CS294-112 Deep Reinforcement Learning HW2: Policy Gradients

Solution by Benjamin Liu@Berkeley

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.

1. Solution to (a):
   Denote $\nabla_\theta \log \pi(a_t|s_t)b(s_t)$ as $g(a_t, s_t; \theta)$; and $\pi_\theta(a_t|s_t)p(s_t|a_{t-1}, s_{t-1})$ as $q(a_t, s_t|a_{t-1}, s_{t-1})$ where $p(s_t|a_{t-1}, s_{t-1})$ is the transition dynamics. W.L.O.G, let's assume a_t, s_t are discrete variable.
   Using the chain rule, we can express $p_\theta(\tau)$ as a product of the state-action marginal (s_t, a_t) and the probability of the rest of the trajectory conditioned on (s_t, a_t). The derivation for the conditional expectation as follows:

$$
\begin{aligned}
&\mathbb{E}_{p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)] \\
&= \sum_{(a_1,s_1)} ... \sum_{(a_t,s_t)} ... \sum_{(a_T,s_T)} g(a_t, s_t; \theta)q(a_T, s_T|a_{T-1}, s_{T-1})...q(a_t, s_t|a_{t-1}, s_{t-1})...q(a_1, s_1) \\
&= \sum_{(a_1,s_1)} ... \sum_{(a_{t-1},s_{t-1})} \sum_{(a_{t+1},s_{t+1})} ... \sum_{(a_T,s_T)} (\sum_{(a_t,s_t)} g(a_t, s_t; \theta)q(a_t, s_t|a_{t-1}, s_{t-1})) \\
&\quad q(a_T, s_T|a_{T-1}, s_{T-1})...q(a_1, s_1) \\
&= \sum_{(a_1,s_1)} ... \sum_{(a_{t-1},s_{t-1})} \sum_{(a_{t+1},s_{t+1})} ... \sum_{a_T,s_T} q(a_T, s_T|a_{T-1}, s_{T-1})...q(a_1, s_1) \\
&\quad \sum_{(a_t,s_t)} g(a_t, s_t; \theta)q(a_t, s_t|a_{t-1}, s_{t-1})
\end{aligned}
$$

And conditioned on a_t:

$$
\begin{aligned}
&\sum_{(a_t,s_t)} g(a_t, s_t; \theta) q(a_t, s_t | a_{t-1}, s_{t-1}) \\
&= \sum_{s_t} \sum_{a_t} \nabla_\theta \log \pi(a_t|s_t) b(s_t) \pi_\theta(a_t|s_t) p(s_t|a_{t-1}, s_{t-1}) \\
&= \sum_{s_t} \sum_{a_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) \nabla_\theta \log \pi(a_t|s_t) \pi_\theta(a_t|s_t) \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) \sum_{a_t} \nabla_\theta \log \pi(a_t|s_t) \pi_\theta(a_t|s_t) \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) \sum_{a_t} \nabla_\theta \pi_\theta(a_t|s_t) \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) \nabla_\theta \sum_{a_t} \pi_\theta(a_t|s_t) \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) \nabla_\theta 1 \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) 0 = 0
\end{aligned}
$$

Hence, $\mathbb{E}_{p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)] = 0$
Similarly, all the arguments can be applied on cases when s_t, a_t are continuous variables.

2. Solution to (b):

   (a) Due to Markov Property of MDP, the future states only depend on the current state and the past is irrelevant.

   (b) With the same notation in (a), consider expectaion over $\tau^* = (s_1, a_1, ..., s_t, a_t)$, and then conditioned on (a_t, s_t)

$$
\begin{aligned}
&\mathbb{E}_{p_\theta(\tau^*)}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)] \\
&= \sum_{(a_1,s_1)} ... \sum_{(a_t,s_t)} g(a_t, s_t; \theta) q(a_t, s_t|a_{t-1}, s_{t-1}) ... q(a_1, s_1) \\
&= \sum_{(a_1,s_1)} ... \sum_{(a_{t-1},s_{t-1})} \sum_{(a_t,s_t)} g(a_t, s_t; \theta) q(a_t, s_t|a_{t-1}, s_{t-1}) q(a_{t-1}, s_t|a_{t-1}, s_{t-1}) ... q(a_1, s_1) \\
&= \sum_{(a_1,s_1)} ... \sum_{(a_{t-1},s_{t-1})} q(a_{t-1}, s_t|a_{t-1}, s_{t-1}) ... q(a_1, s_1) \\
&\quad \sum_{(a_t,s_t)} g(a_t, s_t; \theta) q(a_t, s_t|a_{t-1}, s_{t-1})
\end{aligned}
$$

And again, conditioned on a_t and the same argument in (a):

$$\begin{aligned}
&\sum_{(a_t,s_t)} g(a_t, s_t; \theta) q(a_t, s_t | a_{t-1}, s_{t-1}) \\
&= \sum_{s_t} \sum_{a_t} \nabla_\theta \log \pi(a_t|s_t) b(s_t) \pi_\theta(a_t|s_t) p(s_t|a_{t-1}, s_{t-1}) \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) \sum_{a_t} \nabla_\theta \pi_\theta(a_t|s_t) \\
&= \sum_{s_t} b(s_t) p(s_t|a_{t-1}, s_{t-1}) 0 = 0
\end{aligned}$$

Similarly, all the arguments can be applied on cases when s_t, a_t are continuous variables.