1228 Probabilistic Graphical Models
H. Zhu and J. Boedecker
May 1, 2026

Control as Probabilistic Inference

Contents

1 Introduction

The framework of *reinforcement learning* or *optimal control* provides variants of methods for solving Markov decision problems. In this chapter, we present the basic graphical model that allows us to embed reinforcement learning problems into the framework of probabilistic graphical models.

In the following discussion, we use $s \in \mathcal{S}$ to denote states and $a \in \mathcal{A}$ to denote actions, which may each be discrete of continuous. States evolve according to the stochastic dynamics $p(s_{t+1} \mid s_t, a_t)$, which are in general unknown. We will follow a discrete-time finite-horizon derivation, which horizon T, and omit discount factor for now. A discount γ can be readily incorporated into this framework simply by modifying the transition dynamics, such that any action produces a transition into an absorbing state with probability $1 - \gamma$, and all standard transition probabilities are multiplied by γ. A task in this framework can be defined by a reward function $r(s_t, a_t)$. Solving a task typically involves recovering a policy

$$\pi_\theta(s_t \mid a_t) = p(a_t \mid s_t, \theta),$$

which specifies a distribution over actions conditioned on the state parameterized by some parameter vector θ. A standard reinforcement learning policy search problem is then given by the following optimization problem:

$$\text{maximize} \quad \sum_{t=1}^{T} \mathbf{E}_{(s_t,a_t)\sim p(s_t,a_t|\theta)}\, r(s_t, a_t), \tag{1.1}$$

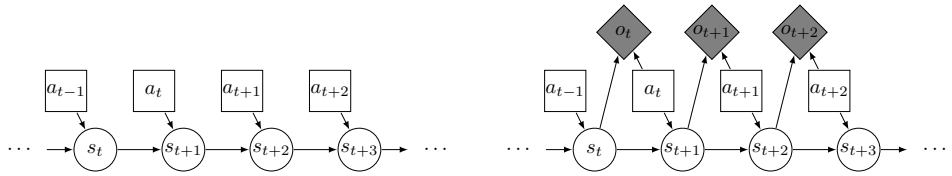

Figure 1 The graphical model for control problems with states and actions (*left*), and with optimality variables (*right*).

with θ being the optimization variable. This problem aims to find a vector of policy parameters θ that maximize the total expected reward $\sum_t r(s_t, a_t)$ of the policy, and the expectation is taken under the policy's trajectory distribution $p(\tau)$, given by

$$p(\tau) = p(s_1, a_1, \ldots, s_T, a_T \mid \theta) = p(s_1) \prod_{t=1}^{T} p(a_t \mid s_t, \theta) p(s_{t+1} \mid s_t, a_t).$$

2 Exact inference

2.1 The graphical model and policy search

To embed the control problem into a graphical model, we can begin simply by modeling the relationship between states, actions, and next states. This relationship is simple, and corresponds to a graphical model with factors of the form $p(s_{t+1} \mid s_t, a_t)$, as shown in figure 1, left. However, this graphical model is insufficient for solving control problems, because it has no notion of rewards or costs. We therefore have to introduce an additional variable into this model, which we will denote $\mathcal{O}$. This additional variable is a binary random variable, where $o_t = 1$ denotes that step t is optimal, and $o_t = 0$ denotes that it is not optimal. We will choose the distribution over this variable to be given by the following equation:

$$p(o_t = 1 \mid s_t, a_t) = \exp(r(s_t, a_t)), \tag{2.1}$$

where $r(s_t, a_t)$ is the reward function and we have assumed without much lose of generality that $r(s_t, a_t) < 0$ for all $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$, so that (2.1) gives a valid probability. The graphical model with these additional variables is summarized in figure 1, right.

In this graphical model, the optimal policy can be written as $p(a_t \mid s_t, o_{1:T} = \mathbf{1})$. (We will write $o_{1:T} = \mathbf{1}$ as $o^*_{1:T}$ in the remainder of the derivation for conciseness.) This distribution is somewhat analogous to $p(a_t \mid s_t, \theta^*)$ with θ^* being the optimal value of θ for (1.1). We can then recover the optimal policy $p(a_t \mid s_t, o^*_{1:T})$ using a standard sum-product inference algorithm, analogously to inference in HMM-style dynamic Bayesian networks. First, note that $o^*_{1:(t-1)}$ is conditionally independent of a_t given s_t, which means that $p(a_t \mid s_t, o^*_{1:T}) = p(a_t \mid s_t, o^*_{t:T})$. Let

