Reinforcement learning for autonomous preparation of Floquet-engineered states: Inverting the quantum Kapitza oscillator

Marin Bukov

Reinforcement Learning to Autonomously Prepare Floquet-Engineered States: Inverting the Quantum Kapitza Oscillator

Marin Bukov[1, *]

[1]*Department of Physics, University of California, Berkeley, CA 94720, USA*
(Dated: November 27, 2018)

I demonstrate the potential of Reinforcement Learning (RL) to prepare quantum states of strongly periodically-driven non-linear single-particle models. The ability of Q-Learning to control systems far away from equilibrium is exhibited by steering the quantum Kapitza oscillator to the Floquet-engineered stable inverted position in the presence of a strong periodic drive within several shaking cycles. The study reveals the potential of the intra-period (micromotion) dynamics, often neglected in Floquet engineering, to take advantage over pure stroboscopic control at moderate drive frequencies. Without any knowledge about the underlying physical system, the algorithm is capable of learning solely from tried protocols and directly from simulated noisy quantum measurement data, and is stable to noise in the initial state, and sources of random failure events in the control sequence. Model-free RL can provide new insights into automating experimental setups for out-of-equilibrium systems undergoing complex dynamics, with potential applications in quantum information, quantum optics, ultracold atoms, trapped ions, and condensed matter.

I. INTRODUCTION

The use of strong periodic modulations to design properties of quantum matter is an established approach from the quantum simulation toolbox. Commonly known as *Floquet engineering* [1–3], these ideas prove essential to realize states of matter inaccessible in conventional materials. Prominent achievements include stabilizing unstable equilibria [4–6] [Fig. 1], dynamical localization and the related dynamically-controlled Mott insulator-superfluid transition in ultracold optical lattices [7–9], the emulation of strong artificial gauge fields [10–20], imprinting topological and spin properties into band insulators [21–29], topological defects [30], quantum magnetism [31–34], spin-orbit coupling [35–37], synthetic dimensions [38–40], and photonic topological insulators [41–43].

The current bottleneck in Floquet engineering is caused by detrimental heating effects, due to uncontrolled energy absorption as a result of a proliferation of Floquet many-body resonances, beyond the inverse-frequency expansion [44–46]. The short-time dynamics of weakly-interacting bosons was shown to be dominated by parametric resonances [47–51]. Fermi's Golden rule was applied to the long-time evolution, and leads to a featureless infinite temperature state [52–54]. Theoretically, for non-integrable many-body lattice systems with bounded on-site Hilbert space dimension, heating was proven to be (at least) exponentially suppressed in the drive frequency [55–57], which allows for the formation of transient long-lived prethermal steady states, ideally suited for Floquet engineering [58–61].

Knowing how to Floquet-engineer an exponentially long-lived state of matter leaves the important open
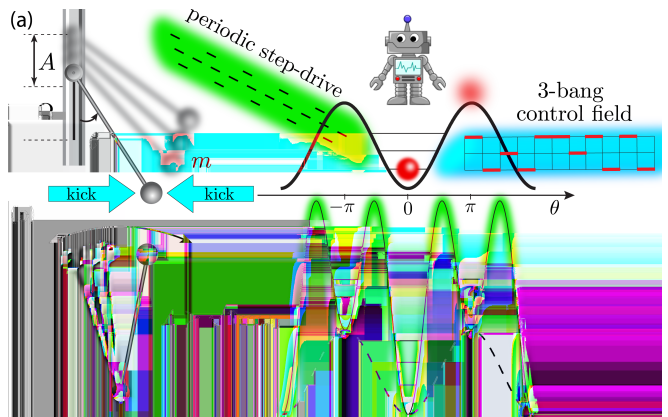

FIG. 1: Prototypical example of Floquet engineering: high-frequency periodic modulation stabilizes the metastable equilibrium at the inverted position of a classical pendulum (left) and a quantum oscillator (right). (a) schematic Floquet control setup displaying the step-periodic drive (green), and the control protocol (cyan). The purpose of this work is to prepare states localized at the inverted position (b) in the presence of strong periodic modulation without knowledge about the physical system, using Reinforcement Learning [Video 1].

problem of how to steer the system in the desired target state. The state-of-the-art approach to manipulate periodically-driven systems is the adiabatic variation of parameters [62–65]. While drive-induced photon absorption avoided crossings in the quasienergy spectrum need to be passed quickly in order to avoid spending much time on resonance and heating up, the state should go slowly, compared to the inverse energy gap, through conventional avoided crossings to suppress excitations [63]. This tension leads to the breakdown of Floquet adiabatic perturbation theory [63, 66, 67], despite the existence of parametrically-controlled windows of applicability, and points towards the necessity to develop new approaches for Floquet control in both single-particle and

---

*Electronic address: mgbukov@berkeley.edu

many-body systems.

Reinforcement Learning (RL) [68] is one of the most promising techniques in Machine Learning [69–71], closely related to optimal [72–79], feedback [80–84] control, and evolutionary algorithms used in quantum chemistry and optics to learn molecular control [85–99]. It is especially well-suited to *autonomously* control systems in the presence of strong drives since it is model-free and robust to imperfections and noise. In physics, RL has been used to navigate thermals [100] and turbulent flows [101], design experiment setups in quantum optics [102], construct molecules with prescribed properties [103], and to control quantum systems [104–111]. Without a physical model, RL was shown to produce comparable results to algorithms from optimal control [106]. Ideas from quantum physics have been suggested to improve RL-related algorithms [112–114]. Despite recent progress, RL's potential remains massively unexplored in physics.

Inspired by Ref. [106], the present work demonstrates the suitability of RL to study the nonequilibrium quantum dynamics of strongly-driven Floquet-engineered states. In a numerical simulation of a quantum experiment, starting with zero knowledge about the system, the RL agent learns how to optimally prepare inverted position states in the Kapitza pendulum from tried protocol configurations [see Video 1]. The algorithm is applied to both the quantum and the classical oscillator. Unlike Ref. [106], the agent learns from quantum (i.e. non-deterministic) measurement data, and is shown to remain robust after adding noise to the initial state, and occasional random failure events in the control sequence. The study shows the advantage of exploiting the micromotion dynamics for control to achieve higher fidelities compared to stroboscopic control.

II. FLOQUET CONTROL PROBLEM

Consider the Hamiltonian of the horizontally kicked quantum Kapitza oscillator

$$
\begin{aligned}
H(t) &= H_0 + H_{\text{drive}}(t) + H_{\text{control}}(t), \\
H_0 &= \frac{p_\theta^2}{2m} - m\omega_0^2 \cos\theta, \\
H_{\text{drive}}(t) &= -\frac{A}{2m}\text{sign}(\cos\Omega t)\left(\sin\theta\, p_\theta + p_\theta \sin\theta\right) \\
&\quad -\frac{A^2}{8m}\left(1 - \text{sign}(\sin 2\Omega t)\right)\cos 2\theta, \\
H_{\text{control}}(t) &= h(t)\sin\theta,
\end{aligned}
\tag{1}
$$

where m and ω_0 are the mass and natural frequency of the oscillator H_0 with position (angle) and (angular) momentum variables obeying $[p_\theta, \theta] = -i$. Applying a strong vertical periodic drive of constant amplitude A and frequency $\Omega = 2\pi/T$ is known to stabilize the metastable inverted position at $\theta = \pi$, a paradigmatic example of Floquet engineering [2] [Fig. 1]. The latter requires the Floquet drive to couple strongly to the system, with an amplitude scaling linearly with Ω in the lab frame [2]. To see how stabilization occurs, and to avoid working at large amplitudes, it is advantageous to adopt the rotating frame description (1), at the expense of introducing a second harmonic in $H_{\text{drive}}(t)$ [cf. App. A]. Therefore, the piece-wise constant drive, designed to speed up numerical simulations, repeats every four steps. The units are chosen such that p_θ, θ, $m\omega_0$, A are all dimensionless, and $\hbar = 1$. Energy is measured in units of ω_0.

For $h(t) \equiv 0$, the dynamics of the uncontrolled Kapitza oscillator at integer multiples of the drive period T (i.e. stroboscopically) is exactly described by the Floquet Hamiltonian $H_F(\Omega)$. In the infinite-frequency limit, taking the period-average of Eq. (1) gives

$$
H_F(\Omega \to \infty, h \equiv 0) = \frac{p_\theta^2}{2m} - m\omega_0^2 \cos\theta - \frac{A^2}{8m}\cos 2\theta. \tag{2}
$$

The periodic drive renormalizes the potential energy of the oscillator [Fig. 1b] and, whenever $A \geq \sqrt{2}m\omega_0$, the potential supports a stable equilibrium at $\theta = \pi$ with frequency of harmonic oscillation $\omega' = \sqrt{A^2/(2m^2) - \omega_0^2}$. Away from the limit $\Omega \to \infty$, finite-frequency corrections can be incorporated using the inverse-frequency expansion [2], yet the exact form of $H_F(\Omega)$ remains unknown. This leads to a modification of the critical amplitude, yet the stabilizing behavior persists qualitatively down to $\Omega \gtrsim 6\omega_0$ for the step-drive.

Turning on the control field $h(t)$ compromises the time-periodicity of $H(t)$, and one can in general no longer rely on Floquet theory. The unknown control field $h(t) \in \{0, \pm 4\}$, $t \in [0, t_f]$, of duration $t_f = N_T T$ is built from a sequence of constant horizontal momentum kicks of duration δt exerted on the oscillator, called bangs, to speed up the efficiency of the RL algorithm. The bounded kick strength reflects possible constraints in experiments with too large control fields; the exact values $\{\pm 4\}$ are chosen to be on the same order of magnitude as the bare oscillator frequency, so that no term dominates the Hamiltonian and the dynamics cannot be studied using perturbation theory. The 0-bang allows to turn off the control field. I further consider drive-commensurate protocols of two types: (i) stroboscopic: $T = \delta t$, and (ii) commensurate non-stroboscopic, $T = 4n\delta t$, $n \in \mathbb{N}$ [Fig. 1]. To control the Kapitza oscillator, non-stroboscopic protocols are chosen $N_T = 15$-driving-cycles long, with 8 steps per cycle (120 bangs). The protocol space contains $3^{120} \sim 10^{57}$ configurations.

The objective of this study is to determine a bang-bang protocol $h(t)$ which finds the system in the ground state $|\psi_i\rangle$ of the non-driven uncontrolled Hamiltonian H_0, and brings it as close as possible to the target state $|\psi_*\rangle$ – the eigenstate of the finite-frequency $H_F(\Omega)$ localized at the inverted position, in a fixed amount of time. The figure of merit is the fidelity $F_h(t_f) = |\langle\psi(t_f)|\psi_*\rangle|^2$ of being in the target state at the end of the control sequence. Note that the amplitude A of the instantly turned on periodic drive is held constant during control and, therefore, there is no natural adiabatic path [in h-space] between the initial
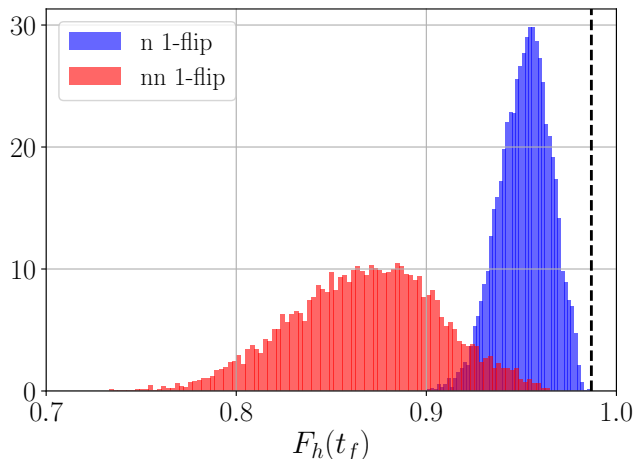
FIG. 2: Distributions of the nearest (n) [i.e. $\{-4, 0\}$ and $\{0, +4\}$] and next-nearest (nn) [$\{-4, +4\}$] 1-flip excitations of the SD protocols [see text]. The dashed vertical line marks the fidelity of the absolute maximum. The data refers to the sample of all local protocol minima which satisfy $F_h(t_f) > 98\%$]. The model parameters are the same as in Fig. 3.

and target states. There is no small parameter to do perturbation theory in, either.

III. COMPLEXITY OF THE DETERMINISTIC KAPITZA CONTROL PROBLEM

At present, I am not aware of an analytical theory to find the optimal protocol or even predict its fidelity in the Kapitza control problem. It is, therefore, advantageous to acquire some intuition about the properties of the control landscape [115], which has recently been shown to exhibit a variety of phase transitions [106, 116, 117] including glassy control phases [118] and symmetry breaking [119] in non-Floquet control problems.

