In: Proc. 2012 Robotics: Science and Systems Conference (R:SS 2012), Sydney, Australia Jul 9-13 (2012).

On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference

Konrad Rawlik*, Marc Toussaint† and Sethu Vijayakumar*
* School of Informatics, University of Edinburgh, UK
† Department of Computer Science, FU Berlin, Germany

***Abstract*—We present a reformulation of the stochastic optimal control problem in terms of KL divergence minimisation, not only providing a unifying perspective of previous approaches in this area, but also demonstrating that the formalism leads to novel practical approaches to the control problem. Specifically, a natural relaxation of the dual formulation gives rise to exact iterative solutions to the finite and infinite horizon stochastic optimal control problem, while direct application of Bayesian inference methods yields instances of risk sensitive control. We furthermore study corresponding formulations in the reinforcement learning setting and present model free algorithms for problems with both discrete and continuous state and action spaces. Evaluation of the proposed methods on the standard Gridworld and Cart-Pole benchmarks verifies the theoretical insights and shows that the proposed methods improve upon current approaches.**

I. Introduction

In recent years the framework of *stochastic optimal control* (SOC) [20] has found increasing application in the domain of planning and control of realistic robotic systems, e.g., [6, 14, 7, 2, 15] while also finding widespread use as one of the most successful normative models of human motion control [23]. In general, SOC can be summarised as the problem of controlling a stochastic system so as to minimise expected cost. A specific instance of SOC is the *reinforcement learning* (RL) formalism [21] which does not assume knowledge of the dynamics or cost function, a situation that may often arise in practice. However, solving the RL problem remains challenging, in particular in continuous spaces [16].

A recent, promising direction in the field has been the application of inference methods [1] to these problems, e.g., [10, 22, 24]. In this context, we introduce a generic formulation of the SOC problem in terms of *Kullback-Leibler* (KL) divergence minimisation. Although the arising KL divergences can, in general, not be minimised in closed form, we provide a natural iterative procedure that results in algorithms that we prove to asymptotically converge to the exact solution of the SOC problem. Specifically, algorithms for both finite and infinite horizon problems are derived and their corresponding formulations in the RL setting are introduced. We show that the latter corresponds to the independently derived result of [5] for the specific case of infinite horizon discrete problems; here, we extend this to problems with continuous actions.

Formulation of SOC problems in terms of KL minimisation has been previously studied by, amongst others, [22], [11] and [10], leading to efficient methods for both stochastic optimal control [22] and RL [7]. However, as we will discuss, these cases make restrictive assumptions about the problem dynamics and costs which can be relaxed under our framework, besides providing a unifying and generic formalism.

Finally, we are able clarify the relation of SOC and the inference control formulation by [24, 17, 26], which allows for arbitrary problems, showing it to be an instance of risk sensitive control. The generalisation of this relation given by our approach makes it possible to apply out of the box inference methods to obtain approximate optimal policies. This is of particular interest in the case of continuous problems – here approximations are unavoidable since explicit representations are often not available.

II. Preliminaries

A. Stochastic Optimal Control

We will consider control problems which can be modeled by a *Markov decision process* (MDP). Using the standard formalism, see also e.g., [21], let $x_t \in \mathbb{X}$ be the state and $u_t \in \mathbb{U}$ the control signals at times $t = 1, 2, \ldots, T$. To simplify the notation, we shall denote complete state and control trajectories $x_{1\ldots T}$,$u_{0\ldots T}$ by $\bar{x}$,$\bar{u}$. Let $P(x_{t+1}|x_t, u_t)$ be the transition probability for moving from x_t to x_{t+1} under control u_t and let $\mathcal{C}_t(x, u) \geq 0$ be the cost incurred per stage for choosing control u in state x at time t. Let policy $\pi(u_t|x_t)$ denote the conditional probability of choosing the control u_t given the state x_t. In particular a deterministic policy is given by a conditional delta distribution, i.e. $\pi(u_t|x_t) = \delta_{u_t=\tau(x_t)}$ for some function τ. The SOC problem consists of finding a policy which minimises the expected cost, i.e., solving

$$\pi^* = \operatorname*{argmin}_{\pi} \left\langle \sum_{t=0}^{T} \mathcal{C}_t(x_t, u_t) \right\rangle_{q_\pi} , \tag{1}$$

where $\langle\cdot\rangle_{q_\pi}$ denotes the expectation with respect to

$$q_\pi(\bar{x}, \bar{u}|x_0) = \pi(u_0|x_0) \prod_{t=1}^{T} \pi(u_t|x_t) P(x_{t+1}|x_t, u_t) , \tag{2}$$

the distribution over trajectories under policy π.

In the case of infinite horizon problems, i.e. we let $T \to \infty$, we will consider the discounted cost setting and specifically assume that $\mathcal{C}_t(x_t, u_t) = \gamma^t \mathcal{C}_\bullet(x_t, u_t)$, where $\mathcal{C}_\bullet$ is a time stationary cost and $\gamma \in [0, 1]$ a discount factor.

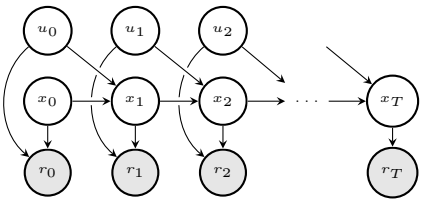


Fig. 1: The graphical model of for the Bayesian formulation of the control problem in the finite horizon case. In the infinite horizon case we obtain a stochastic Markov process.

B. Inference Control Model

A Bayesian inference based approximation of the above control problem can be formulated [24] as illustrated in Fig.1. In addition to the state and control variables of classical SOC, a binary dynamic random task variable r_t is introduced and the task likelihood is related to the classical cost by choosing $P(r_t = 1|x_t, u_t) = \exp\{-\eta \mathcal{C}(x_t, u_t)\}$, where $\eta > 0$ is some constant in analogy with the inverse temperature of a Boltzmann distribution. For some given policy π and assuming the artificial observations $r_{0...T} = 1$, we denote the unnormalised posterior by $p_\pi(\bar{x}, \bar{u})$:

$$\begin{aligned} p_\pi(\bar{x}, \bar{u}) &= P(\bar{x}, \bar{u}, \bar{r} = 1|x_0) \\ &= q_\pi(\bar{x}, \bar{u}) \prod_{t=0}^{T} \exp\{-\eta \mathcal{C}_t(x_t, u_t)\} \ . \end{aligned} \tag{3}$$

C. General Duality

While the Bayesian model has been employed successfully for trajectory planning, e.g., in [24], it's relation to the classical SOC problem remained unclear. Although a specific subset of SOC problems, studied by [11] and [22], can be formulated in a similar Bayesian model, as explicitly done by [10] (we discuss the relation to this work further in III-D3), here, we establish the formal correspondence between the two formalisms in the general case with the following result:

Theorem 1. *Let π^0 be an arbitrary stochastic policy and $\mathbb{D}$ the set of deterministic policies, then the problem*

$$\pi^* = \underset{\pi \in \mathbb{D}}{\operatorname{argmin}} \, \mathrm{KL}\left(q_\pi(\bar{x}, \bar{u}) \| p_{\pi^0}(\bar{x}, \bar{u})\right) \tag{4}$$

is equivalent to the stochastic optimal control problem (1) with cost per stage

$$\hat{\mathcal{C}}_t(x_t, u_t) = \mathcal{C}_t(x_t, u_t) - \frac{1}{\eta} \log \pi^0(u_t|x_t) \ . \tag{5}$$

Proof: see Supplementary Material[1]. ■

As an immediate consequence we may recover any given stochastic optimal control problem with cost $\mathcal{C}_t$ by choosing $\pi^0(\cdot|x)$ to be the uniform distribution over $\mathbb{U}$[2].

[1]Supplementary Material can be found at http://arxiv.org/abs/1009.3958

[2]n.b., formally we require $\mathbb{U}$ to be finite or bounded

III. Iterative Solutions

Although Theorem 1 provides the correspondence between the SOC formulation and the computationally attractive inference control approach, due to the constraint $\pi \in \mathbb{D}$, (4) remains as intractable as the classical formulation via the Bellmann equation. However relaxation of this constraint to allow minimisation over arbitrary stochastic policies provides a closed form solution, and although it does not *directly* lead to an optimal policy, we have the following result:

Theorem 2. *For any $\pi \neq \pi^0$, $\mathrm{KL}\left(q_\pi \| p_{\pi^0}\right) \leq \mathrm{KL}\left(q_{\pi^0} \| p_{\pi^0}\right)$ implies $\langle \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_\pi} < \langle \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_{\pi^0}}$.*

Proof: see Supplementary Material ■

Consequently, with some initial π^0, the iteration

$$\pi^{n+1} \leftarrow \underset{\pi}{\operatorname{argmin}} \, \mathrm{KL}\left(q_\pi \| p_{\pi^n}\right) \ , \tag{6}$$

where π is an arbitrary[3] conditional distribution over u, gives rise to a chain of stochastic policies with ever decreasing expected costs.

We would like to note that the conditions imposed by the above result, in order to guarantee a policy improvement, are relatively weak. By exploiting this, in addition to the iteration arising from (6), we present a relaxation, which satisfy Theorem 2 and leads to practical algorithms for infinite horizon problems, and the related iteration of Bayesian inference which leads to risk-sensitive control.

A. Exact Minimisation - Finite Horizon Problems

The general minimisation in iteration (6) can, as previously indicated, be performed in closed form and the new policy (for derivation, see Supplementary Material), is given by the Boltzmann like distribution,

$$\pi^{n+1}(u_t|x_t) = \exp\{\Psi_t^{n+1}(x_t, u_t) - \bar{\Psi}_t^{n+1}(x_t)\} \ , \tag{7}$$

with energy

$$\begin{aligned} \Psi_t^{n+1}(x_t, u_t) &= \log \pi^n(u_t|x_t) + \log P(r_t = 1|x_t, u_t) \\ &\quad + \int_{x_{t+1}} P(x_{t+1}|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x_{t+1}) \end{aligned} \tag{8}$$

and log partition function

$$\bar{\Psi}_t^{n+1}(x_t) = \log \int_u \exp\{\Psi_t^{n+1}(x_t, u)\} \ . \tag{9}$$

In the finite horizon case, the policy can therefore be computed backwards in time.

