Homework 3

6501 Reinforcement Learning (Spring 2025)

Submission deadline: 11:59pm, March 30

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 $t = 1, 2, \ldots, T$ **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. The hand writing there does not include per-step explanations (because they were given orally in the class), but make sure you explain every step when writing your proof.

Proof. For simplicity, define $z_t = a_t - \mu_{\theta_t}(x_t)$. By the algorithm, we know that z_t is drawn from $\mathcal{N}(0, \sigma^2 I)$.

$$\mathbb{E}_{z_t}[g_t] = \mathbb{E}_{z_t}\left[\frac{r_t(x_t, a_t) - b_t(x_t)}{\sigma^2} z_t\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{z_t}\left[\frac{v_t(x_t)^\top a_t + c_t(x_t) - b_t(x_t)}{\sigma^2} z_t\right] && \text{(by the assumption on } r_t\text{)} \\
&= \mathbb{E}_{z_t}\left[\frac{v_t(x_t)^\top z_t + v_t(x_t)^\top \mu_{\theta_t}(x_t) + c_t(x_t) - b_t(x_t)}{\sigma^2} z_t\right] && (a_t = \mu_{\theta_t}(x_t) + z_t) \\
&= \mathbb{E}_{z_t}\left[\frac{v_t(x_t)^\top z_t}{\sigma^2} z_t\right] && (z_t \text{ is zero-mean conditioned on everything else}) \\
&= \frac{\mathbb{E}_{z_t}[z_t z_t^\top]}{\sigma^2} v_t(x_t) \\
&= v_t(x_t). && (\mathbb{E}_{z_t}[z_t z_t^\top] = \sigma^2 I)
\end{aligned}
$$

□

(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

$$
\|\mathbb{E}_{a_t}[g_t] - \nabla_{\boldsymbol{a}} r_t(x_t, \mu_{\theta_t}(x_t))\| \le \sqrt{\frac{d(d+2)(d+4)}{4}} L\sigma.
$$

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 can repeat similar proof as in (a), but considering the error resulted from approximating $r_t(x_t, \cdot)$ by an affine function. You may want to use Lemma 2 in the appendix.

Proof. Define $v_t(x_t) = \nabla_{\boldsymbol{a}} r_t(x_t, \mu_{\theta_t}(x_t))$ and $c_t(x_t) = r_t(x_t, \mu_{\theta_t}(x_t)) - \nabla_{\boldsymbol{a}} r_t(x_t, \mu_{\theta_t}(x_t))^\top \mu_{\theta_t}(x_t)$. Then the inequality in the hint can be written as

$$
\left| r_t(x_t, a) - \left[c_t(x_t) + v_t(x_t)^\top a\right] \right| \le \frac{L}{2}\|a - \mu_{\theta_t}(x_t)\|^2. \tag{1}
$$

Below, we follow similar calculation as in (a) but incorporate the error term. For simplicity, denote the approximation error as

$$
e_t(x_t, a) \triangleq r_t(x_t, a) - \left[c_t(x_t) + v_t(x_t)^\top a\right]. \tag{2}
$$