To begin with, a quick check shows that using random protocol sequences leads to about 10% fidelity [App. C] – this is expected to reflect the performance of the RL agent during the first episodes of training, when it is still unfamiliar with the behavior of the physical system.

To get an estimate of the magnitude of the reachable fidelities for the chosen set of model parameters, I use Stochastic Descent (SD), initiated from random protocol configurations, and flip the protocol bangs randomly one at a time, until a local minimum of the infidelity landscape is reached [106, 118]. Such minima represent locally optimal protocols, close to which greedy optimization algorithms are likely to get stuck, due to the glassy character of control landscapes [118]. The obtained sample of 10^6 SD-protocols has mean fidelity 87% [App. C]. Out of these, ninety-two local minima protocols have fidelity greater than 98%, with the absolute maximum at 98.68%. The bang sequences of the corresponding protocols are, however, completely different which suggests

that they may occupy deep pockets far way in the control landscape (w.r.t. the Hamming distance) [115].

To test this assertion, starting from each one of the ninety-two best SD minima, I compute the fidelities of all 1-flip variations [called excitations], which fall in two categories: nearest (n) flips are those between protocol values $\{-4, 0\}$ and $\{0, +4\}$, while next-nearest (nn): $\{-4, +4\}$. Figure 2 shows the distributions of one-flip excitations. Intuitively, one expects that the wavefunction cannot undergo drastic changes within the short kicks of strength ($|h|\delta t = 0.314$) during a single bang. Nonetheless, on average the fidelity of 1-flip excitations drops by more than 10%, which points at the rugged profile of the Floquet control landscape. The total number of fidelity evaluations required to obtain the local SD-minima sample is about 10^8 [10^6 runs, with on average 10^2 evaluations each]. SD is a deterministic algorithm [i.e. relies on the exact fidelities to operate]. In this work, I present an *autonomous* RL algorithm which can be coupled to realistic experiments with multiple generic sources of noise.

Challenges in Autonomous Quantum Control.—All information about a quantum state is encoded in its wavefunction. Hence, when developing algorithms for realistic experimental setups, the main challenge for autonomous control arises from the lack of access to the quantum states which are *unmeasurable* mathematical constructs. Invoking Picard-Lindelöf's uniqueness theorem to find the evolution operator $U_h(t, 0)$ which integrates Schrödinger's equation, given a fixed initial state $|\psi_i\rangle$, one can parametrize the accessible states at later times $|\psi(t)\rangle$ by the protocol sequence $h(t)$ up to time t:

$$\begin{aligned} \{|\psi(t)\rangle = U_h(t, 0)|\psi_i\rangle\} &\Leftrightarrow \\ \{h(t') : H[h(t')], |\psi_i\rangle, t' \in [0, t]\} &= \mathcal{S}. \end{aligned} \quad (3)$$

Although this mapping is not one-to-one [a state may be accessible using different protocols], it offers a significant advantage: while quantum states cannot be measured, the applied protocols $h(t)$ can actually be kept track of. Hence, fixing the initial state, one can infer which state the system ought to be in at a later time, from the applied protocol. In fact, the minimal amount of information an autonomous algorithm can have about the controlled quantum system is the applied protocol sequences.

Another challenge in experimental quantum control is the nondeterministic character of projective quantum measurements. Since they destroy the state, one is allowed to measure only once, at the end of the protocol when the system has evolved into the state $|\psi(t_f)\rangle$. Projective measurements are modeled to return 1 with probability set by $F_h(t_f) = |\langle\psi(t_f)|\psi_*\rangle|^2$ if the system is found in the target state $|\psi_*\rangle$, and 0 otherwise [App. B]. Therefore, the algorithm seeks to maximize the fidelity $F_h(t_f)$, whose true value remains unknown at the time of measurement, and is only estimated from the data. During the optimization process, the situation is in fact much more complicated since, as the available protocol family is explored, the true probability $F_h(t_f)$ to determine the

measurement, changes from one protocol to the next, as different protocols in general lead to different final states.

Further challenges arise from (i) the inability to prepare the initial state $|\psi_i\rangle$ with certainty. Experiments only prepare the desired initial state within a given fidelity window. Additionally, (ii) in experiments one has to account for the occasional failure of the control apparatus: even though the algorithm may have requested the protocol sequence $h(t)$, the physical system might experience a slightly modified protocol $h'(t)$ instead. Depending on the sensitivity of the optimal solution, this could lead to drastic changes in the reachable fidelities [106, 118] [see Fig. 2]. Every experiment, (iii), comes with its own imperfections and difficulties: the Hamiltonian $H(t)$, assumed to model the system, is often merely a simplified approximation, which compromises the unconditioned applicability of idealized simulations. Finally, even if all of the above were not present, (iv) there are various constraints imposed by the dynamics of the physical system of interest. In this respect, controlling Floquet systems remains an unsolved challenging problem of nonequilibrium dynamics. It is, thus, desirable to have a versatile algorithm which is capable of dealing with the above scenarios in an efficient way.

IV. REINFORCEMENT LEARNING

In this work, I adopt a modification of Watkins Q-Learning algorithm [120] to study the Kapitza control problem. An RL agent seeks to find an optimal protocol $h(t)$ to prepare the target state $|\psi_*\rangle$, *without knowledge* about the controlled system [138]. To do this, it episodically gains experience and uses it by taking actions to construct bang-bang protocols one bang at a time. These protocols are applied to a simulated quantum system (the environment), returning a reward to the agent: the estimated fidelity $F_h(t_f)$, computed based on the measurement record obtained so far. As the number of training episodes increases, the algorithm progressively correlates protocols with their fidelity, a process referred to as *learning*. Even though in a simulation it is possible to learn from the exact fidelities, in order to better simulate a realistic quantum experiment, for each protocol $h(t)$, the algorithm stores two integers, corresponding to the number m of protocol encounters, and the number n of 1's in the output of the binary quantum measurement. From them, it computes the reward r estimating the mean current fidelity $r = n/m$ of being in the target state (for a given protocol). The error estimate E to be within the 2σ window, controls the number of repetitions used to gather measurement statistics. Hence, during the learning process, the agent has to deal with noisy rewards. These intrinsically quantum features obscure the learning process significantly close to the best attainable fidelities, where it is known that differences in the higher decimal places of the fidelity can play a decisive role [118].

The information the agent has about the controlled system is encoded in the RL state space $\mathcal{S}$, which I define using the correspondence (3). For instance, for a 4-step-long protocol $\{+4\}, \{+4, -4\}, \{+4, -4, 0\}, \{+4, -4, 0, 0\} \in \mathcal{S}$ are all admissible RL states. Note that the size of $\mathcal{S}$ scales exponentially with the number of bangs. While this aggravates the exploration of the state space, it is well within the scope of RL algorithms to learn nearly-optimal policies in complex environments with exponentially-large state spaces, as becomes clear from recent success in mastering video and board games beyond human level [121, 122]. Importantly, this scaling does *not* depend on the Hilbert space dimension of the quantum system which makes the algorithm applicable to large systems [one does need an experiment to simulate their dynamics to provide rewards, though]. In this respect, the RL algorithm is modular: the learning part [which can be chosen insensitive to the Hilbert space dimension] is separate from the quantum mechanics part [which provides the rewards and in a simulation would suffer from the limitations due to large Hilbert space dimensions]. To construct protocols on-the-fly, every time step the agent invokes its knowledge gathered so far to pick an action from the set $\mathcal{A} = \{-4, 0, +4\}$ based on the predicted expected reward. I use an ε-greedy exploration policy, which is attenuated exponentially with the number of episodes [68]. Finally, the reward space is $\mathcal{R} = \{r \in [0, 1] : r = n/m\}$. The algorithm also applies experience replays to enforce exploration around the estimated best-encountered protocol, cf. App. G.

V. AUTONOMOUSLY INVERTING THE QUANTUM KAPITZA OSCILLATOR

The Q-Learning agent is first trained for 10^5 exploration episodes, followed by 10^3 greedy test episodes to examine the stability of the learning process. Due to the probabilistic character of ε-greedy exploration, the algorithm is run for 100 distinct seeds of a pseudo-random number generator, and the results I present show averages, cf. App. E 2. Figure 3a shows that the fidelity $F_h(t_f)$ of being in the target state increases consistently and then saturates at about 80%. Hence, the RL agent is capable of autonomously controlling the Kapitza oscillator by using only information from noisy quantum measurements. Note that the agent slightly overestimates the true fidelity (blue). Yet, it learns the correct noise correlations in the reward [test stage: blue and red shaded areas].

To demonstrate the robustness of RL to noise in the initial state, I draw a Haar-random state $|\phi\rangle$, and consider the noisy initial state $|\psi_i(\eta)\rangle \propto |\psi_i\rangle + \eta|\phi\rangle$, with $\eta = 0.31$ such that $|\langle\psi_i|\psi_i(\eta)\rangle|^2 \approx 0.9$. This results in a small drop of fidelity to about 73%, parametrically controlled by η [Fig. 3b]. Therefore, Q-Learning is stable to small perturbations in the initial condition. Additionally, I expose the RL agent to a stochastic environment,
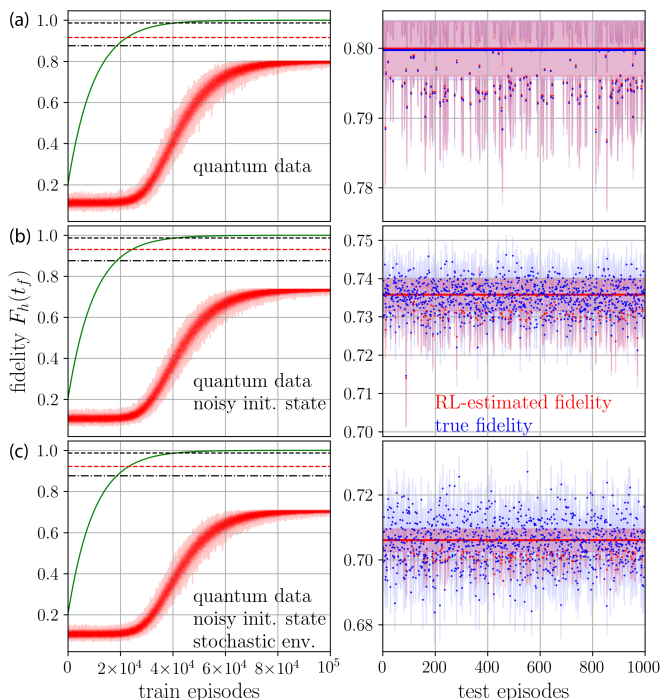
FIG. 3: Running fidelity estimate during Train (left) and Test (right) stages of Q-Learning for the quantum Kapitza oscillator. The three rows show data for stochastic rewards only (top), stochastic rewards and noisy initial condition (middle), and stochastic rewards, noisy initial condition and stochastic environment (bottom). The horizontal black lines show the average fidelity of local infidelity minima computed using one-flip SD (dashed dotted), and the best SD realization (dashed) out of 10^6 local minima samples obtained with *deterministic cost function*. The horizontal red dashed line shows the best-encountered protocol using RL during the Train stage. The RL data points (red) are averaged over 100 seed realizations of the pseudo-random number generator, with the uncertainty window (shaded area) computed using a bootstrapping approach. The green curve shows the exponentially attenuated exploration schedule $\varepsilon(n_{\rm ep})$, normalized within $[0, 1]$ for display purposes: unity corresponds to no exploration. The oscillator parameters are $N_T = 15$ periods with 8 steps each, $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$.

in which every bang is randomly replaced by any of the three available actions with probability $\zeta = 1/120$, mimicking occasional spontaneous failure in the control apparatus. The value $\zeta = 1/120$ is chosen so that one bang of the 120-bang-long control sequence fails on average. This scenario, similar to gate failure in quantum computing, expectedly leads to a further reduction to about 71% [Fig. 3c]. Hence, RL is also capable of learning in stochastic quantum environments. Interestingly, the agent is capable of de-noising the experimental rewards, as indicated by the uniform envelope of the estimated (red) compared to the true (blue) fidelity fluctuations.

