Entropy regularized reinforcement learning using large deviation theory

Argenis Arriojas,[1,*] Jacob Adamczyk,[1] Stas Tiomkin,[2] and Rahul V. Kulkarni[1,†]
[1]*Department of Physics, University of Massachusetts Boston, Boston, Massachusetts 02125, USA*
[2]*Department of Computer Engineering, San Jose State University, San Jose, California 95192, USA*

(Received 4 May 2021; revised 8 June 2022; accepted 1 April 2023; published 10 May 2023)

Reinforcement learning (RL) is an important field of research in machine learning that is increasingly being applied to complex optimization problems in physics. In parallel, concepts from physics have contributed to important advances in RL with developments such as entropy-regularized RL. While these developments have led to advances in both fields, obtaining analytical solutions for optimization in entropy-regularized RL is currently an open problem. In this paper, we establish a mapping between entropy-regularized RL and research in nonequilibrium statistical mechanics focusing on Markovian processes conditioned on rare events. In the long-time limit, we apply approaches from large deviation theory to derive exact analytical results for the optimal policy and optimal dynamics in Markov decision process (MDP) models of reinforcement learning. The results obtained lead to an analytical and computational framework for entropy-regularized RL which is validated by simulations. The mapping established in this work connects current research in reinforcement learning and nonequilibrium statistical mechanics, thereby opening avenues for the application of analytical and computational approaches from one field to cutting-edge problems in the other.

DOI: 10.1103/PhysRevResearch.5.023085

I. INTRODUCTION

The combination of machine learning approaches with concepts and tools from physics has given rise to significant developments in current research [1]. Concepts derived from statistical mechanics have led to important applications in machine learning [2], and recent work has further highlighted the importance of building bridges between the two disciplines [3–5]. Conversely, machine learning approaches such as reinforcement learning (RL) are increasingly being used to address complex optimization problems in diverse fields of physics, ranging from quantum computing and quantum control to adaptive optics [6–11]. While RL approaches are now being widely applied in physics research, there has been less emphasis on using insights and approaches from physics to address open problems in RL. The development of such approaches can lead to important discoveries in RL research as well as provide avenues for the development of novel RL algorithms to solve a diverse range of problems in physics [7,12].

While the connections of machine learning to equilibrium statistical mechanics are well established [2], the interface with nonequilibrium statistical mechanics (NESM) is less explored. Recent work has addressed this gap by developing machine learning approaches with applications to NESM. For example, graph neural network models for estimation of the scaled cumulant generating function for observables in dynamical systems have been developed [13] and evolutionary RL approaches have been used to calculate the likelihood of dynamical large deviations [14]. While RL approaches are thus starting to be applied to study systems of interest in NESM, it is also of interest to explore if insights from NESM can be used to obtain new insights into RL. An example of the latter case arises when considering RL problems that involve optimization over system trajectories with entropy-based regularization [15,16]. This framework, termed maximum entropy RL, or more generally entropy-regularized RL, allows the optimal control problem in RL to be recast as a problem in Bayesian inference. This "control-as-inference" approach involves the introduction of optimality variables such that the *posterior* trajectory distribution, conditioned on optimality, provides the solution to the optimal control problem [15–18]. While this framework has led to several advances, there are open questions relating to the derivation of analytical results that characterize the optimal dynamics.

Recent research in NESM using large deviation theory has developed a framework for analyzing Markovian processes conditioned on rare events [19–23]. In this framework, a generalization of the Doob h-transform [22,24,25] is used to determine the *driven* process: a conditioning-free Markovian process that has the same statistics as the original Markovian process conditioned on a rare event. Similar derivations of the driven or controlled processes have been obtained in previous work using a maximum entropy approach for characterizing nonequilibrium steady states [26–29]. The connection of this framework to RL can be seen by noting that the goal in entropy-regularized RL is to derive the posterior trajectory

*arriojasmaldonado001@umb.edu
†rahul.kulkarni@umb.edu

distribution conditioned on optimality and, in the long-time limit, optimality of the trajectory is a *rare event* for the original dynamics. This commonality of conditioning on rare events suggests that approaches and results from NESM can be used to characterize the optimal control policy for entropy-regularized RL problems. Indeed, recent work has explored connections between entropy-regularized RL and rare trajectory sampling and applied it to a range of problems in physics [12,30]; however an explicit characterization of the optimal controlled processes for general entropy-regularized RL problems has not been derived to date.

In this paper, we develop a mapping between MDP-based entropy-regularized RL and Markovian processes conditioned on rare events in the long-time limit. Using approaches from large deviation theory, we derive exact analytical expressions characterizing trajectory distributions conditioned on optimality. Interestingly, our derivation of these results shows how the generalized Doob h-transform arises naturally from Bayesian inference applied to trajectory distributions. The results obtained lead to analytical expressions for the optimal policy and optimal dynamics in entropy-regularized RL which are validated using simulations. The connections established in this work also lead to an approach for model-free RL and provide avenues for research focusing on the intersection of RL and physics. Specifically, the mapping developed in this work connects RL-based optimization to the estimation of dynamical free energy in NESM [19], thus paving the way for the use of approaches such as deep RL to estimate dynamical free energies in nonequilibrium physics.

II. MARKOV DECISION PROCESS FRAMEWORK

In the following, we provide an overview of the standard Markov decision process (MDP) framework for reinforcement learning. To introduce the formalism, we focus on the finite horizon, undiscounted case with horizon N [15]. Consider a Markov chain with states represented by tuples (s, a), where s is an agent's current state and a is an action taken while in state s. The probability that the agent transitions to state s' after taking action a is denoted by $p(s'|s, a)$. The choice of action a given the agent's current state s is drawn from a policy $\pi(a|s)$, and the corresponding reward collected by the agent is given by the reward function $r(s, a)$.

With the above representation, we can now define probability distributions over trajectories $\tau := \{(s_1, a_1), \ldots, (s_N, a_N)\}$ that are generated by the policy $\pi(a|s)$ and transition probabilities $p(s'|s, a)$. Let $p(s_1)$, $\pi(a|s)$ and $p(s'|s, a)$ denote prior distributions for the initial state, policy, and transition dynamics respectively. The corresponding probability distribution for *uncontrolled* trajectories is given by

$$p(\tau) = p(s_1)\prod_{t=1}^{N} p(s_{t+1}|s_t, a_t)\pi(a_t|s_t). \tag{1}$$

The prior distribution for the transition dynamics corresponds to the system's uncontrolled transition dynamics. In the special case of maximum entropy (MaxEnt) RL, the prior policy is chosen as the uninformative prior, i.e., a uniform distribution over actions.

We now consider the probability distribution for *controlled* trajectories that is generated by a specific policy $\pi_c(a|s)$ and transition dynamics $p_c(s'|s, a)$ that may, in general, be different from the uncontrolled prior distributions. The probability distribution for controlled trajectories is given by

$$p_c(\tau) = p_c(s_1)\prod_{t=1}^{N} p_c(s_{t+1}|s_t, a_t)\pi_c(a_t|s_t). \tag{2}$$

The objective in standard RL is to find the policy $\pi^*(a|s)$ that maximizes the total expected reward. Let $R_\tau = \sum_{t=1}^{N} r(s_t, a_t)$ denote the total reward accumulated over a trajectory τ. Correspondingly, the optimal policy $\pi^*(a|s)$ is given by

$$\pi^*(a|s) = \arg\max_{\pi_c} \mathbb{E}_{p_c(\tau)}[R_\tau]. \tag{3}$$

In entropy-regularized RL, the goal is to determine the decomposition (Eq. 2) for the optimally controlled trajectory distribution $p_c(\tau)$ that maximizes the objective function

$$\mathbb{E}_{p_c(\tau)}[R_\tau] - \frac{1}{\beta}\mathcal{H}(p_c(\tau)||p(\tau)), \tag{4}$$

where β is a regularization parameter corresponding to the inverse temperature. We can see that, in entropy-regularized RL, the standard RL objective function is augmented to include a regularization term $-\frac{1}{\beta}\mathcal{H}(p_c(\tau)||p(\tau))$. This term corresponds to the relative entropy between the controlled trajectory distribution $p_c(\tau)$ and the prior trajectory distribution $p(\tau)$, and is given by the Kullback-Leibler divergence

$$\mathcal{H}(p_c(\tau)||p(\tau)) = \sum_{\tau} p_c(\tau)\ln\frac{p_c(\tau)}{p(\tau)}.$$

This regularization process naturally yields stochastic optimal policies, a desirable feature providing robustness to changes in the problem's dynamics. The role of the β parameter is then to regulate the tradeoff between obtaining a single "greedy" optimal solution and obtaining a collection of solutions with lower returns but improved robustness.