1) Convergence Analysis: Following [12], we bound the progress of the trajectory posterior under policy π^n towards the corresponding distribution under some chosen $\hat{\pi}$, obtaining

Lemma 3. *Let the sequence $\{\pi^n\}$ be generated by (6) and let $\hat{\pi}$ be an arbitrary (stochastic) policy. Then*

$$\begin{aligned} &\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^{n+1}}\right) - \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^n}\right) \\ &\qquad \leq \langle \eta \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_{\hat{\pi}}} - \langle \eta \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_{\pi^{n+1}}} \ . \end{aligned} \tag{10}$$

[3]n.b., formally certain assumptions have to be made to ensure the support of q_π is a subset of the support of p_{π^n}

Proof: See Supplementary Material. ■

Summing the above bound over $0 \dots N$, we can compute the bound

$$\frac{1}{N}\sum_{n=1}^{N+1} \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^n}} \leq \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\hat{\pi}}} + \frac{1}{\eta N}\,\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^0}\right), \quad (11)$$

on the average expected cost of the policies $\pi^1 \dots \pi^{n+1}$. Now, since Theorem 2 guarantees that the expected cost for each π^n is non increasing with n, using (11), we can obtain the following stronger convergence result.

Theorem 4. *Let $\{\pi^n\}$ be a sequence of policies generated by (6), with π^0 s.t. $\pi^0(\cdot|x \in \mathbb{X})$ has support $\mathbb{U}$. Then*

$$\lim_{n\to\infty} \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^n}} = \min_{\pi} \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi}} \ . \quad (12)$$

Proof: See Supplementary Material. ■

B. Asynchronous Updates - Infinite Horizon Problems

In the infinite horizon setting, discussed in II-A, it is easy to show that the time stationary analog of (8) can be obtained as

$$\Psi^{n+1}(x,u) = \log \pi^n(u|x) + \log P(r=1|x,u) + \gamma \int_y P(y|x,u)\bar{\Psi}^{n+1}(y) \ . \quad (13)$$

However, due to the form of $\bar{\Psi}^{n+1}$, this does not yield Ψ^{n+1} in closed form. To obtain a practical solution we make use of the relatively weak conditions given by Theorem 2 for obtaining a lower expected cost, which allow us to consider the minimisation in (6) over some iteration dependent subset $\mathbb{P}^n$ of the set of all (stochastic) policies. Then, Theorem 2 guarantees the expected costs to be non increasing, if for all n, $\pi^n \in \mathbb{P}^n$.

Such iterations admit *asynchronous* updates as an interesting case, i.e., updating one or several time steps of the policy at each iteration in any particular order. Formally, we choose a schedule of time step sets $\mathbb{T}^0, \mathbb{T}^1, \dots$ and let $\mathbb{P}^n = \{\pi : \forall t \notin \mathbb{T}^n, \pi_t = \pi_t^n\}$. Specifically, we will consider the schedule for such updates given by $\mathbb{T}^n = \{0, \dots, n-1\}$, i.e., in each iteration we consider finite horizon problems with increasing horizon. Such a schedule leads to the update $\pi_{t+1}^{n+1} = \pi_t^n$ for all $t > 0$ while the new first step policy, π_0^{n+1}, is of the form (7) and obtained via

$$\Psi_0^{n+1}(x,u) \leftarrow \Psi_0^n(x,u) - \bar{\Psi}_0^n(x) + \log P(r=1|x,u) + \gamma \int_{x'} P(x'|x,u)\bar{\Psi}_0^n(x') \ , \quad (14)$$

hence yielding a practical iteration which has a strong analogy to value iteration, see e.g., [21].

1) Convergence Analysis: Essentially equivalent convergence results to the finite horizon case can be obtained for the asynchronous algorithm (14) in the infinite horizon setting. Informally, we proceed by assuming that the cost is bounded and consider finite horizon problems with growing horizon, bounding the expected cost of the infinite tail. Due to the assumption that the cost is discounted, the expected cost of the tail goes to zero as the horizon increases, leading to a result analogous to Theorem 4 (see Supplementary Material for formal proof).

C. Posterior Policy Iteration

Since our starting point was the relaxation of the relation between SOC and inference control, it is interesting to consider sequential inference of the posterior policy, which is the natural iteration arising in the latter framework. Such an iteration is of particular interest as posterior inference is a well studied problem with a large range of approximate algorithms [1] which could be exploited for practical implementations.

Although unconstrained minimisation of the KL divergence is achieved by the posterior, in our case, the specific form of q_π in (6) is, as can be seen in (2), restricted by the prescribed system dynamics, leading to the results presented in the last sections. Nonetheless, we may consider the iteration

$$\pi^{n+1} = p_{\pi^n}(u_t|x_t) \ , \quad (15)$$

which, as we show (see Supplementary Material), will converge to the policy

$$\tilde{\pi} = \operatorname*{argmin}_{\pi} -\frac{1}{\eta}\log \langle \exp\{-\eta \mathcal{C}_t(\bar{x},\bar{u})\}\rangle_{q_\pi} \ . \quad (16)$$

The objective being minimized is exactly the risk sensitive objective of [8], which has been recently also used in the path integral approach to SOC [3]. In particular, note that for $\eta \to 0$, we obtain the classical risk neutral controls, allowing near optimal policies for arbitrary SOC problems to be computed by iterated Bayesian inference.

D. Relation to Previous Work

1) Dynamic Policy Programming (DPP): The recently introduced DPP algorithm [5] is closely related to the formalism described here. Specifically, while the update equations (14) coincide, we provide a more general view of DPP by deriving it as a special case of the novel result in Theorem 2. In addition, III-A provides the direct extension of DPP to finite horizon problems, while the convergence proofs of III-B extend those given by [5] to continuous state and action spaces.

2) Approximate Inference Control (AICO): The AICO [24] approach to trajectory optimisation shares the same Bayesian Model used as a starting point here (cf. II-B). However, although using local LQG approximations AICO converges to locally optimal trajectories, the relation to the classical SOC problem remained unclear. We not only establish such a formal relation, but also note that AICO can be interpreted as one step of the posterior policy iteration introduced in III-C. More specifically, if one were to use the maximum likelihood policy obtained by AICO one would obtain (approximate) optimal risk seeking controls.

3) Path Integral Control: Let us briefly recall the KL control framework [10], the alternative formulations in [22] being equivalent for our purposes. Choose some free dynamics $\nu_0(x_{t+1}|x_t)$ and let the cost be given as

$$\mathcal{C}(\bar{x}) = \ell(\bar{x}) + \log \frac{\nu(\bar{x})}{\nu_0(\bar{x})}$$

where $\nu(x_{t+1}|x_t)$ is the controlled process under some policy. Then

$$\langle \mathcal{C}(\bar{x}) \rangle_\nu = \mathrm{KL}\left(\nu(\bar{x}) \| \nu_0(\bar{x}) \exp\{-\ell(\bar{x})\}\right) \tag{17}$$

is minimised w.r.t. ν by

$$\nu(x_{1:T}|x_0) = \frac{1}{Z(x_0)} \exp\{-\ell(x_{1:T})\} \nu_0(x_{1:T}|x_0) \tag{18}$$

and one concludes that the optimal control is given by $\nu(x_{t+1}|x_t)$, where the implied meaning is that $\nu(x_{t+1}|x_t)$ is the trajectory distribution under the optimal policy.

Although (18) gives a process which minimises (17), it is not obvious how to compute the actual controls u_t. Specifically when given a model of the dynamics, i.e., $P(x_{t+1}|x_t, u_t)$, and having chosen some ν_0, a non trivial, yet implicitly made, assumption is that there exists a policy implementing the required transitions $\nu(x_{t+1}|x_t)$, i.e., $\exists \pi$ s.t.

$$\mathrm{KL}\left(\int_{u_t} P(x_{t+1}|x_t, u_t)\pi(u_t|x_t) \| \nu(x_{t+1}|x_t)\right) = 0. \tag{19}$$

However, in general, such a π will not exist. This is made very explicit for the discrete MDP case in [22], where it is acknowledged that the method is only applicable if the dynamics are *fully controllable*, i.e., $P(x_{t+1}|x_t, u_t)$ can be brought into any required form by the controls. Although in the same paper, it is suggested that solutions to classical problems can be obtained by continuous embedding of the discrete MDP, such an approach has several drawbacks. For one, it requires solving a continuous problem even for cases which could have been otherwise represented in tabular form, but more importantly such an approach is obviously not applicable to problems which already have continuous state or action spaces.

In the case of problems with continuous states and actions we may consider the specific form

$$\begin{aligned} x_{t+1} &= \mathcal{F}(x_t) + \mathbf{B}(u_t + \xi), \quad \xi \sim \mathcal{N}(0, \mathbf{Q}) \ , \\ \mathcal{C}_t(x_t, u_t) &= \ell(x_t) + u_t^T \mathbf{H} u_t \ , \end{aligned} \tag{20}$$

with $\mathcal{F}, \mathbf{B}$ and ℓ having arbitrary form, but $\mathbf{H}, \mathbf{Q}$ are such that $\mathbf{H} \propto \mathbf{B}^T \mathbf{Q}^{-1} \mathbf{B}$. This is of interest, as it is the discrete time form of the fully controllable continuous time problem which underlies the path integral approach [11]. It also has been claimed, e.g., [10], that, analogously to the continuous time case, the solution of this problem is given by (18). However considering the simple instance of a one step LQG problem, we see that (19) will not hold, as in this case the variance of $P(x_1|x_0, u_0)$ is uncontrolled. Hence ν is *not* the trajectory distribution under the optimal policy. Furthermore it is straightforward to convince oneself that attempting to find the policy implementing the best realisable transition, i.e., relaxation of (19) to

$$\operatorname*{argmin}_{\pi \in \mathbb{D}} \mathrm{KL}\left(\int_{u_t} P(x_{t+1}|x_t, u_t)\pi(u_t|x_t) \| \nu(x_{t+1}|x_t)\right) \ ,$$

does also not lead to the desired result.

