Bayesian Inference Approach for Entropy Regularized Reinforcement Learning with Stochastic Dynamics

Abstract

We develop a novel approach to determine the optimal policy in entropy-regularized reinforcement learning (RL) with stochastic dynamics. For deterministic dynamics, the optimal policy can be derived using Bayesian inference in the control-as-inference framework; however, for stochastic dynamics, the direct use of this approach leads to risk-taking optimistic policies. To address this issue, current approaches in entropy-regularized RL involve a constrained optimization procedure which fixes system dynamics to the original dynamics, however this approach is not consistent with the unconstrained Bayesian inference framework. In this work we resolve this inconsistency by developing an exact mapping from the constrained optimization problem in entropy-regularized RL to a different optimization problem which can be solved using the unconstrained Bayesian inference approach. We show that the optimal policies are the same for both problems, thus our results lead to the exact solution for the optimal policy in entropy-regularized RL with stochastic dynamics through Bayesian inference.

1 MOTIVATION

Reinforcement learning (RL) provides a promising framework for training artificial agents for goal-oriented tasks through trial and error interaction with the environment [Sutton and Barto, 2018, Zhu et al., 2020]. Specifically, the agent receives rewards in the process of solving a task according to a predefined reward function and this interaction informs the agent's behavior policy. The aim is to determine the optimal policy which, in the original formulation of reinforcement learning, maximizes the expected accumulated reward. The problem of RL can be addressed in the model-based [Atkeson and Santamaria, 1997, Boone, 1997, Abbeel and Ng, 2005, Berkenkamp et al., 2017, Asadi et al., 2018, Corneil et al., 2018, Lowrey et al., 2019] or model-free settings [Watkins and Dayan, 1992, Hasselt, 2010, Mnih et al., 2015, Kiumarsi et al., 2018]. In the former case, the agent has access to a model of the environment and in the latter case, it has access only to samples from the environment. Approaches based on RL have led to remarkable successes in robotics [Zhu et al., 2020], board games [Silver et al., 2018, Schrittwieser et al., 2020], and many other fields [Cao et al., 2021, Yu et al., 2019, Charpentier et al., 2021].

A more general framework is entropy-regularized reinforcement learning, which considers reward accumulation with an entropy-based regularization term [Haarnoja et al., 2017, 2018b, Nachum et al., 2017]. The entropic regularization term corresponds to a control cost associated with the control policy (relative to a prior policy) and leads to stochastic optimal policies that are robust to environmental changes [Eysenbach and Levine, 2021] and show improved exploration [Haarnoja et al., 2017]. This generalization of RL towards entropy-regularized RL also makes connections to statistical mechanics, given that the free energy is given by a similar combination of energy and entropy terms. A series of recent works have revealed new connections between non-equilibrium statistical mechanics and entropy-regularized RL, which have led to new algorithms and applications [Rose et al., 2021, Das et al., 2021, Arriojas et al., 2021].

One of the advantages of entropy-regularized RL is that it enables us to recast the problem of reward maximization into a problem of Bayesian Inference [Todorov, 2008, Rawlik et al., 2012, Kappen et al., 2012, Levine, 2018]. This insight brings the rich arsenal of tools in Bayesian inference [Koller and Friedman, 2009] to control and reinforcement learning, motivating the development of the control-as-inference framework [Rawlik et al., 2012, Kappen et al., 2012, Levine, 2018]. An important aspect of this approach is that, in the case of stochastic dynamics, an optimal control-as-inference solution involves inferring both a posterior policy as well as

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

a posterior transition dynamics. Correspondingly, a direct application of this approach for stochastic dynamics leads to the "optimistic agent problem" (see Fig. 1), in which the agent unreasonably assumes for the optimal solution that it can control not only its policy, but also the system dynamics. In practice, system dynamics is typically fixed (e.g. a robot with fixed physical parameters) and not within the agent's control. In such cases, the policy derived using the control-as-inference framework for entropy-regularized RL is sub-optimal. This problem of obtaining the optimal policy using Bayesian inference while imposing the constraint to keep the dynamics fixed is an open problem in entropy-regularized RL, which motivates the current work.

In this work, we take a step towards resolving this problem by showing how to solve the constrained dynamics optimization problem in entropy-regularized RL using a Bayesian inference-based solution. We also develop a model-free algorithm based on the results derived and validate our approach in tabular settings. A simple example illustrating the application of our approach is shown in Fig. 1. Our main contributions include the following:

- a formal mapping of the optimization problem in entropy-regularized RL with fixed (constrained) stochastic dynamics to a different problem for which the optimization is unconstrained with respect to the dynamics. The derived mapping ensures that the optimal policy is identical for the two optimization problems.
- an algorithm for obtaining the optimal policy for entropy-regularized RL with an arbitrary fixed dynamics (i.e. not necessarily constrained to original dynamics) which can also be applied to problems involving distribution shift for system dynamics.

2 RELEVANT WORK

Entropy-regularized RL can be seen as a particular case of the general problem of minimization of a *free energy functional*, wherein energetic quantities such as reward, value, and energy, are combined with entropic quantities such as entropy, cross entropy, and mutual information. Previously, the utility of this combination has been studied from the perspectives of i) cognitive science [Friston, 2009, Friston et al., 2006], ii) information theory [Tishby and Polani, 2011, Tiomkin and Tishby, 2017], iii) control [Mitter and Newton, 2000, Todorov, 2008, Watson et al., 2021], iv) robotics [Toussaint, 2009], v) reinforcement learning [Nachum et al., 2017, Haarnoja et al., 2018b, Levine, 2018]. The preceding is only a short list of prior work that invoke of the free energy formalism, which provides the reader with the big picture and puts the current work in the broader context.