The preceding generalization of standard RL allows one to recast the optimal control problem as an inference problem [15]. This control-as-inference approach involves the introduction of optimality variables $\mathcal{O}_t$ defined such that

$$p(\mathcal{O}_t = 1|s_t, a_t) = \exp[\beta r(s_t, a_t)], \tag{5}$$

The binary random variable $\mathcal{O}_t$ represents the probability that the trajectory is optimal at time step t. The purpose of this definition is that the *posterior* trajectory distribution, obtained by conditioning on $\mathcal{O}_t = 1$ for all t, exactly corresponds to the trajectory distribution generated by optimal control. The optimal control problem in entropy-regularized RL thus becomes equivalent to a problem in Bayesian inference.

Let $\mathcal{O}_{1:N}$ define the event for which all steps in a trajectory τ are optimal, i.e., $\mathcal{O}_{1:N} \doteq \bigcap_{i=1}^{N}(\mathcal{O}_i = 1)$. To make connections to the "statistical mechanics of trajectories" formalism in NESM [19], let us denote by $E_\tau = -R_\tau$ the accumulated *energetic cost* for a trajectory τ. From Bayes's theorem, it follows that the posterior probability distribution for trajectories,

conditioned on $\mathcal{O}_{1:N}$, is given by

$$p(\tau|\mathcal{O}_{1:N}) = \frac{p(\tau)e^{-\beta E_\tau}}{\sum_\tau p(\tau)e^{-\beta E_\tau}}. \tag{6}$$

From the inference perspective, the central problem in entropy-regularized RL is now to determine the posterior distributions for the policy, dynamics, and initial state, *conditioned on optimality*. As noted, these posterior distributions correspond to the solution of the optimal control problem in entropy-regularized RL.

In many practical RL problems, control of system dynamics and initial state distributions is unfeasible. In these cases, the posterior dynamics and initial state distributions must be constrained to exactly match the prior dynamics and initial state distributions and the optimization is carried out by varying the policy alone. We will refer to this approach as the *constrained* optimization approach to entropy-regularized RL. In the constrained optimization problem, the agent only has control over the policy. The optimal trajectory distribution for the constrained problem can therefore be decomposed as

$$p(\tau|\mathcal{O}_{1:N}) = p(s_1)\prod_{t=1}^{N} p(s_{t+1}|s_t, a_t)\pi(a_t|s_t, \mathcal{O}_{1:N}). \tag{7}$$

The preceding (constrained) problem formulation is to be contrasted with the unconstrained optimization problem, where the agent also has control over the transition dynamics and initial state distributions. In this case, the optimal trajectory distribution can be decomposed as [15]

$$p(\tau|\mathcal{O}_{1:N}) = p(s_1|\mathcal{O}_{1:N})\prod_{t=1}^{N} p(s_{t+1}|s_t, a_t, \mathcal{O}_{1:N}) \times \pi(a_t|s_t, \mathcal{O}_{1:N}). \tag{8}$$

In the remainder of the paper, unless otherwise stated, we will focus on the solution of the unconstrained optimization problem in entropy-regularized RL, where the transition dynamics and the initial state distribution are optimized along with the policy. We note that the framework developed in this work also leads to the solution of the constrained optimization problem, which will be shown elsewhere.

III. SOLUTION USING LARGE DEVIATION THEORY

We now proceed to provide an analytical solution to the central problem of entropy-regularized RL in the long-time limit. Without loss of generality [15], we consider reward functions such that the maximum reward is set to zero and we have $r(s, a) \leqslant 0$ for all s, a. In this case, Eq. (5) indicates that, in the long-time limit, optimality of the entire trajectory is a rare event and the problem of determining the posterior policy and dynamics corresponds to conditioning on such a rare event. Research in NESM [20,22] has developed a framework for characterizing Markovian processes conditioned on rare events. In the following, we show how this framework leads to analytical expressions for quantities of interest in entropy-regularized RL. We note that the core of the derivation runs parallel to previous results deriving the Doob h-transform in discrete-time Markov chains [31–34]. In the following, our focus is on applying this framework to obtain new results for entropy-regularized RL.

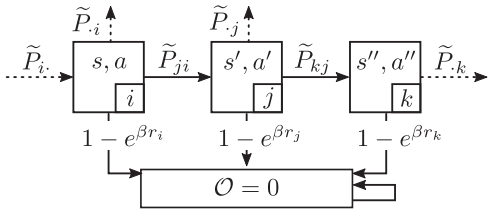

FIG. 1. System dynamics in the extended model with transition matrix $\overline{P}$. Transition $i \to \mathcal{O} = 0$ occurs with probability $1 - e^{\beta r_i}$. The introduction of an absorbing state provides an interpretation for the binary random variable $\mathcal{O}$. Conditioning on optimality (i.e., $\mathcal{O} = 1$) is equivalent to conditioning on nonabsorption.

Let $z = (s, a)$, $z' = (s', a')$ denote two consecutive state-action tuples. We can combine the system dynamics $p(s'|s, a)$ with the fixed prior policy $\pi(a'|s')$ to compose the corresponding transition matrix for the discrete time Markov chain

$$P_{ji} = p(z' = j|z = i) = p(s'|s, a)\pi(a'|s'). \tag{9}$$

Based on the connection to large deviation theory [35], let us define the *tilted transition matrix*

$$\widetilde{P}_{ji} = P_{ji}e^{\beta r_i}, \tag{10}$$

where $r_i = r(z = i) = r(s, a)$ denotes the reward associated to the tuple (s, a). Note that the tilted matrix is not a stochastic matrix and thus it cannot be interpreted as a transition matrix for a Markov chain that conserves probability. To address this issue, we introduce an additional absorbing state for the agent such that the extended transition matrix $\overline{P}$ (as defined below) *is* a stochastic matrix:

$$\overline{P} \equiv \begin{bmatrix} \widetilde{P} & 0 \\ \delta & 1 \end{bmatrix}, \tag{11}$$

where δ is defined such that $\sum_j \overline{P}_{ji} = 1$, i.e., $\delta_i = 1 - e^{\beta r_i}$.

The extended model introduced above provides an interpretation for the optimality variable introduced in Eq. (5) as specifying the probability of nonabsorption (see Fig. 1). Let us consider the system's evolution for N time steps using the transition matrix $\overline{P}$. Imposing the condition $\mathcal{O}_{1:N}$ is equivalent to conditioning on nonabsorption for all N time steps. Thus the optimal trajectory distribution is generated by considering the probability distribution over trajectories generated by $\overline{P}$, conditional on no transitions to the absorbing state for the entire trajectory. This interpretation allows us to make connections to the theory of quasistationary distributions [33,34] which can be used to analyze Markovian processes conditioned on nonabsorption.

