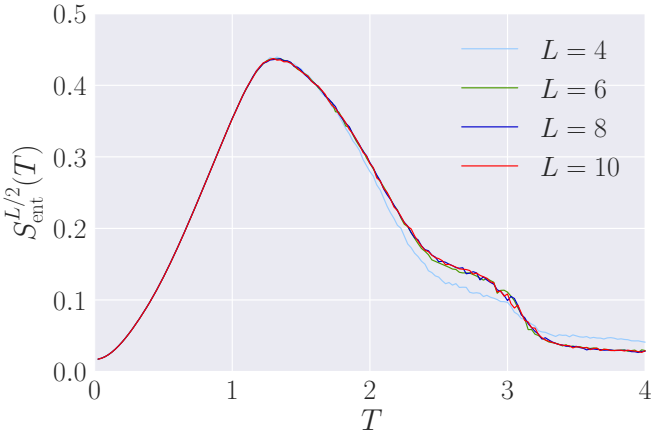
(b) same as (a) with later times zoomed in.



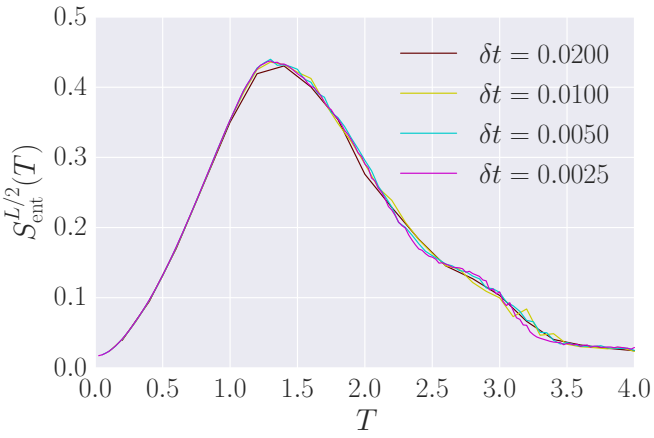
(c) Optimal protocol many-body fidelity: linear scale. Convergence is reached at $L \geq 6$.



(d) Optimal protocol many-body fidelity: logarithmic scale. Convergence is reached at $L \geq 6$.



(e) Entanglement entropy of the half chain as a function of the system size L.



(f) Entanglement entropy of the half chain as a function of the protocol step size δt, for $L = 10$.

FIG. 5: Finite system-size L scaling of the order parameter $q(T)$[top], the many-body fidelity $F_h(T)$ [middle] and the entanglement entropy $S_{\text{ent}}^{L/2}$ for protocol step size $\delta t = 0.0025$.

IV. VIDEO MATERIAL

This paper is accompanied by eight short videos, labeled Video 1 through 8. The videos are available as part of the Supplemental Material published together with the paper. A legend for the videos can be found on https://mgbukov.github.io/RL_movies/.

[1] P. Doria, T. Calarco, and S. Montangero, Phys. Rev. Lett. **106**, 190501 (2011), URL https://link.aps.org/doi/10.1103/PhysRevLett.106.190501.

[2] T. Caneva, T. Calarco, and S. Montangero, Phys. Rev. A **84**, 022326 (2011), URL https://link.aps.org/doi/10.1103/PhysRevA.84.022326.

[3] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrueggen, and S. J. Glaser, Journal of Magnetic Resonance **172**, 296 (2005), ISSN 1090-7807, URL http://www.sciencedirect.com/science/article/pii/S1090780704003696.

[4] S. Machnes, U. Sander, S. J. Glaser, P. de Fouquières, A. Gruslys, S. Schirmer, and T. Schulte-Herbrüggen, Phys. Rev. A **84**, 022305 (2011), URL https://link.aps.org/doi/10.1103/PhysRevA.84.022305.

[5] R. R. Agundez, C. D. Hill, L. C. L. Hollenberg, S. Rogge, and M. Blaauboer, Phys. Rev. A **95**, 012317 (2017), URL https://link.aps.org/doi/10.1103/PhysRevA.95.012317.

[6] P. De Fouquieres, S. Schirmer, S. Glaser, and I. Kuprov, Journal of Magnetic Resonance **212**, 412 (2011), URL https://www.sciencedirect.com/science/article/pii/S1090780711002552?via%3Dihub.

[7] M. Tomka, T. Souza, S. Rosenberg, and A. Polkovnikov, arXiv p. arXiv:1606.05890 (2016), URL http://arxiv.org/abs/arXiv:1606.05890.

[8] D. Sels and A. Polkovnikov, Proceedings of the National Academy of Sciences **114**, E3909 (2017).

[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, Nat Phys **13**, 330 (2017), ISSN 1745-2473, letter, URL http://dx.doi.org/10.1038/nphys3967.

[10] C. Jarzynski, S. Deffner, A. Patra, and Y. Subasi, Phys. Rev. E **95**, 032122 (2017), URL https://link.aps.org/doi/10.1103/PhysRevE.95.032122.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2017), URL https://webdocs.cs.ualberta.ca/~sutton/book/the-book-2nd.html.

[12] T. Castellani and A. Cavagna, Journal of Statistical Mechanics: Theory and Experiment **2005**, P05012 (2005), URL http://stacks.iop.org/1742-5468/2005/i=05/a=P05012.

[13] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002), URL http://science.sciencemag.org/content/297/5582/812.

[14] G. C. Hegerfeldt, Phys. Rev. Lett. **111**, 260501 (2013), URL http://link.aps.org/doi/10.1103/PhysRevLett.111.260501.

[15] M. Kolodrubetz, D. Sels, P. Mehta, and A. Polkovnikov, Physics Reports **697**, 1 (2017), URL https://www.sciencedirect.com/science/article/pii/S0370157317301989?via%3Dihub.

[16] M. Bukov, A. G. R. Day, P. Weinberg, A. Polkovnikov, P. Mehta, and D. Sels, Phys. Rev. A **97**, 052114 (2018), URL https://link.aps.org/doi/10.1103/PhysRevA.97.052114.