The most relevant prior work to the current research is in the setting of RL [Rawlik et al., 2012, Nachum et al., 2017, Haarnoja et al., 2018b, Levine, 2018]. In particular, the

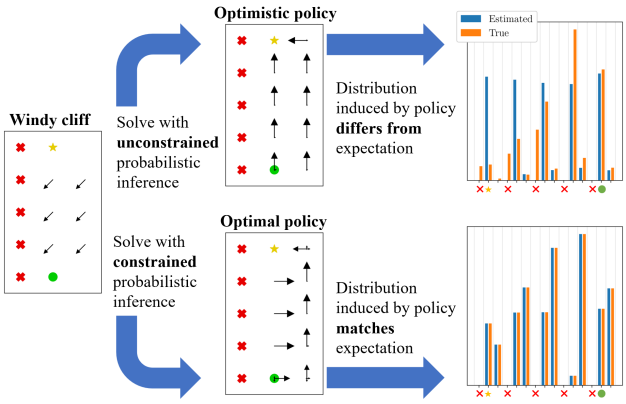

Figure 1: Demonstration of the optimistic agent problem in a cliff environment with stochastic dynamics. Left: The maze layout showing the force of wind in six states. At each time step the agent must choose a direction to walk and the wind may push in one direction with some probability which is determined by wind direction and intensity. Here there is 35% chance to move left, 35% to move down and 30% for no move due to wind. Each time step has fixed penalization $r = -1$. Red crosses represent traps with $r = -5$, and the golden star is the goal with $r = 0$. The MDP is such that the agent transitions to the start state (green circle) after stepping into a trap or the goal. Center: Policies computed with our proposed biasing method (bottom) and without biases (top), with $\beta = 50$. Right: The corresponding state visitation distribution for each policy. The optimistic agent fails to predict how often it will fall off the cliff, while the optimal solution has realistic expectations.

question that motivates our work, i.e. how to find the optimal policy using the framework of control-as-inference for the case of stochastic dynamics, has been clearly discussed in [Levine, 2018]. As noted in [Levine, 2018] the standard solution to this question within the formalism of control-as-inference results in a policy that leads to risk-taking behaviour which is undesirable. In this work we develop a novel mapping that leads to the derivation of an exact solution for entropy-regularized RL within the framework of control-as-inference in the general case of stochastic dynamics.

3 PRELIMINARIES

In this section, we overview the standard setting of Markov decision processes (MDP) in RL. Then, we discuss its extension to entropy-regularized RL, providing both the classical perspective of control-as-inference and the free energy perspective. The latter emphasises the usefulness of the properties of free energy for the derivation of the optimal solution. Then, we overview an existing analytical solution (in the long-time limit) for the general case of entropy-regularized RL with *unconstrained* stochastic dynamics [Arriojas et al., 2021], which we apply in Section 4 to solve the general case

of *constrained/fixed* stochastic dynamics.

3.1 MARKOV DECISION PROCESSES IN RL

In the following, we introduce the notation for the standard MDP formulation for RL [Puterman, 2014]. We will focus on the undiscounted finite horizon version with horizon T. The state of the system is denoted by $s \in \mathcal{S}$ and actions are denoted by $a \in \mathcal{A}$. The action of the agent is specified by the policy function $\pi(a|s)$ which represents the probability of choosing action a, given that the state is s. The initial state distribution is given by $\mu(s)$. Without any loss of generality, we take μ to be deterministic. The dynamics is determined by the state-transition function $p(s'|s,a)$ which denotes the probability of transitioning to state s' given that action a was chosen when in state s. The reward function $r(s,a)$ specifies the reward received after choosing action a in state s.

The objective in standard RL is to find the optimal policy that maximizes expected rewards collected by the agent, i.e.

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right], \tag{1}$$

where the expectation is taken over the possible trajectories generated by following π, and subject to the problem's dynamics. The summation represents the sequence of steps that form a trajectory.

3.2 ENTROPY-REGULARIZED RL

In entropy-regularized RL [Haarnoja et al., 2018a, Levine, 2018], the preceding objective function is modified to include an entropic regularization term, such that the optimal policy is given by

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t) - \frac{1}{\beta}\log\left(\frac{\pi(a_t|s_t)}{\pi^0(a_t|s_t)}\right)\right] \tag{2}$$

where β is an inverse temperature parameter and π^0 denotes the prior policy distribution. In the special case of maximum entropy RL (MaxEnt RL), the prior policy is taken to be the uniform distribution over actions [Levine, 2018]. In the above formulation, it is implicit that the system dynamics remains fixed to the original dynamics $p(s'|s,a)$ and the optimization is over the policy distribution π. Furthermore, we note that, without any loss of generality, we will consider reward functions such that $r(s,a) \leq 0$, since a constant offset for the reward function for all state-action pairs does not impact the optimal policy [Levine, 2018].