For the dynamics generated by $\overline{P}$, given an initial state-action pair i, the probability of transitioning to state-action pair j after taking N steps is given by $[\widetilde{P}^N]_{ji}$. In the following, we assume that $\widetilde{P}$ is a primitive matrix, meaning that the corresponding dynamics is irreducible and aperiodic. In this case, the Perron-Frobenius theorem implies that $\widetilde{P}$ has a unique dominant eigenvalue ρ with a corresponding unique right eigenvector $\mathbf{v}$ (with $v_i > 0$) and a unique left eigenvector $\mathbf{u}$ (with $u_i > 0$). The normalization of the eigenvectors is chosen

such that $\sum_i v_i = 1$ and $\sum_i u_i v_i = 1$ [33]. Furthermore since $\widetilde{P}$ is substochastic (column sums between 0 and 1), we must have $\rho < 1$ and so we define $\theta > 0$ such that $\rho = e^{-\beta\theta}$.

We now consider the limit of large N, for which, using the spectral decomposition of $\widetilde{P}$, we have

$$[\widetilde{P}^N]_{ji} \approx e^{-\beta\theta N} u_i v_j \tag{12}$$

Furthermore, let $e^{-\beta\xi}$ denote the magnitude of the next dominant eigenvalue. Then the convergence of the preceding equation is exponential in N, i.e., the condition determining the long-time limit corresponds to $e^{-N\beta(\xi-\theta)} \ll 1$.

Now the probability that a trajectory starting with state-action pair $z_1 = (s_1, a_1)$ is optimal for N steps is given by

$$P(\mathcal{O}_{1:N}|z_1 = i) = \sum_j [\widetilde{P}^N]_{ji} \approx e^{-\beta\theta N} u_i. \tag{13}$$

This result can be used to derive the posterior distribution over trajectories conditioned on optimality. Typically, the difficulty in deriving expressions for the posterior distribution stems from estimating the partition sum in the denominator of Eq. (6). However, we note that the partition sum is given by $P(\mathcal{O}_{1:N}) = \sum_i p(z = i) P(\mathcal{O}_{1:N}|z = i)$ and thus can be estimated using the results derived.

To derive expressions for the posterior dynamics and state distributions conditioned on optimality, we define, consistent with the terminology in NESM, the *driven transition matrix*

$$[P_d]_{ji} = p(z' = j|z = i, \mathcal{O}_{1:N}). \tag{14}$$

This definition implies that the driven transition matrix is the generator of the Markov chain corresponding to the optimal dynamics. In the long-time limit, we obtain that the driven matrix is given by (see Appendix B 1)

$$[P_d]_{ji} = \frac{\widetilde{P}_{ji} u_j}{e^{-\beta\theta} u_i}, \tag{15}$$

which recovers the expression for the driven model as a generalized Doob h-transform in recent work in NESM [20,22,23]. It is interesting to note that our analysis recovers this result based on Bayesian inference of the posterior trajectory distribution.

The result for the driven matrix can be used to derive the following expressions for the optimal dynamics, policy, and initial state-action pair distributions (see Appendices B 2 and B 3)

$$p(s'|s, a, \mathcal{O}_{1:N}) = \frac{p(s'|s, a) e^{\beta r(s,a)}}{e^{-\beta\theta} u(s, a)} \sum_{a'} u(s', a') \pi(a'|s'), \tag{16}$$

$$\pi(a|s, \mathcal{O}_{1:N}) = \frac{u(s, a)\pi(a|s)}{\sum_{a'} u(s, a')\pi(a'|s)}, \tag{17}$$

$$p(s_1, a_1|\mathcal{O}_{1:N}) = \frac{p(s_1, a_1) u(s_1, a_1)}{\sum_{(s'_1, a'_1)} p(s'_1, a'_1) u(s'_1, a'_{e1})}. \tag{18}$$

The preceding equations, which are among the main results of this paper, show that in the long-time limit the optimal dynamics can be completely characterized by the dominant eigenvalue and the corresponding left eigenvector of the tilted matrix $\widetilde{P}$. While previous work has shown how a special class of MDPs are linearly solvable [36,37], our results show that linear solutions can be obtained for more general MDP models in the long-time limit.

The significance of this result is that it provides a closed-form solution for the central problem of entropy-regularized RL [stated in Eq. (8)]. For the case of deterministic dynamics, the results show that the optimal dynamics is unchanged from the original dynamics and the optimal policy is determined by the left eigenvector $\mathbf{u}$. For the case of stochastic dynamics, the results allow us to determine how the original dynamics must be controlled to obtain the optimal dynamics.

IV. VALUE FUNCTIONS AND STATISTICAL MECHANICS

The results derived for the optimal dynamics can be used to derive analytical expressions for optimal value functions in entropy-regularized RL [also called soft value functions [15] and denoted by $Q(s, a)$ and $V(s)$] and to make further connections to statistical mechanics. The optimal value function $Q(s, a)$ represents the expected future return to be collected, given that action a is taken from the initial state s, and the optimal dynamics and policy are followed thereafter. Note that this expected future return includes the penalization given by the entropic cost term $\beta^{-1}\mathcal{H}$ [see Eq. (4)]. Specifically $Q(s, a)$ is obtained by maximizing the average return over the controlled trajectory distribution: $\mathbb{E}_{p_c(\tau|s,a)}[R_\tau] - \frac{1}{\beta}\mathcal{H}(p_c(\tau|s, a)||p(\tau|s, a))$. Note that, if we instead consider the energetic costs over trajectories (i.e., $E_\tau = -R_\tau$), the problem of maximizing average returns is equivalent to the problem of minimizing average costs: $\mathbb{E}_{p_c(\tau|s,a)}[E_\tau] + \frac{1}{\beta}\mathcal{H}(p_c(\tau|s, a)||p(\tau|s, a))$, in correspondence with Eq. (4). In the following, we show how this optimization problem can be solved by connecting to the free energy concept from statistical mechanics.

To find the optimal value function, we need to consider the trajectory distribution corresponding to optimal control. Conditioned on the first step $z_1 = (s, a)$, the optimal trajectory distribution is given by

$$p(\tau|s, a, \mathcal{O}_{1:N}) = \frac{1}{Z_p(s, a)} p(\tau|s, a) e^{-\beta E_\tau}, \tag{19}$$

where $Z_p(s, a) = \sum_\tau p(\tau|s, a) e^{-\beta E_\tau}$ can be regarded as the partition function corresponding to the nonequilibrium free energy function

$$F_p(s, a) = -\frac{1}{\beta} \ln Z_p(s, a). \tag{20}$$

We note that the free energy defined above corresponds to the lower bound of the entropy-regularized RL objective, representing the minimized expected total cost with both energetic and entropic contributions [38],

$$F_p(s, a) \leqslant \mathbb{E}_{p_c(\tau|s,a)}[E_\tau] + \frac{1}{\beta}\mathcal{H}(p_c(\tau|s, a)||p(\tau|s, a)),$$

and equality is attained when the controlled trajectory distribution is given by Eq. (19). Thus the problem of minimizing the expected costs, or equivalently maximizing the expected return, is solved by the free energy, and correspondingly we obtain $Q(s, a) = -F_p(s, a)$. In other words, the function that maximizes the expected total returns in entropy-regularized

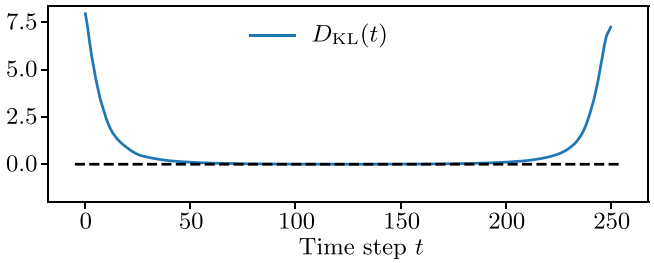

FIG. 2. Comparison of the optimal state-action pair distribution and its approximation using the Perron eigenvectors of the tilted matrix $\widetilde{P}$, as a function of time step t, with $N = 250$. The Kullback-Leibler divergence between the exact, time-dependent distribution and the bulk/stationary distribution estimated using Eq. (23) is shown. The plot shows that the ratio $\frac{u(s_t,a_t)v(s_t,a_t)}{p(s_t,a_t|\mathcal{O}_{1:T})} \approx 1$ in the "bulk" region of the trajectory.

