Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), Beijing, 2013

On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference (Extended Abstract)*

Konrad Rawlik
School of Informatics
University of Edinburgh

Marc Toussaint
Inst. für Parallele und Verteilte Systeme
Universität Stuttgart

Sethu Vijayakumar
School of Informatics
University of Edinburgh

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.

1 Introduction

The primary aim of this work is of theoretical nature and illustrated in Fig. 1. Recently, a series of new algorithms for solving SOC problems has been proposed, which, in one way or another, draw on the *duality* between SOC and probabilistic inference. Figure 1 summarizes some exemplary work, ranging from approaches utilising Expectation Maximization for solving POMDPs to efficient Reinforcement Learning methods. Each of these algorithms demonstrates – in its specific domain – the benefits of transferring methodologies from the realm of probabilistic inference to solving SOC problems.

Our work first provides a common theoretical foundation of these methods by showing that they are special cases of a *general duality*. This is in the tradition of previous formulations of such a general duality [Kappen, 2005; Todorov, 2009]. While the dual reformulation of SOC problems does not directly allow for analytical solutions, it leads to iterative solutions that provide an alternative to the classical iterative SOC solvers. We introduce two classes of such iterative solutions, Ψ-Iterations and PPI.

In the case of Ψ-learning we can prove global convergence and derive model-free Reinforcement-Learning versions which are interestingly related to standard temporal difference learning. The existing methods eNAC [Peters and Schaal, 2008], DPP [Azar *et al.*, 2011] and REPS [Peters *et al.*, 2010] can be discussed as special cases of Ψ-learning.

*The paper on which this extended abstract is based was the recipient of the best paper runner-up award of 2012 Robotics: Science and Systems [Rawlik *et al.*, 2012]

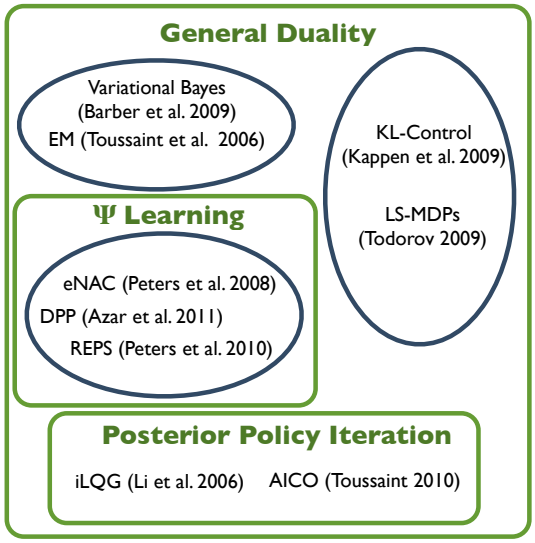

Figure 1: Summary of relations between this work and previously proposed approaches.

Posterior policy iteration is a second type of iterative solution, which directly allows for application of standard Bayesian inference methods. The motivation here is to draw on the rich variety of efficient approximate inference methods for structured graphical models, e.g., to handle hierarchical, hybrid or multi-agent systems. The existing methods AICO of [Toussaint, 2009] and iLQG by [Li and Todorov, 2006], but also risk sensitive control as described by [Marcus *et al.*, 1997] can be discussed as special cases of PPI.

2 Preliminaries

2.1 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., [Sutton and Barto, 1998], 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 $\mathcal{J}(\pi)$, i.e., solving

$$\pi^* = \operatorname*{argmin}_{\pi} \mathcal{J}(\pi) = \operatorname*{argmin}_{\pi} \mathbb{E}_{q_\pi}\left[\sum_{t=0}^{T} \mathcal{C}_t(x_t, u_t)\right] , \quad (1)$$

where the expectation is taken with respect to

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

the distribution over trajectories under policy π.

2.2 Inference Control Model

A Bayesian inference based approximation of the above control problem can be formulated [Toussaint, 2009] as illustrated in Fig. 2. 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 $r_{0\ldots T} = 1$, we denote the un-normalised posterior by $p_\pi(\bar{x}, \bar{u})$:

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

with $Z = P(\bar{r} = 1|x_0)$.

2.3 General Duality

While the Bayesian model has been employed successfully for trajectory planning, see, e.g., [Toussaint, 2009], it's general relation to the classical SOC problem remained unclear. Although a specific subset of SOC problems, studied by [Kappen, 2005] and [Todorov, 2009], can be formulated in a similar Bayesian model, as explicitly done by [Kappen *et al.*, 2009], we establish the formal correspondence between the two formalisms in the general case.