However, for problems of the specific form (20), a closer relation between Theorem 1 and (17) does indeed exist. To illustrate this, we write the KL divergence of Theorem 1 in terms of the state trajectory ($\bar{x}$) marginals as

$$\begin{aligned} \mathrm{KL}\left(q_\pi(\bar{x}, \bar{u}) \| p_{\pi^0}(\bar{x}, \bar{u})\right) &= \mathrm{KL}\left(q_\pi(\bar{x}) \| \nu(\bar{x})\right) \\ &\quad - \left\langle \sum m_t^T \mathbf{Q}^{-1} \mathbf{B} u_t - \frac{1}{2} u_t^T \mathbf{H} u_t) \right\rangle_{q_\pi(\bar{x}, \bar{u})} \ , \end{aligned}$$

where $m_t = x_{t+1} - x_t - \mathcal{F}(x_t)$. Furthermore, since for a deterministic policy, i.e. $\pi(u_t|x_t) = \delta_{u_t = \tau(x_t)}$,

$$\langle m_t \rangle_{q_\pi} = \langle \mathbf{B} u_t \rangle_{q_\pi} = \mathbf{B} \tau(x_t) \ ,$$

the second term is zero under the condition required, i.e., $\mathbf{H} = 2\mathbf{B}^T \mathbf{Q}^{-1} \mathbf{B}$, and analogous to (17), it is sufficient to consider the distributions over state trajectories only.

In conclusion, for discrete time problems, the work of [10, 22] constitutes special cases of Theorem 1, which either assume fully controllable dynamics or where the control trajectories can be marginalised from Theorem 1.

4) Expectation Maximisation: Several suggestions for mapping the SOC problem onto a maximum likelihood problem and using *Expectation Maximization* (EM) have been recently made in the literature, e.g., [25], and going further back, the probability matching approach [4, 19] has also close links with EM. Considering (6), the proposed approach has a close relation to the free energy view of EM. Given a free energy

$$\begin{aligned} F(\tilde{q}, \pi) &= \log P(\bar{r} = 1; \pi) - \mathrm{KL}\left(\tilde{q} \| p_\pi\right) &(21) \\ &= \langle \log P(\bar{r} = 1, \bar{x}, \bar{u}; \pi) \rangle_{\tilde{q}} + H(\tilde{q}) \ , &(22) \end{aligned}$$

EM alternates between minimizing $\mathrm{KL}\left(\tilde{q} \| p_\pi\right)$ w.r.t. $\tilde{q}$ in (21) and maximising the free energy w.r.t. the potentially infinite parameter vector π in (22). Our iteration of (6) deviates from this standard EM in that the KL-minimization in (6) is w.r.t. a *constrained* $\tilde{q}$, namely one which can be generated by a control π. The M-step is then trivially assigning the new π to the one corresponding to $\tilde{q}$. The constraint E-step departs from standard EM but is a special case of the alternating minimisation procedures of [9]. Importantly however, unlike the previously mentioned EM based approaches which can only guarantee convergence to a local extremum, we have demonstrated algorithms with guaranteed convergence to the global optimum.

IV. Reinforcement Learning

We now turn to the RL setting [21], where one aims to learn a good policy given only samples from the transition probability and associated incurred costs. As RL usually considers the discounted cost infinite horizon setting we concentrate on this case, with the understanding that equivalent steps can

be taken in the finite horizon case. We note that for any given x, u the update of (14) can be written as an expectation w.r.t. the transition probability $P(y|x,u)$, and hence, may be approximated from a set of sampled transitions. In particular given a single sample $(x, u, \mathcal{R}, y)$ of a transition from x to y under control u, obtaining reward $\mathcal{R} = \log P(r = 1|x, u)$, we may perform the approximate update

$$\Psi(x,u) \leftarrow \Psi(x,u) + \alpha \left[\mathcal{R} + \gamma\bar{\Psi}(y) - \bar{\Psi}(x)\right] \ , \qquad (23)$$

with α a learning rate and for trajectory data applying such an update individually for each tuple $(x_t, u_t, \mathcal{R}_t, x_{t+1})$.

A. Relation to Classical Algorithms

Before proceeding let us highlight certain similarities and differences between (23) and two classical algorithms, $\mathcal{Q}$-learning and TD(0) [21].

The $\mathcal{Q}$-learning algorithm learns the state-action value function. We note that Ψ has certain similarities to a $\mathcal{Q}$ function, in the sense that a higher value of Ψ for a certain control in a given state indicates that the control is *'better'* – in fact, for the optimal controls the $\mathcal{Q}$ function and Ψ converge to the same absolute value (see Supplementary Material). However, unlike the $\mathcal{Q}$ function, which also converges to the expected cost for the sub-optimal controls, Ψ goes to $-\infty$ for sub-optimal actions. A potentially more insightful difference between the two algorithm is the nature of updates employed. The $\mathcal{Q}$-learning algorithm uses updates of the form

$$\mathcal{Q}(x,u) \leftarrow \mathcal{Q}(x,u) + \alpha \left[\mathcal{R} + \gamma \max_{u'} \mathcal{Q}(y,u') - \mathcal{Q}(x,u)\right] \ ,$$

where α is a learning rate. Note that it employs information from the current command and the single best future command under current knowledge. The proposed algorithm on the other hand uses a soft-max operation by employing $\bar{\Psi}$, averaging over information about the future according to the current belief about the control distribution, hence taking uncertainty arising from, e.g., sampling into account.

On the other hand, the TD(0) algorithm, which learns through value function approximation, has updates of the form

$$\mathcal{V}(x) = \mathcal{V}(x) + \alpha \left[\mathcal{R} + \gamma\mathcal{V}(y) - \mathcal{V}(x)\right] \ ,$$

with α again a learning rate. Since it can be shown that $\bar{\Psi}$ converges to the value function of the optimal policy (cf. Supplementary Material), the proposed update converges towards the TD(0) update for samples generated under the optimal policy. In particular, while TD(0) is an on-policy method and learns the value function of the policy used to generate samples, the proposed method learns the value function of the optimal policy directly.

B. RL with continuous states and actions

One needs to use parametric representations [21] to store Ψ when tabular means are no longer viable or efficient, as is the case with high dimensional, large discrete [5] or continuous state and control spaces. Similar to numerous previous approaches, e.g., [5], we used a linear basis function model to approximate Ψ, i.e.,

$$\Psi(x,u) \approx \tilde{\Psi}(x,u,\mathbf{w}) = \sum_{m=0}^{M} w_m \phi(x,u) \qquad (24)$$

where $\phi_i : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$ are a set of given basis functions and $\mathbf{w} = (w_1, \ldots, w_M)$ is the vector of parameters that are optimised. For such an approximation, and given a set of samples $(x_{1\ldots K}, u_{1\ldots K}, \mathcal{R}_{1\ldots K}, y_{1\ldots K})$, the updates (8) and (23) can be written in matrix notation as

$$\Phi\mathbf{w}^{n+1} = \Phi\mathbf{w}^{n} + \mathbf{z} \ , \qquad (25)$$

where Φ is the $K \times M$ matrix with entries $\Phi_{i,j} = \phi_i(x_j, u_j)$ and $\mathbf{z}$ is the vector with elements

$$\mathbf{z}_k = \gamma\bar{\Psi}(y_k) + \mathcal{R}_k - \bar{\Psi}(x_k) \ . \qquad (26)$$

This suggests the update rule of the form

$$\mathbf{w} \leftarrow \mathbf{w} + (\Phi^T\Phi)^{-1}\Phi^T\mathbf{z} \ . \qquad (27)$$

The choice of basis functions is somewhat complicated by the need to evaluate the log partition function of the policy $\bar{\Psi}$, i.e. $\log \int_u \exp\{\tilde{\Psi}(x,u)\}$, when forming the vector $\mathbf{z}$. In cases where $\mathbb{U}$ is a finite set, arbitrary basis functions can be chosen as the integral reduces to a finite sum. However, for problems with infinite (or continuous) control spaces, bases need to be chosen such that the resulting integral is analytically tractable, i.e. the partition function of the stochastic policy can be evaluated. One class of such basis sets is given by those $\tilde{\Psi}(x,u,\mathbf{w})$ that can be brought into the form

$$\tilde{\Psi}(x,u,\mathbf{w}) = -\frac{1}{2}u^T\mathbf{A}(x,\mathbf{w})u + u^T\mathbf{a}(x,\mathbf{w}) + a(x,\mathbf{w}) \ , \qquad (28)$$

where $\mathbf{A}(x,\mathbf{w})$ is a positive definite matrix-valued function, $\mathbf{a}(x,\mathbf{w})$ is a vector-valued function and $a(x,\mathbf{w})$ a scalar function. For such a set, the integral is of the Gaussian form and the closed form solution

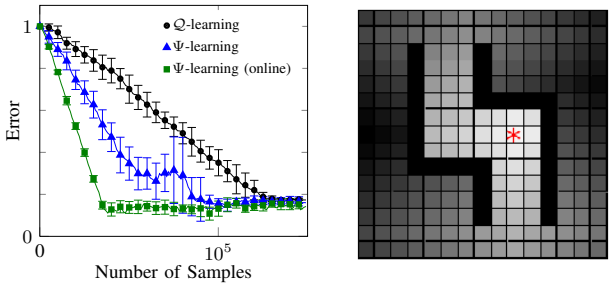
$$\log \int_u \exp\{\tilde{\Psi}\} = -\log|\mathbf{A}| - \frac{1}{2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + a + constant \qquad (29)$$

is obtained. This gives us a recipe to employ basis functions that lead to tractable computations and the policy can be computed as $\pi(u|x,\mathbf{w}) = \mathcal{N}(u|\mathbf{A}^{-1}\mathbf{a}, \mathbf{A}^{-1})$.