RL $[Q(s,a)]$ is given by $e^{\beta Q(s,a)} = Z_p(s,a) = p(\mathcal{O}_{1:N}|s,a)$, consistent with [15]. This result, in combination with Eq. (13) and the definition of the state-dependent value function $V(s)$, $e^{\beta V(s)} = \sum_a \pi(a|s) e^{\beta Q(s,a)}$, yields the relations

$$\beta Q(s,a) = -\beta\theta N + \ln u(s,a), \tag{21}$$

$$\beta V(s) = -\beta\theta N + \ln \sum_a \pi(a|s) u(s,a). \tag{22}$$

Thus the value functions in entropy-regularized RL can be obtained using the dominant eigenvalue and the left Perron eigenvector of the tilted matrix $\widetilde{P}$. These results have been validated by comparing with the dynamic programming solution for entropy-regularized RL (see Appendix D). The significance of the preceding equations is that they provide a mapping between problems of interest in NESM and entropy-regularized RL such that approaches from one field can be used to solve problems in the other. For example, using the derived equations, function approximators, a popular tool in deep reinforcement learning for estimating value functions [39], can potentially be used as a method for calculating the left and right dominant eigenvectors of the tilted generator in NESM.

Besides the value functions, other quantities of interest in RL can also be obtained using the Perron-Frobenius eigenvalue and the corresponding eigenvectors, as previously noted in diverse systems of interest [20,33,40,41]. For example, in the long-time limit the right eigenvector gives the probability of observing a state-action pair conditioned on optimality: $p(s_t, a_t|\mathcal{O}_{1:t-1}) = v(s_t, a_t)$. Using Eq. (18), for t such that $t \to \infty$ and $(N-t) \to \infty$ (i.e., the "bulk" region of the trajectory), we also have (see Appendix B 4)

$$p(s_t, a_t|\mathcal{O}_{1:N}) \approx u(s_t, a_t) v(s_t, a_t). \tag{23}$$

We note that $u(s,a)v(s,a)$ represents the components of the dominant right eigenvector of the driven matrix P_d, i.e., the components of the steady-state distribution over state-action pairs generated by the driven dynamics.

As shown in Fig. 2, the exact optimal state-action pair distribution is in excellent agreement with the approximation obtained using the steady-state distribution of the driven dynamics, for time t in the "bulk" region of the trajectory (i.e., far from the extremities at $t = 0$ and $t = N$). Given that the steady-state distribution over state-action pairs is a quantity of significant interest in RL applications such as Inverse RL [42], the result obtained in Eq. (23) can significantly impact the computations involved in such RL approaches.

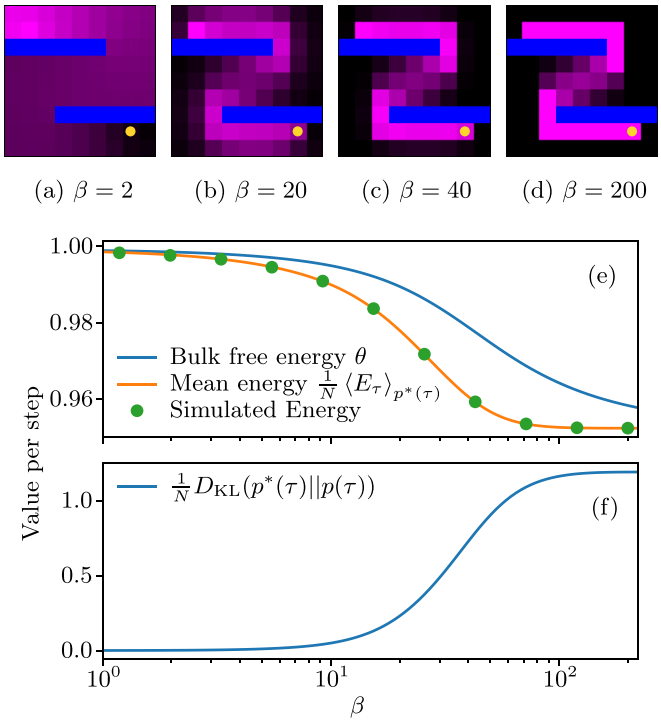

FIG. 3. Results for a 9 by 9 maze, trajectory length $N = 10^4$. Panels (a)–(d) show how state occupation frequencies (derived from the optimal trajectory distribution) change with temperature. Panels (e) and (f) show the mean energetic costs, and relative entropy per time step as functions of β.

To further validate the theory presented, we consider the "grid-world" setting shown in Figs. 3(a)–3(d) in which an agent can take actions a by deterministically moving up, down, left, or right. The state s of the agent is simply the grid cell in which it resides. The agent's task is to navigate to the only rewarding state: the goal, indicated by the yellow circle. The initial state of the agent is in the top left part of the maze. The shading of states represents the steady-state distribution $\sum_a u(s,a)v(s,a)$ for various values of the control parameter, β. We note that, as $\beta \to \infty$, the agent acts greedily by not deviating from the shortest path, that is, the most probable trajectories are those with higher rewards. This observed behavior reveals the role of the β parameter, which is to control the preference of the agent to purely minimize energy (maximize rewards) in exchange for stochasticity.

In the limit of large N, we have $\frac{F(s,a)}{N} \to \theta$, which can be interpreted as the "bulk" free energy per time step. Furthermore, we can also obtain approximations for quantities of interest such as the mean energetic cost per time step, through the steady state distribution in Eq. (23), resulting in the following expression:

$$\frac{1}{N}\mathbb{E}[E_\tau] = -\sum_{s,a} u(s,a)v(s,a)r(s,a). \tag{24}$$

As shown in Fig. 3(e), the preceding equation is in excellent agreement with results from simulations. We further note that as the inverse temperature parameter β is varied, the optimal trajectory distribution switches from primarily minimizing entropic costs at high temperatures (low β) to primarily minimizing energetic costs at low temperatures (high β). The approach developed therefore not only enables us to obtain the value functions of interest in entropy-regularized RL, but to also derive analytical expressions for the energetic and entropic contributions, which were previously unavailable.

V. u-θ LEARNING

The framework developed shows how several quantities of interest in entropy-regularized RL can be obtained using the dominant eigenvalue and the corresponding left eigenvector of the tilted matrix. In the following, we show how these quantities can be obtained in a *model-free* setting (that is, without explicit knowledge of the dynamics and rewards) by allowing the agent to collect experience by randomly exploring using the original transition dynamics.

By taking the sum over the columns of the driven matrix in Eq. (15), we note that the left eigenvector elements can be written as an expectation value over the original transition dynamics. Correspondingly, the dominant eigenvalue and left eigenvector can be obtained through a learning process based on the following equation:

$$u(s,a)e^{-\beta\theta} = e^{\beta r(s,a)}\mathbb{E}_{\sim p(s',a'|s,a)}[u(s',a')]. \tag{25}$$

The corresponding update equations for learning $u(s,a)$ and θ are

$$u(s,a) \leftarrow (1-\alpha)u(s,a) + \alpha\frac{e^{\beta r(s,a)}}{e^{-\beta\theta}}u(s',a'), \tag{26}$$

$$e^{-\beta\theta} \leftarrow (1-\alpha_\theta)e^{-\beta\theta} + \alpha_\theta e^{\beta r(s,a)}\frac{u(s',a')}{u(s,a)}, \tag{27}$$

where α and α_θ are their respective learning rates [43]. Further refinements of the algorithm outlined above can be developed following the connections to learning algorithms for risk-sensitive control [44]. Note that the prior policy is used for sampling actions during the training process [see Eq. (25)]. Thus this model-free approach to RL, which we term u-θ learning, is fundamentally an *off-policy* approach [45] wherein the optimal policy is obtained via system exploration using the prior policy. Our simulations (see Appendix D) indicate that optimal policies obtained using this method are in excellent agreement with the corresponding results obtained using dynamic programming [46] on the soft Bellmann backup equation. Appendix C shows how the soft Bellmann backup equation arises from the definition of the tilted matrix $\widetilde{P}$.