In the following we will distinguish between the unknown control policy π and a prior policy π^0. We derive statements about the KL divergence $\mathrm{KL}\left(q_\pi \| p_{\pi^0}\right)$ – intuitively we think of q_π as the *controlled process* which is *not* conditioned on costs (as defined in (2)), and p_{π^0} as the *posterior process*, which is conditioned on costs but generated via a potentially uninformed policy π^0 (as defined in (3)). The dual problem will be to find a control policy π such that the controlled process q_π matches the posterior process p_{π^0}. The following result establishes the basic relation between such a KL-divergence and SOC:

Proposition 1. *Let π^0 and π be an arbitrary stochastic policies, then the following identities hold*

$$\begin{aligned} \mathrm{KL}(q_\pi \| p_{\pi^0}) =& Z + \eta \mathcal{J}(\pi) + \mathbb{E}_{q_\pi(\bar{x})}\left[\mathrm{KL}\left(\pi \| \pi^0\right)\right] \quad (4) \\ =& Z + \eta \mathcal{J}(\pi) - \mathbb{E}_{q_\pi(\bar{x}, \bar{u})}\left[\log \pi^0(\bar{u}|\bar{x})\right] \\ & - \mathbb{E}_{q_\pi(\bar{x})}\left[H(\pi)\right] , \quad (5) \end{aligned}$$

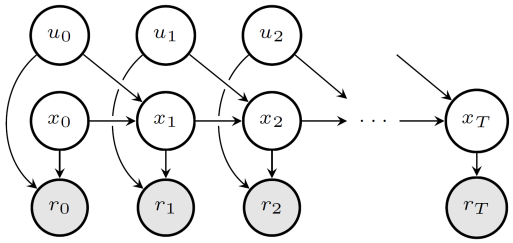

Figure 2: The graphical model of for the Bayesian formulation of the control problem in the finite horizon case. In the infinite horizon case we obtain a stochastic Markov process.

where $Z = \log P(\bar{r} = 1; \pi^0)$.

Proof. n.b. proofs may be found in the full paper [Rawlik *et al.*, 2012]. □

The presented identities are interesting in several respects: Equation (4) tells us that finding an unconstrained policy $\pi^* = \operatorname{argmin}_\pi \mathrm{KL}\left(q_\pi \| p_{\pi^0}\right)$ is a compromise between minimized expected costs $\eta \mathcal{J}(\pi)$ and choosing π similar to the prior policy π^0. In particular, in the limit $\eta \to \infty$ the expected cost term dominates and we retrieve a solution to the SOC problem. Further, when choosing the prior policy π^0 uniform the term $\mathbb{E}_{q_\pi(\bar{x}, \bar{u})}\left[\log \pi^0(\bar{u}|\bar{x})\right]$ in (5) becomes constant and π^* is a compromise between minimized expected costs $\eta \mathcal{J}(\pi)$ and *maximizing* the policy's entropy $\mathbb{E}_{q(\bar{x})}\left[H(\pi)\right]$. This hints at a relation to risk-sensitive control, which we will discuss in more detail in Section 3.2.

The following corollary is a direct consequence of these identities.

Corollary (General duality)**.** *Let π^0 be an arbitrary stochastic policy and $\mathbb{D}$ the set of deterministic policies, then the problem*

$$\pi^* = \operatorname*{argmin}_{\pi \in \mathbb{D}} \mathrm{KL}\left(q_\pi(\bar{x}, \bar{u}) \| p_{\pi^0}(\bar{x}, \bar{u})\right) , \quad (6)$$

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

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

3 Iterative Solutions

Although the above corollary provides the correspondence between the SOC formulation and the computationally attractive inference control approach, due to the constraint $\pi \in \mathbb{D}$, (6) 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:

Proposition 2 (Monotonicity)**.** *For any $\pi \neq \pi^0$,*

$$\mathrm{KL}\left(q_\pi \| p_{\pi^0}\right) \leq \mathrm{KL}\left(q_{\pi^0} \| p_{\pi^0}\right) \Longrightarrow \mathcal{J}(\pi) < \mathcal{J}(\pi^0) .$$

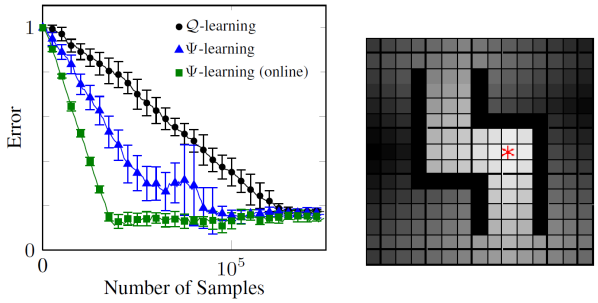

Figure 3: Results from the Gridworld problem. *(left)* Evolution of the mean error in (9) averaged over 10 trials with error bars indicating the s.d. *(right)* Optimal value function (white low expected cost - black high expected cost) of the problem. Obstacles are black and the target state is indicated by *.