To investigate the ability of the algorithm to navigate noisy environments, I also trained the agent in a noise-less deterministic environment, but tested it in a noisy

stochastic setup, cf. App. E 1. The test-stage true fidelity fluctuates about the same value, as if the agent was trained in a noisy stochastic environment. This behavior likely originates from the intrinsic exploration noise built in ε-greedy policy. The inability to exploit the knowledge in the presence of initial state noise is presumably a consequence of the RL state space definition (3). This reveals a potential drawback: if the system is initiated in a sufficiently different state, the gained knowledge is not immediately exploitable, and the agent takes time to explore again [App. E 1]. This is, however, a feature of the current choice for the state-action space, rather than of the algorithm.

I could not distinguish any significant features in the best protocol sequences [cf. App. C], which suggests that the bang-bang family might not be naturally suitable for Floquet control problems. Nonetheless, one can visualize the dynamics of the real-space probability distribution of the oscillator. I consider three stages: (i) the Floquet system is subject to the best RL protocol in the presence of the Floquet drive. Once the control stage is over, (ii) I keep the Floquet-drive on with $h \equiv 0$, before (iii) both the Floquet drive and the control are turned off ($h \equiv 0$, $A = 0$) and the system evolves under H_0, see Video 1 [best-encountered RL] and Video 2 [best SD 1-flip local minimum]. Once can observe the complexity of preparing entire local probability distributions with a single global control field, as becomes clear from the large fluctuations in fidelity between the short bangs. Note that the RL agent seems to first push the real-space weight of the wavefunction clockwise [Video 1], before the final state is eventually reached from the opposite counter-clockwise direction. This is reminiscent of the classical problem with scarce control resources (mountain car paradigm in RL [68]) where, in the short time available, one might decide to push the pendulum one way to convert energy from the drives into potential energy which, with the help of gravity, can then be unleashed to reach the inverted position from the other side. The quantum character of the dynamics likely determines the precise nontrivial bang sequence to keep the structure of the wavepacket during the evolution. This classical-like behavior is intriguing, since the quantum nature of the dynamics is clearly exhibited during stage (iii), where the overlap with the target state remains large even when the oscillator is not controlled or driven, as a consequence of $|\psi_*\rangle$ having a finite overlap with the excited eigenstate of H_0 corresponding to the meta-stable classical inverted state.

Last, the study also confirms the clear superiority of non-stroboscopic Floquet control (8 bangs per cycle) over stroboscopic control (1 bang per cycle). Figure 4 shows the learned saturation fidelity for several moderate frequencies in both cases. As expected, since the stroboscopic protocols can also be viewed as non-stroboscopic ones, the stroboscopic optimal fidelity is always a lower bound on the non-stroboscopic one. However, the smaller stroboscopic family ($3^{15} \sim 10^7$ protocols) can be explored more efficiently which leads to seemingly better RL per-
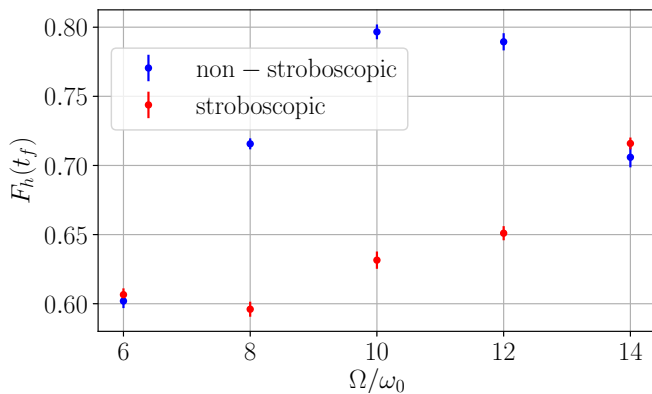
FIG. 4: RL agent performs better in non-stroboscopic (8 bangs per cycle) than stroboscopic (1 bang per cycle) control at moderate drive frequencies. See Fig. 3 for the parameters.

formance at high drive frequencies. Even though non-strobscopic dynamics is often neglected in Floquet engineering, it can offer advantages in Floquet control, and should not be easily dismissed [123, 124].

VI. DISCUSSION

The best SD-protocol out of the family of local minima has 98.6% fidelity [Fig. 3, dashed black line], better than the average learned RL-fidelity. It is also superior to the best protocol encountered by the RL agent during training at 91.6% [Fig. 3, dashed red line]. Why could the agent not learn any of these protocols? One possibility is that it estimated their true fidelities poorly, and decided to ignore them. Another suggests the existence of very deep pockets in the infidelity landscape, which are unstable to noise, naturally present in the exploration schedule [Fig. 2]. RL is designed to find stable solutions, even if they are further from the global minimum as measured by the cost function. Last but not least, the RL fidelities do improve with increasing the number of training episodes to 10^6 – still a tiny fraction of the total RL state space size, cf. App. E 3. In this respect, notice that the data in Fig. 3 is shown for 10^5 fidelity evaluations, as opposed to 10^8 evaluations for SD [139].

The control setup considered in this paper comes in contrast to typical Floquet control problems, which support an adiabatic path in parameter space between the initial and target states, e.g. by slowly turning on the drive amplitude. As a result, comparing the numerically obtained protocols to analytical predictions is a formidable challenge. However, this challenge does not arise from the specific nonadiabatic setup alone. Even though adiabatic perturbation theory has been extended to periodically-driven systems, Floquet resonances, leading to gaps in the quasienergy spectrum along the adiabatic trajectory, are known to result in the breakdown of Floquet adiabaticity [63, 66, 67]. In contrast, this is not a problem in static (i.e. non-Floquet) systems, where

RL has been applied to a many-body spin chain to prepare ground states, adiabatically connected by the control field [106]: it was found that, at short durations, the functional form of optimal protocols differs significantly from that of adiabatic protocols, but can still be understood within the analytical framework of shortcuts to adiabaticity. Extending such ideas to Floquet systems is, to be best of my knowledge, an open problem, where RL and optimal control could provide useful insights.

One might also raise a valid objection that experiments currently cannot project the system to exact Floquet eigenstates, and hence using this particular target state makes the study not immediately applicable to realistic experimental setups. It is, however, possible to target a quasi-Gaussian state, localized at the inverted position, $\langle\theta|\psi_*\rangle \propto \exp(-(\omega')^{1/4}\cos\theta)$. This gives qualitatively similar results, cf. App. D.

It is also important to mention that there exist alternative algorithms that can be used to study Floquet control. Examples include GRAPE [75], CRAB [76], VQE [78, 79, 125], QAOA [126–128], and Lyapunov-based feedback control [129–132]. Whereas some of them require to have a model for the system under control, RL is model-free and can be applied in situations where the Hamiltonian of the system (more generally, the dynamics of the environment) is unknown. In the current study, I make use of a model only to provide the training data.

Quite generally, bang-bang protocols also come with an experimental limitation, posed by the bandwidth of pulse generators. Even though they constitute a convenient theory starting point, it would be interesting to parametrize the protocols by Fourier components, an idea underlying the CRAB algorithm [76]. Such protocols can be resonant with the drive, and the corresponding control process likely admits a Floquet description. I should emphasize that the use of bang-bang protocols is *not* at all a requirement imposed by RL algorithms, and certain RL algorithms (e.g. Policy Gradient) can even be applied to learn continuous control fields.

The major bottleneck in using RL to control realistic experiments is set by sample efficiency. In the present implementation, I repeat each protocol 100 times [not shown in Fig. 3] to estimate its fidelity from the quantum measurement data. While this slows down the learning process, I stress that this is an intrinsic feature of all quantum measurements, unrelated to RL. In a large class of platforms, such as cold atoms, this can be alleviated by measuring multiple copies of the system simultaneously. Theoretically, the problem could also be mitigated by employing techniques from statistical inference, or a suitable pre-training procedure.

Even though the obtained fidelities are model-dependent and do not carry over to other control problems, the Q-Learning algorithm is universal in the sense that the RL agent, starting with no prior knowledge, learns only from its actions [protocols] and the stochastic reward. This is analogous to playing video or board games without seeing the game configuration, but only

the (noisy) score. Hence, Q-Learning is agnostic on the fine details of the controlled system and can be applied to any model, even classical ones, as I demonstrate using the classical Kapitza pendulum, see Video 3 and App. F.

VII. OUTLOOK

Every experiment comes with its own imperfections which obscure the physics of interest. Building a theory to describe them all in detail is often a formidable setup-dependent task, and requires considerable efforts. In the era of machine learning and automation, it is desirable to develop autonomous algorithms to delegate the tedious task of exploring the fine details of experiments to computers, and RL emerges as a natural candidate. It is currently an open question whether experimental imperfections can be turned into features, and exploited for the purpose of control.

In this respect, RL presents a set of promising algorithms, capable of simultaneously dealing with various sources of uncertainty and noise, even in highly complex far-from-equilibrium scenarios with no available analytical description. Out of a variety of RL algorithms [68], it is not clear which ones are best suited for controlling quantum systems away from equilibrium. In the current study, I chose Q-Learning, because it is *off-policy*, i.e. one can use data, generated when the policy of the agent was suboptimal, to improve the current policy. Further advantages are expected to be offered by Deep Learning [107, 108, 110, 133, 134], especially in the search of an efficient compressed representation of the state-action space, which is one way to incorporate continuous pro-

tocols. Deep Learning allows the agent to generalize and evaluate the value of previously unseen protocols, but also brings in difficulties associated with uncontrolled approximations and the absence of convergence guarantees for the algorithm. I verified that the tabular Q-Learning algorithm used in this paper is convergent.

This work represents a pioneering step in introducing RL to control quantum systems far away from equilibrium. Whereas it is difficult to a priori assess the suitability of RL for nonequilibrium many-body control, recent work successfully applied RL algorithms to control *static* [i.e. non-Floquet] chaotic many-body spin chains [106, 107]. The higher complexity of many-body control may as well require to cast the RL problem as a partially-observable Markov decision process. While still at the beginning of this quest, the present proof-of-principle theoretical study already hints towards the applicability of RL to a large class of problems in quantum dynamics, and will hopefully spawn more research in this exciting new direction.

[1] N. Goldman and J. Dalibard, Phys. Rev. X **4**, 031027 (2014).

[2] M. Bukov, L. D'Alessio, and A. Polkovnikov, Advances in Physics **64**, 139 (2015).

[3] A. Eckardt, Rev. Mod. Phys. **89**, 011004 (2017).

[4] P. L. Kapitza, Soviet Phys. JETP **21**, 588 (1951).

[5] A. Lerose, J. Marino, A. Gambassi, and A. Silva, arXiv preprint arXiv:1803.04490 (2018).

[6] R. Citro, E. G. Dalla Torre, L. DAlessio, A. Polkovnikov, M. Babadi, T. Oka, and E. Demler, Annals of Physics **360**, 694 (2015).

[7] A. Eckardt, C. Weiss, and M. Holthaus, Phys. Rev. Lett. **95**, 260404 (2005).

[8] C. Sias, H. Lignier, Y. P. Singh, A. Zenesini, D. Ciampini, O. Morsch, and E. Arimondo, Phys. Rev. Lett. **100**, 040404 (2008).

[9] A. Zenesini, H. Lignier, O. Morsch, D. Ciampini, and E. Arimondo, Laser Physics **20**, 1182 (2010).

[10] D. Jaksch and P. Zoller, Ann. Phys. **315**, 52 (2005).

[11] J. Struck, C. Ölschläger, R. Le Targatn, P. Soltan-Panahi, A. Eckardt, M. Lewenstein, P. Windpassinger, and K. Sengstock, Science **333 (6045)**, 996 (2011).

[12] J. Struck, C. Ölschläger, M. Weinberg, P. Hauke, J. Si-

monet, A. Eckardt, M. Lewenstein, K. Sengstock, and P. Windpassinger, Phys. Rev. Lett. **108**, 225304 (2012).

[13] P. Hauke, O. Tieleman, A. Celi, C. Ölschläger, J. Simonet, J. Struck, M. Weinberg, P. Windpassinger, K. Sengstock, M. Lewenstein, and A. Eckardt, Phys. Rev. Lett. **109**, 145301 (2012).

[14] J. Struck, M. Weinberg, C. Ölschläger, P. Windpassinger, J. Simonet, K. Sengstock, R. Höppner, P. Hauke, A. Eckardt, M. Lewenstein, and L. Mathey, Nature Physics **9**, 738 (2013).

[15] M. Aidelsburger, M. Atala, M. Lohse, J. T. Barreiro, B. Paredes, and I. Bloch, Phys. Rev. Lett. **111**, 185301 (2013).

[16] H. Miyake, G. A. Siviloglou, C. J. Kennedy, W. C. Burton, and W. Ketterle, Phys. Rev. Lett. **111**, 185302 (2013).