Let us now consider the preceding optimization problem from the trajectory perspective. Let $\tau := \{(s_t, a_t)\}_{t=0}^{T}$ denote a trajectory and (with a slight abuse of notation) let $p(\tau)$ denote the corresponding trajectory distribution with the dynamics fixed to the original system dynamics. Given the initial state distribution $\mu(s_1)$ and a control policy $\pi(a_t|s_t)$, the trajectory distribution can be expressed as

$$p(\tau) = \mu(s_1)\prod_{t=1}^{T} p(s_{t+1}|s_t, a_t)\pi(a_t|s_t). \tag{3}$$

When the control policy is taken to be the prior policy $\pi^0(a_t|s_t)$, the corresponding prior trajectory distribution will be denoted by $p_0(\tau)$. Furthermore, let us denote the energy of a trajectory τ as

$$E(\tau) = -\sum_{t=1}^{T} r(s_t, a_t).$$

It is readily seen that the optimization problem in Eqn. (2) is equivalent to determining the trajectory distribution $p(\tau)$ that *minimizes* the objective function:

$$J\left[p(\tau)\right] = \mathbb{E}_{\tau\sim p(\tau)}\left[E(\tau)\right] + \frac{1}{\beta}\mathcal{H}(p(\tau)|p_0(\tau)) \tag{4}$$

where $\mathcal{H}(p(\tau)|p_0(\tau))$ denotes the relative entropy between the prior and controlled trajectory distributions:

$$\mathcal{H}(p(\tau)|p_0(\tau)) = \sum_{\tau} p(\tau)\log\frac{p(\tau)}{p_0(\tau)}$$

3.3 CONTROL-AS-INFERENCE APPROACH

To connect to the control-as-inference approach, let us consider a general controlled trajectory distribution denoted by $q(\tau)$. In contrast with $p(\tau)$ in Eqn. (3), for $q(\tau)$ the system's transition dynamics is not constrained to be the same as the original dynamics. In this more general setting, the objective function is the same as in Eqn. (4) but with $p(\tau)$ replaced by the unconstrained trajectory distribution $q(\tau)$. We will refer to the problem of minimizing this objective $J\left[q(\tau)\right]$ as the *unconstrained* optimization problem.

The solution of the unconstrained optimization problem is related to the concept of free energy. Given a prior trajectory distribution $p_0(\tau)$, the corresponding free energy is defined as

$$F \doteq -\frac{1}{\beta}\log\mathcal{Z}, \text{ where } \mathcal{Z} \doteq \sum_{\tau} p_0(\tau)e^{-\beta E(\tau)}.$$

The connection to the unconstrained optimization problem is given by the relationship [Mitter and Newton, 2000, Todorov, 2008, Theodorou and Todorov, 2012]:

$$F = \inf_{q(\tau)}\left[\langle E\rangle_q + \frac{1}{\beta}\mathcal{H}(q(\tau)|p_0(\tau))\right]. \tag{5}$$

Therefore the free energy F above yields the solution to the unconstrained optimization problem.

Furthermore, the corresponding optimal trajectory distribution $q(\tau) = q^*(\tau)$ is given by [Mitter and Newton, 2000, Todorov, 2008, Theodorou and Todorov, 2012]

$$q^*(\tau) = \frac{p_0(\tau)e^{-\beta E(\tau)}}{\sum_\tau p_0(\tau)e^{-\beta E(\tau)}} \tag{6}$$

We will refer to the preceding result for the optimal trajectory distribution as the *inference approach solution*.

This result provides insight into the control-as-inference framework. This approach [Ziebart et al., 2010, Toussaint, 2009, Levine, 2018] involves the introduction of the binary random variable $\mathcal{O}_t$ such that

$$p(\mathcal{O}_t = 1|s_t, a_t) = \exp(\beta r(s_t, a_t)) \tag{7}$$

This choice is motivated by the observation that, conditioned on optimality (i.e. $\mathcal{O}_t = 1$ for all t), the posterior trajectory distribution $p(\tau|\mathcal{O}_{1:T})$ exactly corresponds to the optimal control distribution in Eqn. (6). Correspondingly, the posterior policy derived using this Bayesian approach is the optimal policy for the unconstrained optimization problem.

3.4 ENTROPY-REGULARIZED RL VIA UNCONSTRAINED OPTIMISATION

One of the advantages of the inference approach solution is that, in the long-time limit, it is possible to derive analytical expressions for the optimal policy and optimal dynamics. Recent work [Arriojas et al., 2021], using approaches from large deviation theory, has shown how the optimal dynamics and policy can be expressed in terms of the Perron-Frobenius eigenvalue ($e^{-\theta}$) and corresponding left eigenvector ($u(s, a)$) of a sub-stochastic matrix ($\widetilde{P}$) whose elements are given by

$$\widetilde{P}_{(s',a'),(s,a)} = p(s'|s, a)\pi^0(a'|s')e^{\beta r(s,a)}$$

Using this framework, it can be shown [Arriojas et al., 2021] that the posterior (i.e. optimal) transition dynamics p^* is related to the original transition dynamics by:

$$p^*(s'|s, a) \propto p(s'|s, a)e^{\beta V^*(s')} \tag{8}$$

where $V^*(s)$ is the optimal value function and the proportionality constant (for each s, a) is determined by normalization.

Let us now consider the *constrained optimization* problem, with the objective function defined by Eqn. (4), i.e. the transition dynamics is fixed to the original dynamics $p(s'|s, a)$. For the case of deterministic transition dynamics, the solution to the constrained optimization problem is provided by the inference approach solution. This can be seen from Eqn. (8), which shows that, for the case of deterministic dynamics, the optimal dynamics is the same as the original dynamics. However, for the case of stochastic dynamics, the same result indicates that the optimal dynamics is, in general, different from the original dynamics. Thus, the constraint that the optimal dynamics is the same as the original dynamics is satisfied by the inference approach solution for the case of deterministic dynamics but not for stochastic dynamics.

3.4.1 The optimistic agent problem in the inference approach solution

The results for the inference approach solution outlined in Eqn. (8) define the posterior transition dynamics that is necessary to achieve optimal control. Although this result can be useful in scenarios where the transition dynamics can be controlled, in many cases such control is not feasible. In such cases, the resulting policy derived from the inference approach is no longer optimal, since the agent optimistically expects that unfavorable transitions are unlikely [Levine, 2018] (see Fig. (1)).

