Proceedings of the Twenty-Third International Conference on Automated Planning and Scheduling

Optimal Control as a Graphical Model Inference Problem

Hilbert J. Kappen and **Vicenç Gómez**
Donders Institute for Brain Cognition and Behaviour
Radboud University Nijmegen
6525 EZ Nijmegen, The Netherlands

Manfred Opper
Department of Computer Science
D-10587 Berlin, TU Berlin, Germany

Abstract

In this paper we show the identification between stochastic optimal control computation and probabilistic inference on a graphical model for certain class of control problems. We refer to these problems as Kullback-Leibler (KL) control problems. We illustrate how KL control can be used to model a multi-agent cooperative game for which optimal control can be approximated using belief propagation when exact inference is unfeasible.

Control as Kullback-Leibler Minimization

Stochastic optimal control theory deals with the problem to compute an optimal set of actions to attain some future goal. With each action and each state a cost is associated and the aim is to minimize the total future cost. Examples are found in many contexts such as motor control for robotics or planning and scheduling tasks. The most common approach to compute the optimal control is through the Bellman equation. However, its direct application is limited for high dimensional or continuous systems due to the large size of the state space.

In (Kappen, Gómez, and Opper 2012), we have shown the equivalence between stochastic optimal control computation and probabilistic inference on a graphical model for a certain class of non-linear stochastic optimal control problems introduced in (Todorov 2007). For this class of problems, the control cost can be written as a KL divergence between p and some interaction terms. The optimal control is expressed as a probability distribution p over future trajectories given the current state.

Let q denote a discrete time Markov process on a discrete state space. We will refer to this process as uncontrolled (or free) dynamics. Such process defines the *allowed* state transitions. Consider the class of control problems for which the cost C of a Markov process p can be written as a sum of two terms: a *control cost* term and a *state dependent* expected cost of future states that are visited within a time horizon T:

$$C = \mathrm{KL}(p\|q) + \langle R\rangle_p, \tag{1}$$

with $\mathrm{KL}(p\|q) = \sum_\tau p(\tau)\log\frac{p(\tau)}{q(\tau)}$ the Kullback-Leibler divergence and $p(\tau)$, $q(\tau)$ distributions over sequences of states (trajectories) indexed by τ according to the Markov processes p and q, respectively.

The optimal cost for a given state x_0 and time $t = 0$ is found by minimizing C with respect to p subject to the normalization constraint $\sum_\tau p(\tau|x_0) = 1$. The result of this KL minimization yields the "Boltzmann distribution"

$$p(\tau|x_0) = \frac{1}{Z(x_0)}q(\tau|x_0)\exp(-\sum_{t=1}^{T}R(x_t)) \tag{2}$$

where the partition function $Z(x_0)$ is a normalization constant. In other words, the optimal control solution is the (normalized) product of the uncontrolled dynamics and the exponentiated state dependent costs. It is a distribution that avoids states of high R, at the same time deviating from q as little as possible. Note that since q is a first order Markov process, p in Equation (2) is a first order Markov process as well.

The optimal cost $C(x_0, p)$ can be expressed in closed form in terms of the known quantities q and R

$$\begin{aligned} C(x_0,p) &= -\log Z(x_0) \\ &= -\log\sum_\tau q(\tau|x_0)\exp(-\sum_{t=1}^{T}R(x_t)) \end{aligned} \tag{3}$$

that is the expectation value of the exponentiated state dependent costs under the uncontrolled dynamics q.

Computation of the optimal control in the current state x_0 and time $t = 0$ is given by the marginal of the probability distribution over future trajectories

$$p(x_1|x_0) = \sum_{x_{2:T}} p(x_{1:T}|x_0). \tag{4}$$

which is a standard probabilistic inference problem, with p given by Equation (2). For tractable instances, it can be solved exactly by backward message passing or using the Junction Tree method (JT) (van den Broek, Wiegerinck, and Kappen 2008). Alternatively, we can apply a number of well-known approximation methods, such as belief propagation (BP) (Murphy, Weiss, and Jordan 1999). We refer to this class of problems as KL control problems.

KL control problems can be viewed as a generalization of continuous space time stochastic optimal control problems (Kappen 2005) that have been successfully applied in robotics (Theodorou, Buchli, and Schaal 2010).

Example: Multi-Agent Cooperative Game

We illustrate the KL control using a variant of the stag hunt game, a prototype game of social conflict between personal risk and mutual benefit (Skyrms 2004). The original game consists of two hunters that can either hunt a hare by themselves giving a small reward, or cooperate to hunt a stag and getting a bigger reward, see table 1. Both stag hunting (*payoff* equilibrium, top-left) and hare hunting (*risk-dominant* equilibrium, bottom-right) are *Nash equilibria*.

	Stag	Hare
Stag	$\mathbf{3}, \mathbf{3}$	$0, 1$
Hare	$1, 0$	$\mathbf{1}, \mathbf{1}$

Table 1: Payoff matrix for the two player stag-hung game.

We define the KL-stag-hunt game as a multi-agent version of the original stag hunt game where M agents live in a 2D grid of N locations and can move to adjacent locations. The grid also contains hares and stags at certain fixed locations. The game is played for a finite time T and at each time-step all the agents can move.