[17] M. Atala, M. Aidelsburger, M. Lohse, J. T. Barreiro, B. Paredes, and I. Bloch, Nature Physics **10**, 588 (2014).

[18] N. Goldman, G. Juzeliunas, P. Ohberg, and I. B. Spielman, Rep. Prog. Phys. **77**, 126401 (2014).

[19] C. J. Kennedy, W. C. Burton, W. C. Chung, and W. Ketterle, Nature Physics **11**, 859 (2015).

[20] P. Roushan, C. Neill, A. Megrant, Y. Chen, R. Babbush, R. Barends, B. Campbell, Z. Chen, B. Chiaro, A. Dunsworth, *et al.*, Nature Physics **13**, 146 (2017).

[21] T. Oka and H. Aoki, Phys. Rev. B **79**, 081406 (2009).

[22] T. Kitagawa, T. Oka, A. Brataas, L. Fu, and E. Demler, Phys. Rev. B **84**, 235108 (2011).

[23] M. S. Rudner, N. H. Lindner, E. Berg, and M. Levin, Phys. Rev. X **3**, 031005 (2013).

[24] G. Jotzu, M. Messer, R. Desbuquois, M. Lebrat, T. Uehlinger, D. Greif, and T. Esslinger, Nature **515**, 237 (2014).

[25] M. Aidelsburger, M. Lohse, C. Schweizer, M. Atala, J. T. Barreiro, S. Nascimbène, N. R. Cooper, I. Bloch, and N. Goldman, Nature Physics **11**, 162 (2015).

[26] N. Fläschner, B. S. Rem, M. Tarnowski, D. Vogel, D.-S. Lühmann, K. Sengstock, and C. Weitenberg, Science **352**, 1091 (2016).

[27] M. Tarnowski, F. N. Ünal, N. Fläschner, B. S. Rem, A. Eckardt, K. Sengstock, and C. Weitenberg, arXiv preprint arXiv:1709.01046 (2017).

[28] M. Aidelsburger, S. Nascimbene, and N. Goldman, Comptes Rendus Physique (2018).

[29] G. Jotzu, M. Messer, F. Görg, D. Greif, R. Desbuquois, and T. Esslinger, Phys. Rev. Lett. **115**, 073002 (2015).

[30] M. Tarnowski, M. Nuske, N. Fläschner, B. Rem, D. Vogel, L. Freystatzky, K. Sengstock, L. Mathey, and C. Weitenberg, Phys. Rev. Lett. **118**, 240403 (2017).

[31] A. Eckardt, P. Hauke, P. Soltan-Panahi, C. Becker, K. Sengstock, and M. Lewenstein, EPL **89**, 10010 (2010).

[32] J. Mentink, K. Balzer, and M. Eckstein, Nature communications **6**, 6708 (2015).

[33] U. Meyer, G. Haack, C. Groth, and X. Waintal, Phys. Rev. Lett. **118**, 097701 (2017).

[34] F. Görg, M. Messer, K. Sandholzer, G. Jotzu, R. Desbuquois, and T. Esslinger, Nature **553**, 481 (2018).

[35] B. M. Anderson, I. B. Spielman, and G. Juzeliunas, Phys. Rev. Lett. **111**, 125301 (2013).

[36] V. Galitski and I. B. Spielman, Nature **494**, 49 (2013).

[37] K. Jiménez-García, L. J. LeBlanc, R. A. Williams, M. C. Beeler, C. Qu, M. Gong, C. Zhang, and I. B. Spielman, Phys. Rev. Lett. **114**, 125301 (2015).

[38] A. Celi, P. Massignan, J. Ruseckas, N. Goldman, I. B. Spielman, G. Juzeliunas, and M. Lewenstein, Phys. Rev. Lett. **112**, 043001 (2014).

[39] B. Stuhl, H.-I. Lu, L. Aycock, D. Genkina, and I. Spielman, Science **349**, 1514 (2015).

[40] M. Mancini, G. Pagano, G. Cappellini, L. Livi, M. Rider, J. Catani, C. Sias, P. Zoller, M. Inguscio, M. Dalmonte, *et al.*, Science **349**, 1510 (2015).

[41] M. C. Rechtsman, J. M. Zeuner, Y. Plotnik, Y. Lumer, D. Podolsky, F. Dreisow, S. Nolte, M. Segev, and A. Szameit, Nature **496**, 196 (2013).

[42] M. Hafezi, Phys. Rev. Lett. **112**, 210405 (2014).

[43] S. Mittal, J. Fan, S. Faez, A. Migdall, J. M. Taylor, and M. Hafezi, Phys. Rev. Lett. **113**, 087403 (2014).

[44] M. Bukov, M. Heyl, D. A. Huse, and A. Polkovnikov, Phys. Rev. B **93**, 155132 (2016).

[45] P. W. Claeys, S. De Baerdemacker, O. E. Araby, and J.-S. Caux, Phys. Rev. Lett. **121**, 080401 (2018).

[46] S. E. T. Arze, P. W. Claeys, I. P. Castillo, and J.-S. Caux, arXiv preprint arXiv:1804.10226 (2018).

[47] M. Bukov, S. Gopalakrishnan, M. Knap, and E. Demler, Phys. Rev. Lett. **115**, 205301 (2015).

[48] S. Lellouch, M. Bukov, E. Demler, and N. Goldman, Phys. Rev. X **7**, 021015 (2017).

[49] S. Lellouch and N. Goldman, "Parametric instabilities in resonantly-driven bose-einstein condensates," https://arxiv.org/abs/1711.08832 (2017), arXiv:1711.08832 .

[50] J. Näger, K. Wintersperger, M. Bukov, S. Lellouch, E. Demler, U. Schneider, I. Bloch, N. Goldman, and M. Aidelsburger, arXiv preprint arXiv:1808.07462 (2018).

[51] T. Boulier, J. Maslek, M. Bukov, C. Bracamontes, E. Magnan, S. Lellouch, E. Demler, N. Goldman, and J. V. Porto, arXiv preprint arXiv:1808.07637 (2018).

[52] T. Bilitewski and N. R. Cooper, Phys. Rev. A **91**, 033601 (2015).

[53] T. Bilitewski and N. R. Cooper, Phys. Rev. A **91**, 063611 (2015).

[54] M. Reitter, J. Näger, K. Wintersperger, C. Sträter, I. Bloch, A. Eckardt, and U. Schneider, Phys. Rev. Lett. **119**, 200402 (2017).

[55] D. A. Abanin and Z. Papić, Annalen der Physik **529** (2017).

[56] T. Mori, T. Kuwahara, and K. Saito, Phys. Rev. Lett. **116**, 120401 (2016).

[57] T. Kuwahara, T. Mori, and K. Saito, Annals of Physics **367**, 96 (2016).

[58] S. A. Weidinger and M. Knap, Scientific reports **7**, 45382 (2017).

[59] F. Peronaci, M. Schiró, and O. Parcollet, arXiv preprint arXiv:1711.07889 (2017).

[60] O. Howell, P. Weinberg, D. Sels, A. Polkovnikov, and M. Bukov, arXiv preprint arXiv:1802.04910 (2018).

[61] M. Messer, K. Sandholzer, F. Görg, J. Minguzzi, R. Desbuquois, and T. Esslinger, arXiv preprint arXiv:1808.00506 (2018).

[62] S. Guérin and H. Jauslin, Advances in Chemical Physics **125**, 147 (2003).

[63] P. Weinberg, M. Bukov, L. D'Alessio, A. Polkovnikov, S. Vajna, and M. Kolodrubetz, Physics Reports **688**, 1 (2017).

[64] V. Novičenko, E. Anisimovas, and G. Juzeliunas, Phys. Rev. A **95**, 023615 (2017).

[65] W. W. Ho and D. A. Abanin, arXiv preprint arXiv:1611.05024 (2016).

[66] D. W. Hone, R. Ketzmerick, and W. Kohn, Phys. Rev. A **56**, 4045 (1997).

[67] A. Eckardt and M. Holthaus, Phys. Rev. Lett. **101**, 245302 (2008).

[68] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, Vol. 1 (MIT press Cambridge, 1998).

[69] C. M. Bishop, *Pattern recognition and machine learning* (springer, 2006).

[70] V. Dunjko and H. J. Briegel, Reports on Progress in Physics **81**, 074001 (2018).

[71] P. Mehta, M. Bukov, C. H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, arXiv:1803.08823 (2018).

[72] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, Science **288**, 824 (2000).

[73] E. Todorov, Bayesian brain: probabilistic approaches to neural coding , 269 (2006).

[74] S. J. Glaser, U. Boscain, T. Calarco, C. P. Koch, W. Köckenberger, R. Kosloff, I. Kuprov, B. Luy,

S. Schirmer, T. Schulte-Herbrüggen, *et al.*, The European Physical Journal D **69**, 279 (2015).

[75] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrueggen, and S. J. Glaser, Journal of Magnetic Resonance **172**, 296 (2005).

[76] T. Caneva, T. Calarco, and S. Montangero, Phys. Rev. A **84**, 022326 (2011).

[77] J. Sørensen, M. Aranburu, T. Heinzel, and J. Sherson, Physical Review A **98**, 022119 (2018).

[78] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. Obrien, Nature communications **5**, 4213 (2014).

[79] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New Journal of Physics **18**, 023023 (2016).

[80] A. C. Doherty and K. Jacobs, Phys. Rev. A **60**, 2700 (1999).

[81] G. G. Gillett, R. B. Dalton, B. P. Lanyon, M. P. Almeida, M. Barbieri, G. J. Pryde, J. L. O'Brien, K. J. Resch, S. D. Bartlett, and A. G. White, Phys. Rev. Lett. **104**, 080503 (2010).

[82] H. M. Wiseman, Phys. Rev. A **49**, 2133 (1994).

[83] H. Rabitz, New Journal of Physics **11**, 105030 (2009).

[84] F. Frank, T. Unden, J. Zoller, R. S. Said, T. Calarco, S. Montangero, B. Naydenov, and F. Jelezko, npj Quantum Information **3**, 48 (2017).

[85] R. S. Judson and H. Rabitz, Phys. Rev. Lett. **68**, 1500 (1992).

[86] T. Baumert, T. Brixner, V. Seyfried, M. Strehle, and G. Gerber, Applied Physics B: Lasers and Optics **65**, 779 (1997).

[87] D. Yelin, D. Meshulach, and Y. Silberberg, Optics letters **22**, 1793 (1997).

[88] B. J. Pearson, J. L. White, T. C. Weinacht, and P. H. Bucksbaum, Phys. Rev. A **63**, 063412 (2001).

[89] A. M. Weiner, Review of scientific instruments **71**, 1929 (2000).

[90] R. Bartels, S. Backus, E. Zeek, L. Misoguti, G. Vdovin, I. Christov, M. Murnane, and H. Kapteyn, Nature **406**, 164 (2000).

[91] D. Meshulach and Y. Silberberg, Nature **396**, 239 (1998).

[92] J. L. Herek, W. Wohlleben, R. J. Cogdell, D. Zeidler, and M. Motzkus, Nature **417**, 533 (2002).

[93] T. Weinacht, J. White, and P. Bucksbaum, The Journal of Physical Chemistry A **103**, 10166 (1999).

[94] C. J. Bardeen, V. V. Yakovlev, K. R. Wilson, S. D. Carpenter, P. M. Weber, and W. S. Warren, Chemical Physics Letters **280**, 151 (1997).

[95] R. J. Levis, G. M. Menkir, and H. Rabitz, Science **292**, 709 (2001).

[96] A. Assion, T. Baumert, M. Bergt, T. Brixner, B. Kiefer, V. Seyfried, M. Strehle, and G. Gerber, Science **282**, 919 (1998).

[97] P. Nuernberger, D. Wolpert, H. Weiss, and G. Gerber, Proceedings of the National Academy of Sciences **107**, 10366 (2010).

[98] T. Brixner and G. Gerber, ChemPhysChem **4**, 418 (2003).

[99] C. Winterfeldt, C. Spielmann, and G. Gerber, Rev. Mod. Phys. **80**, 117 (2008).

[100] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, Proceedings of the National Academy of Sciences **113**, E4877 (2016).

[101] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, Phys. Rev. Lett. **118**, 158004 (2017).

[102] A. A. Melnikov, H. P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel, Proceedings of the National Academy of Sciences , 201714936 (2018).

[103] M. Popova, O. Isayev, and A. Tropsha, Science advances **4**, eaap7885 (2018).

[104] D. Dong, C. Chen, T.-J. Tarn, A. Pechen, and H. Rabitz, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **38**, 957 (2008).