$$\beta(s_t, a_t) = p(o^*_{t:T} \mid s_t, a_t)$$

being the backward message of state-action pairs (s_t, a_t), which denotes the probability that a trajectory can be optimal for time steps from t to T if it begins in state s_t with the action a_t. Slightly overloading the notation, we will also introduce the message

$$\beta(s_t) = p(o^*_{t:T} \mid s_t),$$

which denotes the probability that the trajectory from t to T is optimal if it begins in state s_t. We can recover the state-only message from the state-action message by integrating out the action:

$$\beta(s_t) = p(o^*_{t:T} \mid s_t) = \int_{\mathcal{A}} p(o^*_{t:T} \mid s_t, a_t) p(a_t \mid s_t)\ da_t = \int_{\mathcal{A}} \beta(s_t, a_t) p(a_t \mid s_t)\ da_t,$$

where the factor $p(a_t \mid s_t)$ is the *action prior*. Note that it is not conditioned on $o^*_{1:T}$ in any way: it does not denote the probability of an optimal action, but simply the prior probability of actions. The graphical model in figure 1 doesn't actually contain this factor, and we can assume that $p(a_t \mid s_t) = 1/|\mathcal{A}|$ without losing of generality, since any non-uniform priors can be incorporated instead into (2.1) via the reward function. The recursive message passing algorithm for computing $\beta(s_t, a_t)$ proceeds from the last time step $t = T$ backward through time to $t = 1$. In the base case, we note that $\beta(s_T, a_T)$ is simply proportional to $\exp(r(s_T, a_T))$ according to (2.1), since there is only one factor to consider. The recursive case is then given as following:

$$\beta(s_t, a_t) = p(o^*_{t:T} \mid s_t, a_t) = \int_{\mathcal{S}} \beta(s_{t+1}) p(s_{t+1} \mid s_t, a_t) p(o^*_t \mid s_t, a_t)\ ds_{t+1}. \tag{2.2}$$

From these backward messages, we can then derive the optimal policy $p(a_t \mid s_t, o^*_{t:T})$ as:

$$\begin{aligned} p(a_t \mid s_t, o^*_{t:T}) &= \frac{p(s_t, a_t \mid o^*_{t:T})}{p(s_t \mid o^*_{t:T})} = \frac{p(o^*_{t:T} \mid s_t, a_t) p(a_t \mid s_t) p(s_t)}{p(o^*_{t:T} \mid s_t) p(s_t)} \\ &\propto \frac{p(o^*_{t:T} \mid s_t, a_t)}{p(o^*_{t:T} \mid s_t)} = \frac{\beta(s_t, a_t)}{\beta(s_t)}, \end{aligned} \tag{2.3}$$

where the order of conditioning in the third step is flipped by using Bayes' rule, and cancelling the factor of $p(o^*_{t:T})$ that appears in both the numerator and denominator. The term $p(a_t \mid s_t)$ disappears, since we previously assumed it was a uniform distribution.

2.2 Connection to Bellman equations

The intuition about (2.3) can be recovered by considering what these equations are doing in log space. To that end, we will introduce the log-space messages as

$$Q(s_t, a_t) = \log \beta(s_t, a_t),$$

and

$$V(s_t) = \log \beta(s_t).$$

The use of Q and V here is not accidental: the log-space messages correspond to 'soft' variants of the state and state-action value functions. First, consider the marginalization over actions in log-space:

$$V(s_t) = \log \int_{\mathcal{A}} \exp(Q(s_t, a_t))\ da_t.$$

When the values of $Q(s_t, a_t)$ are large, the above equation resembles a hard maximum over a_t. That is, for large $Q(s_t, a_t)$, we have

$$V(s_t) = \log \int_{\mathcal{A}} \exp(Q(s_t, a_t))\ da_t \approx \max_{a_t} Q(s_t, a_t).$$