In conclusion, we have established a mapping between entropy-regularized RL and recent research in NESM using large deviation theory. The results derived include analytical expressions for quantities of interest in RL and lead to a learning algorithm for model-free RL. The results obtained have thus established a framework for analyzing optimization problems using entropy-regularized RL, and generalizations of this approach hold promise for obtaining solutions to a broader range of optimization problems in physics and machine learning.

ACKNOWLEDGMENTS

The authors acknowledge funding support from the NSF through Award No. DMS-1854350.

APPENDIX A: IMPLEMENTATION DETAILS

For the purposes of testing and validation we have developed an implementation of the method using Python, and used the Gym environment framework developed by OpenAI [47]. Since we focus on discrete state-action spaces, we shall work with the FrozenLake Gym environment, which we have modified to meet our needs regarding the transition dynamics and reward structure.

The code's implementation includes model-based and model-free solutions, along with example scripts to use our method. The code is made available as Supplemental Material in this publication [48], and as a Github repository [49].

APPENDIX B: DRIVEN DYNAMICS AND OPTIMAL DISTRIBUTIONS

1. Driven dynamics

The probability distribution for trajectories, $\tau_{1:T} = (z_1, z_2, \ldots, z_T)$ with $z_t = (s_t, a_t)$, conditioned on optimality is given by [see Eq. (6)]

$$p(\tau_{1:T}|\mathcal{O}_{1:T}) = \frac{p(\tau_{1:T},\mathcal{O}_{1:T})}{p(\mathcal{O}_{1:T})} = \frac{p(\tau)e^{-\beta E_\tau}}{\sum_\tau p(\tau)e^{-\beta E_\tau}},$$

For notational convenience, let $z_t = (s_t, a_t) = i$ and $z_{t+1} = (s'_{t+1}, a'_{t+1}) = j$ denote two consecutive state-action tuples in the trajectory $\tau_{1:T}$, with $1 \leqslant t < T$. The corresponding elements of the driven and tilted matrices are, by definition,

$$[P_d]_{ji} = p(s'_{t+1}, a'_{t+1}|s_t, a_t, \mathcal{O}_{1:T}),$$
$$[\widetilde{P}]_{ji} = p(s'_{t+1}, a'_{t+1}|s_t, a_t)e^{\beta r(s_t,a_t)},$$

From the above equations, it can be seen that the tilted matrix is time independent whereas the driven matrix will, in general, depend on the time index t. In the following, we consider the long-time limit $(T-t) \to \infty$. In this case, we will see that the driven matrix is independent of the time index t.

Let us divide the trajectory $\tau_{1:T}$ into two parts such that $\tau_{1:t-1} = (z_1, z_2, \ldots, z_{t-1})$ and $\tau_{t:T} = (z_t, z_{t+1}, \ldots, z_T)$. We will first focus on $\tau_{t:T}$ in the limit $(T-t) = N \to \infty$. Using the definition of the driven matrix, we have

$$p(\tau_{t+2:T}, z_{t+1} = j|z_t = i, \mathcal{O}_{t:T}) = p(\tau_{t+2:T}|z_{t+1} = j, \mathcal{O}_{t+1:T})[P_d]_{ji} \tag{B1}$$

Using Eq. (6), the left-hand side of Eq. (B1) can also be expressed as

$$p(\tau_{t+2:T}, z_{t+1} = j|z_t = i, \mathcal{O}_{t:T}) = \frac{p(\tau_{t+2:T}, \mathcal{O}_{t+1:T}|z_{t+1} = j)}{p(\mathcal{O}_{t:T}|z_t = i)}[\widetilde{P}]_{ji}. \tag{B2}$$

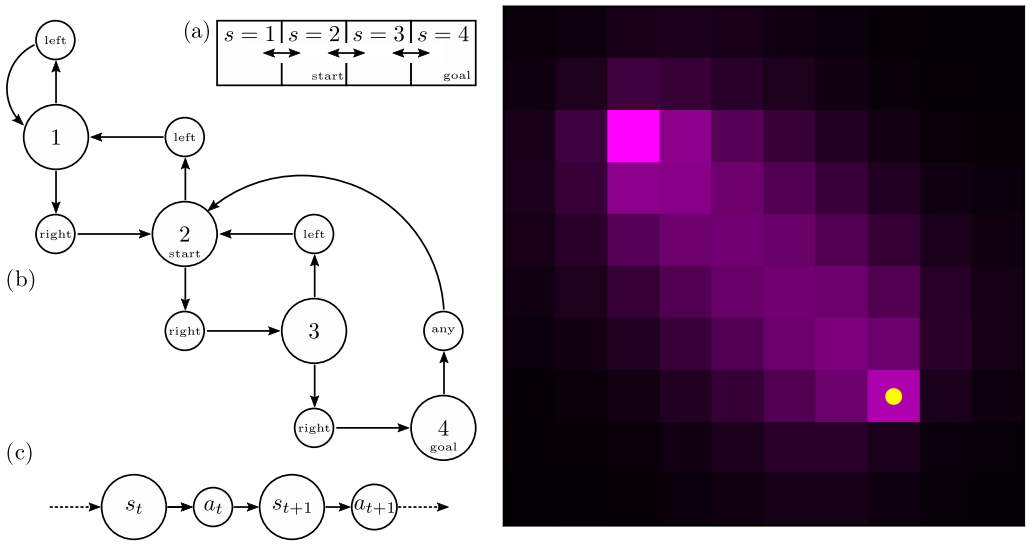

FIG. 4. (a) A four-state one-dimensional (1D) maze with two actions available to navigate. (b) The corresponding graphical model. Here we model an irreducible Markov chain by making the agent return to the initial state after reaching the goal state. (c) Part of the Markov chain at time step t. Right: representation of the stationary distribution resulting from the optimal dynamics, for a larger 10 by 10 2D maze and four available actions: left, down, right, up.

In Eq. (B1) using the substitution

$$p(\tau_{t+2:T}|z_{t+1}=j,\mathcal{O}_{t+1:T})=\frac{p(\tau_{t+2:T},\mathcal{O}_{t+1:T}|z_{t+1}=j)}{p(\mathcal{O}_{t+1:T}|z_{t+1}=j)},$$

and comparing with Eq. (B2), we get

$$[P_d]_{ji}=\frac{[\widetilde{P}]_{ji}\,p(\mathcal{O}_{t+1:T}|z_{t+1}=j)}{p(\mathcal{O}_{t:T}|z_t=i)}. \tag{B3}$$

Taking the long-time limit and approximating the tilted transition matrix using the dominant contribution,

$$P(\mathcal{O}_{t:T}|z_t=i)=\sum_j[\widetilde{P}^N]_{ji}=e^{-\beta\theta N}u_i,$$

$$P(\mathcal{O}_{t+1:T}|z_{t+1}=j)=\sum_k[\widetilde{P}^{N-1}]_{kj}=e^{-\beta\theta(N-1)}u_j. \tag{B4}$$