[105] C. Chen, D. Dong, H.-X. Li, J. Chu, and T.-J. Tarn, IEEE transactions on neural networks and learning systems **25**, 920 (2014).

[106] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Phys. Rev. X **8**, 031086 (2018).

[107] M. August and J. M. Hernández-Lobato, arXiv preprint arXiv:1802.04063 (2018).

[108] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Phys. Rev. X **8**, 031084 (2018).

[109] X.-M. Zhang, Z.-W. Cui, X. Wang, and M.-H. Yung, Phys. Rev. A **97**, 052333 (2018).

[110] M. Y. Niu, S. Boixo, V. Smelyanskiy, and H. Neven, arXiv preprint arXiv:1803.01857 (2018).

[111] F. Albarran-Arriagada, J. C. Retamal, E. Solano, and L. Lamata, arXiv:1803.05340 (2018).

[112] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, Phys. Rev. X **4**, 031002 (2014).

[113] T. Sriarunothai, S. Wölk, G. S. Giri, N. Fries, V. Dunjko, H. J. Briegel, and C. Wunderlich, arXiv preprint arXiv:1709.01366 (2017).

[114] J. Clausen and H. J. Briegel, arXiv preprint arXiv:1601.07358 (2016).

[115] H. A. Rabitz, M. M. Hsieh, and C. M. Rosenthal, Science **303**, 1998 (2004).

[116] A. P. Solon and J. M. Horowitz, Phys. Rev. Lett. **120**, 180605 (2018).

[117] M. Larocca, P. M. Poggi, and D. A. Wisniacki, Journal of Physics A: Mathematical and Theoretical **51**, 385305 (2018).

[118] A. G. Day, M. Bukov, P. Weinberg, P. Mehta, and D. Sels, arXiv preprint arXiv:1803.10856 (2018).

[119] M. Bukov, A. G. R. Day, P. Weinberg, A. Polkovnikov, P. Mehta, and D. Sels, Phys. Rev. A **97**, 052114 (2018).

[120] C. J. C. H. Watkins and P. Dayan, Machine Learning **8**, 279 (1992).

[121] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, Nature **518**, 529 (2015).

[122] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, nature **529**, 484 (2016).

[123] M. Bukov and A. Polkovnikov, Phys. Rev. A **90**, 043613 (2014).

[124] R. Desbuquois, M. Messer, F. Görg, K. Sandholzer, G. Jotzu, and T. Esslinger, Phys. Rev. A **96**, 053602 (2017).

[125] E. F. Dumitrescu, A. J. McCaskey, G. Hagen, G. R. Jansen, T. D. Morris, T. Papenbrock, R. C. Pooser, D. J. Dean, and P. Lougovski, Phys. Rev. Lett. **120**, 210501 (2018).

[126] E. Farhi, J. Goldstone, and S. Gutmann, arXiv preprint arXiv:1411.4028 (2014).

[127] W. W. Ho, C. Jonay, and T. H. Hsieh, arXiv preprint arXiv:1810.04817 (2018).

[128] W. W. Ho and T. H. Hsieh, arXiv preprint arXiv:1803.00026 (2018).

[129] S. Cong and F. Meng, The Scientific World Journal **2013** (2013).

[130] S. Grivopoulos and B. Bamieh, *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, **1**, 434 (2003).

[131] S. Zhao, H. Lin, J. Sun, and Z. Xue, *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, , 3811 (2009).

[132] S. Qamar, S. Cong, and B. Riaz, *Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), 2017 IEEE International Conference on*, , 48 (2017).

[133] W. Huggins, P. Patel, K. B. Whaley, and E. M. Stoudenmire, arXiv preprint arXiv:1803.11537 (2018).

[134] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[135] P. Weinberg and M. Bukov, SciPost Phys. **2**, 003 (2017).

[136] P. Weinberg and M. Bukov, arXiv preprint arXiv:1804.06782 (2018).

[137] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, Advances in Physics **65**, 239 (2016).

[138] Note that the agent is presented with the target state, and does not discover it.

[139] without counting non-exploratory repetitions required by RL to collect measurement statistic

Appendix A: Simulating the Dynamics of the Kapitza Oscillator

This section motivates the choice of Hamiltonian for the Kapitza oscillator [Eq. (1), main text], and discusses similarities and differences with the original Kapitza pendulum due to the periodic step drive.

The Kapitza oscillator with mass m and natural frequency ω_0 is governed by the Hamiltonian [2]

$$H_{\text{lab}}(t) = H_0 + A\Omega f(t)\cos\theta, \quad H_0 = \frac{p_\theta^2}{2m} - m\omega_0^2\cos\theta, \tag{A1}$$

where A and $\Omega = 2\pi/T$ are the (dimensionless) amplitude and the frequency of the periodic drive, and $f(t+T) = f(t)$ is a T-periodic function with zero period-average. In the infinite-frequency limit, one may expect that, since the drive averages to zero, the system is effectively governed by the non-driven oscillator H_0. However, since the strength of the drive-system coupling scales linearly with the drive frequency Ω, this naïve picture breaks down, and the effective infinite-frequency Floquet Hamiltonian $H_F(\Omega \to \infty) \neq H_0$.

In Ref. [2], a generic way to circumvent this problem was suggested, by going to a rotating frame:

$$H_{\text{rot}}(t) = V^\dagger(t)H_{\text{lab}}(t)V(t) - iV^\dagger(t)\partial_t V(t), \quad V(t) = \exp\left(-i\Delta(t)\cos\theta\right), \quad \Delta(t) = A\Omega\int^t dt' f(t'). \tag{A2}$$

This transformation removes the linear with Ω scaling of the amplitude in the time-integrated drive $\Delta(t)$, and facilitates the computation of the infinite-frequency Floquet Hamiltonian, as I now briefly revisit. Fundamentally, changing the reference frames can be seen as a resummation of an entire subseries of the inverse-frequency expansion [1–3]. A straightforward calculation yields

$$H_{\text{rot}}(t) = H_0 + \frac{1}{2m}\Delta(t)[\sin\theta, p_\theta]_+ - \frac{1}{2m}\Delta^2(t)\sin^2\theta, \tag{A3}$$

where $[\cdot, \cdot]_+$ denotes the anti-commutator. For instance, specializing to $f(t) = A\Omega\sin(\Omega t)$ gives $\Delta(t) = -A\cos(\Omega t)$ and $\Delta^2(t) = A^2/2(1 + \cos 2\Omega t)$. Since in this rotating frame the drive couplings (i.e. the amplitudes) of $\Delta(t)$ and $\Delta^2(t)$ are independent of Ω, the infinite-frequency Floquet Hamiltonian (up to a constant) can be computed by taking the time-average:

$$H_F(\Omega \to \infty) = H_0 - \frac{A^2}{8m}\cos 2\theta. \tag{A4}$$

Next to the free oscillator H_0, it contains an extra potential-energy term, which is responsible for stabilizing the inverted position of the oscillator at $\theta = \pi$ for high-enough frequencies. Note that this change of frames changes the micromotion (i.e. intra-period) evolution, but not the Floquet Hamiltonian; hence, the stabilizing property of the dynamics is left intact.

To set up an efficient simulator for quantum dynamics which produces the data from which the RL agent learns, it is advantageous to consider periodic step-drives. These multi-harmonic analogues of the monochromatic drive allow to circumvent solving Schrödinger's equation using ODE integrators, and reduce simulation time. Typically, multi-harmonic drives do not change the structure of the Floquet Hamiltonian, i.e. they preserve the stabilization effect: this can be seen with the help of the inverse-frequency expansion where the operator structure decouples from the time-ordered integrals and is, thus, independent of the specific choice of drive. However, there can be subtleties, which I now discuss.

A first guess would be to use a periodic-step time dependence for the lab-frame drive $f(t)$, with $\Delta(t)$ the corresponding continuous periodic zig-zag function. However, choosing a step-drive in the lab frame results in a more complicated time average $\int_0^T \Delta^2(t)dt \sim \Omega^{-2}$, which eliminates the stabilizing $\cos 2\theta$ term in the infinite-frequency Floquet Hamiltonian, and compromises the engineering property of the Floquet drive.

This observation suggests to use step drives in the rotating frame. I propose the following periodic step-drive Hamiltonian:

$$H_{\text{rot}}^{\text{step}}(t) = H_0 - \frac{A}{2m}\text{sign}\cos\Omega t\,[\sin\theta, p_\theta]_+ - \frac{A^2}{8m}\left(1 - \text{sign}\sin 2\Omega t\right)\cos 2\theta. \tag{A5}$$

Certainly, Eq. (A5) has the correct infinite-frequency limit. However, the choice of the time-periodic step-functions in Eq. (A5) comes at a price, and a few remarks are in order: (i) Notice that the time-dependence of the $\cos 2\theta$-term is not equal to the square of the time-dependence in front of the $[\sin\theta, p_\theta]_+$ term with this choice of drives. Hence, although the stroboscopic dynamics of the Hamiltonian (A5) at infinite-frequencies coincides with that of the original

Kapitza oscillator (A1), this is not necessarily the case at finite frequencies. (ii) Even though the Hamiltonian in Eq. (A5) does not describe the original dynamics of the lab-frame oscillator, the Floquet Hamiltonian associated with Eq. (A5) also supports a stable equilibrium at high enough frequencies. Thus, the Floquet engineering properties of the Hamiltonian (A5) are the same as in the original Kapitza oscillator, although the corresponding finite-frequency Floquet Hamiltonians differ. (iii) Sacrificing the equivalence of the micromotion dynamics by going from the lab to the rotating frame (and further by using step-drives), permits to also adjust the relative phases of the two time-dependent drives in Eq. (A5). The term $1/2 \, (1 - \mathrm{sign} \sin 2\Omega t)$ allows to use a minimum of four (instead of eight) steps per driving cycle. Keeping the periodic drive commensurate with the bang-bang control further facilitates the simulation of the dynamics. In turn, this enables reaching longer evolution times at a small computational cost. (iv) If one adds additional terms to the lab-frame Hamiltonian (A1), which depend on the position θ only, then they remain unaffected by the transformation to the rotating frame, and thus can simply be added to Eq. (A5). Such a term is given, e.g., by the horizontal displacement operator $\sin\theta$, to which the control field couples [see main text]. Hence, Eq. (A5) preserves the control properties of the original oscillator.

For these practical reasons, the simulator of quantum dynamics used to provide the data for the RL agent uses the Hamiltonian (A5). I use the (angular) momentum basis with 21 states in the Hilbert space, such that the low-energy initial and target wavefunctions remain marginally affected by increasing the number of states.

Appendix B: Simulating Quantum Measurements

In this section, I motivate the specific choice of binary quantum measurement used to simulate a setup close to realistic experiments. Suppose one had access to all observables in the quantum Kapitza oscillator, and one could readily measure any combination of them. To determine if the state $|\psi(t_f)\rangle$ at the end of the protocol is the Floquet eigenstate localized at the inverted position, one would proceed as follows. Since one cannot measure states, but only observables, one has to measure the hermitian operator corresponding to the Floquet Hamiltonian H_F. A projective quantum measurement then returns probabilistically the n-th eigenvalue of H_F with probability $|\langle\psi(t_f)|n_F\rangle|^2$, where $H_F|n_F\rangle = \varepsilon_F^n|n_F\rangle$. Let us reconcile this with the binary measurement defined in the main text: if, in a fixed outcome, n coincides with the target state at the inverted position, this corresponds to the binary output 1, and in all other cases – to 0.

Note that for many-body systems, the probability $|\langle\psi(t_f)|n_F\rangle|^2$ is likely to be exponentially small for almost all states, due to the exponentially large (with the system size) dimension of the Hilbert space. In such cases, it will be infeasible to successfully target a specific many-body state. This is related to the fact that the fidelity, being a probability, is a microscopic quantity, while in many-body systems the measurable quantities are observables (and their densities). For instance, in Ref. [106] it was demonstrated that in certain many-body control problems, one can successfully target microscopic states, such that the normalized logarithmic fidelity $-L^{-1} \log F_h(t_f)$ remains finite as $L \to \infty$. Another argument, based on typicality of many-body states [137], shows that even though the eigenstates of generic observables are orthogonal by definition, within a small eigenvalue shell they share the same macroscopic properties, such as expectation values of observables, up to exponentially suppressed finite-size corrections. This raises the question whether it is possible to target macroscopic properties of many-body systems using RL and quantum control, to be addressed in future studies.

Appendix C: Control Protocols Learned by the RL Agent

In this section I discuss the protocols learned by the RL agent. Even though I do not yet fully understand the physics behind the best RL protocols, certain features present themselves worthy of attention.