For smaller values of $Q(s_t, a_t)$, the maximum is soft. Hence, we can refer to V and Q as soft optimal value functions and Q-functions, respectively. We can also consider the backup in (2.2) in log-space. In the case of deterministic dynamics, this backup is given by

$$Q(s_t, a_t) = r(s_t, a_t) + V(s_{t+1}),$$

which exactly corresponds to the Bellman optimality equations. However, when the dynamics are stochastic, the backup is given by

$$\begin{aligned} Q(s_t, a_t) &= r(s_t, a_t) + \log \int_{\mathcal{S}} p(s_{t+1} \mid s_t, a_t) \exp(V(s_{t+1}))\ ds_{t+1} \\ &= r(s_t, a_t) + \log \mathop{\mathbf{E}}_{s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)} \exp(V(s_{t+1})). \end{aligned} \tag{2.4}$$

The backup (2.4) is peculiar, since it does not consider the expected value at the next state, but a 'softmax' over the next expected value. Intuitively, this produces Q-functions that are optimistic: if among the possible outcomes for the next state there is one outcome with a very high value, it will dominate the backup, even when there are other possible states that might be likely and have extremely low value. This creates risk-seeking behavior: if an agent behaves according to this Q-function, it might take actions that have extremely high risk, so long as they have some non-zero probability of a high reward.

3 Approximate inference

3.1 Maximum entropy control

Given the graphical model in figure 1, and recall that we consider the distribution of the optimality variable $\mathcal{O}$ to be given by

$$p(o_t^* \mid s_t, a_t) = \exp(r(s_t, a_t)).$$

We then obtain the posterior distribution over trajectories τ when we condition on $o_t = 1$ for all $t = 1, \ldots, T$, *i.e.*, all actions are optimal:

$$p(\tau \mid o_{1:T}^*) \propto p(\tau, o_{1:T}^*) = p(s_1) \prod_{t=1}^{T} p(o_t^* \mid s_t, a_t) p(s_{t+1} \mid s_t, a_t)$$

$$= p(s_1) \prod_{t=1}^{T} \exp(r(s_t, a_t)) p(s_{t+1} \mid s_t, a_t)$$
$$= \left(p(s_1) \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t) \right) \exp \left(\sum_{t=1}^{T} r(s_t, a_t) \right). \quad (3.1)$$

Suppose we are given some policy π_θ parameterized by θ, the distribution over trajectories τ can be written as

$$p_\theta(\tau) = p(s_1) \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t) \pi_\theta(a_t \mid s_t). \quad (3.2)$$

Obviously, the optimal policy π^* has to result in a $p^*(\tau)$ according to (3.2) that match exactly to the optimal posterior trajectory distribution $p(\tau \mid o^*_{1:T})$ in (3.1). We can therefore view the inference process as minimizing the *KL-divergence* between $p_\theta(\tau)$ and $p(\tau \mid o^*_{1:T})$, which corresponds to the following optimization problem:

$$\text{minimize} \quad D_{\text{kl}}(p_\theta(\tau) \parallel p(\tau \mid o^*_{1:T})),$$

where θ is the optimization variable and the objective is given by:

$$D_{\text{kl}}(p_\theta(\tau) \parallel p(\tau \mid o^*_{1:T})) = - \mathop{\mathbf{E}}_{\tau \sim p_\theta(\tau)} (\log p(\tau \mid o^*_{1:T}) - \log p_\theta(\tau)).$$

Negating both sides and substituting in the equations for $p_\theta(\tau)$ and $p(\tau \mid o^*_{1:T})$, we get

$$\begin{aligned} -D_{\text{kl}}(p_\theta(\tau) \parallel p(\tau \mid o^*_{1:T})) &= \mathop{\mathbf{E}}_{\tau \sim p_\theta(\tau)} \Bigg(\log p(s_1) + \sum_{t=1}^{T} (\log p(s_{t+1} \mid s_t, a_t) + r(s_t, a_t)) \\ &\quad - \log p(s_1) - \sum_{t=1}^{T} (\log p(s_{t+1} \mid s_t, a_t) + \log \pi_\theta(a_t \mid s_t)) \Bigg) \\ &= \mathop{\mathbf{E}}_{\tau \sim p_\theta(\tau)} \left(\sum_{t=1}^{T} r(s_t, a_t) - \log \pi_\theta(a_t \mid s_t) \right) \\ &= \sum_{t=1}^{T} \mathop{\mathbf{E}}_{(s_t, a_t) \sim p_\theta(s_t, a_t)} (r(s_t, a_t) - \log \pi_\theta(a_t \mid s_t)) \\ &= \sum_{t=1}^{T} \mathop{\mathbf{E}}_{(s_t, a_t) \sim p_\theta(s_t, a_t)} r(s_t, a_t) + \sum_{t=1}^{T} \mathop{\mathbf{E}}_{s_t \sim p_\theta(s_t)} \mathcal{H}(\pi_\theta(s_t)), \end{aligned} \quad (3.3)$$

where $\mathcal{H}(\pi_\theta(s_t))$ denotes the entropy of policy π_θ at state s_t. Therefore, minimizing the KL-divergence corresponds to maximizing the expected reward and the expected policy entropy, in contrast to the standard control objective in (1.1), which only maximizes reward. This type of control objective is sometimes referred to as *maximum entropy reinforcement learning* or *maximum entropy control*.