We formulate the problem as a KL control problem. The uncontrolled dynamics factorizes among the agents. It allows an agent to stay on the current position or move to an adjacent position (if possible) with equal probability, thus performing a random walk on the grid. The state dependent cost $R(x_T)/\lambda$ defines the profit when at the end time T, two (or more) agents are at the location of a stag, or individual agents are at a hare location (Kappen, Gómez, and Opper 2012). Figure 1 shows the associated graphical model.

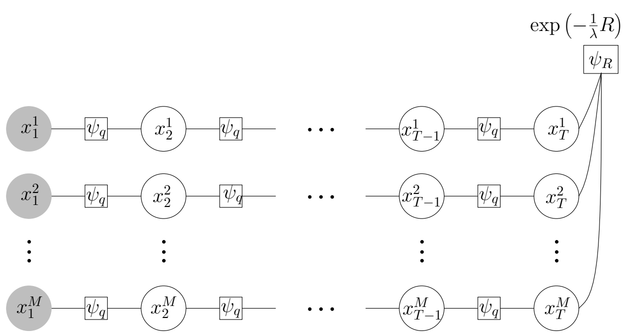

Figure 1: Factor graph representation of the KL-stag-hunt problem. Circles denote variables (states of the agents at a given time-step) and squares denote factors. There are two types of factors: the ones corresponding to the uncontrolled dynamics ψ_q and one corresponding to the state cost ψ_R that couples all agents states. Initial configuration in gray denotes the agents "clamped" to their initial positions.

Computing the exact solution using the JT method becomes unfeasible even for small number of agents, since the joint state space scales as N^M. Belief propagation (BP) algorithm is an alternative approximate algorithm that has polynomial time and space complexity an can be run on an extended factor graph where the factor ψ_R is decomposed.

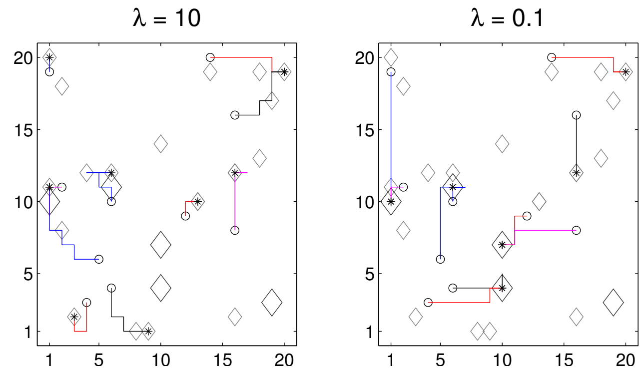

Figure 2: Approximate inference KL-stag-hunt using BP in a large grid for $M = 10$ hunters. Hares and stags are denoted using small and big diamonds respectively. Initial and final positions of the hunters are denoted using small circles and asterisks respectively. $T = 10$ time-steps. The trajectories obtained from BP are drawn in continuous lines. (**Left**) Risk dominant control is obtained for $\lambda = 10$, where all hunters go for a hare. (**Right**) Payoff dominant control is obtained for $\lambda = 0.1$. In this case, all hunters cooperate to capture the stags except the ones on the upper-right corner, who are too far away from the stag to reach it in ten steps.

Figure 2 shows results obtained using BP for $\lambda = 10$ and $\lambda = 0.1$. Trajectories are random realizations calculated from the BP estimated marginals at factor nodes. For high λ, each hunter catches a hare. In this case, the cost function is dominated by the KL term. For sufficiently small λ, the $R(x_T)/\lambda$ term dominates and hunters cooperate and organize in pairs to catch stags. Thus λ can be seen as a parameter that determines whether the optimal control strategy is risk-dominant or payoff-dominant.

This example shows how KL control can be used to model a complex multi-agent cooperative game. The graphical model representation of the problem allows to use approximate inference methods like BP that provide an efficient and good approximation of the control for large systems where exact inference is not feasible.

References

Kappen, H. J.; Gómez, V.; and Opper, M. 2012. Optimal control as a graphical model inference problem. *Mach. Learn.* 87:159–182.

Kappen, H. J. 2005. Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.* 95(20):200201.

Murphy, K.; Weiss, Y.; and Jordan, M. 1999. Loopy belief propagation for approximate inference: An empirical study. In *UAI'99*, 467–47. San Francisco, CA: Morgan Kaufmann.

Skyrms, B., ed. 2004. *The Stag Hunt and Evolution of Social Structure*. Cambridge, MA, USA: Cambridge University Press.

Theodorou, E. A.; Buchli, J.; and Schaal, S. 2010. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *ICRA 2010*, 2397–2403. IEEE Press.

Todorov, E. 2007. Linearly-solvable markov decision problems. In *NIPS 19*. Cambridge, MA: MIT Press. 1369–1376.

van den Broek, B.; Wiegerinck, W.; and Kappen, H. J. 2008. Graphical model inference in optimal control of stochastic multi-agent systems. *J. Artif. Int. Res.* 32(1):95–122.