Figure 5 (left column) shows the best protocol out of a family of 10^6 local minima obtained using 1-flip SD, and the corresponding instantaneous fidelity evolution. The middle column shows the best encountered protocol by the RL agent during the train stage, according to the estimated fidelity [recall that the measurement is noisy and the agent only gets an estimate of the true value]. I checked that, for this seed realization, the agent in fact learned this protocol and was following it during the test stage. The right column shows the best true-fidelity protocol encountered during the train stage. Despite notable similarities at early times between the two protocols, there are small differences suggesting that the agent has hard time estimating the true value of a protocol close to optimality, due to the presence of noise in the reward. Additionally, notice that the instantaneous fidelities are not monotonic: they first rise and drop at intermediate times before they shoot up for the final values. In fact, the rise appears during a stage dominated by the +4 bang mode. This is reminiscent of the agent pushing the pendulum (on average) in one direction in order in the second stage to make use of the gained gravitational energy to overcome the potential barrier and eventually reach the inverted position from the other side in the time allotted. Indeed, this naïve classical picture
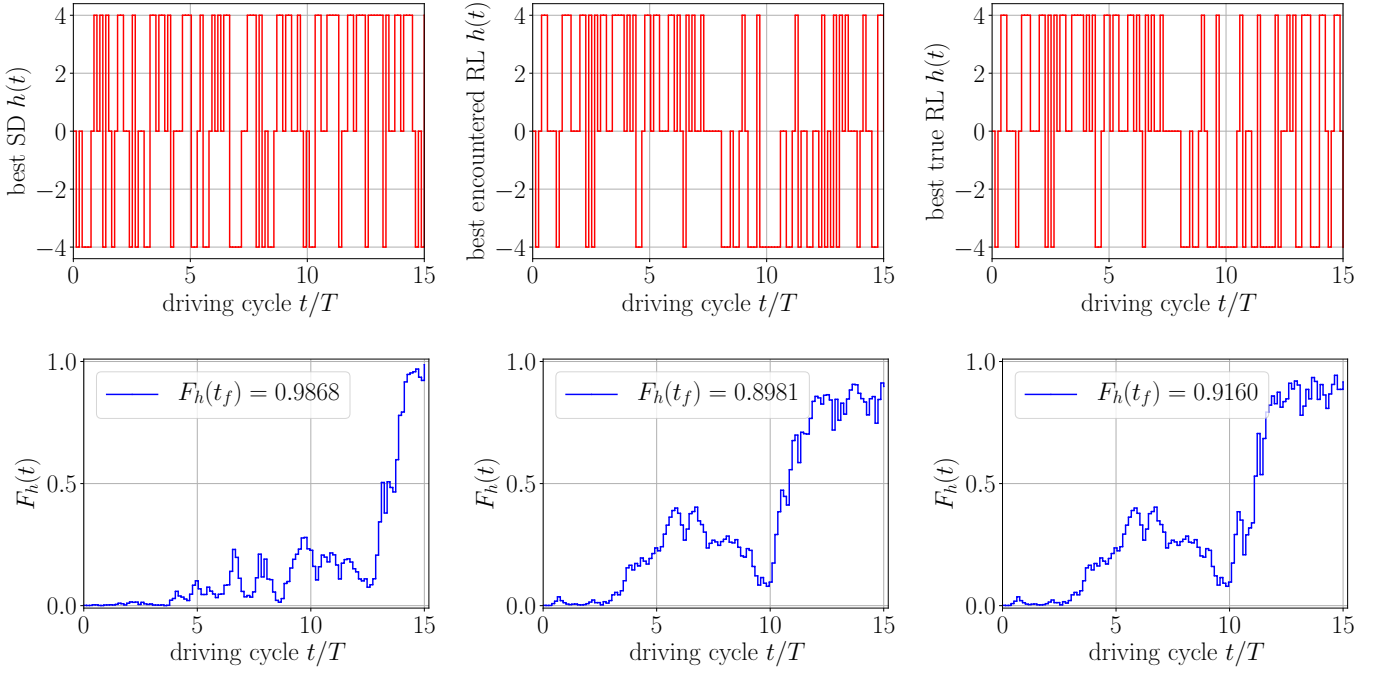
FIG. 5: Protocols (upper row) and the corresponding fidelities (lower row) against time. The oscillator parameters are $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$. The protocols contain 8 steps per period for a total of $N_T = 15$ periods.
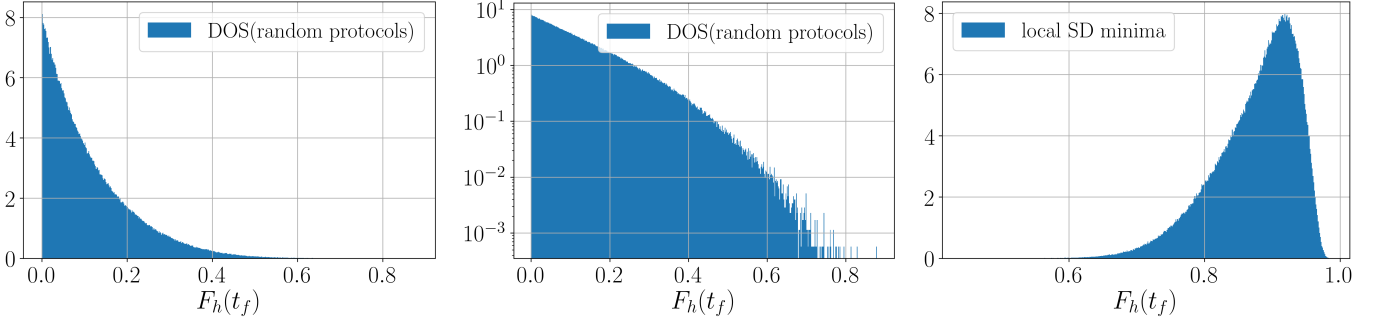

FIG. 6: Fidelity distributions for random protocols on linear (left), and a semi-log (middle) scale, and the sample of local SD-minima (right). The oscillator parameters are $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$. The protocols contain 8 steps per period for a total of $N_T = 15$ periods.

is confirmed by the protocol visualizations [see Video 1]. The quantum nature of the dynamics is most likely hidden in the non-trivial character of the bang-sequence.

While it is hard to make precise sense of these protocol patterns, one can gain insights into the complexity of Kapitza bang-bang control as an optimization problem. Figure 6 (left panel) shows the histogram of 10^6 randomly chosen protocols. As observed using the train curves in RL, the mean fidelity of a random protocol is about 10%, which is consistent. This distribution represents the density of states (DOS) in protocol space [118]. It decays at least exponentially with fidelity (middle panel). This distribution comes in strong contrast to that of 1-flip local SD minima in the infidelity landscape (right panel), which is heavily sifted towards the high-fidelities of interest. When pushed to 10^6 training episodes, the RL agent learns on average a fidelity which is consistent with the mean of this distribution, cf. Fig. 11 [top right panel].
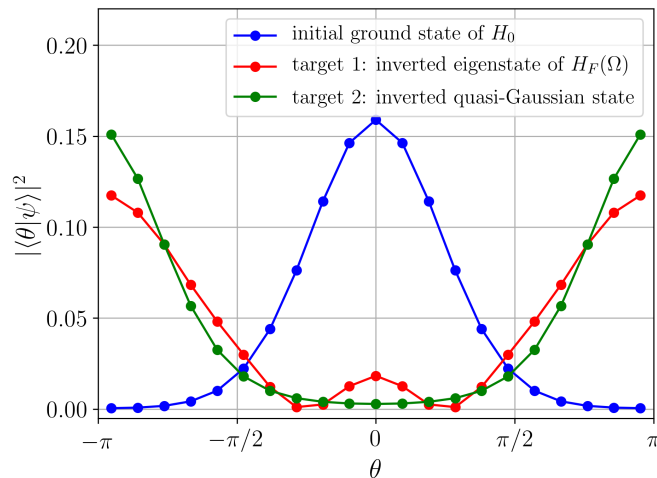
FIG. 7: Initial state and target states at the inverted position for $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$.

Appendix D: Learning a Quasi-Gaussian Target State

As I explained in Sec. B, targeting an exact Floquet eigenstate requires the ability to measure the Floquet Hamiltonian. Unfortunately, often this is beyond the capabilities of present-day experiments.

At the same time, however, the main reason behind the interest in that particular Floquet eigenstate, is precisely its feature to be localized at the inverted position. One might then wonder how the RL agent would perform, if required to target a simpler state with the same property. To test this, I use as an alternative target state the quasi-Gaussian state $\langle\theta|\psi_*\rangle \propto \exp(-a^{-1/4}\cos\theta)$ [Fig. 7], with oscillator length $a = (\omega')^{-1/4}$ in the harmonic approximation set by the infinite-frequency effective potential at the inverted position: $\omega' = \sqrt{A^2/(2m^2) - \omega_0^2}$. Such a state can be emulated easily as the ground state of some static Hamiltonian, and is thus much more easily accessible compared to the exact Floquet eigenstate. Figure 8 shows that this does not introduce any additional difficulties for the RL algorithm. As a matter of fact, the obtained fidelities are slightly higher, compared to targeting the exact Floquet eigenstate, cf. Fig. 3 in the main text.

Appendix E: Q-Learning in Noisy and Stochastic Environments

1. Numerical Experiment: does the Q-Learning Agent Learn Specifics of the Stochastic Environment?

Training the Q-Learning agent with noise in the initial state and in a stochastic environment, raises the question whether it is capable of learning the details of such uncertainty sources and exploit them to its advantage in the learning process. To this end, I consider the following numerical experiment: the agent is trained on a noise-free deterministic environment [with the only uncertainty in the reward, as a result of the quantum measurement], but subsequently tested on a noisy stochastic environment [i.e. with additional occasional random failures in the bangs of the protocols]. The performance during the test stage should then be compared to the case where the agent was also trained in a noisy stochastic environment [$\eta = 0.31$, $\zeta = 1/120$, see main text]. In which scenario does the agent perform better?

To answer this, I distinguish between two quantities during the test stage: the estimated fidelity expected by the agent (red) and the true fidelity of the protocol (blue). Figure 9 clearly shows that, after learning in a noise-free deterministic environment, the agent erroneously learns to expect a higher fidelity, compared to the true fidelity associated with the learned protocol.

There are two important conclusions from this numerical experiment. (i) Notice that the true fidelity in the train stage in Fig. 9 (blue dots) is about the same as the expected fidelity had the agent been trained in the presence of uncertainty [see Fig. 3 (red dots), main text]. This suggests that the agent does not learn to exploit any additional features of the environment using this state-action space parametrization. This means that the RL algorithm is intrinsically robust to noise. The most likely reason for this lies in the stochastic ε-greedy exploration schedule used in Q-Learning, cf. Sec. G. (ii) Recall that the RL state space definition depends on the initial state $|\psi_i\rangle$. As the
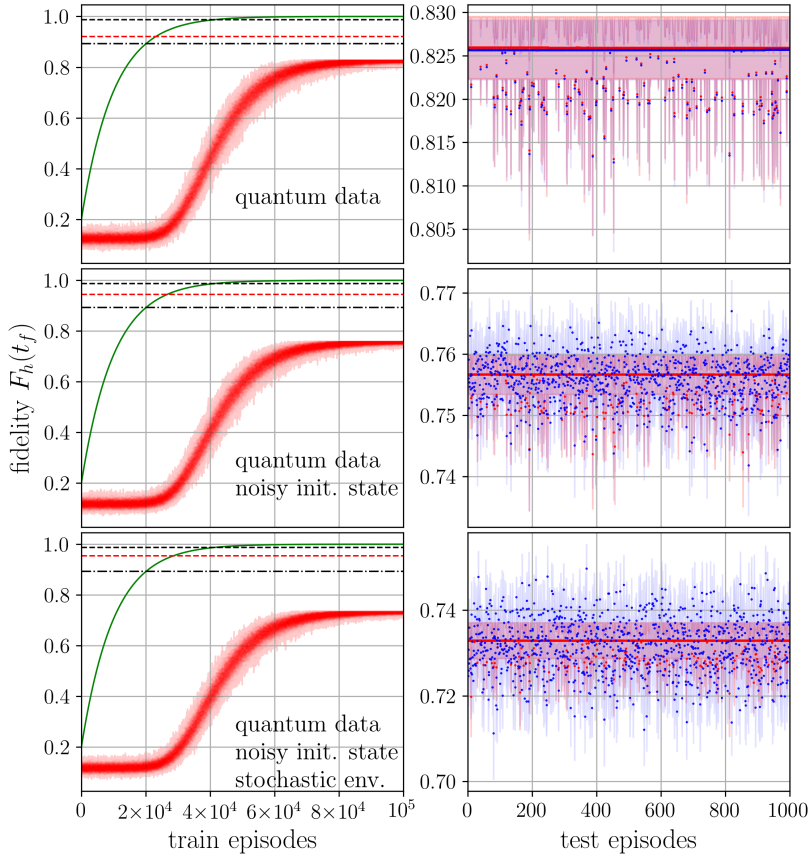
FIG. 8: Train and test stages for the quasi-Gaussian target state $\langle\theta|\psi_*\rangle \propto \exp(-a^{-1/4}\cos^2\theta)$, localized at the inverted position. The details are the same as in Fig. 3, main text.

noise induces a change in the initial state, the algorithm learns the average fidelity over an ensemble of noisy initial states. This reveals the weakness of the current state-action choice to generalize to arbitrary initial conditions, which comes as a trade-off to the capability to learn solely from the actions taken. This behavior can be explained by the observation that the agent is not presented with any information about the uncertainty in the environment during learning: e.g., if the agent was told retroactively once a bang in a protocol had randomly failed, it might be possible to learn to 'correct' or 'counteract' this change. This behavior will be explored in future studies.