3.2 Connection to variational inference

One way to interpret the objective function (3.3) is as a particular type of structured variational inference. In structured variational inference, our goal is to approximate some distribution $p(x)$ with another, potentially simpler distribution $q(x)$. Typically, $q(x)$ is taken to be some tractable factorized distribution, such as a product of conditional distributions connected in a chain or tree, which leads itself to tractable exact inference. In our case, we aim to approximate $p(\tau \mid o^*_{1:T})$, given by

$$p(\tau \mid o^*_{1:T}) = \left(p(s_1) \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t) \right) \exp\left(\sum_{t=1}^{T} r(s_t, a_t) \right),$$

with the distribution

$$q(\tau) = q(s_1) \prod_{t=1}^{T} q(s_{t+1} \mid s_t, a_t) q(a_t \mid s_t). \tag{3.4}$$

Let $q(s_1) = p(s_1)$ and $q(s_{t+1} \mid s_t, a_t) = p(s_{t+1} \mid s_t, a_t)$, then $q(\tau)$ is exactly the distribution $p_\theta(\tau)$ from (3.2) with $q(a_t \mid s_t) = \pi_\theta(a_t \mid s_t)$. In structured variational inference, approximate inference is performed by optimizing the *variational lower bound* (also called the *evidence lower bound*). Recall that our evidence here is that $o_t = 1$ for all $t = 1, \ldots, T$, thus the variational lower bound is given by

$$\begin{aligned}
\log p(o^*_{1:T}) &= \log \iint p(o^*_{1:T}, s_{1:T}, a_{1:T}) \, ds_{1:T} da_{1:T} \\
&= \log \iint p(o^*_{1:T}, s_{1:T}, a_{1:T}) \frac{q(s_{1:T}, a_{1:T})}{q(s_{1:T}, a_{1:T})} \, ds_{1:T} da_{1:T} \\
&= \log \mathop{\mathbf{E}}_{(s_{1:T}, a_{1:T}) \sim q(s_{1:T}, a_{1:T})} \left(\frac{p(o^*_{1:T}, s_{1:T}, a_{1:T})}{q(s_{1:T}, a_{1:T})} \right) \\
&\geq \mathop{\mathbf{E}}_{(s_{1:T}, a_{1:T}) \sim q(s_{1:T}, a_{1:T})} (\log p(o^*_{1:T}, s_{1:T}, a_{1:T}) - \log q(s_{1:T}, a_{1:T})),
\end{aligned}$$

where the last inequality holds because of Jensen's inequality. Substituting the terms $p(o^*_{1:T}, s_{1:T}, a_{1:T})$ and $q(s_{1:T}, a_{1:T})$ according to (3.1) and (3.4), the bound reduces to

$$\log p(o^*_{1:T}) \geq \mathop{\mathbf{E}}_{(s_{1:T}, a_{1:T}) \sim q(s_{1:T}, a_{1:T})} \left(\sum_{t=1}^{T} r(s_t, a_t) - \log q(a_t \mid s_t) \right)$$

up to an additive constant. Optimizing this objective with respect to the policy $q(a_t \mid s_t)$ corresponds exactly to the objective in (3.3).