Substituting in Eq. (B3) we find that the driven matrix is given by the Doob h-transform [see Eq. (15)]:

$$[P_d]_{ji}=\frac{\widetilde{P}_{ji}u_j}{e^{-\beta\theta}u_i}.$$

2. Optimal policy

To derive the optimal policy, we begin with the observation

$$p(s_t,a_t|\mathcal{O}_{t:T})=\frac{p(s_t,a_t)p(\mathcal{O}_{t:T}|s_t,a_t)}{\sum_{(s_t,a_t)}p(s_t,a_t)p(\mathcal{O}_{t:T}|s_t,a_t)}. \tag{B5}$$

Using the approximation in Eq. (B4), we can rewrite Eq. (B5) as

$$p(s_t,a_t|\mathcal{O}_{t:T})=\frac{p(s_t,a_t)u(s_t,a_t)}{\sum_{s_t,a_t}p(s_t,a_t)u(s_t,a_t)}. \tag{B6}$$

Note that the preceding equation is valid for times t such that $(T-t)\gg 1$. In particular, it can be applied for the initial time-step to obtain the optimal initial state-action pair distribution result derived in the main text. For general t, the optimal state distribution can be obtained from Eq. (B6) as

$$p(s_t|\mathcal{O}_{t:T})=\frac{\sum_{a_t}p(s_t,a_t)u(s_t,a_t)}{\sum_{s_t,a_t}p(s_t,a_t)u(s_t,a_t)}.$$

From the preceding equations, we see that, in the long-time limit $(T-t)\to\infty$, the optimal state-action pair distribution is time independent. Therefore, using these equations and suppressing the time index, we obtain that the optimal policy is given by

$$p(a|s,\mathcal{O}_{1:T})=\frac{p(a|s)u(s,a)}{\sum_a p(a|s)u(s,a)},$$

$$\pi^*(a|s)=\frac{\pi(a|s)u(s,a)}{\sum_a\pi(a|s)u(s,a)}, \tag{B7}$$

where $\pi^*(a|s)$ denotes the optimal policy and $\pi(a|s)$ is the prior policy.

3. Optimal transition dynamics

To derive the optimal transition dynamics, we first write Eq. (15) as

$$p(s',a'|s,a,\mathcal{O}_{1:T})=\frac{p(s',a'|s,a)e^{\beta r(s,a)}u(s',a')}{e^{-\beta\theta}u(s,a)},$$

$$\pi^*(a'|s')p^*(s'|s,a)=\frac{\pi(a'|s')p(s'|s,a)e^{\beta r(s,a)}u(s',a')}{e^{-\beta\theta}u(s,a)}. \tag{B8}$$

By substituting the optimal policy in Eq. (B7) into Eq. (B8), we find that the optimal transition dynamics is given by

$$p^*(s'|s,a)=\frac{p(s'|s,a)e^{\beta r(s,a)}}{e^{-\beta\theta}u(s,a)}\sum_{a'}\pi(a'|s')u(s',a').$$

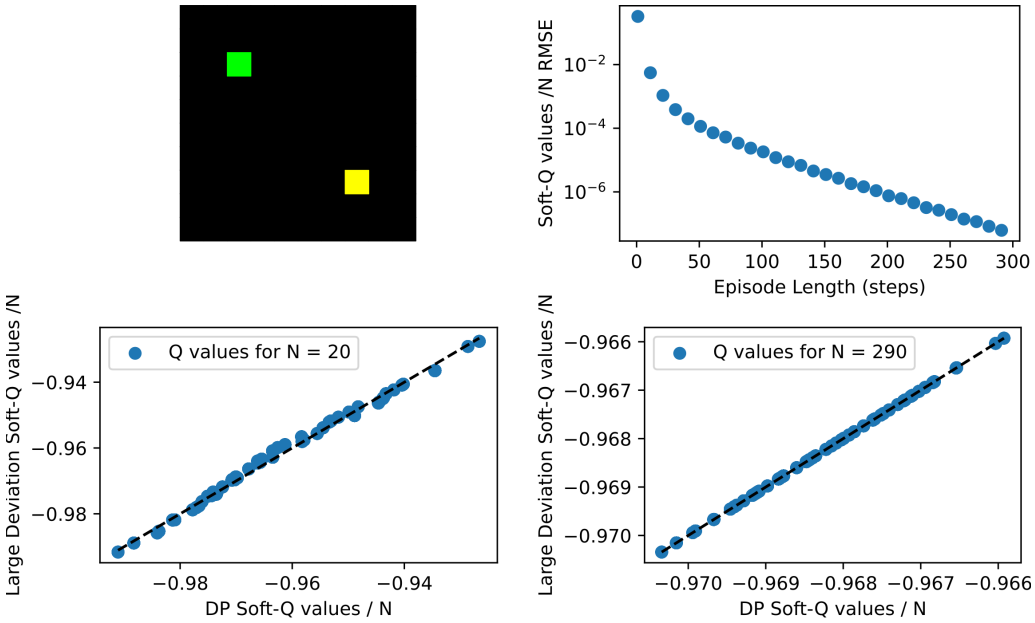

FIG. 5. Comparison of the soft-Q values computed by the large deviation approach vs. the dynamic programming solution. Top left: The 10 by 10 empty maze used for the plots in this figure. Top right: Root mean squared deviations of Q values between the large deviation and dynamic programming solutions, as a function of trajectory length. Bottom left: 20 step trajectories. Bottom right: 290 step trajectories. Here we can see perfect correlation between both solutions, for long enough trajectories.

4. Optimal steady-state distribution

Now we consider the initial part of the trajectory $\tau_{1:t-1}$. Consider

$$p(z_t = j|z_1 = i, \mathcal{O}_{1:t-1}) = \frac{p(z_t = j, \mathcal{O}_{1:t-1}|z_1 = i)}{p(\mathcal{O}_{1:t-1}|z_1 = i)}.$$

In the limit $t \to \infty$, using the Perron-Frobenius theorem and Eq. (B4), we get

$$p(z_t = j|z_1 = i, \mathcal{O}_{1:t-1}) = \frac{e^{-\beta\theta N} u_i v_j}{e^{-\beta\theta N} u_i} = v_j.$$

Thus, the optimal state-action pair distribution at time t is time-independent and independent of the initial state-action pair distribution. This distribution is given by the right eigenvector of the tilted matrix, and is referred to as the quasistationary distribution [33].

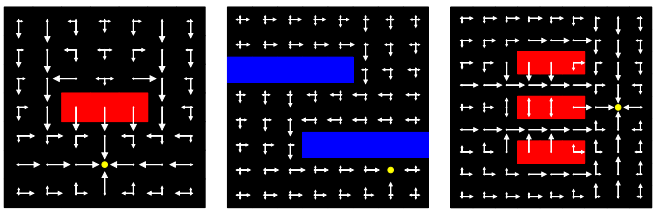

FIG. 6. Optimal policies for three different mazes, obtained from the dominant eigenvalue's corresponding left eigenvector of the tilted transition matrix. In these examples, the size of an arrow is proportional to the probability of taking a step in that direction. Blue squares represent hard walls, i.e., the agent is not allowed to step on them. Each step taken by the agent incurs a penalization ($r = -1$). When on a red square, there is a higher penalization ($r = -1.5$) and the agent is allowed to continue its trajectory. The goal state is depicted by the yellow circle, for which there is no penalization ($r = 0$) and the agent will be replaced at the initial state, regardless of the action taken.