2. Single Run vs. Average Performance

Even in deterministic setups, the Q-Learning algorithm, cf. Sec. G, contains intrinsic noise due to the ε-greedy exploration schedule used during the train stage. Therefore, the algorithm is run for 100 independent realizations of the pseudo-random number generator, and the graphs in this paper show averages. The deviation from the mean is computed using a bootstrapping approach (shaded area). Figure 10 shows the worst (left) and the best (middle) runs, and compares them to the average fidelity performance (middle).

3. Dependence on the Number of Training Episodes

It is curious to study how the agent's learning capabilities change as the number of training episodes increases. As noted in the main text, the huge protocol space contains $3^{120} \sim 10^{57}$ protocol configurations. Additionally, the fidelity histograms, cf. Fig. 6 (left panel), show that most states have very poor fidelities. In the main text I showed data for up to 10^5 training episodes. Figure 11 shows that one can achieve a reasonable improvement by increasing the training episodes by an order of magnitude. I do not consider it appropriate to push the Q-Learning algorithm to its limits, since the maximum number of training episodes in realistic experimental setups is set by the sample
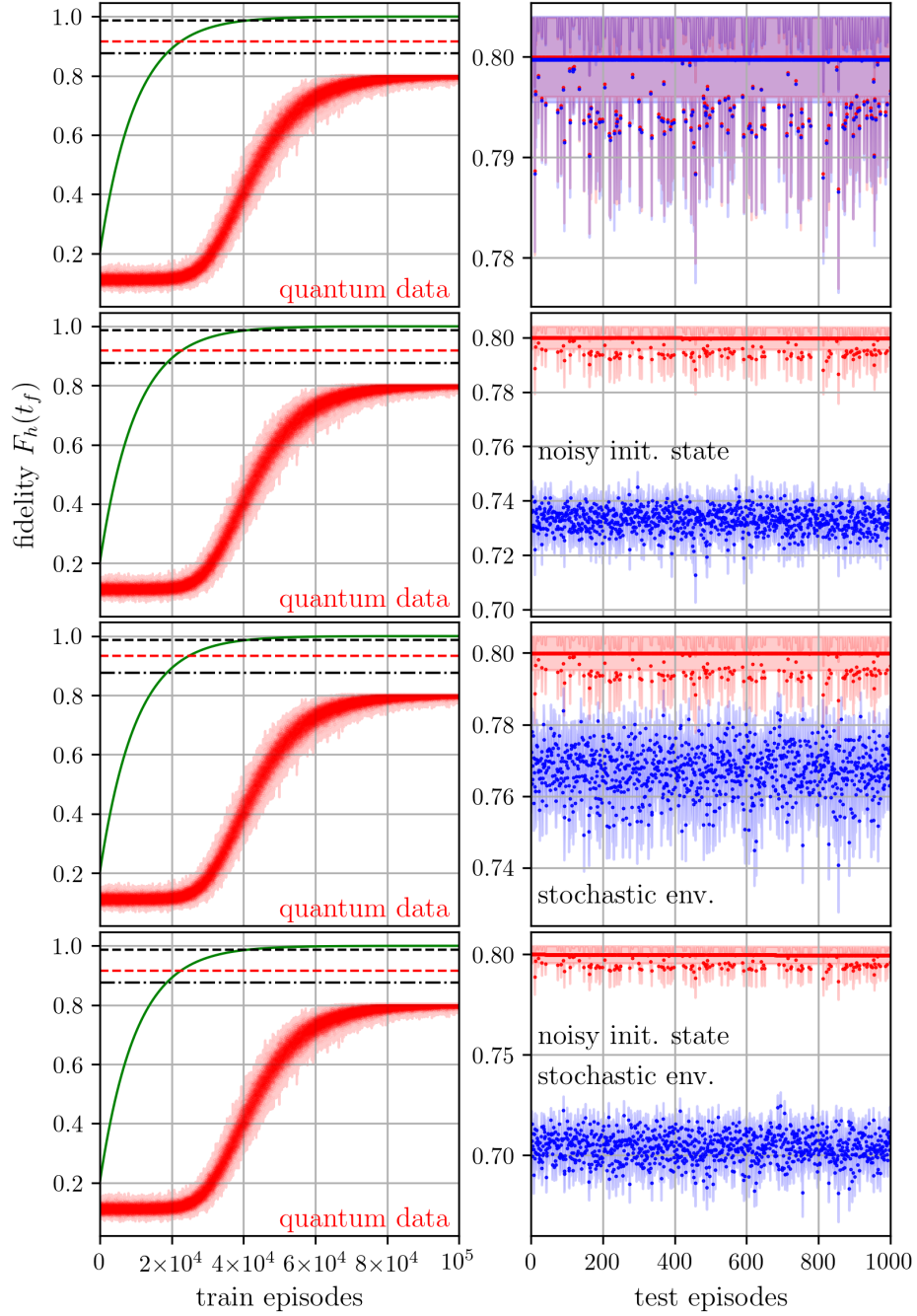
FIG. 9: Numerical experiment training the agent using binary quantum data in a deterministic environment starting from a fixed initial state, while testing on noisy initial states and/or stochastic environment. The red data shows the agent's estimated fidelity, while the true fidelity is shown in blue. The mismatch between the agent's estimate (red) and the true fidelity (blue) arises due to the absence of noise and stochasticity during the Train stage. The left side (train curves) shows the same data for better comparison. The parameters are the same as in Fig. 3 of the main text.
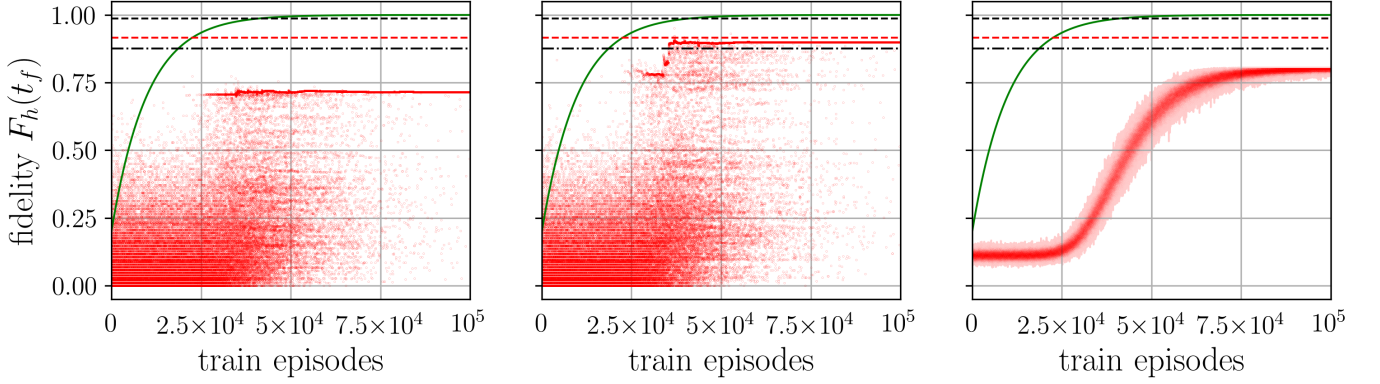
FIG. 10: Training behavior as a function of the number of Train episodes: worst run (let), best run (middle) and average performance (right) over 100 seed realizations of the pseudo-random number generator. The oscillator parameters are $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$. The protocols contain 8 steps per period for a total of $N_T = 15$ periods. The target is the Floquet eigenstate. The initial state is noise-free and the environment is deterministic.

efficiency. Instead, I believe that the algorithm can be made more useful if it is appropriately improved in sample efficiency instead.

Appendix F: Reinforcement Learning to Invert the Classical Kapitza Pendulum

Last but not least, I demonstrate the versatility and universality of the Q-Learning algorithm by applying it to the classical Kapitza pendulum. For the sake of a better comparison with the quantum Kapitza oscillator, I choose the same time-dependent Hamiltonian in the rotating frame:

$$H(t) = \frac{p_\theta^2}{2m} - m\omega_0^2 \cos\theta - \frac{A}{m} \left(\text{sign}\cos\Omega t\right) p_\theta \sin\theta - \frac{A^2}{8m} \left(1 - \text{sign}\sin 2\Omega t\right) \cos 2\theta + h(t)\sin\theta, \qquad \text{(F1)}$$

where p_θ and θ are classical conjugate variables, and $h(t)$ is the bang-bang control field. The dynamics of the pendulum is governed by Hamilton's equations of motion:

$$\dot{\theta} = \frac{1}{m} \left(p_\theta - A\left(\text{sign}\cos\Omega t\right)\sin\theta\right),$$

$$\dot{p_\theta} = -m\omega_0^2\sin\theta + \frac{A}{m}\left(\text{sign}\cos\Omega t\right)p_\theta\cos\theta - \frac{A^2}{4m}\left(1 - \text{sign}\sin 2\Omega t\right)\sin 2\theta - h(t)\cos\theta. \qquad \text{(F2)}$$

The initial state is chosen as $\theta(t = 0) = 0.01$ and $p_\theta(t = 0) = 0$. The target state is the inverted position at $\theta = \pi$. The finite value of the initial angle breaks the symmetry of the optimal protocol [i.e. reaching the target clockwise and counter-clockwise].

To define the reward for the agent, note that simply reaching the target is not enough to assure a stable orbit at the inverted position after the control sequence is over, if the momentum at the end of the protocol is large enough to cause spin-over. Hence, a good cost function should penalize large final momenta. I thus (empirically) choose the following reward:

$$r(\theta, p_\theta) = \frac{1}{\pi^2}\left[(\theta(t_f) + \pi)\text{mod}(2\pi) - \pi\right]^2 - 4|p_\theta(t_f)|^2, \qquad r \in [-\infty, 1]. \qquad \text{(F3)}$$

In classical systems, measurements are deterministic. However, they might still be noisy. To take this into account in RL, I add Gaussian noise to the values of the position and momentum with zero mean and variance $\sigma = 0.05$. This leads to uncertain rewards. Similar to the quantum case, I also consider cases in which (i) the initial state is noisy, by adding Gaussian noise in the initial condition with zero mean and variance $\eta = 0.1$, and (ii) there are occasional failure events in the control bangs. This works in the same way as for the quantum oscillator.