V. Experiments

A. Gridworld - Analytical Infinite Horizon RL

We start by evaluating the proposed algorithm (23) on a problem used in [22], with finite state and action spaces, which allows a tabular representation of Ψ. The state space is given by a $N \times N$ grid (see Fig. 2(b)) with some obstacles. The control can move the state to any adjacent ones not occupied by an obstacle and the move succeeds with a probability of 0.8. Additionally, a set $\mathbb{A} \subseteq \mathbb{X}$ of absorbing target states was defined and the agent incurs a cost of 1 at all states other than the target, i.e., $C(x,u) = \delta_{x \notin \mathbb{A}}$ with δ the Kronecker delta. The cost was not discounted. We benchmark performance against tabular $\mathcal{Q}$-learning [21]. Both algorithms were given



(a)

(b)

Fig. 2: Results from the Gridworld problem. (a) Evolution of the mean error in (30) averaged over 10 trials with error bars indicating the s.d. (b) Optimal value function (white low expected cost - black high expected cost) of the problem. Obstacles are black and the target state is indicated by *.

data from episodes generated with controls sampled from an uninformed policy. Once a target state was reached, or if the target wasn't reached within 100 steps, the state was reset randomly. The learning rate for Q-learning decayed as $\alpha = c/(c+k)$ with k the number of transitions sampled and c a constant which was optimised manually. Representative results are illustrated in Fig. 2. We plot the approximation error

$$e_{\mathcal{J}} = \frac{\max_x |\mathcal{J}(x) - \hat{\mathcal{J}}(x)|}{\max_x \mathcal{J}(x)} \tag{30}$$

between the true value function $\mathcal{J}$, obtained by value iteration, and it's estimate $\hat{J}$, given by $\bar{\Psi}$ and $\max_u Q(x,u)$ respectively. Both algorithms achieved the same error at convergence, but the proposed algorithm (Ψ-learning) consistently required fewer samples than Q-learning for convergence – this is consistent with the discussion in IV-A. We additionally considered a online variant of Ψ-learning where the controls are sampled from the policy given by the current Ψ, i.e. $\pi(u|x) = \exp\{\Psi(x,u) - \bar{\Psi}(x)\}$. As expected, the online version outperformed sampling using an uninformed policy. The aim of this evaluation, besides providing a sanity check to the working of the algorithm, was to illustrate that the proposed method provides similar performance advantages as obtained for the restricted class of problems considered in [22], despite working in the product space of states and actions, as necessitated by considering the unrestricted SOC problem.

B. Cart-Pole System

We now move on to problems with continuous state and action spaces which will make approximations necessary, demonstrating that the theoretical results presented in III can lead to practical algorithms. Specifically we will consider, both, an approximate inference approach for implementing the posterior policy iteration of III-C on a finite horizon problem and the basis function based approach, discussed in IV-B, to the RL version of the asynchronous updates for infinite horizon problems derived in III-B.

We have chosen the classical Cart-Pole problem [21], which has been repeatedly used as a benchmark in reinforcement learning, e.g., [18]. This plant, illustrated in Fig. 3a, consists of a inverted pendulum which is mounted on a cart and is controlled by exerting forces on the latter. Formally, the state space is given by $\mathbf{x} = (x, \dot{x}, \theta, \dot{\theta})$, with x the position of the cart, θ the pendulum's angular deviation from the upright position and $\dot{x}, \dot{\theta}$ their respective temporal derivatives. Neglecting the influence of friction, the continuous time dynamics of the state are given by

$$\begin{aligned} \ddot{\theta} =& \frac{g\sin(\theta) + \cos(\theta)\left[-c_1 u - c_2 \dot{\theta}^2 \sin(\theta)\right]}{\frac{4}{3}l - c_2\cos^2(\theta)} \\ \ddot{x} =& c_1 u + c_2\left[\dot{\theta}^2\sin(\theta) - \ddot{\theta}\cos(\theta)\right] \end{aligned} \tag{31}$$

with $g = 9.8m/s^2$ the gravitational constant, $l = 0.5m$ the pendulum length and constants $c_1 = (M_p + M_c)^{-1}$ and $c_2 = lM_p(M_p+M_c)^{-1}$ where $M_p = 0.1kg, M_c = 1kg$ are the pendulum and cart masses, respectively. The control interval was $0.02s$ and the dynamics were simulated using the fourth order Runga-Kutta method. Stochasticity was introduced by adding zero mean Gaussian noise, with small diagonal covariance, to the new state. These settings correspond to those used by comparative evaluations in [18].

1) Model Based Posterior Policy Iteration: First, we consider a finite horizon optimal control problem, assuming we have access to both the plant dynamics and cost function. The exact algorithm of III-A does not lend itself easily to this setting, due to the intractable integrals arising in (8) as a consequence of the nonlinear dynamics – although we note that by taking local LQG approximations of the problem, closed form updates can be derived. However, we demonstrate that by using standard approximate inference techniques, the alternative posterior policy iteration in III-C can yield good approximate optimal policies.

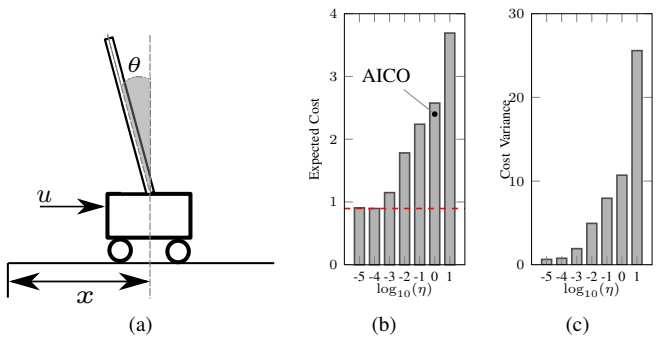
Specifically we consider the swing up task in which the pendulum has to be moved from a hanging down to an upright position and balanced. The per-step cost for this task is given by

$$\mathcal{C}_t(x_t, u_t) = \omega_1\theta^2 + \omega_2\dot{\theta}^2 + \omega_3 u_t^2 \quad \forall t \in [0, T] \ , \tag{32}$$

where ω is a vector of weights. The time horizon was $T = 3s$, but note that, since a cost is incurred in each time step for pendulum positions away from rest in the upright position, a rapid swing up followed by holding is encouraged.

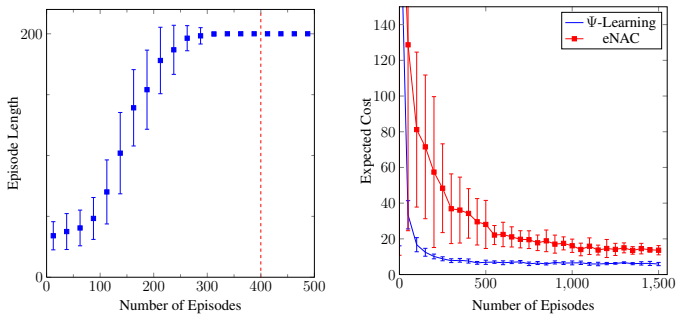
As the posterior $p_\pi(\bar{x}, \bar{u})$ is not tractable in this setting, we use an extended Kalman smoother [20] to estimate a Gaussian approximation to the full posterior, leading to a Gaussian posterior policy. As a consequence of the Gaussian approximation and inference method chosen, inference is required to be performed only once, for $p_{\pi^0}(\bar{x}, \bar{u})$, and the eventual result of the iteration (15) can be obtained as the linear policy given by the mean of the posterior policy.

In Fig. 3, we plot the expected costs and the cost variances, both estimated by sampling under the obtained policies, for different values of the parameter η. For reference, we also



(a) (b) (c)

Fig. 3: Results for model based approximate posterior policy iteration on the Cart-Pole swing-up task. (a) Schematic of the pole on cart plant used in the experiments. (b) Expected cost achieved by policies obtained for different values of the parameter η. Red dashed line indicates expected cost of policy obtained using iLQG. All values estimated from 1000 trajectories sampled using the respective policy. (c) Variance of the costs achieved by the same policies as in (b).



(a) (b)

Fig. 4: Results for RL with continuous state and action spaces. (a) Length of training episodes, averaged over blocks of 25 episodes, for Ψ-Learning, when initialized with an uninformed policy. The dashed red line indicates the point at which initial policies for the results in the subsequent comparison experiment were picked. Error bars indicate s.d. (b) Comparison of evolution of the expected cost between eNAC and Ψ-Learning. Both methods are initialised with the same stabilising policies (cf. (a)) and results averaged over 10 trials with error bars indicating s.d.

show the expected cost from the policy obtained using the iLQG algorithm [13] which also computes an approximately optimal linear policy. We first observe that as predicted by III-C, η acts to control the risk seeking behavior of the policy, and for increasing values of η the cost variance increases substantially. Furthermore, we note that the choice of $\eta = 1$, which, as discussed, corresponds to the AICO setting, leads to results substantially different from the case of classical (risk neutral) optimal control. However reducing η leads rapidly to policies obtained by approximate inference which exhibit similar performance to those obtained by classical approximate methods.

2) RL with approximations: To evaluate the RL approach proposed in IV-B we consider the balancing task, following closely the procedures in [18], where this task was used for evaluation of policy gradient methods.