Then we have

$$
\begin{aligned}
&\|\mathbb{E}_{z_t}[g_t] - v_t(x_t)\|^2 \\
&= \left\| \mathbb{E}_{z_t}\left[\frac{r_t(x_t, a_t) - b_t(x_t)}{\sigma^2} z_t\right] - v_t(x_t) \right\|^2 \\
&= \left\| \mathbb{E}_{z_t}\left[\frac{v_t(x_t)^\top a_t + c_t(x_t) - b_t(x_t)}{\sigma^2} z_t\right] + \mathbb{E}_{z_t}\left[\frac{\color{red}{e_t(x_t, a_t)}}{\sigma^2} z_t\right] - v_t(x_t) \right\|^2 && \text{(by the definition in (2))} \\
&= \left\| \mathbb{E}_{z_t}\left[\frac{e_t(x_t, a_t)}{\sigma^2} z_t\right] \right\|^2 && \text{(by the same calculation as in (a))} \\
&\le \mathbb{E}_{z_t}\left[\left\| \frac{e_t(x_t, a_t)}{\sigma^2} z_t \right\|^2\right] && \text{(Jensen's inequality)} \\
&= \mathbb{E}_{z_t}\left[\frac{e_t(x_t, a_t)^2}{\sigma^4} \|z_t\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}_{z_t}\left[\frac{1}{\sigma^4}\left(\frac{L}{2}\|z_t\|^2\right)^2\|z_t\|^2\right] && \text{(by (1) we have } |e_t(x_t,a_t)| \leq \tfrac{L}{2}\|z_t\|^2\text{)}\\
&= \frac{L^2}{4\sigma^4}\mathbb{E}_{z_t}[\|z_t\|^6]\\
&= \frac{d(d+2)(d+4)}{4}L^2\sigma^2. && \text{(by Lemma 2)}
\end{aligned}
$$

□

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 may want to use is $e^u \approx 1+u$ for $u\approx 0$.

Proof. By the definition of π_θ, we have

$$
\begin{aligned}
\frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_t}(a_t|x_t)} &= \exp\left(-\frac{\|a_t-\mu_\theta(x_t)\|^2}{2\sigma^2}+\frac{\|a_t-\mu_{\theta_t}(x_t)\|^2}{2\sigma^2}\right)\\
&\approx 1-\frac{\|a_t-\mu_\theta(x_t)\|^2}{2\sigma^2}+\frac{\|a_t-\mu_{\theta_t}(x_t)\|^2}{2\sigma^2} && \text{(using the approximation rule in the hint)}\\
&= 1+\frac{\langle 2a_t-\mu_\theta(x_t)-\mu_{\theta_t}(x_t),\mu_\theta(x_t)-\mu_{\theta_t}(x_t)\rangle}{2\sigma^2}\\
&\approx 1+\frac{\langle a_t-\mu_{\theta_t}(x_t),\mu_\theta(x_t)-\mu_{\theta_t}(x_t)\rangle}{\sigma^2}. && \text{(using } \theta\approx\theta_t\text{)}
\end{aligned}
$$

Below, we use the notation "$u \equiv_\theta v$" to indicate that $u-v$ is a function unrelated to θ. With the approximation above, we have

$$
\begin{aligned}
&\frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_t}(a_t|x_t)}(r_t(x_t,a_t)-b_t(x_t))\\
&\approx\left(1+\frac{\langle a_t-\mu_{\theta_t}(x_t),\mu_\theta(x_t)-\mu_{\theta_t}(x_t)\rangle}{\sigma^2}\right)(r_t(x_t,a_t)-b_t(x_t))\\
&\equiv_\theta \frac{\langle a_t-\mu_{\theta_t}(x_t),\mu_\theta(x_t)-\mu_{\theta_t}(x_t)\rangle}{\sigma^2}(r_t(x_t,a_t)-b_t(x_t))\\
&= \langle\mu_\theta(x_t)-\mu_{\theta_t}(x_t),g_t\rangle.
\end{aligned}
$$

On the other hand,

$$
\mathrm{KL}\left(\pi_\theta(\cdot|x_t),\pi_{\theta_t}(\cdot|x_t)\right) = \frac{1}{2\sigma^2}\|\mu_\theta(x_t)-\mu_{\theta_t}(x_t)\|^2
$$

because they are two multivariate Gaussians with the same covariance matrix. Combining everything above proves the approximate equivalence. □

(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.