The universality of the state-action representation makes the Q-Learning algorithm agnostic on the physical system it is applied to. Thus, I apply the same algorithm to the classical Kapitza pendulum, see Fig. 12. The best RL protocol for the case of noisy reward but noiseless initial state in a deterministic environment is shown in Video 3.
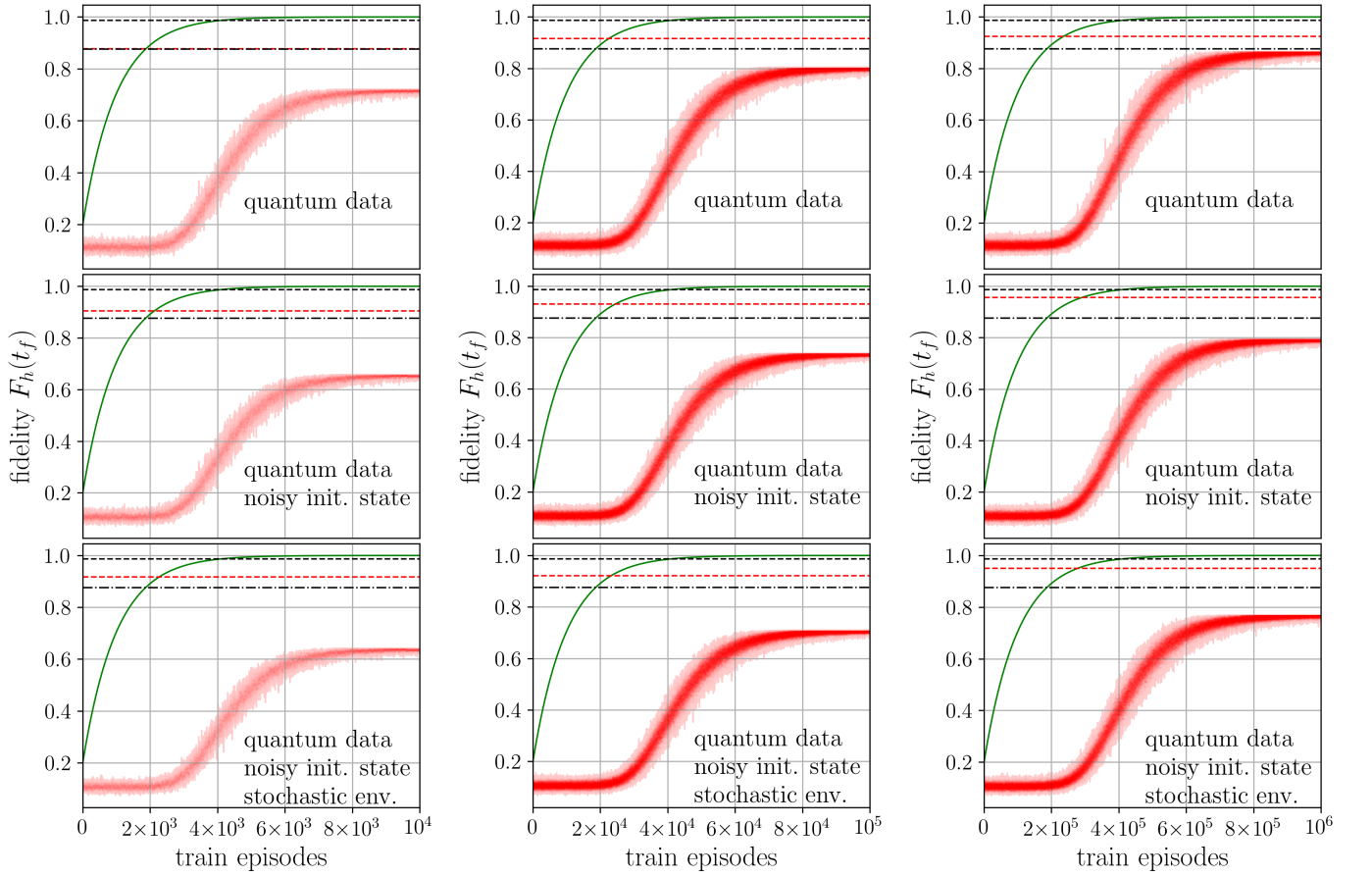
FIG. 11: Training behavior as a function of the number of Train episodes: 10^4(left), 10^5 (middle), and 10^6(right). The oscillator parameters are $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$. The protocols contain 8 steps per period for a total of $N_T = 15$ periods. The target is the Floquet eigenstate.

Appendix G: Q-Learning Algorithm for Autonomous Quantum Control

Below, I provide a trivial extension of Watkins tabular Q-Learning [68] which uses as a reward the noisy running estimate of the fidelity $F_h(t_f)$, and learns from experience replays.

To study the Floquet control problem, I apply the version of a tabular Q-Learning algorithm [68] with eligibility trace depth parameter $\lambda = 0.6$. For the Q-Learning update rule, I use a small learning rate of $\alpha = 0.1$ in order to account for the running stochastic reward (the fidelity), estimated from binary quantum measurements as $r = m/n$. Here m is the number of $+1$ measurement outcomes, and n is the total number of measurements for a fixed protocol [see main text]. To gain measurement statistics, each protocol is repeated 100 times every time it is encountered, until the error estimate to be within the 2σ-window, $E = 2\sqrt{r(1-r)/n}$, becomes less than 1%. During this repetition stage, no updates of the Q-function take place. To choose actions, the algorithm uses an ε-greedy policy: the best action [according to the current Q-function] is taken with probability $1 - \varepsilon(n_{\rm ep})$, or else a random action is chosen with probability $\varepsilon(n_{\rm ep})$. The exploration schedule $\varepsilon(n_{\rm ep})$ is attenuated exponentially according to

$$\varepsilon(n_{\rm ep}) = (\varepsilon_i - \varepsilon_f)\, \exp\left(-\frac{10 n_{\rm ep}}{N_{\rm episodes}}\right) + \varepsilon_f, \tag{G1}$$

with $\varepsilon_i = 10$ and $\varepsilon_f = 50$ (chosen empirically). The larger $\varepsilon(n_{\rm ep})$, the less the RL agent explores [see green curves in all learning plots, where $\varepsilon(n_{\rm ep})$ is normalized within $[0, 1]$ for display purposes]. The current episode and the total number of train episodes are denoted by $n_{\rm ep}$ and $N_{\rm episodes}$, respectively. In order to help the agent explore the exponentially large RL state space efficiently, I keep track of the best encountered protocol w.r.t. the current fidelity estimate, and replay it every 100 episodes for 200 times, thereby updating the Q-function.

Algorithm 1, describes the pseudo-code for the RL algorithm used to obtain the results in the main text. Familiarity
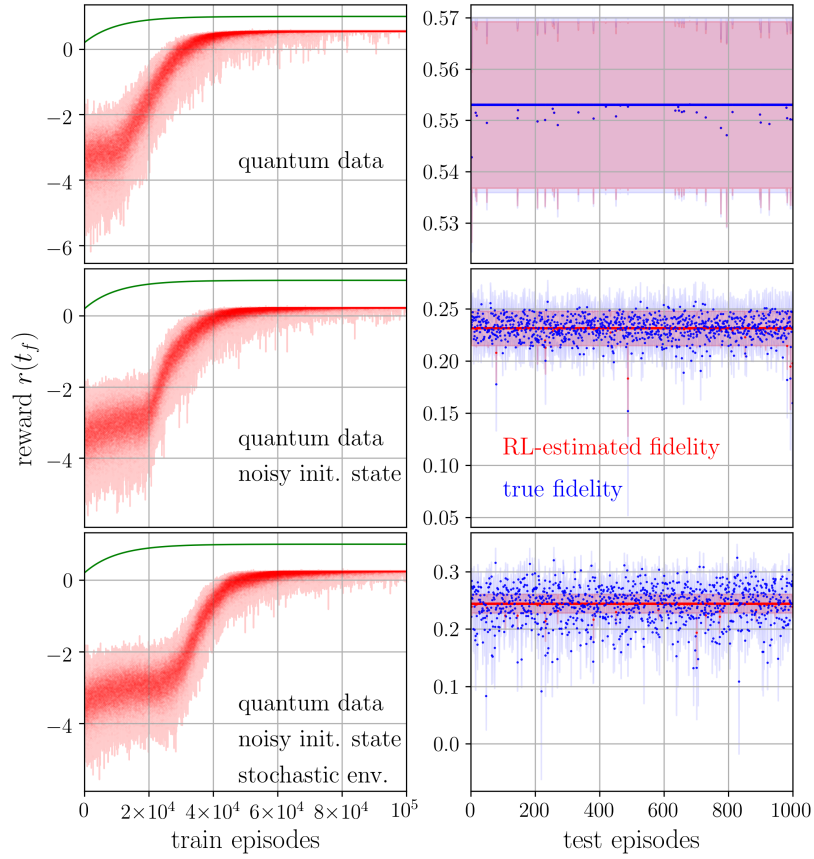
FIG. 12: Training behavior as a function of the number of Train episodes for the classical Kapitza pendulum. The oscillator parameters are $\Omega/\omega_0 = 10$, $A = 2$ and $m\omega_0 = 1$. The protocols contain 8 steps per period for a total of $N_T = 4$ periods. The target is $\theta_* = \pi$ and $p_* = 0$. See Video 3 for a visualization of the best-encountered RL protocol.

with the original Watkins' Q-Learning algorithm and its extension TD(λ), see e.g. Ref. [68], is helpful to facilitate understanding. It is straightforward to extend Algorithm 1 to Deep Learning.

Appendix H: Video Simulations of the RL-Controlled Kapitza Oscillator

Legends for all three movies is available on https://mgbukov.github.io/RL_kapitza/.

Algorithm 1 Q-Learning with nondeterministic rewards [quantum measurements]

1: **procedure** Q-LEARNING
2:     initialize an empty $Q(s,a)$ function for all states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}(s)$
3:     initialize an empty registry $R(h) = (m, n, E)$ for all protocol sequences h, number of protocol encounters $m \in \mathbb{N}$, number of positive quantum measurement outcomes $n \in \mathbb{N}$, and statistical error estimate $E \in \mathbb{R}$
4:     initialize `best_encountered_actions` arbitrarily and find the corresponding `best_encountered_protocol`
5:     initialize `best_encountered_return`$= -1$
6:     **repeat** for every episode:
7:         set value of ε-greedy exploration according to some schedule
8:         run ε-greedy `QL_episode` in `explore` mode
9:         **if** statistical error estimate E of return for most recent protocol h is within some threshold **then**
10:             repeat protocol to collect data and improve statistics (Q-function is not updated)
11:             update registry $R(h)$.
12:         run `QL_episode` in `repeat` mode (Q-function is updated)
13:         run `update_best_encountered` routine
14:         **if** episode is scheduled for replay **then**
15:             **if** statistical error estimate E of return for best protocol h is within some threshold **then**
16:                 **repeat**
17:                     run greedy `QL_episode` in `replay` mode with unit learning rate $\alpha = 1$ (Q-function is updated)
18:                     `best_encountered_return` $\leftarrow r$
19:                 **until** a number of replay episodes is exhausted
20:     **until** number of episodes is exhausted
21:
22:

1: **function** QL_EPISODE(MODE)
2:     reset environment into initial state $S = S_0$
3:     $q \leftarrow Q(S_0, :)$ (compute Q-function value of initial state for all available actions $a \in \mathcal{A}(s)$)
4:     $\text{trace}(s,a) \equiv 0$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
5:     **repeat** for each step in the episode:
6:         run `choose_action(mode)` to get action A from state S using policy derived from Q (e.g., ε-greedy)
7:         take action A, environment goes to new state S'
8:         $\text{trace}(S, A) \leftarrow \alpha$: fire eligibility trace
9:         set $\delta_t \leftarrow -q(A)$
10:         **if** S is terminal **then**
11:             compute current estimate of return r
12:             $\delta_t \leftarrow \delta_t + r$
13:             $Q(s,a) \leftarrow Q(s,a) + \delta_t \, \text{trace}(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
14:             **goto** next episode
15:         $q \leftarrow Q(S', :)$ (find Q-function for all actions in state S')
16:         $\delta_t \leftarrow \delta_t + \max_a q(a)$ (find action that maximizes q)
17:         $Q(s,a) \leftarrow Q(s,a) + \delta_t \, \text{trace}(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
18:         $\text{trace}(s,a) \leftarrow \lambda \, \text{trace}(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
19:     **until** episode is complete
20:
21:

1: **function** CHOOSE_ACTION(MODE)
2:     **if** mode is `repeat` **then**
3:         $A \leftarrow$ action taken in previous run at this time step
4:     **else if** mode is `explore` **then**
5:         compute available actions $\mathcal{A}(S)$ from current state S
6:         compute $A_{\text{greedy}} \leftarrow$ one of the (possibly many) actions that maximize $Q(S, :)$
7:         choose $A \leftarrow A_{\text{greedy}}$ with probability ε (depending on exploration schedule), otherwise $A \leftarrow$ a random available action
8:         **if** A is not A_{greedy} **then**
9:             $\text{trace}(s,a) \leftarrow 0$, reset trace for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
10:     **else if** mode is `replay` **then**
11:         $A \leftarrow$ `best_encountered` action at this time step

1: **function** UPDATE_BEST_ENCOUNTERED
2:     **if** `return_best_encountered`$<$current return r **then**
3:         **if** statistical error estimate E of return for most recent protocol h is within some threshold **then**
4:             `return_best_encountered` $\leftarrow r$
5:             overwrite `actions_best_encountered` with most recent actions sequence
6:             overwrite `state_best_encountered` with most recent protocol