The task, which consists of stabilising the pendulum in the upright position while simultaneously keeping the cart at the center of the track, had the cost function

$$\mathcal{C}_\bullet(x,u) = \begin{cases} 0 & \text{if } (x,\theta) \text{ in target set} \\ \omega_\theta \theta^2 + \omega_x x^2 & \text{else} \end{cases}, \quad (33)$$

where the target was given by $x \in [-0.05m, 0.05m]$ and $\theta \in [-0.05rad, 0.05rad]$ and the discount rate was $\gamma = 0$. We chose this cost as we found it to give better results for uniformed initial policies, for which the piece wise constant cost of [18] provided little information.

The linear policy learned in [18] corresponds to a second order polynomial basis for Ψ in the proposed method (Ψ-Learning). Specifically we used the basis set

$$\{u^2, ux, u\dot{x}, u\theta, u\dot{\theta}, x^2, x\dot{x}, x\theta, x\dot{\theta}, \dot{x}^2, \dot{x}\theta, \dot{x}\dot{\theta}, \theta^2, \theta\dot{\theta}, \dot{\theta}^2\}$$

which is of the form (28) and indeed only constitutes an approximation to the true Ψ as the problem is non-LQG.

Episodes were sampled with starting states drawn such that $\theta \in [-0.2rad, 0.2rad]$ and $x \in [-0.5m, 0.5m]$ and controls were sampled from the stochastic policy given by the current parameters. During training, episodes were terminated if the plant left the acceptable region $\theta \in [-0.2rad, 0.2rad]$ and $x \in [-0.5m, 0.5m]$ or after 200 steps. Policy parameters were updated every 10 episodes and every 5 updates policies were evaluated by sampling 50 episodes of 500 step length using the mean of the policy. All results were averaged over 10 trials. The learning rate parameter for policy gradient methods was adjusted manually for best results.

Despite the change in cost function, like [18], we were not able to reliably obtain good policies from uninformed initialisation when using policy gradient methods. Our method on the other hand, when initialised with an uninformed policy, i.e., zero mean and a variance of 10, was able to learn a stabilising policy within 400 training episodes. This is illustrated in Fig. 4a where the average length of training episodes is shown. In order to be able to compare to the *episodic Natural Actor Critic* (eNAC) method, which produced the best result in [18], we used the policies obtained by Ψ-Learning after 400 training episodes as initial policies. By this stage, the average expected cost of the policies was 239.35 compared to the initial cost which had been of the order 3×10^5. Fig. 4b shows the evolution of the expected cost for both methods with such an initialisation and as can be seen Ψ-Learning outperformed eNAC both in terms of convergence speed and attained expected cost.

As the quality of the obtained policy will depend on how well the basis set can approximate the true Ψ, we also considered a more complex set of bases. Specifically, while keeping $\mathbf{A}$ in (28) a set of non-zero constant basis functions,

we represented $\mathbf{a}(\mathbf{x}, \mathbf{w})$ and $a(\mathbf{x}, \mathbf{w})$ using the general and commonly used squared exponential bases which are of the form

$$\phi(\mathbf{x}) = \exp\{-(\mathbf{x} - m_\phi)^T \Sigma_\phi (\mathbf{x} - m_\phi)\} \tag{34}$$

with center m_ϕ and metric Σ_ϕ. The centers were sampled randomly from a region given by the acceptable region specified earlier and $\dot{x} \in [-1m/s, 1m/s], \dot{\theta} \in [-1rad/s, 1rad/s]$ and Σ_ϕ was chosen to be diagonal. For this setting we were not able to obtain good policies using eNAC, while in the case of Ψ-Learning this choice did not outperform the polynomial basis, yielding a best policy with expected cost 26.4.

VI. Conclusion

We have presented a general relation between stochastic optimal control problems and minimisation of KL divergences of the form (4). This allowed us to derive iterative algorithms for obtaining both risk neutral and risk sensitive optimal controls for finite and infinite horizon MDPs. We show that these algorithms, although instances of generalised EM procedures, enjoy guaranteed convergence to the global optimum. Further, we discuss the connections of our work to previous approaches in this area, highlighting that many of these arise in our formulation as special cases which either require restrictions on the class of problems (e.g., [22, 10]), or for which the relation to SOC was previously unclear (e.g., [24]). The formalism is then extended to the model free RL setting in both the finite and infinite horizon case. In the case of finite state and action spaces, using a tabular representation, we obtain an exact algorithm with interesting relations to $\mathcal{Q}$- and TD(0) learning. We also present an approximation, based on basis function representations, which extends [5] to problems with continuous state and action spaces.

Our approach is verified in the discrete setting and we highlight the novel aspects of our work in experiments on a problem with continuous states and actions in the form of the standard Cart-Pole benchmark. On the one hand we show that, by employing standard out of the box approximate inference methods, optimal policies can be computed for model based finite horizon problems, improving the shortcomings of [24]. On the other hand, we consider an infinite horizon problem in the model free RL setting, demonstrating that the proposed approximate algorithm shows performance competitive with the well established eNAC algorithm. We also provide a recipe for selecting appropriate basis functions that lead to efficient, tractable solutions.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] D. Braun, M. Howard, and S. Vijayakumar. Exploiting variable stiffness in explosive movement tasks. In *R:SS*, 2011.

[3] J.L. van den Broek, W.A.J.J. Wiegerinck, and H.J. Kappen. Risk sensitive path integral control. In *UAI*, 2010.

[4] P. Dayan and G. E. Hinton. Using EM for reinforcement learning. *Neural Computation*, 9:271–278, 1997.

[5] A.M. Gheshlaghi et al. Dynamic policy programming with function approximation. In *AISTATS*, 2011.

[6] D. Mitrovic et al. Optimal feedback control for anthropomorphic manipulators. In *ICRA*, 2010.

[7] E. A. Theodorou et al. Learning policy improvements with path integrals. In *AISTATS*, 2010.

[8] S.I. Marcus et al. Risk sensitive markov decision processes. *Systems and control in the 21st century*, 1997.

[9] A. Gunawardana and W. Byrne. Convergence theorems for generalized alternating minimization procedures. *J. of Machine Learning Research*, 6:2049–2073, 2005.

[10] B. Kappen, V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. arXiv:0901.0633v2, 2009.

[11] H. J. Kappen. Path integrals and symmetry breaking for optimal control theory. *J. of Statistical Mechanics: Theory and Experiment*, page 11011ff, 2005.

[12] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–64, 1997.

[13] W. Li and E. Todorov. An iterative optimal control and estimation design for nonlinear stochastic system. In *CDC*, 2006.

[14] D. Mitrovic, S. Klanke, and S. Vijayakumar. Adaptive optimal control for redundantly actuated arms. In *SAB*, 2008.

[15] J. Nakanishi, K. Rawlik, and S. Vijayakumar. Stiffness and temporal optimization in periodic movements: An optimal control approach. In *IROS*, 2011.

[16] J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *Humanoids*, 2003.

[17] K. Rawlik, M. Toussaint, and S. Vijayakumar. An approximate inference approach to temporal optimization in optimal control. In *NIPS*, 2010.

[18] M. Riedmiller, J. Peters, and S. Schaal. Evaluation of policy gradient methods and variants on the cart-pole benchmark. In *IEEE ADPRL*, 2007.

[19] P. N. Sabes and M. I. Jordan. Reinforcement learning by probability matching. In *NIPS*, 1996.

[20] R. F. Stengel. *Optimal Control and Estimation*. Dover Publications, 1986.

[21] R.S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, 1998.

[22] E. Todorov. Efficient computation of optimal actions. *PNAS*, 106:11478–11483, 2009.

[23] E. Todorov and M. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5:1226–1235, 2002.

[24] M. Toussaint. Robot trajectory optimization using approximate inference. In *ICML*, 2009.

[25] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. In *ICML*, 2006.

[26] D. Zarubin, V. Ivan, M. Toussaint, T. Komura, and S. Vijayakumar. Hierarchical motion planning in topological representations. In *R:SS*, 2012.

On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference - Supplementary Material

Konrad Rawlik
University of Edinburgh
Edinburgh, UK
Email: k.c.rawlik@ed.ac.uk

Marc Toussaint
FU Berlin
Berlin, Germany

Sethu Vijayakumar
University of Edinburgh
Edinburgh, UK

Contents

1 Proofs and Derivation from the main text

1.1 Proofs of Duality and General Iterative Procedure (cf. II-C & III)

Theorem 1 (see also II-C in the main text)**.** *Let π^0 be an arbitrary stochastic policy and $\mathbb{D}$ the set of deterministic policies, then the problem*

$$\pi^* = \operatorname*{argmin}_{\pi\in\mathbb{D}} \mathrm{KL}\left(q_\pi \| p_{\pi^0}\right)$$

is equivalent to the stochastic optimal control problem (1) with cost per stage

$$\hat{\mathcal{C}}_t(x_t, u_t) = \mathcal{C}_t(x_t, u_t) - \frac{1}{\eta}\log\pi^0(u_t|x_t)$$

Proof. Let $\pi_t(u_t\|x_t) = \delta_{u_t=\tau_t(x_t)}$, for some function τ, then

$$\begin{aligned}
\mathrm{KL}\left(q_\pi \| p_{\pi^0}\right) &= \log P(\bar{r}=1) + \int_{\bar{x}} \mathrm{d}\bar{x} \int_{\bar{u}} \mathrm{d}\bar{u}\; q_\pi(\bar{x},\bar{u}) \log\frac{q_\pi(\bar{x},\bar{u})}{q_{\pi^0}(\bar{x},\bar{u})} \\
&\quad + \int_{\bar{x}} \mathrm{d}\bar{x} \int_{\bar{u}} \mathrm{d}\bar{u}\; q_\pi(\bar{x})\; \pi(\bar{u}|\bar{x}) \sum_{t=0}^{T} \log\frac{1}{\exp\{-\eta\mathcal{C}_t(x_t,u_t)\}} \qquad (33) \\
&= \log P(\bar{r}=1|x_0;\pi^0) + \mathrm{KL}\left(q_\pi(\bar{x},\bar{u}) \| q_{\pi^0}(\bar{x},\bar{u})\right) \\
&\quad + \int_{\bar{x}} \mathrm{d}\bar{x} \int_{\bar{u}} \mathrm{d}\bar{u}\; q_\pi(\bar{x})\; \delta_{\bar{u}=\tau(\bar{x})} \sum_{t=0}^{T} \eta\mathcal{C}_t(x_t,u_t) \qquad (34) \\
&= \log P(\bar{r}=1|x_0;\pi^0) + \mathrm{KL}\left(q_\pi(\bar{x},\bar{u}) \| q_{\pi^0}(\bar{x},\bar{u})\right) \\
&\quad + \int_{\bar{x}} \mathrm{d}\bar{x}\; q_\pi(\bar{x}) \sum_{t=0}^{T} \eta\mathcal{C}_t(x_t,\tau_t(x_t)) \,. \qquad (35)
\end{aligned}$$