The preceding equations have shown the equality $p(z_t = j|\mathcal{O}_{1:t-1}) = v_j$. To obtain the steady-state distribution of the optimal dynamics, we need to derive an expression for $p(z_t = j|\mathcal{O}_{1:T})$. To proceed, we split the trajectory in a similar way as above:

$$p(z_t = j|\mathcal{O}_{1:t-1}, \mathcal{O}_{t:T}) = \frac{p(\mathcal{O}_{t:T}|z_t = j, \mathcal{O}_{1:t-1})p(z_t = j|\mathcal{O}_{1:t-1})}{\sum_k p(\mathcal{O}_{t:T}|z_t = k, \mathcal{O}_{1:t-1})p(z_t = k|\mathcal{O}_{1:t-1})}.$$

Furthermore, using

$$p(\mathcal{O}_{t:T}|z_t = j, \mathcal{O}_{1:t-1}) = p(\mathcal{O}_{t:T}|z_t = j) = e^{-\beta\theta N} u_i$$

in combination with $p(z_t = j|\mathcal{O}_{1:t-1}) = v_j$, we get

$$p(z_t = j|\mathcal{O}_{1:T}) = \frac{e^{-\beta\theta N} u_j v_j}{e^{-\beta\theta N} \sum_k u_k v_k} = u_j v_j.$$

Thus the optimal state-action pair distribution in the "bulk" region of the trajectory [i.e., the times t such that $t \to \infty$ and $(T - t) \to \infty$] is time-independent and is given by the Hadamard product of the left and right eigenvectors of the tilted matrix. It is readily verified that this distribution also corresponds to the steady-state distribution of the driven matrix P_D.

APPENDIX C: DERIVATION OF SOFT BELLMAN BACKUP EQUATIONS

Recall that we write the indices $i = (s, a)$ and $j = (s', a')$ for two consecutive steps, and the transition matrix is

$$P_{ji} = p(s', a'|s, a) = p(s'|s, a)\pi(a'|s').$$

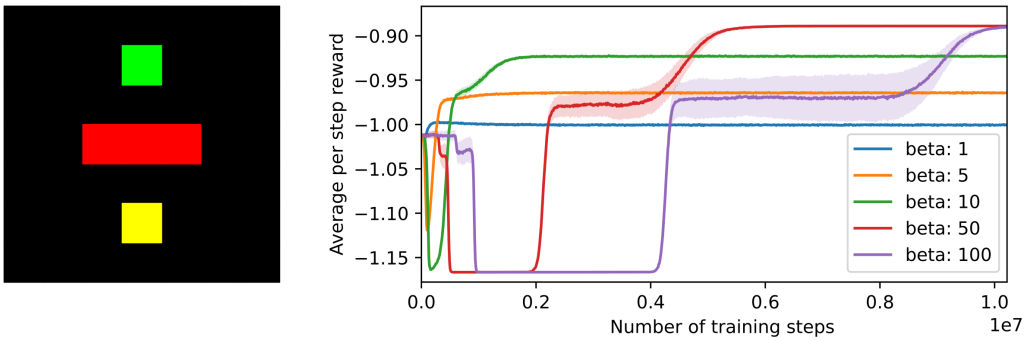

FIG. 7. Training evolution of u-θ learning agents for five different temperatures as a function of the training progress. Lower temperature agents take longer to converge. Note that the optimal greedy policy is recovered at the lowest temperatures.

From Markov chain theory, when given a transition matrix P, we interpret $[P^N]_{ji}$ as the probability of arriving at j after N steps, given that we start from i. Since the transition matrix P is a stochastic matrix, we have that $\sum_j [P^N]_{ji} = 1$. For large N, P^N leads to the stationary distribution for the corresponding Markov process.

Let us now consider the tilted transition matrix

$$\widetilde{P}_{ji} = e^{\beta r_i} P_{ji},$$

which represents a substochastic transition matrix. As pointed out in the main text, we can expand the graphical model with an extra state in such a way that we obtain a proper stochastic transition matrix. This extra state is an absorbing state, and any trajectory that reaches it is regarded as suboptimal.

We can write the probability of remaining optimal after taking N steps in the Markov chain as the probability of nonabsorption,

$$p(\mathcal{O}_{1:N}|s,a) = \sum_j [\widetilde{P}^N]_{ji}.$$

The preceding equation represents the so-called backward messages [15]. Using this we can write a recursive relation which then leads to the soft Bellman backup equation

$$\begin{aligned} p(\mathcal{O}_{1:N}|s,a) &= \sum_j \sum_m [\widetilde{P}^{N-1}]_{jm} e^{\beta r_i} P_{mi} \\ &= e^{\beta r(s,a)} \sum_{s',a'} p(s',a'|s,a) p(\mathcal{O}_{2:N}|s',a'). \end{aligned}$$

Now, using the definitions of the soft value functions in entropy-regularized RL [15],

$$\begin{aligned} \beta Q(s,a) &= \ln p(\mathcal{O}_{1:N}|s,a), \\ \beta V(s) &= \ln \sum_a \pi(a|s) \exp[\beta Q(s,a)], \end{aligned}$$

we obtain, consistent with the result derived in [15], the following soft backup equation:

$$\begin{aligned} Q(s,a) &= r(s,a) + \frac{1}{\beta} \ln \mathbb{E}\{\exp[\beta Q(s',a')]\} \\ &= r(s,a) + \frac{1}{\beta} \ln \left[\sum_{s'} p(s'|s,a) \exp[\beta V(s')] \right], \end{aligned}$$

where the expectation is taken with respect to the uncontrolled dynamics: the prior policy and the original transition dynamics.

APPENDIX D: EXPERIMENTAL VALIDATION

In order to validate the analytical framework proposed in the main text and derived here, we defined a series of grid-world mazes for which a complete dynamics model is available, i.e., all available states, actions and transition dynamics are known beforehand. We modified the OpenAI Gym environment "FrozenLakeEnv" [47], which has a discrete state-action space. Our version of this environment provides control over the reward function, stochastic behavior, and an

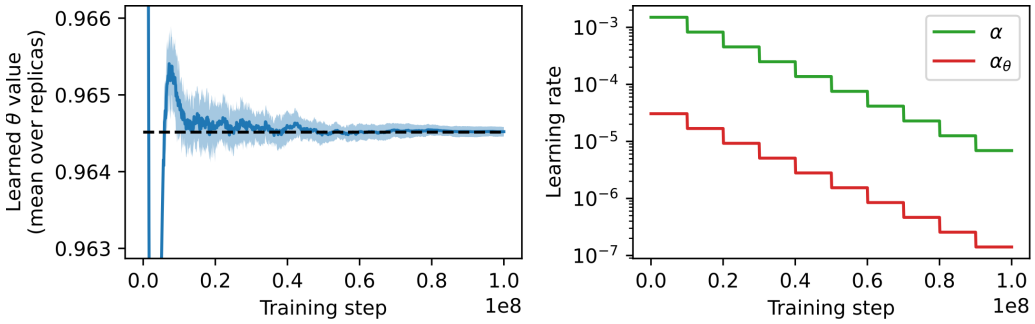

FIG. 8. Validation of solution by u-θ learning algorithm. The maze used is the same as in Fig. 7, with $\beta = 10$ and trajectory length $N = 1000$ steps. Left: Convergence of the θ parameter learned by the agent towards the target value as computed by the model-based version. The curve plots the mean values over 32 replicas, and the shaded area is the standard deviation. Right: Learning rate schedules used to learn the θ parameter.

option to define a cyclic mode that results in irreducible MDPs (see Fig. 4). With this setup, we are able to compute the optimal solution for the objective function in entropy-regularized RL. The resulting soft-Q value function has been compared with the dynamic programming result, which is obtained by directly computing the soft-Q and soft-V value functions at every step (see Fig. 5). Figure 6 shows three examples of mazes and corresponding optimal policies. In the figure we see how the policy can successfully steer the agent toward the goal state, while avoiding *dangerous* states.

Here we provide some details about the validation of the model-free version of our method (u-θ learning). The approach consists of a temporal difference method [see Eqs. (26) and (27) in the main text]. Validation of the algorithm has been performed by comparing to the exact solution as computed by dynamic programming. In Fig. 7 we show solutions to the displayed maze for several temperatures, as a function of training progress. Fig. 8 examines the learned parameters and their comparison with dynamic programming.