3.3 Obtaining the optimal policy

To maximize the maximum entropy control objective

$$-D_{\mathrm{kl}}(p_\theta(\tau) \parallel p(\tau \mid o^*_{1:T})) = \sum_{t=1}^{T} \mathop{\mathbf{E}}_{(s_t, a_t) \sim p_\theta(s_t, a_t)} (r(s_t, a_t) - \log \pi_\theta(a_t \mid s_t))$$

$$= \sum_{t=1}^{T} \mathop{\mathbf{E}}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} r(s_t,a_t) + \sum_{t=1}^{T} \mathop{\mathbf{E}}_{s_t\sim p_\theta(s_t)} \mathcal{H}(\pi_\theta(s_t)),$$

we have to derive the backward messages from an optimization perspective as a dynamic programming algorithm. We will begin with the base case of optimizing $\pi(s_T \mid a_T)$, which consists in maximizing

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(s_T,a_T)\sim p_\theta(s_T,a_T)} (r(s_T,a_T) - \log \pi_\theta(a_T \mid s_T)) \\
&\quad = \mathop{\mathbf{E}}_{(s_T,a_T)\sim p_\theta(s_T,a_T)} \left(\log \frac{\exp(r(s_T,a_T))}{\exp(V(s_T))} - \log \pi_\theta(a_T \mid s_T) + V(s_T) \right) \\
&\quad = \mathop{\mathbf{E}}_{s_T\sim p_\theta(s_T)} \left(-D_{\mathrm{kl}} \left(\pi_\theta(s_T) \,\middle\|\, \frac{1}{\exp(V(s_T))} \exp(r(s_T)) \right) + V(s_T) \right),
\end{aligned}$$

where the last equality holds from the definition of KL-divergence, and $\exp(V(s_T))$ is the normalizing constant for $\exp(r(s_T))$ with respect to a_T, *i.e.*,

$$V(s_T) = \log \int_{\mathcal{A}} \exp(r(s_T,a_T))\ da_T.$$

Since we know that the KL-divergence is minimized when the two arguments represent the same distribution, the optimal policy is given by

$$\pi_\theta(a_T \mid s_T) = \exp(r(s_T,a_T) - V(s_T)).$$

The recursive case can then computed as following: for a given time step t, $\pi_\theta(a_t \mid s_t)$ must maximize two terms:

$$\mathop{\mathbf{E}}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} (r(s_t,a_t) - \log \pi_\theta(a_t \mid s_t)) + \mathop{\mathbf{E}}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} \left(\mathop{\mathbf{E}}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)} V(s_{t+1}) \right)$$

$$= \mathop{\mathbf{E}}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} \left(r(s_t,a_t) + \mathop{\mathbf{E}}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)} V(s_{t+1}) - \log \pi_\theta(a_t \mid s_t) \right)$$

$$= \mathop{\mathbf{E}}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} \left(\log \frac{\exp(r(s_t,a_t) + \mathbf{E}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)} V(s_{t+1}))}{\exp(V(s_t))} - \log \pi_\theta(a_t \mid s_t) + V(s_t) \right)$$

$$= \mathop{\mathbf{E}}_{s_t\sim p_\theta(s_t)} \left(-D_{\mathrm{kl}} \left(\pi_\theta(s_t) \,\middle\|\, \frac{1}{\exp(V(s_t))} \exp(Q(s_t)) \right) + V(s_t) \right), \tag{3.5}$$

where we now define

$$Q(s_t,a_t) = r(s_t,a_t) + \mathop{\mathbf{E}}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)} V(s_{t+1}),$$

and

$$V(s_t) = \log \int_{\mathcal{A}} \exp(Q(s_t,a_t))\ da_t,$$

which corresponds to the standard Bellman optimality equations with a soft maximization for the value function. Choosing

$$\pi_\theta(a_t \mid s_t) = \exp(Q(s_t, a_t) - V(s_t)),$$

we again see that the KL-divergence evaluates to zero, where the objective function (3.5) is maximized. This means that we recover a Bellman backup operator that uses the expected value of the next state, rather than the optimistic estimate we saw in (2.4), which provides a solution to the practical problem of risk-seeking policies.

Bibliography

This chapter is mostly based on [Lev18]. Interested readers can also refer to [Att03, TS06, Tou09, BT12, KGO12, HR17] for more theoretical and empirical discussion about control as inference problems.

References

[Att03] H. Attias. Planning by probabilistic inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 9–16. PMLR, 2003.

[BT12] M. Botvinick and M. Toussaint. Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488, 2012.

[HR17] C. Hoffmann and P. Rostalski. Linear optimal control on factor graphs — A message passing perspective. *IFAC-PapersOnLine*, 50(1):6314–6319, 2017.

[KGO12] H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87:159–182, 2012.

[Lev18] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv*, 1805.00909, 2018.

[Tou09] M. Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1049–1056. Association for Computational Learning, 2009.

[TS06] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 945–952. Association for Computational Learning, 2006.