Furthermore the divergence between the controlled process, q_π, and prior process, q_{π^0} is

$$\mathrm{KL}\left(q_\pi(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right) = \int_{\bar{x}} \mathrm{d}\bar{x} \int_{\bar{u}} \mathrm{d}\bar{u}\; q_\pi(\bar{x},\bar{u}) \sum_{t=0}^{T} \log \frac{\delta_{u_t=\tau_t(x_t)}}{\pi^0(u_t|x_t)} \tag{36}$$

$$= -\int_{\bar{x}} \mathrm{d}\bar{x}\; q_\pi(\bar{x}) \sum_{t=0}^{T} \log \pi^0(\tau_t(x_t)\|x_t)\ . \tag{37}$$

Hence,

$$\mathrm{KL}\left(q_\pi\|p_{\pi^0}\right) = \log P(\bar{r}=1|x_0;\pi^0) + \eta \left\langle \sum_{t=0}^{T} \left[\mathcal{C}_t(x_t,\tau_t(x_t)) - \frac{1}{\eta}\log \pi^0(\tau_t(x_t)\|x_t)\right]\right\rangle_{q_\pi}\ , \tag{38}$$

and as $\log P(\bar{r}=1|x_0;\pi^0)$ is constant w.r.t. π, the result follows. ■

Theorem 2 (see also III in the main text)**.** *For any $\pi \neq \pi^0$,*

$$\mathrm{KL}\left(q_\pi\|p_{\pi^0}\right) \leq \mathrm{KL}\left(q_{\pi^0}\|p_{\pi^0}\right) \Longrightarrow \left\langle \mathcal{C}(\bar{x},\bar{u})\right\rangle_{q_\pi} < \left\langle \mathcal{C}(\bar{x},\bar{u})\right\rangle_{q_{\pi^0}}\ .$$

Proof. Expanding the KL divergences we have

$$\begin{aligned}\mathrm{KL}\left(q_\pi(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right) &- \left\langle \log P(r_t=1|\bar{x},\bar{u})\right\rangle_{q_\pi(\bar{x},\bar{u})} + \log P(\bar{r}=1\|x_0;\pi^0) \\ &\leq \mathrm{KL}\left(q_{\pi^0}(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right) - \left\langle \log P(\bar{r}=1|\bar{x},\bar{u})\right\rangle_{q_{\pi^0}(\bar{x},\bar{u})} + \log P(\bar{r}=1\|x_0;\pi^0)\ .\end{aligned} \tag{39}$$

Subtracting $\log P(\bar{r}=1\|x_0;\pi^0)$ on both sides and noting that $\mathrm{KL}\left(q_{\pi^0}(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right)=0$, we obtain

$$\mathrm{KL}\left(q_\pi(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right) - \left\langle \log P(\bar{r}=1|\bar{x},\bar{u})\right\rangle_{q_\pi(\bar{x},\bar{u})} \leq -\left\langle \log P(\bar{r}=1|\bar{x},\bar{u})\right\rangle_{q_{\pi^0}(\bar{x},\bar{u})}\ . \tag{40}$$

and as $\log P(\bar{r}=1|\bar{x},\bar{u}) = -\eta\mathcal{C}(\bar{x},\bar{u})$

$$\mathrm{KL}\left(q_\pi(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right) + \left\langle \eta\mathcal{C}(\bar{x},\bar{u})\right\rangle_{q_\pi(\bar{x},\bar{u})} \leq \left\langle \eta\mathcal{C}(\bar{x},\bar{u})\right\rangle_{q_{\pi^0}(\bar{x},\bar{u})}\ . \tag{41}$$

Hence, as $\eta \geq 0$ and $\mathrm{KL}\left(q_\pi(\bar{x},\bar{u})\|q_{\pi^0}(\bar{x},\bar{u})\right) \geq 0$ with equality iff $\pi=\pi^0$, the result follows. ■

1.2 Derivation of updates in III-A

The form of the updates can be derived by induction. In particular consider the policy of time $T-1$, π_{T-1}. Applying Lemma 12 with $a = u_{T-1}|x_{T-1}$, $b = x_T$ and $P(c=\hat{c}|b) = \exp\{-\eta\mathcal{C}_T(x_T)\}$ leads to the base case. For the inductive step we observe that we may write the KL divergence in a recursive form as

$$\begin{aligned}\mathrm{KL}\left(q_{\pi^{n+1}}(\bar{x},\bar{u})\|p_{\pi^n}(\bar{x},\bar{u})\right) = \int_{u_0} \pi^{n+1}(u_0|x_0) \Bigg[\log \frac{\pi^{n+1}(u_0|x_0)}{\pi^n(u_0|x_0)P(r_0|x_0,u_0)} \\ + \int_{x_1} P(x_1|x_0,u_0)\,\mathrm{KL}\left(q_{\pi^{n+1}}(x_{1:T},u_{1:T})\|p_{\pi^n}(x_{1:T},u_{1:T})\right)\Bigg]\end{aligned} \tag{42}$$

We can now apply Lemma 12 recursively with $a = u_t|x_t$, $b = x_{t+1}$ and

$$P(c=\hat{c}|b) = P(r_t|x_t,u_t)\exp\{-\bar{\Psi}_{t+1}^{n+1}(x_{t+1})\} \tag{43}$$

and the updates of the form (8) in III-A follow.

1.3 Proof of Convergence of Exact Updates (cf. III-A1)

The convergence proof is completed by the following proofs of the two propositions given in the main text.

Lemma 3 (see also III-A1 in the main text)**.** *Let the sequence* $\{\pi^n\}$ *be generated by (6) and let* $\hat{\pi}$ *be an arbitrary (stochastic) policy. Then*

$$\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^{n+1}}\right) - \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^n}\right) \leq \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\hat{\pi}}(\bar{x}, \bar{u})} - \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^{n+1}}(\bar{x}, \bar{u})}$$

Proof. Let $\hat{\pi}$ be an arbitrary policy and consider

$$\begin{aligned}
&\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^{n+1}}\right) - \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^n}\right) && (44)\\
&\quad = \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \log \frac{q_{\pi^n}}{q_{\pi^{n+1}}} && (45)\\
&\quad = \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \log \prod_{t=0}^{T} \frac{\pi^n(u_t|x_t)}{\pi^{n+1}(u_t|x_t)} && (46)\\
&\quad = \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \sum_{t=0}^{T} \log \frac{\pi^n(u_t|x_t)}{\pi^n(u_t|x_t) \exp\{-\eta \mathcal{C}_t(x_t, u_t) + \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x') - \bar{\Psi}_t^{n+1}\}} && (47)\\
&\quad = \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \sum_{t=0}^{T} \left[\eta \mathcal{C}_t(x_t, u_t) - \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x') + \bar{\Psi}_t^{n+1}\right] && (48)\\
&\quad = \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \sum_{t=0}^{T} \eta \mathcal{C}_t(x_t, u_t) + \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \sum_{t=0}^{T} \left[\bar{\Psi}_t^{n+1} - \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x')\right] . && (49)\\
& && (50)
\end{aligned}$$

Now

$$\begin{aligned}
&\int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \sum_{t=0}^{T} \left[\bar{\Psi}_t^{n+1} - \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x')\right]\\
&\quad = \sum_{t=0}^{T} \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \bar{\Psi}_t^{n+1}(x_t) - \sum_{t=0}^{T} \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x') && (51)\\
&\quad = \int_{x_0} P(x_0) \bar{\Psi}_0^{n+1}(x_0) + \sum_{t=1}^{T} \int_{x_{0:t}, u_{0:t}} q_{\hat{\pi}}(x_{0:t-1}, u_{0:t-1}) \int_{x_t} P(x_t|x_{t-1}, u_{t-1}) \bar{\Psi}_t^{n+1}(x_t)\\
&\qquad - \sum_{t=0}^{T} \int_{x_{0:t}, u_{0:t}} q_{\hat{\pi}}(x_{0:t}, u_{0:t}) \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{t+1}^{n+1}(x') && (52)\\
&\quad = \int_{x_0} P(x_0) \bar{\Psi}_0^{n+1}(x_0) - \int_{\bar{x},\bar{u}} q_{\hat{\pi}}(\bar{x}, \bar{u}) \int_{x'} P(x'|x_t, u_t) \bar{\Psi}_{T+1}^{n+1}(x') && (53)\\
&\quad = \int_{x_0} P(x_0) \bar{\Psi}_0^{n+1}(x_0) && (54)
\end{aligned}$$

and hence

$$\begin{aligned}
\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^{n+1}}\right) - \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^n}\right) &= \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\hat{\pi}}(\bar{x}, \bar{u})} + \int_{x_0} P(x_0) \bar{\Psi}_0^{n+1} && (55)\\
&\leq \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\hat{\pi}}(\bar{x}, \bar{u})} - \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^{n+1}}(\bar{x}, \bar{u})} , && (56)
\end{aligned}$$

where in the final line we used the bound from Lemma 11. ■

Theorem 4. *(see also III-A1 in the main text) Let* $\{\pi^n\}$ *be a sequence of policies generated by (6), with* π^0 *s.t.* $\pi^0(\cdot|x \in \mathbb{X})$ *has support* $\mathbb{U}$*. Then*

$$\lim_{n \to \infty} \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}} = min_{\pi} \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi}} \tag{57}$$