An additional perspective on the optimistic agent problem comes from considering the backup equations for the optimal soft value functions (assuming a prior policy $\pi^0(a|s)$) [Haarnoja et al., 2018b]

$$Q(s, a) = r(s, a) + \sum_{s'} p(s'|s, a)V(s'), \tag{9}$$

$$V(s) = \sum_a \pi^*(a|s)\left[Q(s, a) - \frac{1}{\beta}\log\frac{\pi^*(a|s)}{\pi^0(a|s)}\right]. \tag{10}$$

Here the constraint is implicitly imposed in Eqn. (9), where the original dynamics is directly used. Note that the optimism problem does not arise when we consider the equations above. However, when we consider the inference-based approach we get the following backup equations [Levine, 2018]

$$Q(s, a) = r(s, a) + \frac{1}{\beta}\log\sum_{s'} p(s'|s, a)e^{\beta V(s')}, \tag{11}$$

$$V(s) = \frac{1}{\beta}\log\sum_a \pi^0(a|s)e^{\beta Q(s,a)}. \tag{12}$$

We note that the backup equation for $Q(s, a)$ in the inference approach (Eqn. (11)) is equivalent to Eqn. (9) only for the case of deterministic dynamics. For stochastic dynamics, the averaging over exponentiated future rewards [Levine, 2018, Levine and Koltun, 2013] in Eqn. (11) is the source of optimistic behavior by the agent.

In summary, two sources of the optimistic agent problem for stochastic dynamics in the inference approach to entropy-regularized RL are: 1) averaging over exponentiated rewards in the value function computation and 2) posterior transition

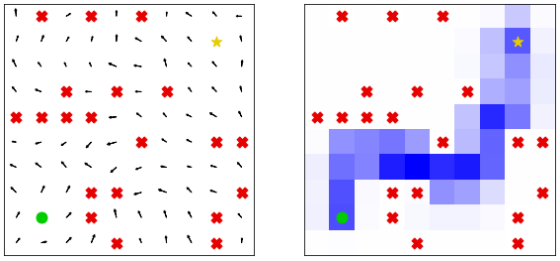

Figure 2: Left: an example maze with traps placed randomly, and wind field blowing in random directions and intensity. Wind dynamics is similar to that of Fig. 1. Each time step has fixed penalization $r = -1$. Red crosses represent traps with $r = -2$, and the golden star is the goal with $r = 0$. The MDP is such that the agent transitions to the start state (green circle) after stepping into a trap or the goal. Right: Shows the state distribution induced by the optimal policy which is computed using the proposed method. This result has been validated by comparison to the ground truth solution computed with value/policy iteration.

dynamics being different from the original dynamics. To resolve this problem, we develop an approach that ensures that (i) the posterior transition dynamics is fixed to the original dynamics *and* (ii) the backup equations, even though they involve averaging over exponentiated future rewards, reduce to the entropy-regularized RL backup equations Eqns. (9) and (10). Note that there can be other sources of optimistic behavior in finite-horizon control-as-inference approaches (as discussed in [Watson et al., 2021]), however these issues do not apply for the current formulation and thus are not considered in this work.

Given that the unconstrained entropy-regularized RL problem for stochastic dynamics can be solved exactly using the inference approach, we ask if it is possible to similarly solve the constrained entropy-regularized RL problem. In the next section, we present an approach to solve the constrained entropy-regularized RL problem through a transformation of the unconstrained approach.

4 CONSTRAINED OPTIMIZATION VIA UNCONSTRAINED INFERENCE

The core idea underlying our approach for *constrained optimization via unconstrained inference* is outlined in the following. The results from previous section show that, for the case of stochastic dynamics, the unconstrained inference approach solution leads to posterior dynamics that differs from the original dynamics. This implies that, if we want to use the unconstrained inference approach to obtain the solution for constrained entropy-regularized RL, it has to be applied to a *different* problem. Our approach is to determine the parameters for this different problem such that the optimal policy for the unconstrained inference problem is identical to the optimal policy of the original constrained optimization problem.

4.1 MAPPING TO CONSTRAINED OPTIMIZATION

Let us begin by considering the general controlled trajectory distribution denoted by $q(\tau)$. In contrast with $p(\tau)$ in Eqn. (3), the system dynamics is not constrained to be the same as the original dynamics for $q(\tau)$. As noted in the preceding section, the corresponding *unconstrained* objective function $J\left[q(\tau)\right]$ is minimized by the inference approach solution.

In the following, we will show how, for specific parameter choices for the dynamics and reward function, the unconstrained objective function $J\left[q(\tau)\right]$ exactly reduces to objective function for constrained entropy-regularized RL $J([p(\tau)])$ (Eqn. (4)).

