7. Control as probabilistic inference

- Exact inference
  - The graphical model and policy search
  - Connection to Bellman equations

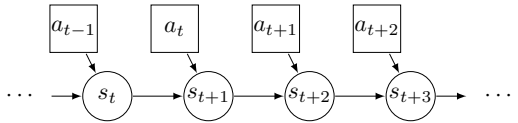
- Approximate inference
  - Maximum entropy control
  - Connection to variational inference
  - Obtaining the optimal policy

Outline

- Exact inference
  - The graphical model and policy search
  - Connection to Bellman equations

- Approximate inference
  - Maximum entropy control
  - Connection to variational inference
  - Obtaining the optimal policy

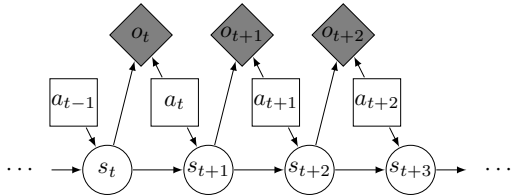
Optimal control problems



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

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

The graphical model



- $\mathcal{O}$: binary random variable, $o_t = 1 \implies$ step t is optimal

$$p(o_t = 1 \mid s_t, a_t) = \exp(r(s_t, a_t))$$

  - assume $r(s_t, a_t) < 0$ for all $s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$

Policy search

target: find the optimal policy $p(a_t \mid s_t, o_{1:T} = \mathbf{1})$

- we will denote $o_{1:T} = \mathbf{1}$ as $o^*_{1:T}$ subsequently for simplicity
- according to the Markov property of the system: $p(a_t \mid s_t, o^*_{1:T}) = p(a_t \mid s_t, o^*_{t:T})$

backward messages

- state-action message

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

- state-only message

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

Policy search

- recover $\beta(s_t)$ from $\beta(s_t, a_t)$:

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

  - $p(a_t \mid s_t)$: action prior, assumed to be uniform, i.e., $p(a_t \mid s_t) = 1/|\mathcal{A}|$

- recursive expression

$$\begin{aligned}\beta(s_t, a_t) &= p(o^*_{t:T} \mid s_t, a_t) \\ &= \begin{cases} \exp(r(s_T, a_T)) & t = T \\ \int_{\mathcal{S}} \beta(s_{t+1}) p(s_{t+1} \mid s_t, a_t) p(o^*_t \mid s_t, a_t)\ ds_{t+1} & t < T \end{cases}\end{aligned}$$

Policy search

optimal policy

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

- $p(a_t \mid s_t)$ disappears since it's assumed to be uniform

Connection to Bellman equations

backward messages in log-space

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

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

- marginalization over actions:

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

  - $V(s_t) \approx \max_{a_t} Q(s_t, a_t)$ for large $Q(s_t, a_t)$

Connection to Bellman equations

backups in log-space

$$\beta(s_t, a_t) = \int_{\mathcal{S}} \beta(s_{t+1}) p(s_{t+1} \mid s_t, a_t) p(o_t^* \mid s_t, a_t) \; ds_{t+1}$$

- deterministic dynamics: soft Bellman optimality equations

$$Q(s_t, a_t) = r(s_t, a_t) + V(s_{t+1}) = r(s_t, a_t) + \log \int_{\mathcal{A}} \exp(Q(s_{t+1}, a_{t+1})) \; da_{t+1}$$

- stochastic dynamics:

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

  - optimistic Q-functions, creating risk-seeking behavior

Outline

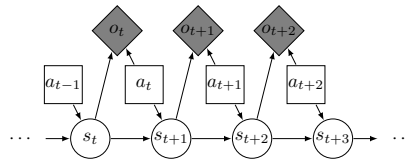
- Exact inference
  - The graphical model and policy search
  - Connection to Bellman equations

- Approximate inference
  - Maximum entropy control
  - Connection to variational inference
  - Obtaining the optimal policy

Maximum entropy control

- posterior distribution over trajectories τ given that all actions are optimal:

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



$$p(o_t = 1 \mid s_t, a_t) = \exp(r(s_t, a_t))$$

- distribution over trajectories τ given some policy π_θ:

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

Maximum entropy control

the inference problem

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

- the optimal policy π^* has to result in a $p^*(\tau)$ that match exactly to the optimal posterior trajectory distribution $p(\tau \mid o^*_{1:T})$
- $D_{\mathrm{kl}}(p_\theta(\tau) \parallel p(\tau \mid o^*_{1:T})) = -\mathbf{E}_{\tau \sim p_\theta(\tau)}(\log p(\tau \mid o^*_{1:T}) - \log p_\theta(\tau))$

Maximum entropy control

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

- $\mathcal{H}(\pi_\theta(s_t))$: the entropy of policy π_θ at state s_t
- minimizing the KL-divergence equals to maximizing the expected reward and the expected policy entropy

Connection to variational inference

variational inference

- approximate some distribution $p(x)$ with another, potentially simpler distribution $q(x)$
- $q(x)$ is taken to be some tractable factorized distribution, which leads itself to tractable exact inference
- approximate inference is performed by optimizing the **variational lower bound** (also called the **evidence lower bound**).

Connection to variational inference

- target distribution

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

- approximate distribution

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

  - $q(s_1) = p(s_1)$
  - $q(s_{t+1} \mid s_t, a_t) = p(s_{t+1} \mid s_t, a_t)$
  - $q(a_t \mid s_t) = \pi_\theta(a_t \mid s_t)$

Connection to variational inference

- variational lower bound given evidence $o_t = 1$ for all $t = 1, \dots, T$:

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

  - the inequality holds because of Jensen's inequality
  - optimizing $\log p(o^*_{1:T})$ equals to optimizing $D_{\mathrm{kl}}(p_\theta(\tau) \parallel p(\tau \mid o^*_{1:T}))$

Obtaining the optimal policy

$$\text{maximize (over } \theta) \quad \sum_{t=1}^{T} \mathbf{E}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} (r(s_t,a_t) - \log \pi_\theta(a_t \mid s_t))$$

dynamic programming

- the base case:

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

  - $V(s_T) = \log \int_{\mathcal{A}} \exp(r(s_T,a_T))\, da_T$: normalizing constant
  - optimal policy: $\pi_\theta(a_T \mid s_T) = \exp(r(s_T,a_T) - V(s_T))$

Obtaining the optimal policy

- the recursive case:

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

  - $Q(s_t, a_t) = r(s_t, a_t) + \mathbf{E}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)} V(s_{t+1})$
  - $V(s_t) = \log \int_{\mathcal{A}} \exp(Q(s_t, a_t))\ da_t$
  - optimal policy: $\pi_\theta(a_t \mid s_t) = \exp(Q(s_t, a_t) - V(s_t))$