Proof. Summing the bound of Lemma 3 over $n = 0..N$ we have

$$\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^N}\right) - \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^0}\right) \leq N \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\hat{\pi}}(\bar{x},\bar{u})} - \sum_{n=1}^{N+1} \left\langle \eta \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}(\bar{x},\bar{u})} \tag{58}$$

and hence

$$\sum_{n=1}^{N+1} \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}(\bar{x},\bar{u})} \leq N \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\hat{\pi}}(\bar{x},\bar{u})} + \frac{1}{\eta} \left[\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^0}\right) - \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^N}\right)\right] \tag{59}$$

$$\leq N \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\hat{\pi}}(\bar{x},\bar{u})} + \frac{1}{\eta N} \mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^0}\right) \tag{60}$$

where the last line follows from $\mathrm{KL}\left(q_{\hat{\pi}} \| q_{\pi^N}\right) \geq 0$. Noting that $\hat{\pi}$ was chosen arbitrarily we may now choose $\hat{\pi} = \pi^* = \mathrm{argmin}_\pi \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_\pi}$ so that we have

$$\frac{1}{N} \sum_{n=1}^{N+1} \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}(\bar{x},\bar{u})} \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^*}(\bar{x},\bar{u})} + \frac{1}{N\eta} \mathrm{KL}\left(q_{\pi^*} \| q_{\pi^0}\right) \ . \tag{61}$$

Note that as the lhs side is the average expected cost over $\pi^1 \ldots \pi^{N+1}$ there exists some $n \in 1 \ldots N$ s.t.

$$\left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^{N+1}}(\bar{x},\bar{u})} \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}(\bar{x},\bar{u})} \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^*}(\bar{x},\bar{u})} + \frac{1}{\eta N} \mathrm{KL}\left(q_{\pi^*} \| q_{\pi^0}\right) \ , \tag{62}$$

with the first inequality following from Theorem 2.

Now, as by assumption on π^0, $\mathrm{KL}\left(q_{\pi^*} \| q_{\pi^0}\right) < \infty$, for any $\epsilon > 0$ there exists a N_ϵ s.t. $\frac{1}{N_\epsilon \eta} \mathrm{KL}\left(q_{\pi^*} \| q_{\pi^0}\right) < \epsilon$ and

$$\left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^{N_\epsilon}}(\bar{x},\bar{u})} \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^*}(\bar{x},\bar{u})} + \epsilon \tag{63}$$

which gives the required result. ■

1.4 Proof of Convergence for Asynchronous Updates (cf. III-B1)

Essentially equivalent results to those for the finite horizon case can be obtained for the asynchronous algorithm (14) in the infinite horizon setting. In general we assume that the cost is bounded and consider finite horizon problems with growing horizon, bounding the expected cost of the infinite tail. As we assume that the cost is discounted the expected cost of the tail goes to zero as the horizon increases.

More specifically we assume the cost is bounded, then $\exists \bar{\mathcal{C}}$ s.t. $\forall \pi \ \bar{\mathcal{C}} \geq \left\langle \sum_t \gamma^t \mathcal{C}_\bullet(x_t, u_t) \right\rangle_{q_\pi}$. For notational convenience we shall also assume $\eta = 1$. Then we first show that

Theorem 5. *Let $\{\pi^i\}$ be a sequence of policies generated by (14) and let $\hat{\pi}$ be an arbitrary (stochastic) policy, then*

$$\lim_{N\to\infty} -\frac{1}{N} \sum_{i=0}^{N} \bar{\Psi}^{i+1}(x) \leq \left\langle \sum_{t=0}^{\infty} \gamma^t \mathcal{C}_\bullet(x_t, u_t) \right\rangle_{q_{\hat{\pi}}} \tag{64}$$

The proof is by induction on the time horizon using the following two lemmas. The base case is given by

Lemma 6. *For any $\epsilon > 0$ there exists N_ϵ s.t. for all $N > N_\epsilon$*

$$-\frac{1}{N} \sum_{i=0}^{N} \Psi^i(x_0) \leq \left\langle \mathcal{C}_\bullet(x_0, u_0) \right\rangle_{q_{\hat{\pi}}} + \gamma \bar{\mathcal{C}} + \epsilon \tag{65}$$

Proof. Consider

$$\mathrm{KL}\left(\hat{\pi}\|\pi^{n+1}\right) - \mathrm{KL}\left(\hat{\pi}\|\pi^{n}\right) = \int_u \hat{\pi}(u|x) \log \frac{\pi^n}{\pi^{n+1}} \tag{66}$$

$$= \int_u \hat{\pi}(u|x) \log \exp\{\mathcal{C}(x,u) - \gamma \int_y P(y|x,u)\bar{\Psi}^i(y) + \bar{\Psi}^{i+1}(x)\} \tag{67}$$

$$= \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) - \gamma \int_y P(y|x,u)\bar{\Psi}^i(y)\right] + \bar{\Psi}^{i+1}(x) \tag{68}$$

$$\leq \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) + \gamma\bar{\mathcal{C}}\right] + \bar{\Psi}^{i+1}(x) \tag{69}$$

Summing the bound over $i = 1..N$ we have

$$\mathrm{KL}\left(\hat{\pi}\|\pi^{N}\right) - \mathrm{KL}\left(\hat{\pi}\|\pi^{0}\right) \leq N \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) + \gamma\bar{\mathcal{C}}\right] + \sum_{i=0}^{N} \bar{\Psi}^{i+1}(x) \tag{70}$$

and hence

$$\frac{1}{N}\sum_{i=0}^{N} \bar{\Psi}^{i+1}(x) \leq \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) + \gamma\bar{\mathcal{C}}\right] + \frac{1}{N}\mathrm{KL}\left(\hat{\pi}\|\pi^{0}\right) \ . \tag{71}$$

■

The following inductive step completes the proof of Theorem 5.

Lemma 7. *Assume for a given T and any $\epsilon > 0$ there exists N_ϵ s.t. for all $n > N_\epsilon$*

$$-\frac{1}{N}\sum_{n=0}^{N} \bar{\Psi}^{n}(x) \leq \left\langle \sum_{t=0}^{T} \gamma^t C(x_t,u_t) + \gamma^T \bar{C} \right\rangle_{q_{\hat{\pi}}} + \epsilon \tag{72}$$

then for any $\delta > 0$ there exists N_δ s.t. for all $n > N_\delta$

$$-\frac{1}{N}\sum_{n=0}^{N} \bar{\Psi}^{n}(x) \leq \left\langle \sum_{t=0}^{T+1} \gamma^t \mathcal{C}_\bullet(x_t,u_t) + \gamma^{T+1} \bar{C} \right\rangle_{q_{\hat{\pi}}} + \delta \tag{73}$$

Proof. Consider

$$\mathrm{KL}\left(\hat{\pi}\|\pi^{n+1}\right) - \mathrm{KL}\left(\hat{\pi}\|\pi^{n}\right) = \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) - \gamma \int_y P(y|x,u)\bar{\Psi}^n(y)\right] + \bar{\Psi}^{n+1}(x) \tag{74}$$

$$\tag{75}$$

Summing the bound over $i = 1..N$ we have

$$\mathrm{KL}\left(\hat{\pi}\|\pi^{N}\right) - \mathrm{KL}\left(\hat{\pi}\|\pi^{0}\right) \leq \sum_{n=0}^{N} \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) - \gamma \int_y P(y|x,u)\bar{\Psi}^n(y)\right] + \sum_{n=0}^{N} \bar{\Psi}^{n+1}(x) \tag{76}$$

and therefore

$$-\frac{1}{N}\sum_{n=0}^{N} \bar{\Psi}^{n+1}(x) \leq \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) - \gamma \int_y P(y|x,u) \frac{1}{N}\sum_{n=0}^{N} \bar{\Psi}^{n}(y)\right] + \frac{1}{N}\mathrm{KL}\left(\hat{\pi}\|\pi^{0}\right) \tag{77}$$

$$\leq \int_u \hat{\pi}(u|x) \left[\mathcal{C}_\bullet(x,u) - \gamma \int_y P(y|x,u) \frac{1}{N} \left\langle \sum_{t=0}^{T} \gamma^t \mathcal{C}_\bullet(x,u) + \gamma^T \bar{\mathcal{C}} \right\rangle_{q_{\hat{\pi}}} + \epsilon\right] \tag{78}$$

$$+ \frac{1}{T}\mathrm{KL}\left(\hat{\pi}\|\pi^{0}\right) \tag{79}$$

$$= \left\langle \sum_{t=0}^{T+1} \gamma^t \mathcal{C}_\bullet(x,u) + \gamma^{T+1} \bar{\mathcal{C}} \right\rangle_{q_{\hat{\pi}}} + \frac{1}{T}\mathrm{KL}\left(\hat{\pi}\|\pi^{0}\right) \ . \tag{80}$$

■

Using the above result we may now show:

Theorem 8. *Let the cost be bounded and let π^n be a sequence generated policies with π^0 s.t. $\forall x$ $\mathrm{KL}\left(\pi^*(\cdot|x)\|\pi^0(\cdot|x)\right) < \infty$ then*

$$\lim_{n\to\infty} \left\langle \sum_{t=0}^{\infty} \gamma^t \mathcal{C}_\bullet(x_t, u_t) \right\rangle_{q_{\pi^n}} = min_\pi \left\langle \sum_{t=0}^{\infty} \gamma^t \mathcal{C}_\bullet(x_t, u_t) \right\rangle_{q_\pi} \tag{81}$$

Proof. As $\hat{\pi}$ in Theorem 5 is arbitrary we may choose the tightest bound given by

$$\hat{\pi} = \pi^* = \operatorname*{argmin}_\pi \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_\pi} \ , \tag{82}$$

where we use the notation $\mathcal{C}(\bar{x}, \bar{u}) = \sum_{t=0}^{\infty} \gamma^t \mathcal{C}_\bullet(x_t, u_t)$. Now as for a given x_0

$$\left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}} \leq -\bar{\Psi}^n(x_0) \tag{83}$$

we have

$$\lim_{n\to\infty} \frac{1}{N} \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}} \leq \lim_{n\to\infty} -\frac{1}{N} \sum_{i=0}^{N} \bar{\Psi}^{n+1}(x_0) \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^*}} \tag{84}$$

As the lhs is the average expected cost over $\pi^1 \dots \pi^N$ there exists $n \in 1 \dots N$ s.t.

$$\left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^{N+1}}} \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^n}} \leq \left\langle \mathcal{C}(\bar{x}, \bar{u}) \right\rangle_{q_{\pi^*}} \tag{85}$$

with the first inequality following from Theorem 2. Noting that by the definition of π^*, i.e., (82), the rhs is also a lower bound gives the required result. ■

1.5 Proof of Convergence for Posterior Policy Iteration (cf. III-C)

The following results establish asymptotic behavior of the posterior policy iteration as given in III-C.

