Homework 3

6501 Reinforcement Learning (Spring 2025)

Submission deadline: 11:59pm, March 27

Latex template can be accessed here.

1 Gradient Estimators in Continuous Action Spaces

In this problem, we consider the following algorithmic framework (Algorithm 1) for continuous action sets. For simplicity, we assume the action set is the entire $\mathbb{R}^d$ (unconstrained).

Algorithm 1 Policy update framework for continuous action sets

---

Parameter: σ.
Initialize a neural network $\mu_\theta : \mathcal{X} \to \mathbb{R}^d$, where $\mathcal{X}$ is the space of contexts, and d is the dimension of the action set.
Let θ_1 be the initial weights.
for $t = 1, 2, \dots, T$ **do**
  Receive context x_t.
  Sample $a_t \sim \mathcal{N}(\mu_{\theta_t}(x_t), \sigma^2 I)$.
  Receive $r_t(x_t, a_t)$.
  Obtain θ_{t+1} from θ_t and the reward feedback (there could be different ways to perform this update).

Let $b_t : \mathcal{X} \to \mathbb{R}$ be an arbitrary time-varying baseline function, and let g_t be the one-point gradient estimator constructed as the following:

$$g_t = \frac{1}{\sigma^2}(a_t - \mu_{\theta_t}(x_t))(r_t(x_t, a_t) - b_t(x_t)).$$

Below, we use $\nabla_{\boldsymbol{a}} r_t$ to denote the gradient of r_t with respect its second argument (i.e., action). That is, for any (x_0, a_0), $\nabla_{\boldsymbol{a}} r_t(x_0, a_0) = \nabla_a r_t(x_0, a)|_{a=a_0}$.

(a) (5%) Assume that $r_t(x_t, \cdot)$ is an affine function under any context x_t. In other words, there exist $v_t(x_t) \in \mathbb{R}^d$ and $c_t(x_t) \in \mathbb{R}$ such that

$$\forall a, \qquad r_t(x_t, a) = c_t(x_t) + v_t(x_t)^\top a.$$

Prove that g_t is an unbiased gradient estimator, i.e., $\mathbb{E}_{a_t}[g_t] = v_t(x_t)$, where $\mathbb{E}_{a_t}[\cdot]$ denotes the expectation over the randomness of a_t.
Hint: We did this proof in Page 17 of this slide under a slightly different setting and notation. You only need to repeat that proof with slight adaptation.

(b) (5%) Assume that $r_t(x_t, \cdot)$ is an L-smooth function under any context x_t. Prove that the bias of g_t satisfies

$$|\mathbb{E}_{a_t}[g_t] - \nabla_{\boldsymbol{a}} r_t(x_t, \mu_{\theta_t}(x_t))| \le L\sigma^2.$$

Hint: A function $f : \mathbb{R}^d \to \mathbb{R}$ is called L-smooth if for any a, b, $\|\nabla f(a) - \nabla f(b)\| \le L\|a - b\|$. This means that the gradient changes slowly, and thus we can locally approximate a smooth function by an affine function. Indeed, using Lemma 1, we are able to bound

$$\left| r_t(x_t, a) - \underbrace{\left[r_t(x_t, \mu_{\theta_t}(x_t)) + \nabla_{\boldsymbol{a}} r_t(x_t, \mu_{\theta_t}(x_t))^\top (a - \mu_{\theta_t}(x_t)) \right]}_{\text{Taylor expansion up to the first-order term}} \right| \le \frac{L}{2}\|a - \mu_{\theta_t}(x_t)\|^2.$$

Therefore, you only need to repeat similar proof as in (a), but considering the error resulted from approximating $r_t(x_t, \cdot)$ by an affine function.

The following two questions do not rely on the results of (a) and (b), so you can work on them without first working out (a) and (b). Define policy π_θ as

$$\pi_\theta(a|x) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \exp\left(-\frac{\|a - \mu_\theta(x)\|^2}{2\sigma^2} \right).$$

This is essentially the policy being executed in Algorithm 1.

(c) (5%) Show that the unclipped and unbatched PPO update

$$\theta_{t+1} \leftarrow \operatorname*{argmax}_\theta \left\{ \frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_t}(a_t|x_t)}(r_t(x_t, a_t) - b_t(x_t)) - \frac{1}{\eta}\mathrm{KL}\left(\pi_\theta(\cdot|x_t), \pi_{\theta_t}(\cdot|x_t)\right) \right\}$$

is approximately equivalent to

$$\theta_{t+1} \leftarrow \operatorname*{argmax}_\theta \left\{ \langle \mu_\theta(x_t) - \mu_{\theta_t}(x_t), g_t \rangle - \frac{1}{2\eta\sigma^2}\|\mu_\theta(x_t) - \mu_{\theta_t}(x_t)\|^2 \right\}$$

when η is close to zero (thus $\theta_{t+1} \approx \theta_t$).
Hint: It suffices to show that the expressions in the two $\operatorname{argmax}\{\cdot\}$'s are approximately equal or off by a constant unrelated to θ. The approximation you will need is $\exp(u) \approx 1 + u$ for $u \in \mathbb{R}$ close to zero.

(d) (5%) Show that the PG update

$$\theta_{t+1} \leftarrow \theta_t + \eta \nabla_\theta \log \pi_\theta(a_t|x_t)\Big|_{\theta=\theta_t}(r_t(x_t, a_t) - b_t(x_t))$$

is approximately equivalent to

$$\theta_{t+1} \leftarrow \operatorname*{argmax}_\theta \left\{ \langle \mu_\theta(x_t) - \mu_{\theta_t}(x_t), g_t \rangle - \frac{1}{2\eta}\|\theta - \theta_t\|^2 \right\}$$

when η is close to zero (thus $\theta_{t+1} \approx \theta_t$).
Hint: The approximation you will need is $f_{\theta'}(x) - f_\theta(x) \approx (\theta' - \theta)^\top \nabla_\theta f_\theta(x)$ for $\theta' \approx \theta$ and for function $f_\theta : \mathcal{X} \to \mathbb{R}$ that is smooth in θ.

(c) and (d) verify again that PPO and PG differ in the distance measure they use to regularize the policy updates.

A Appendix

Lemma 1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then for any a, b,*

$$\left| f(a) - \left[f(b) + \nabla f(b)^\top (a - b) \right] \right| \le \frac{L}{2} \|a - b\|^2.$$

Proof. By Taylor's theorem, there exists a' that lies in the line segment between a and b such that

$$f(a) - f(b) = \nabla f(b)^\top (a - b) + \frac{1}{2}(a - b)^\top \nabla^2 f(a')(a - b)$$

The smoothness assumption implies that $\left|(a - b)^\top \nabla^2 f(a')(a - b)\right| \le L\|a - b\|^2$ and thus the desired inequality. $\square$