We begin by considering a modified unconstrained problem with a *biased* transition dynamics and *biased* reward function, $p_b(s'|s,a)$ and $r_b(s,a)$, respectively, which are given by

$$p_b(s'|s,a) = b(s'|s,a)\, p(s'|s,a) \tag{13}$$

$$r_b(s,a) = r(s,a) + \delta(s,a) \leq 0 \;\forall\, s,a, \tag{14}$$

with $b(s'|s,a) > 0$ s.t. $\sum_{s'} p_b(s'|s,a) = 1$. For a given choice of biasing functions, we can express the corresponding unconstrained objective function, using Eqn. (5), as:

$$F = \inf_{q(\tau)} \left[\langle E \rangle_q - \langle \delta \rangle_q + \frac{1}{\beta} \left\langle \log \frac{1}{b} \right\rangle_q + \frac{1}{\beta} \mathcal{H}(q(\tau)|p_0(\tau)) \right] \tag{15}$$

with $\delta(\tau) \doteq \sum_{t=0}^{T} \delta(s_t, a_t)$, $b(\tau) \doteq \prod_{t=0}^{T} b(s_{t+1}|s_t, a_t)$, $\left\langle \log \frac{1}{b} \right\rangle_q = \sum_\tau q(\tau) \log \frac{1}{b(\tau)}$, and $\langle \delta \rangle_q = \sum_\tau q(\tau)\delta(\tau)$.

Since we want the inference approach solution to be identical to the solution for constrained entropy-regularized RL, the first condition is that the optimal dynamics for the biased model should be the same as the original dynamics. Using Eqn. (8), the condition that the optimal dynamics for the biased model must be the same as the original dynamics imposes the constraint equation

$$\forall s, a: \; b(s'|s,a) \propto e^{-\beta V_b(s')} \tag{16}$$

with the proportionality constant determined by normalization of the distribution function for transition dynamics.

We can interpret the above equation as follows: for any given choice of biased reward function $r_b(s,a)$, this equation determines the biasing function for the dynamics $b(s'|s,a)$ which is such that the optimal dynamics for the biased problem is the same as the original dynamics. Thus for each choice of $r_b(s,a)$ for which the above equation has a solution, we have identified biased dynamics parameters which satisfy the constraint on the dynamics. Now we can ask the following question:

Within this set of biased dynamics and biased reward functions, can we identify the choice of reward function which gives rise to the same optimal policy as constrained entropy-regularized RL?

Remarkably, we can derive a simple constraint equation that answers this question. The basic insight is that we need a condition such that the objective function for the biased unconstrained problem becomes identical to the objective function for constrained entropy-regularized RL. Correspondingly, we focus on the case where the cost contributions due to b and δ cancel each other out in Eqn. (15). This can be achieved by choosing $\delta(s,a)$ such that

$$\begin{aligned}\beta\delta(s,a) &= -\sum_{s'} p(s'|s,a)\log b(s'|s,a) \\ &= D_{\mathrm{KL}}(p(\cdot|s,a)||p_b(\cdot|s,a))\end{aligned} \tag{17}$$

As before, let $p(\tau)$ denote the trajectory distributions subject to the constraint that the dynamics is fixed to the original dynamics of the problem, such that the variation among different trajectory distributions is entirely due to the policy π. After applying both constraints in Eqns. (16) and (17), Eqn. (15) gets simplified to

$$F_q = \inf_{p(\tau)}\left[\langle E\rangle_p + \frac{1}{\beta}\mathcal{H}(p(\tau)|p_0(\tau))\right], \tag{18}$$

which is the free energy objective to be minimized for the *constrained* problem (Eqn. (4)). In both cases, the optimization is to be carried out by varying the policy π, and the preceding derivation shows that for every policy π, the corresponding objective function (i.e. sum of energetic and entropic costs) is the same for constrained and the unconstrained optimization problems, for a specific choice of biasing functions. Correspondingly, the optimal policy distribution is identical for the two problems. Thus we have shown that, assuming the constraint Eqns. (16) and (17) can be solved, the constrained optimization problem is identical to an *unconstrained* optimization problem for biased dynamics and biased reward function, which can then be solved using the inference approach.

4.2 EQUIVALENCE OF BACKUP EQUATIONS

The previous section has derived conditions which, when satisfied, lead to the solution of the constrained entropy-regularized RL problem using the inference approach. It is instructive to consider the equivalence between the two optimization problems by considering the corresponding backup equations.

Using Eqn. (8), the inference approach backup equation (Eqn. (11)) can be recast as

$$\begin{aligned}Q(s,a) = \; & r(s,a) + \sum_{s'} p^*(s'|s,a)V(s') \\ & - \frac{1}{\beta}D_{\mathrm{KL}}(p^*(\cdot|s,a)||p(\cdot|s,a)),\end{aligned} \tag{19}$$

Comparing with Eqn. (9), we see that the two equations are equivalent only when the optimal dynamics $p^*(s'|s,a)$ is the same as the original dynamics $p(s'|s,a)$ and this is true only for the case of deterministic dynamics.

When we consider the biased version of the unconstrained problem, Eqn. (19) becomes

$$\begin{aligned}Q(s,a) = \; & r(s,a) + \sum_{s'} p^*(s'|s,a)V(s') \\ & + \delta(s,a) - \frac{1}{\beta}D_{\mathrm{KL}}(p^*(\cdot|s,a)||p_b(\cdot|s,a)).\end{aligned} \tag{20}$$

We now consider biasing functions that satisfy the constraint in Eqn. (17) which, when substituted in Eqn. (20), gives

$$\begin{aligned}Q(s,a) = r(s,a) & + \sum_{s'} p^*(s'|s,a)V(s') \\ & - \frac{1}{\beta}D_{\mathrm{KL}}(p^*(\cdot|s,a)||p(\cdot|s,a)) \\ & + \frac{1}{\beta}\sum_{s'}\left[p^*(s'|s,a) - p(s'|s,a)\right]\log b(s'|s,a).\end{aligned} \tag{21}$$

Finally, we note that the condition in Eqn. (16) imposes the constraint $p^* = p$ (i.e. the biased optimal dynamics is the same as the original dynamics), using which Eqn. (21) turns into Eqn. (9). This shows that by solving the biased unconstrained optimization problem in this framework, with the bias parameters chosen to satisfy the constraint equations, we effectively solve the original, constrained version of entropy-regularized RL.