Theorem 9. *Let $\{\pi^n\}$ be a sequence of policies generated by (15), then*

$$\pi^n \to \operatorname*{argmin}_\pi -\frac{1}{\eta} \log \left\langle \exp\{-\eta \mathcal{C}_t(\bar{x}, \bar{u})\} \right\rangle_{q_\pi} \tag{86}$$

Proof. We may write the policy in terms of a suitable distribution over deterministic policies τ and in particular $\pi^n \propto \int P(u_t|x_t, \tau(\cdot)) P^n(\tau(\cdot))$ where $P(u_t|x_t, \tau(\cdot)) = \delta_{u_t = \tau(x_t)}$. With this notation the iteration becomes

$$P^{n+1}(\tau(\cdot)) = \frac{1}{Z} P(\bar{r} = 1|\tau(\cdot)) P^n(\tau(\cdot)) \tag{87}$$

with Z a normalisation constant. Expanding from P^0 for n iterations we therefore have

$$P^n(\tau(\cdot)) \propto \left[P(\bar{r} = 1|\tau(\cdot))\right]^n P^0(\tau(\cdot)) \tag{88}$$

and hence for $n \to \infty$, $P(\bar{r} = 1|\tau(\cdot))$ dominates and $P^n(\tau)$ converges to the delta at the maximum of

$$P(\bar{r} = 1|\tau(\cdot)) = \left\langle \exp\{-\eta \mathcal{C}_t(\bar{x}, \bar{u})\} \right\rangle_{q_\tau} \ . \tag{89}$$

As log is strictly monotonic and $\eta > 0$ this establishes the result. ■

1.6 Asymptotic behavior of Ψ & $\bar{\Psi}$

In IV-A discussing the relation to $\mathcal{Q}$-learning and TD(0) specific, claims are made about the asymptotic values of Ψ and $\bar{\Psi}$. Convergence in the absolute value of $\bar{\Psi}$ to the value function follows directly from Lemma 11 and the convergence results of III-A. As we also have shown that π^n converges to the optimal policy it follows that for any sub optimal action u in state x , $\pi^n(u|x) \to 0$ which, as $\pi^n \propto \exp\{\Psi(x, u)\}$ (cf. equation (7)), implies $\Psi(x, u) \to -\infty$.

2 Other Lemmas

The following lemmas although not mentioned in the main text are referenced by the preceding proofs. In the following we use $\mathcal{KL}\left(q_\pi \| p_{\tilde{\pi}}\right)$ to denote the KL divergence with unnormalised second argument, i.e.,

$$\mathcal{KL}\left(q_\pi \| p_{\tilde{\pi}}\right) = \int_{\bar{x},\bar{u}} q_\pi(\bar{x},\bar{u}) \log \frac{q_\pi(\bar{x},\bar{u})}{P(\bar{x},\bar{u},\bar{r}=1)} \ . \tag{90}$$

Lemma 10. *Let $\{\pi^n\}$ be a sequence generated by (6), then*

$$\mathcal{KL}\left(q_{\pi^{n+1}} \| p_{\pi^n}\right) = -\int_{x_0} P(x_0)\bar{\Psi}_0^{n+1}(x_0) \tag{91}$$

Proof. This follows by application of Lemma 12. ■

Lemma 11. *Let $\{\pi^n\}$ be a sequence of policies generated by (6), then*

$$\langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^{n+1}}} \leq -\int_{x_0} P(x_0)\bar{\Psi}_0^{n+1}(x_0) \leq \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^n}} \tag{92}$$

Proof. We have

$$\mathcal{KL}\left(q_{\pi^{n+1}} \| p_{\pi^n}\right) = \mathrm{KL}\left(q_{\pi^{n+1}} \| q_{\pi^n}\right) + \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^{n+1}}} \geq \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^{n+1}}} \ , \tag{93}$$

where the inequality follows from $\mathrm{KL}\left(q_{\pi^{n+1}} \| q_{\pi^n}\right) \geq 0$. Also from (6) we have

$$\mathrm{KL}\left(q_{\pi^{n+1}} \| p_{\pi^n}\right) \leq \mathrm{KL}\left(q_{\pi^n} \| p_{\pi^n}\right) \tag{94}$$

hence

$$\begin{aligned}
\mathcal{KL}\left(q_{\pi^{n+1}} \| p_{\pi^n}\right) &\leq \mathcal{KL}\left(q_{\pi^n} \| p_{\pi^n}\right) && (95)\\
&\leq \mathcal{KL}\left(q_{\pi^n}(\bar{x},\bar{u}) \| q_{\pi^n}(\bar{x},\bar{u})\right) - \langle \log P(\bar{r}=1|\bar{x},\bar{u})\rangle_{q_{\pi^n}(\bar{x},\bar{u})} && (96)\\
&\leq \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^n}(\bar{x},\bar{u})} \ . && (97)
\end{aligned}$$

It follows that

$$\langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^{n+1}}} \leq \mathcal{KL}\left(q_{\pi^{n+1}} \| p_{\pi^n}\right) \leq \langle \mathcal{C}(\bar{x},\bar{u})\rangle_{q_{\pi^n}(\bar{x},\bar{u})} \tag{98}$$

and by Lemma 10 the result follows. ■

Lemma 12. *Let a, b, c be random variables with joint $P(a,b,c) = P(a)P(b|a)P(c|b,a)$ and $\mathbb{P}$ the set of distributions over a, then*

$$P(a)\exp\{\int_b P(b|a)\log P(c=\hat{c}|b)\} \propto \operatorname*{argmin}_{q\in\mathbb{P}} \mathrm{KL}\left(q(a)P(b|a) \| P(a,b|c=\hat{c})\right) \tag{99}$$

and

$$-\log \int_a P(a)\exp\{\int_b P(b|a)\log P(c=\hat{c}|b)\} = \min_{q\in\mathbb{P}} \mathcal{KL}\left(q(a)P(b|a) \| P(a,b|c=\hat{c})\right) \ . \tag{100}$$

Proof. We form the Lagrangian

$$\begin{aligned}
\mathcal{L} &= \mathrm{KL}\left(q(a)P(b|a) \| P(a,b|c=\hat{c})\right) + \lambda\left[\int_a q(a) - 1\right] && (101)\\
&\cong \int_{a,b} q(a)P(b|a)\log\frac{q(a)P(b|a)}{P(a)P(b|a)P(c=\hat{c}|b)} + \lambda\left[\int_a q(a) - 1\right] && (102)\\
&= \int_a q(a)\log\frac{q(a)}{P(a)} - \int_{a,b} q(a)P(b|a)\log P(c=\hat{c}|b) \ , && (103)
\end{aligned}$$

where we use $\cong$ to indicate equality up to an additive constant. Setting the partial derivatives w.r.t. $q(a)$ to 0 gives

$$0 = \log \frac{q(a)}{P(a)} + 1 - \int_b P(b|a) \log P(c = \hat{c}|b) + \lambda \tag{104}$$

$$= \log \frac{q(a)}{\mathcal{Z}(\lambda) P(a) \exp\{\int_b P(b|a) \log P(c = \hat{c}|b)\}} , \tag{105}$$

where $\mathcal{Z}$ is a function of the Lagrange multiplier. The result in (99) now directly follows and more specifically the minimizer is

$$q^*(a) = \frac{P(a) \exp\{\int_b P(b|a) \log P(c = \hat{c}|b)\}}{\int_a P(a) \exp\{\int_b P(b|a) \log P(c = \hat{c}|b)\}} . \tag{106}$$

Substituting q^* into the KL divergence, we have

$$\mathcal{KL}\left(q^*(a)\, P(b|a) \| \; P(a, b \| c = \hat{c})\right) \tag{107}$$

$$= \int_a q^*(a) \log \frac{q^*(a)}{P(a)} - \int_{a,b} q^*(a) P(b|a) \log P(c = \hat{c}|b) \tag{108}$$

$$= \int_a q^*(a) \log \frac{\exp\{\int_b P(b|a) \log P(c = \hat{c}|b)\}}{Z} - \int_a q^*(a) \int_b P(b|a) \log P(c = \hat{c}|b) \tag{109}$$

$$= \int_a q^*(a) \int_b P(b|a) \log P(c = \hat{c}|b) + \int_a q^* a \log \frac{1}{Z} - \int_a q^*(a) \int_b P(b|a) \log P(c = \hat{c}|b) \tag{110}$$

$$= \int_a q^*(a) \log \frac{1}{Z} \tag{111}$$

$$= -\log Z , \tag{112}$$

with $Z = \int_a P(a) \exp\{\int_b P(b|a) \log P(c = \hat{c}|b)\}$. ■