Consequently, with some initial π^0, the iteration

$$\pi^{n+1} \leftarrow \operatorname*{argmin}_{\pi} \mathrm{KL}\left(q_\pi \| p_{\pi^n}\right) , \tag{7}$$

where π is an arbitrary[1] conditional distribution over u, gives rise to a chain of stochastic policies with ever decreasing expected costs. Note that our discussion of equation (4) suggest that the convergence rate of such an iteration increases with η, as the expected cost term becomes more dominating.

Note however, that the conditions imposed by the above result, in order to guarantee a policy improvement, are very weak. By exploiting this, in addition to the iteration arising from (7), we present in the following a relaxation, which satisfies Proposition 2 and leads to practical algorithms for infinite horizon problems, and the related iteration of Bayesian inference which leads to risk-sensitive control.

3.1 Ψ-Iterations

We first examine specific instances of iterations of the form (7). Specifically we show that in the finite horizon problem, a closed form solutions to the iterates can be obtained. We subsequently study a class of approximations to (7), which eventually allows us to extend the results of finite horizon case to the the discounted infinite horizon setting. In summary we obtain the following results. In both cases the iterates π^{n+1} take the general form of a Boltzmann like distribution

$$\pi^{n+1}(u_t|x_t) = \exp\{\Psi^{n+1}(x_t, u_t) - \bar{\Psi}^{n+1}(x_t)\} , \tag{8}$$

with energy Ψ and log partition function

$$\bar{\Psi}^{n+1}(\cdot) = \log \int_{u_t} \Psi^{n+1}(\cdot, u_t) ,$$

where the specific update for the two cases take the forms

- **Finite Horizon:**

$$\Psi_t^{n+1}(x_t, u_t) = \log \pi^n(u_t|x_t) - \eta \mathcal{C}_t(x_t, u_t) + \mathbb{E}_{x_{t+1}|x_t,u_t}\left[\bar{\Psi}_{t+1}^{n+1}(x_{t+1})\right]$$

- **Discounted Infinite Horizon:**

$$\Psi^{n+1}(x, u) = \log \pi^n(u|x) - \eta \mathcal{C}_t(x, u) + \gamma \mathbb{E}_{y|x,u}\left[\bar{\Psi}^n(y)\right]$$

where γ is the discount rate. We refer to these methods collectively as Ψ-Iterations and demonstrate that both of the above cases enjoy convergence to the globally optimal policy.

For practical application of the above iterations we propose Monte Carlo based evaluation of the required expectations with in a Reinforcement Learning approach, additionally introducing a suitable basis function expansion of Ψ^{n+1} for problems with large or continuous state and control spaces.

Example

We illustrate the Reinforcement Learning algorithm on a problem used by [Todorov, 2009], 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. 3) 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 [Sutton and Barto, 1998].

Both algorithms were given 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. 3. We plot the approximation error

$$e_{\mathcal{J}} = \frac{\max_x |\mathcal{J}(x) - \hat{\mathcal{J}}(x)|}{\max_x \mathcal{J}(x)} \tag{9}$$

between the true value function $\mathcal{J}$, obtained by value iteration, and it's estimate $\hat{J}$, which can be shown to be 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. 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.

3.2 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 [Bishop, 2006] which could be exploited for practical implementations.