Proof. Since for any v and θ_t, the maximizer of $\langle \theta - \theta_t, v \rangle - \frac{1}{2\eta}\|\theta - \theta_t\|^2$ is $\theta = \theta_t + \eta v$, the PG update is equivalent to

$$\theta_{t+1} = \operatorname*{argmax}_\theta \left\{ \left\langle \theta - \theta_t, \Big[\nabla_\theta \log \pi_\theta(a_t|x_t)\Big]_{\theta=\theta_t} \right\rangle (r_t(x_t, a_t) - b_t(x_t)) - \frac{1}{2\eta}\|\theta - \theta_t\|^2 \right\}. \tag{3}$$

By the definition of π_θ, we have

$$\begin{aligned}
\Big[\nabla_\theta \log \pi_\theta(a_t|x_t)\Big]_{\theta=\theta_t} &= \left[\nabla_\theta \left(-\frac{\|a_t - \mu_\theta(x_t)\|^2}{2\sigma^2} - \frac{d}{2}\log(2\pi\sigma^2) \right)\right]_{\theta=\theta_t} \\
&= \left[(\nabla_\theta \mu_\theta(x_t)) \frac{a_t - \mu_\theta(x_t)}{\sigma^2}\right]_{\theta=\theta_t} && \text{(chain rule)} \\
&= \Big[\nabla_\theta \mu_\theta(x_t)\Big]_{\theta=\theta_t} \frac{a_t - \mu_{\theta_t}(x_t)}{\sigma^2}.
\end{aligned}$$

Notice that $\big[\nabla_\theta \mu_\theta(x_t)\big]$ is a $d_\theta \times d$ Jacobian matrix where d_θ is the dimension of θ and d is the dimension of the actions. Therefore, the objective in (3) can be written as

$$\begin{aligned}
&\frac{r_t(x_t, a_t) - b_t(x_t)}{\sigma^2} (\theta - \theta_t)^\top \big[\nabla_\theta \mu_\theta(x_t)\big]_{\theta=\theta_t} (a_t - \mu_{\theta_t}(x_t)) - \frac{1}{2\eta}\|\theta - \theta_t\|^2 \\
&= \frac{r_t(x_t, a_t) - b_t(x_t)}{\sigma^2} (\theta - \theta_t)^\top \Big[\nabla_\theta \big\langle \mu_\theta(x_t), a_t - \mu_{\theta_t}(x_t) \big\rangle\Big]_{\theta=\theta_t} - \frac{1}{2\eta}\|\theta - \theta_t\|^2 \\
&\approx \frac{r_t(x_t, a_t) - b_t(x_t)}{\sigma^2} \big\langle \mu_\theta(x_t) - \mu_{\theta_t}(x_t), a_t - \mu_{\theta_t}(x_t) \big\rangle - \frac{1}{2\eta}\|\theta - \theta_t\|^2 \\
&\qquad\qquad \text{(using the approximation given in the hint)} \\
&= \big\langle \mu_\theta(x_t) - \mu_{\theta_t}(x_t), g_t \big\rangle - \frac{1}{2\eta}\|\theta - \theta_t\|^2.
\end{aligned}$$

This shows the approximate equivalence. □

(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,*

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

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 $\left|(a-b)^\top \nabla^2 f(a')(a-b)\right| \leq L\|a-b\|^2$ and thus the desired inequality. □

Lemma 2. *Let $X \in \mathbb{R}^d$ be a multivariate Gaussian following $X \sim \mathcal{N}(0, I_d)$. Then*

$$\mathbb{E}\left[\|X\|^6\right] = d(d+2)(d+4).$$

Proof. Since $X \in \mathbb{R}^d$ follows the standard Gaussian, $\|X\|^2$ follows the chi-square distribution with degree d [1]. Then by [2], $\mathbb{E}[(\|X\|^2)^3]$ can be calculated as $\prod_{k=0}^{2}(d+2k) = d(d+2)(d+4)$. □

References

[1] ProofWiki. Definition:Chi-Squared Distribution. `https://proofwiki.org/wiki/Definition:Chi-Squared_Distribution`.

[2] ProofWiki. Raw Moment of Chi-Squared Distribution. `https://proofwiki.org/wiki/Raw_Moment_of_Chi-Squared_Distribution`.