In summary, the inference approach backup equations for *biased* dynamics reduce to the backup equations for constrained entropy-regularized RL for the optimal policy, thereby showing that both approaches lead to the same soft-value functions $Q(s,a)$ and $V(s)$.

4.3 OPTIMIZATION FOR ARBITRARY TARGET DYNAMICS

The approach developed in previous sections can be generalized to the case where the transition dynamics is constrained to some arbitrary target distribution $\hat{p}(s'|s,a)$ (not necessarily the original dynamics). Specifically, the constrained optimization problem now corresponds to backup equations as given in Eqns. (9) and (10), but with original dynamics

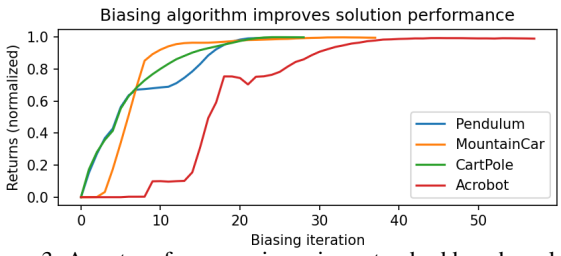

Figure 3: Agent performance in various standard benchmark dynamics as a function of iterations in the biasing process by the proposed method. At iteration 0, no biases are applied and the obtained solution corresponds to the optimistic agent (the existing optimal solution by inference). Transition dynamics are obtained from discretized state observations. Then, Eqn. (24) is used to find the optimal solutions. See Algorithm (1) and Table (2) in Appendix A for more details.

$p(s'|s,a)$ replaced by $\hat{p}(s'|s,a)$. This situation can be relevant when the agent's original dynamics either changes due to some failures or can be changed by the agent to specific target dynamics, corresponding to which we would like to determine the optimal policy.

It is readily seen that this scenario, which corresponds to a distribution shift in the transition dynamics, can be addressed by modifying the constraint equations (Eqns. (16) and (17)) as follows

$$b(s'|s,a) \propto \frac{\hat{p}(s'|s,a)}{p(s'|s,a)} e^{-\beta V_b(s')} \tag{22}$$

$$\beta\delta(s,a) = D_{\mathrm{KL}}(\hat{p}(\cdot|s,a)||p_b(\cdot|s,a)). \tag{23}$$

4.4 ALGORITHMS AND EXPERIMENTAL VALIDATION

In order to determine the optimal policy in constrained entropy-regularized RL using the inference approach, we need to determine the corresponding biased dynamics and rewards. We have developed a procedure to determine the biasing functions $b(s'|s,a)$ and $\delta(s,a)$ through an iterative approach which receives $\pi^0(a|s)$, $p(s'|s,a)$, $\hat{p}(s'|s,a)$, and $r(s,a)$, and calculates $b(s',a,s)$ and $\delta(s,a)$ by iteratively solving the constraint equations. Details are provided in Algorithm (1). The basic idea of the algorithm is to iteratively solve the unconstrained MDP problem, while updating the biasing functions for dynamics and rewards through Eqn. (24). The algorithm implements a fixed-point iteration method on the biasing functions b and $\delta(b)$ (see Eqs. (22) and (23)), such that

$$p_b^{(n+1)}(s'|s,a) = \frac{1}{C}\hat{p}(s'|s,a)e^{-\beta V_b^{(n)}(s')} \tag{24}$$

where $V_b^{(n)}$ is computed for the biased problem with $p_b^{(n)}$ and $r_b^{(n)}$; and C is a normalization constant.

Algorithm 1 Find Biases for dynamics and rewards

Parameters: inverse temperature β, update rate α
Input: $\pi^0(a|s)$, $p(s'|s,a)$, $\hat{p}(s'|s,a)$, $r(s,a)$
Output: $b(s'|s,a)$, $\delta(s,a)$, Δ
1. Initialize $b(s'|s,a) \leftarrow 1$ and $\delta(s,a) \leftarrow 0$
2. Initialize $p_b \leftarrow p$ and $r_b \leftarrow r$
repeat
  3. $V, p^* \leftarrow$ SolveMDP(β, r_b, p_b, π^0)
  4. $p_b(s'|s,a) \leftarrow \hat{p}(s'|s,a)e^{-\beta V(s')}$ See Eqn. (24)
  5. Normalize $p_b(s'|s,a)$
  6. $\delta(s,a) \leftarrow \beta^{-1} D_{\mathrm{KL}}(\hat{p}(\cdot|s,a)||p_b(\cdot|s,a))$ See Eqn. (23)
  7. $\Delta \leftarrow \max_{(s,a)}[r(s,a) + \delta(s,a)]$
  8. $r_b(s,a) \leftarrow r(s,a) + \delta(s,a) - \Delta$
until convergence $p^* \to \hat{p}$
function SOLVEMDP(β, $r(s,a)$, $p(s'|s,a)$, $\pi^0(a|s)$)
  a. $\widetilde{P}(s',a'|s,a) \leftarrow \pi^0(a'|s')p(s'|s,a)\exp(\beta r(s,a))$
  b. get dominant eigenvalue $e^{-\theta}$ and left eigenvector u
  c. compute $e^{\beta V(s)} \leftarrow \sum_a \pi^0(a|s)u(s,a)$
  d. $p^*(s'|s,a) \leftarrow$ from Eqn. (8)
  return: V, p^*
end function

Convergence is tested by computing the KL divergences between optimal and target dynamics. The convergence is considered attained when the following condition is true:

$$\max_{(s,a)} [D_{\mathrm{KL}}(p^*(\cdot|s,a)||\hat{p}(\cdot|s,a))] < 10^{-6}$$

We have tested this algorithm for various environments as summarized at Figs. (2) and (3). The experimental details are provided in the Appendix. To test the algorithm on the scenario of an arbitrary target dynamics, we set out a model-based proof-of-concept experiment where a prior transition dynamics is defined for which the optimal policy can be obtained. We then introduce a change in the dynamics representing a failure mode in the agent. In the example presented in Fig. (4), the agent can no longer walk directly towards the goal, but can still take advantage of the wind field to move in the desired direction. With this setting we were able to find the optimal policy for the altered dynamics by following the procedure outlined to determine the corresponding biases to the prior transition dynamics and reward function.

