CS294-112 Deep Reinforcement Learning HW2: Policy Gradients

due September 19th 2018, 11:59 pm

Problem 1. State-dependent baseline: In lecture we saw that the policy gradient is unbiased if the baseline is a constant with respect to τ (Equation **??**). The purpose of this problem is to help convince ourselves that subtracting a state-dependent baseline from the return keeps the policy gradient unbiased. Using the law of iterated expectations show that the policy gradient is still unbiased if the baseline b is function of a state at a particular timestep of τ (Equation **??**). Please answer the questions below in LaTeXin your report.

(a) Note that by linearity of expectation the objective can be written as:

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[r(\tau) \right] \\ &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{(s_t, a_t) \sim p(s_t, a_t)} \left[r(s_t, a_t) \right] \end{aligned}$$

when we subtract the baseline $b(s_t)$, the objective becomes:

$$= \sum_{t=1}^{T} \mathbb{E}_{(s_t, a_t) \sim p(s_t, a_t)} \left[r(s_t, a_t) - b(s_t) \right].$$

Please show that

$$\nabla_\theta \sum_{t=1}^{T} \mathbb{E}_{(s_t, a_t) \sim p(s_t, a_t)} \left[b(s_t) \right] = 0.$$

(b) Solution to (a):
Assume a_t and s_t are discrete variables. The trajectory follows $p(s_t, a_t)$. The policy is

$pi_\theta(a_t|s_t)$. The state transition probability is $p(s_t|s_{t-1}, a_{t-1})$. Notice that given at time t, s_{t-1} and a_{t-1} are known

$$\begin{aligned}
J(\theta) &= \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\theta(s_t,a_t)} \left[b(s_t)\right] \\
&= \sum_{t=1}^{T} \sum_{s_t} \sum_{a_t} b(s_t)\pi_\theta(a_t|s_t)p(s_t|s_{t-1}, a_{t-1}) \\
&= \sum_{t=1}^{T} \sum_{s_t} b(s_t)p(s_t|s_{t-1}, a_{t-1}) \sum_{a_t} \pi_\theta(a_t|s_t) \\
&= \sum_{t=1}^{T} \sum_{s_t} b(s_t)p(s_t|s_{t-1}, a_{t-1})
\end{aligned}$$

Above is independent of θ, then

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_{t=1}^{T} \sum_{s_t} b(s_t)p(s_t|s_{t-1}, a_{t-1}) = 0$$

(c) An alternative approach is to look at the entire trajectory and consider a particular timestep $t^* \in [1, T-1]$ (the timestep T case would be very similar to part (a)).

   (a) We can exploit the conditional independency structure of $\pi_\theta(\tau) = p(s_1, a_1, ..., s_T, a_T)$ and use the law of iterated expectations to break Equation 1 into two expectations, where the the outer expectation is over $(s_1, a_1, ..., a_{t^*-1}, s_{t^*})$, and the inner expectation is over the rest of the trajectory, conditioned on $(s_1, a_1, ..., a_{t^*-1}, s_{t^*})$. Explain why, for the inner expectation, conditioning on $(s_1, a_1, ..., a_{t^*-1}, s_{t^*})$ is equivalent to conditioning only on s_{t^*}.

   (b) Using the iterated expectation described above, show that

$$\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta(\tau)} \left[b\left(s_{t^*}\right)\right] = 0. \tag{1}$$

(d) Solution to (c):
Denote τ^* as $(s_1, a_1, ..., a_{t^*-1}, s_{t^*})$
Denote τ^c as $(a_{t^*}, s_{t^*+1}, ..., a_T, s_T)$

$$\begin{aligned}
&\mathbb{E}_{\tau\sim p_\theta(\tau)} \left[b\left(s_{t^*}\right)\right] = \mathbb{E}_{\tau} \left[b\left(s_{t^*}\right)\right] \\
&= \sum_{\tau^C} \sum_{\tau^*} b(s_{t^*})p(\tau^*)p(\tau|\tau^*)
\end{aligned}$$

Notice that $p(\tau|\tau^*) = p(\tau|s_{t^*}) = p(\tau^C)$ due to Markov Property of the MDP.

$$= \sum_{\tau^C} \sum_{\tau^*} b(s_{t^*}) p(\tau^*) p(\tau^C)$$

Notice that $p(\tau^*) = p(s_{t^*}|s_{t-1}, a_{t-1}) p(s_{t-1}, a_{t-1}, ..., s_1)$

$$\sum_{\tau^*} b(s_{t^*}) p(\tau^*) =$$