CS294-112 Deep Reinforcement Learning HW2: Policy Gradients

Ninh DO - SID#25949105

Problem 1. State-dependent baseline:

$$\sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t) \right) \right] = 0. \tag{1}$$

(a) Please show equation 1 by using the law of iterated expectations, breaking $\mathbb{E}_{\tau \sim p_\theta(\tau)}$ by decoupling the state-action marginal from the rest of the trajectory.

Given $p_\theta(\tau) = p_\theta(s_t, a_t) p_\theta(\tau / s_t, a_t | s_t, a_t)$, we write:

$$\sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t) \right) \right]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{p_\theta(s_t, a_t)} \left[\mathbb{E}_{p_\theta(\tau / s_t, a_t | s_t, a_t)} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t) \right) \right] \right]$$

$$= \sum_{t=1}^{T} \int_{s_t} p_\theta(s_t, a_t) \int_{a_t} p_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t) \right) da_t ds_t$$

(since $p_\theta(\tau / s_t, a_t | s_t, a_t) = p_\theta(a_t|s_t)$)

$$= \sum_{t=1}^{T} \int_{s_t} p_\theta(s_t, a_t) \int_{a_t} \nabla_\theta \pi_\theta(a_t|s_t) \left(b(s_t) \right) da_t ds_t$$

(since $p_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) = \nabla_\theta \pi_\theta(a_t|s_t)$, p_θ and π_θ are the same)

$$= \sum_{t=1}^{T} \int_{s_t} p_\theta(s_t, a_t) b(s_t) \nabla_\theta \int_{a_t} \pi_\theta(a_t|s_t) da_t ds_t$$

$$= \sum_{t=1}^{T} \int_{s_t} p_\theta(s_t, a_t) b(s_t) \nabla_\theta 1 da_t ds_t = 0$$

(since $\int_{a_t} \pi_\theta(a_t|s_t) da_t = 1, \nabla_\theta 1 = 0$)

(b) Alternatively, we can consider the structure of the MDP and express $p_\theta(\tau)$ as a product of the trajectory distribution up to s_t (which we denote as $(s_{1:t}, a_{1:t-1})$) and the trajectory distribution after s_t conditioned on the first part (which we denote as $(s_{t+1:T}, a_{t:T}|s_{1:t}, a_{1:t-1})$):

(a) 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^*}.

Since the Markov chain is memoryless, the current state/action only depends on its most recent action/state.

(b) Please show equation 1 by using the law of iterated expectations, breaking $\mathbb{E}_{\tau\sim p_\theta(\tau)}$ by decoupling trajectory up to s_t from the trajectory after s_t.

Given

$$\begin{aligned} p_\theta(\tau) &= p_\theta(s_{1:t}, a_{1:t-1})p_\theta(s_{t+1:T}, a_{t:T}|s_{1:t}, a_{1:t-1}) \\ &= p_\theta(s_{1:t}, a_{1:t-1})p_\theta(s_{t+1:T}, a_{t:T}|s_t) \end{aligned}$$

We write:

$$\sum_{t=1}^{T} \mathbb{E}_{\tau\sim p_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t)\right)\right]$$

$$= \mathbb{E}_{p_\theta(s_{1:t^*}, a_{1:t^*-1})} \left[\mathbb{E}_{p_\theta(s_{t^*+1:T}, a_{t^*:T}|s_{t^*})} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t)\right)\right]\right]$$

$$= \mathbb{E}_{p_\theta(s_{1:t^*}, a_{1:t^*-1})} \left[\mathbb{E}_{p_\theta(s_{t^*+1:T}, a_{t^*:T}|s_{t^*})} \left[\sum_{t=t^*}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \left(b(s_t)\right)\right]\right]$$

(truncating the head of sum because probabilily distribution is $p_\theta(s_{t^*+1:T}, a_{t^*:T}|s_{t^*})$)

$$= \mathbb{E}_{p_\theta(s_{1:t^*}, a_{1:t^*-1})} \left[\mathbb{E}_{p_\theta(s_{t^*+1:T}, a_{t^*:T}|s_{t^*})} \left[\nabla_\theta \log \prod_{t=t^*}^{T} \pi_\theta(a_t|s_t) \left(b(s_t)\right)\right]\right]$$

$$= \mathbb{E}_{p_\theta(s_{1:t^*}, a_{1:t^*-1})} \left[\mathbb{E}_{p_\theta(s_{t^*+1:T}, a_{t^*:T}|s_{t^*})} \left[\nabla_\theta \log \pi_\theta(s_{t^*+1:T}, a_{t^*:T}|s_t) \left(b(s_{t^*})\right)\right]\right]$$

$$= \int_{s_t} p_\theta(s_{1:t^*}, a_{1:t^*-1}) \int_{a_t} p_\theta(s_{t^*+1:T}, a_{t^*:T}|s_{t^*}) \nabla_\theta \log \pi_\theta(s_{t^*+1:T}, a_{t^*:T}|s_t) \left(b(s_t)\right) da_t ds_t$$

$$= \int_{s_t} p_\theta(s_{1:t^*}, a_{1:t^*-1}) \int_{a_t} \nabla_\theta \pi_\theta(s_{t^*+1:T}, a_{t^*:T}|s_t) \left(b(s_{t^*})\right) da_t ds_t$$

$$= \int_{s_t} p_\theta(s_{1:t^*}, a_{1:t^*-1}) b(s_{t^*}) \nabla_\theta \int_{a_t} \pi_\theta(s_{t^*+1:T}, a_{t^*:T}|s_t) da_t ds_t$$

$$= \int_{s_t} p_\theta(s_{1:t^*}, a_{1:t^*-1}) b(s_{t^*}) \nabla_\theta 1 da_t ds_t = 0$$

(since $\int_{a_t} \pi_\theta(s_{t^*+1:T}, a_{t^*:T}|s_t) da_t = 1, \nabla_\theta 1 = 0$)