Finally, a model-free version that works in the tabular setting has been developed, wherein the biasing functions are learned through experience, along with the intermediate policies (see Appendix B and Algorithm (2)). The environment used is the same shown in Figure (1) (windy cliff environment). Our approach utilizes a single experience dataset collected from the original dynamics and the prior policy (uniform policy) throughout the whole process, making it an off-policy approach. Figure (5) shows the performance evaluation during the training process for several biasing iterations. As more iterations are completed, the optimistic

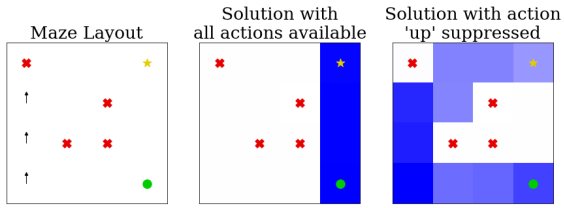

Figure 4: A windy environment used to test the feasibility of the method for forcing a target transition dynamics, different from the initial/prior transition dynamics. Left: The maze layout. Center: the solution to the original problem. Right: the new optimal solution to the modified problem where the action 'up' has been suppressed from the transition dynamics.

behavior is removed. The proposed approach successfully leads to the optimal policy for constrained optimization.

5 DISCUSSION

Control-as-inference is a powerful formalism for solving control problems using tools from Bayesian inference. Previously, the advantage of this formalism has been demonstrated by generalization of existing methods and derivation of new sophisticated algorithms. However, for the case of stochastic dynamics, this framework could not be directly applied to obtain the optimal solution for entropy-regularized RL. This work closes this gap in the field and provides a novel approach to the problem. Our solution can provide an alternative to standard approaches based on structured variational inference [Levine, 2018]. In general, such approaches provide variational bounds, whereas our results show that there is a mapping to a problem that has an exact solution.

The proposed solution not only adds to the formalism of control-as-inference by providing an analytical solution in the general case, but it also opens doors for new research directions and applications. For example, our method enables us to calculate the optimal policy and to chose optimal stochastic dynamics from a set of possible dynamics. A particular application of such optimal choice of dynamics can be, for example, hierarchical control where the upper level (manager) signals to the lower level (worker) to change dynamics (e.g., to update system dynamics to different frictions coefficients and/or different control gains). Another natural application can be self-recovering robots from failures, for which the ability to find a policy that works well under distribution shift of the dynamics would be useful. The results derived provide a novel approach for addressing such issues.

The scope of this work is to develop a novel probabilistic inference-based solution to entropy-regularized RL with stochastic dynamics, which we demonstrate in various model-based and model-free environments. We defer to future work the extension towards high-dimensional continuous spaces via function approximators. Another avenue for future work is a study of the theoretical properties of the iterative coupled equations for determining the biasing functions b and δ. We do not yet have a theoretical analysis for their convergence, but we do provide empirical evidence for various stochastic dynamics models.

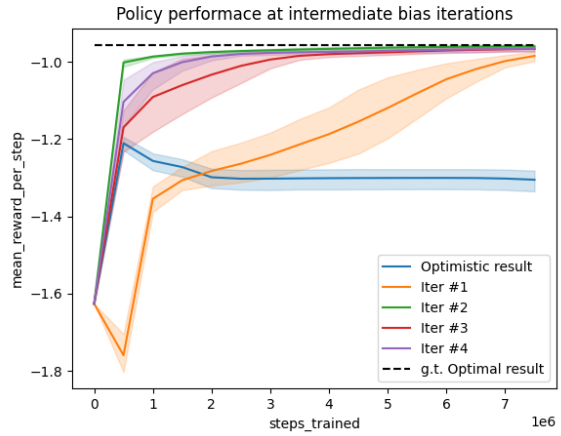

Figure 5: Progression of the learning process in a model-free setting. Biases are learnt from experience along with the policies. The environment used is the same as in Figure (1). An initial policy is learnt from experience, which results in an optimistic agent. Then biases are successively learnt and a new policies obtained. As expected, the optimistic policy has sub-optimal performance. After learning the biases, the approach recovers the optimal policy.

Finally, we note that the approach developed in this work can be applied more generally (i.e beyond entropy-regularized RL) as outlined in the following. Consider a setting wherein the solution to an unconstrained optimization problem is readily accessible (e.g. via Bayesian inference), however the problem of interest requires constrained optimization. Our approach considers a broader class of optimization problems which, for a specific parameter choice, reduce to the original system of interest. We then ask the question: Can we determine a (different) set of parameters such that a) the optimal solution to the unconstrained optimization problem satisfies the constraints of the original optimization problem, and b) there is a one-to-one mapping between objective functions for the two optimization problems? It will be of interest to see if the approach presented here for solving constrained optimization problems by mapping them to unconstrained problems that can be analyzed via inference can also be applied to other settings involving a more general class of objective functions [Hazan et al., 2019, Zhang et al., 2020].