[1] S. D. Sarma, D.-L. Deng, and L.-M. Duan, Machine learning meets quantum physics, Phys. Today **72**(3), 48 (2019).

[2] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019).

[3] P. Mehta and D. J. Schwab, An exact mapping between the variational renormalization group and deep learning, arXiv:1410.3831.

[4] H. W. Lin, M. Tegmark, and D. Rolnick, Why does deep and cheap learning work so well?, J. Stat. Phys. **168**, 1223 (2017).

[5] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[6] M. Ostaszewski, L. M. Trenkwalder, W. Masarczyk, E. Scerri, and V. Dunjko, Reinforcement learning for optimization of variational quantum circuit architectures, Adv. Neural Inf. Process. Syst. **34**, 18182 (2021).

[7] A. Barr, W. Gispen, and A. Lamacraft, Quantum ground states from reinforcement learning, in *Proceedings of The First Mathematical and Scientific Machine Learning Conference* Vol. 107 (PMLR, 2020), pp. 635–653, https://proceedings.mlr.press/v107/barr20a.html.

[8] A. Bolens and M. Heyl, Reinforcement Learning for Digital Quantum Simulation, Phys. Rev. Lett. **127**, 110502 (2021).

[9] Y.-H. Zhang, P.-L. Zheng, Y. Zhang, and D.-L. Deng, Topological Quantum Compiling with Reinforcement Learning, Phys. Rev. Lett. **125**, 170501 (2020).

[10] S. Borah, B. Sarma, M. Kewming, G. J. Milburn, and J. Twamley, Measurement-Based Feedback Quantum Control with Deep Reinforcement Learning for a Double-Well Nonlinear Potential, Phys. Rev. Lett. **127**, 190403 (2021).

[11] J. Nousiainen, C. Rajani, M. Kasper, and T. Helin, Adaptive optics control using model-based reinforcement learning, Opt. Express **29**, 15327 (2021).

[12] D. C. Rose, J. F. Mair, and J. P. Garrahan, A reinforcement learning approach to rare trajectory sampling, New J. Phys. **23**, 013013 (2021).

[13] J. Yan and G. M. Rotskoff, Physics-informed graph neural networks enhance scalability of variational nonequilibrium optimal control, J. Chem. Phys. **157**, 074101 (2022).

[14] S. Whitelam, D. Jacobson, and I. Tamblyn, Evolutionary reinforcement learning of dynamical large deviations, J. Chem. Phys. **153**, 044113 (2020).

[15] S. Levine, Reinforcement learning and control as probabilistic inference: Tutorial and review, arXiv:1805.00909.

[16] H. J. Kappen, V. Gómez, and M. Opper, Optimal control as a graphical model inference problem, Mach. Learn. **87**, 159 (2012).

[17] E. Todorov, General duality between optimal control and estimation, in *2008 47th IEEE Conference on Decision and Control* (IEEE, Piscataway, NJ, 2008), pp. 4286–4292.

[18] K. Rawlik, M. Toussaint, and S. Vijayakumar, On stochastic optimal control and reinforcement learning by approximate inference, in *Proceedings of Robotics: Science and Systems VIII* (MIT Press, Cambridge, 2012).

[19] J. P. Garrahan, R. L. Jack, V. Lecomte, E. Pitard, K. van Duijvendijk, and F. van Wijland, First-order dynamical phase transition in models of glasses: an approach based on ensembles of histories, J. Phys. A: Math. Theor. **42**, 075007 (2009).

[20] R. L. Jack and P. Sollich, Large deviations and ensembles of trajectories in stochastic models, Prog. Theor. Phys. Suppl. **184**, 304 (2010).

[21] R. Chetrite and H. Touchette, Nonequilibrium Microcanonical and Canonical Ensembles and Their Equivalence, Phys. Rev. Lett. **111**, 120601 (2013).

[22] R. Chetrite and H. Touchette, Nonequilibrium markov processes conditioned on large deviations, Ann. Henri Poincaré **16**, 2005 (2015).

[23] R. Chetrite and H. Touchette, Variational and optimal control representations of conditioned and driven processes, J. Stat. Mech.: Theory Exp. (2015) P12001.

[24] H. D. Miller, A convexity property in the theory of random variables defined on a finite Markov chain, Ann. Math. Stat. **32**, 1260 (1961).

[25] D. Simon, Construction of a coordinate Bethe ansatz for the asymmetric simple exclusion process with open boundaries, J. Stat. Mech.: Theory Exp. (2009) P07017.

[26] R. M. L. Evans, Rules for Transition Rates in Nonequilibrium Steady States, Phys. Rev. Lett. **92**, 150601 (2004).

[27] R. M. L. Evans, Detailed balance has a counterpart in nonequilibrium steady states, J. Phys. A: Math. Gen. **38**, 293 (2005).

[28] A. Simha, R. M. L. Evans, and A. Baule, Properties of a nonequilibrium heat bath, Phys. Rev. E **77**, 031117 (2008).

[29] R. L. Jack and R. M. L. Evans, Absence of dissipation in trajectory ensembles biased by currents, J. Stat. Mech.: Theory Exp. (2016) 093305.

[30] A. Das, D. C. Rose, J. P. Garrahan, and D. T. Limmer, Reinforcement learning of rare diffusive dynamics, J. Chem. Phys. **155**, 134105 (2021).

[31] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*, Cambridge Mathematical Library (Cambridge University Press, Cambridge, UK, 2000).

[32] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, (American Mathematical Society, Providence, 2017).

[33] S. Méléard and D. Villemonais, Quasi-stationary distributions and population processes, Probab. Surv. **9**, 340 (2012).

[34] E. A. van Doorn and P. K. Pollett, Quasi-stationary distributions for discrete-state models, Eur. J. Oper. Res. **230**, 1 (2013).

[35] H. Touchette, A basic introduction to large deviations: Theory, applications, simulations, arXiv:1106.4146.

[36] E. Todorov, Linearly-solvable Markov decision problems, in *Advances in Neural Information Processing Systems, Vol. 19*, edited by B. Schölkopf, J. Platt, and T. Hoffman (MIT Press, Cambridge, 2007).

[37] E. Todorov, Efficient computation of optimal actions, Proc. Natl. Acad. Sci. USA **106**, 11478 (2009).

[38] E. A. Theodorou, Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations, Entropy **17**, 3352 (2015).

[39] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, Playing atari with deep reinforcement learning, arXiv:1312.5602.

[40] C. Giardinà, J. Kurchan, and L. Peliti, Direct Evaluation of Large-Deviation Functions, Phys. Rev. Lett. **96**, 120603 (2006).

[41] T. Nemoto, F. Bouchet, R. L. Jack, and V. Lecomte, Population-dynamics method with a multicanonical feedback control, Phys. Rev. E **93**, 062123 (2016).

[42] J. Fu, K. Luo, and S. Levine, Learning robust rewards with adverserial inverse reinforcement learning, in International Conference on Learning Representations (2018), https://openreview.net/forum?id=rkHywl-A-.

[43] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, 2018).

[44] V. S. Borkar, Learning algorithms for risk-sensitive control, in *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, 2010 (unpublished).

[45] S. Levine, A. Kumar, G. Tucker, and J. Fu, Offline reinforcement learning: Tutorial, review, and perspectives on open problems, arXiv:2005.01643.

[46] R. Bellman, The theory of dynamic programming, Bull. Am. Math. Soc. **60**, 503 (1954).

[47] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, OpenAI Gym, arXiv:1606.01540.

[48] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.5.023085 for example Python code to replicate results.

[49] https://github.com/argearriojas/2023-EntRegRL.