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

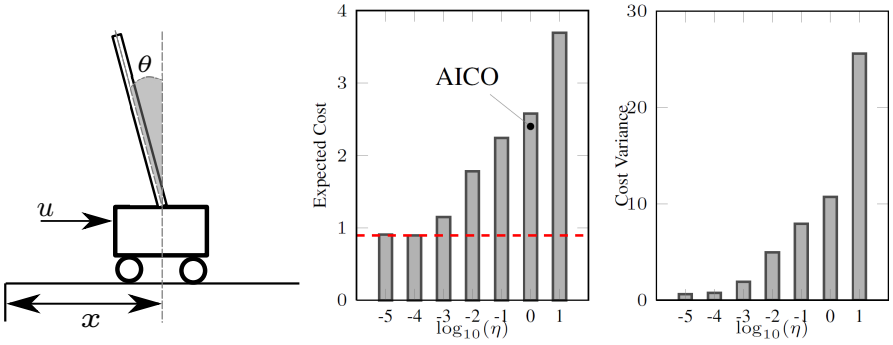

Figure 4: Results for model based approximate posterior policy iteration on the Cart-Pole swing-up task. *(left)* Schematic of the pole on cart plant used in the experiments. *(middle)* Expected cost achieved by policies obtained for different values of the parameter η. Dashed line indicates expected cost of policy obtained using iLQG. All values estimated from 1000 trajectories sampled using the respective policy. *(right)* Variance of the costs achieved by the same policies as for the expected costs of the central plot.

Although unconstrained minimisation of the KL divergence is achieved by the posterior, in our case, the specific form of q_π in (7) is, as can be seen in (2), restricted by the prescribed system dynamics, leading to the Ψ-Iterations presented in the previous section. Nonetheless, we may consider the iteration

$$\pi^{n+1} = p_{\pi^n}(u_t|x_t) \;, \tag{10}$$

which, as we show, will converge to the policy

$$\tilde{\pi} = \operatorname*{argmin}_{\pi} -\frac{1}{\eta} \log \mathbb{E}_{q_\pi} \left[\exp\{-\eta \mathcal{C}_t(\bar{x}, \bar{u})\}\right] \;. \tag{11}$$

The objective being minimized is exactly the risk sensitive objective of [Marcus *et al.*, 1997], which has been recently also used in the path integral approach to SOC [Broek *et al.*, 2010]. 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.

The proposed iteration can be seen as a generalisation of the AICO framework of [Toussaint, 2009] and our results provide the previously lacking formal interpretation of this formulation.

Example

We consider the classical Cart-Pole plant [Sutton and Barto, 1998], illustrated in Fig. 4, and consisting of a inverted pendulum which is mounted on a cart and is controlled by exerting forces on the latter. The task is 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

$$\mathcal{C}_t(x_t, u_t) = \omega_1 \theta^2 + \omega_2 \dot{\theta}^2 + \omega_3 u_t^2 \quad \forall t \in [0, T] \;, \tag{12}$$

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. The required solution to the inference problem arising from PPI was obtained using an extended Kalman filter, leading to an linear policy solution.

In Fig. 4, 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 show the expected cost from the policy obtained using the iLQG algorithm [Li and Todorov, 2006] which also computes an approximately optimal linear policy. We first observe that as predicted, η acts to control the risk seeking behaviour 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.

4 Conclusion

We have present a general relation between stochastic optimal control problems and minimisation of KL divergences of the form (6). This allowed us to derive iterative algorithms for obtaining both risk neutral and risk sensitive optimal controls for finite and infinite horizon MDPs. In the main paper [Rawlik *et al.*, 2012] we show that these algorithms enjoy guaranteed convergence to the global optimum and also propose efficient implementations in the Reinforcement Learning setting for both finite and continuous domains. 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., [Todorov, 2009; Kappen *et al.*, 2009]), or for which the relation to SOC was previously unclear (e.g., [Toussaint, 2009]). Finally we provide experimental validation of our theoretical results.

References

[Azar *et al.*, 2011] M. G. Azar, V. Gomez, and H. J. Kappen. Dynamic policy programming with function approximation. In *Proc. of 14th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2011)*, 2011.

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

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

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

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

[Li and Todorov, 2006] W. Li and E. Todorov. An iterative optimal control and estimation design for nonlinear stochastic system. In *Proc. of 45th IEEE Conference on Decision and Control*, 2006.

[Marcus *et al.*, 1997] S.I. Marcus, E. Fernndez-Gaucherand, D. Hernndez-Hernndez, S. Coraluppi, and P. Fard. Risk sensitive markov decision processes. *Systems and control in the 21st century*, 1997.

[Peters and Schaal, 2008] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(79):1180 – 1190, 2008.

[Peters *et al.*, 2010] Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proc. of. 24th AAAI Conference on Artificial Intelligence*, 2010.

[Rawlik *et al.*, 2012] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Proc. of Robotics: Science and Systems VIII (R:SS)*, 2012.

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

[Todorov, 2009] E. Todorov. Efficient computation of optimal actions. *Proc. of the National Academy of Sciences*, 106:11478–11483, 2009.

[Toussaint, 2009] M. Toussaint. Robot trajectory optimization using approximate inference. In *Proc. of the 26nd Int. Conf. on Machine Learning (ICML 2009)*, 2009.