SOFTWARE AND DATA

We make source code immediately available at `https://github.com`, which can be used to reproduce all the results obtained.

References

Pieter Abbeel and Andrew Ng. Learning first-order markov models for control. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005. URL https://proceedings.neurips.cc/paper/2004/file/02f657d55eaf1c4840ce8d66fcdaf90c-Paper.pdf.

Argenis Arriojas, Stas Tiomkin, and Rahul V Kulkarni. Closed-form analytical results for maximum entropy reinforcement learning. *arXiv preprint arXiv:2106.03931*, 2021.

Kavosh Asadi, Evan Cater, Dipendra Misra, and Michael L. Littman. Towards a simple approach to multi-step model-based reinforcement learning, 2018.

C.G. Atkeson and J.C. Santamaria. A comparison of direct and model-based reinforcement learning. In *Proceedings of International Conference on Robotics and Automation*, volume 4, pages 3557–3564 vol.4, 1997. doi: 10.1109/ROBOT.1997.606886.

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 908–919, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

G. Boone. Efficient reinforcement learning: model-based acrobot control. In *Proceedings of International Conference on Robotics and Automation*, volume 1, pages 229–234 vol.1, 1997. doi: 10.1109/ROBOT.1997.620043.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Zhong Cao, Shaobing Xu, Huei Peng, Diange Yang, and Robert Zidek. Confidence-aware reinforcement learning for self-driving cars. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12, 2021. doi: 10.1109/TITS.2021.3069497.

Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38, 2021.

Dane Corneil, Wulfram Gerstner, and Johanni Brea. Efficient model-based deep reinforcement learning with variational state tabulation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1049–1058. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/corneil18a.html.

Avishek Das, Dominic C Rose, Juan P Garrahan, and David T Limmer. Reinforcement learning of rare diffusive dynamics. *arXiv preprint arXiv:2105.04321*, 2021.

Benjamin Eysenbach and Sergey Levine. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. *arXiv*, Mar 2021. URL https://arxiv.org/abs/2103.06257v1.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of physiology-Paris*, 100 (1-3):70–87, 2006.

Alborz Geramifard, Christoph Dann, Robert H. Klein, William Dabney, and Jonathan P. How. Rlpy: A value-function-based reinforcement learning framework for education and research. *Journal of Machine Learning Research*, 16(46):1573–1578, 2015. URL http://jmlr.org/papers/v16/geramifard15a.html.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement Learning with Deep Energy-Based Policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, Jul 2017. URL http://proceedings.mlr.press/v70/haarnoja17a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018b.

Hado Hasselt. Double q-learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates,

Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/091d584fced301b442654dd8c23b3fc9-Paper.pdf.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2681–2691. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hazan19a.html.

Hilbert J. Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Mach. Learn.*, 87(2):159–182, May 2012. ISSN 1573-0565. doi: 10.1007/s10994-012-5278-7.

Bahare Kiumarsi, Kyriakos G. Vamvoudakis, Hamidreza Modares, and Frank L. Lewis. Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6):2042–2062, 2018. doi: 10.1109/TNNLS.2017.2773458.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. *arXiv*, May 2018. URL https://arxiv.org/abs/1805.00909v3.

Sergey Levine and Vladlen Koltun. Guided policy search. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1–9, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/levine13.html.

Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, and Igor Mordatch. Plan online, learn offline: Efficient learning and exploration via model-based control. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Byey7n05FQ.

Sanjoy K Mitter and NJ Newton. The duality between estimation and control. *Published in Festschrift for A. Bennoussan*, 2000.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. ISSN 1476-4687. doi: 10.1038/nature14236.

Andrew W Moore. *Efficient Memory-Based Learning for Robot Control*. PhD thesis, University of Cambridge, 1990.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *arXiv preprint arXiv:1702.08892*, 2017.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

Dominic C Rose, Jamie F Mair, and Juan P Garrahan. A reinforcement learning approach to rare trajectory sampling. *New Journal of Physics*, 23(1):013013, 2021.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, Dec 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-03051-4.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362 (6419):1140–1144, Dec 2018. ISSN 0036-8075. doi: 10.1126/science.aar6404.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In D. Touretzky, M. C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996. URL https://proceedings.neurips.cc/paper/1995/file/8f1d43620bc6bb580df6e80b0dc05c48-Paper.pdf.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018. ISBN 978-0-26203924-6. URL https://mitpress.mit.edu/books/reinforcement-learning-second-edition.

Evangelos A Theodorou and Emanuel Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *2012 ieee 51st ieee conference on decision and control (cdc)*, pages 1466–1473. IEEE, 2012.

Stas Tiomkin and Naftali Tishby. A unified bellman equation for causal information and value in markov decision processes. *arXiv preprint arXiv:1703.01585*, 2017.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

Emanuel Todorov. General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292. IEEE, 2008.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pages 1049–1056, 2009.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Mach. Learn.*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698.

Joe Watson, Hany Abdulsamad, Rolf Findeisen, and Jan Peters. Stochastic control through approximate bayesian input inference. *CoRR*, abs/2105.07693, 2021. URL https://arxiv.org/abs/2105.07693.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey. *arXiv preprint arXiv:1908.08796*, 2019.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/30ee748d38e21392de740e2f9dc686b6-Paper.pdf.

Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real world robotic reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJe2syrtvS.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, pages 1255–1262, 2010. URL https://icml.cc/Conferences/2010/papers/28.pdf.