STAT 184, PROBLEM SET 3

Matthew Qu
Due: November 10, 2022
matthewqu@college.harvard.edu

0. (**Collaborators and Acknowledgements**)
Collaborators: Kevin Huang, Katherine Zhou
Acknowledgements:

1. (**Linear Quadratic Regulator**)

Proof. I claim that $V_h^*(s) = s^\top P_h s + p_h^\top s + b_h$ is the form for the value function V_h^* for all $h = 0, \dots, H$, where $P_h \in \mathbb{R}^{d\times d}$, $p_h \in \mathbb{R}^d$, and $b_h \in \mathbb{R}$. First, note that trivially, $V_H^*(s) = 0$, which is in the form above for $P_H = \mathbf{0}_{d\times d}$, $p_H = \mathbf{0} \in \mathbb{R}^d$, and $b_H = 0$. Now, assume in our inductive hypothesis that $V_{h+1}^*(s) = s^\top P_{h+1} s + p_{h+1}^\top s + b_{h+1}$. Then, we have

$$\begin{aligned}
Q_h^*(s,a) &= c(s,a) + \mathbb{E}[V_{h+1}^*(s')] \\
&= s^\top Q s + a^\top R a + s^\top M a + q^\top s + r^\top a + b + \mathbb{E}[V_{h+1}^*(As + Ba + m)] \\
&= s^\top (Q + A^\top P_{h+1} A) s + a^\top (R + B^\top P_{h+1} B) a + s^\top (M + 2A^\top P_{h+1} B) a \\
&+ (q^\top + 2m^\top P_{h+1} A + p_{h+1}^\top A) s + (r^\top + 2m^\top P_{h+1} B + p_{h+1}^\top B) a \\
&+ (b + b_{h+1} + m^\top P_{h+1} m + m^\top p_{h+1}),
\end{aligned}$$

after collecting terms and noting that

$$\mathbb{E}[V_{h+1}^*(As + Ba + m)] = (As + Ba + m)^\top P_{h+1} (As + Ba + m) + (As + Ba + m)^\top p_{h+1} + b_{h+1}$$

by the inductive hypothesis. Now, we take the gradient of this expression with respect to a to get

$$\nabla_a Q_h^*(s,a) = 2(R + B^\top P_{h+1} B) a + (M^\top + 2B^\top P_{t+1} A) s + (r + 2B^\top P_{h+1} m + B^\top p_{h+1}).$$

Setting this equal to 0 and solving for a yields the optimal policy at time h as a linear function of s. We have $\pi_h^*(s) = -K_h s - k_h$, where

$$\begin{aligned}
K_h &= \frac{1}{2}(R + B^\top P_{h+1} B)^{-1} (M^\top + 2B^\top P_{t+1} A) \\
k_h &= \frac{1}{2}(R + B^\top P_{h+1} B)^{-1} (r + 2B^\top P_{h+1} m + B^\top p_{h+1}).
\end{aligned}$$

Now, we know that $V_h^*(s) = Q_h^*(s, \pi_h^*(s))$, so we plug in $a = -K_h s - k_h$ into the above expression. After collecting terms, we are left with

$$V_h^*(s) = s^\top (U + K_h^\top V K_h + W K_h) s + (2k_h^\top V K_h - k_h^\top W^\top + X - Y K_h) s + (k_h^\top V k_h - Y k_h + Z),$$

where

$$\begin{aligned}
U &:= Q + A^\top P_{h+1} A \\
V &:= R + B^\top P_{h+1} B \\
W &:= M + 2A^\top P_{h+1} B \\
X &:= q^\top + 2m^\top P_{h+1} A + p_{h+1}^\top A \\
Y &:= r^\top + 2m^\top P_{h+1} B + p_{h+1}^\top B \\
Z &:= b + b_{h+1} + m^\top P_{h+1} m + m^\top p_{h+1}.
\end{aligned}$$

We can see that $V_h^*(s)$ is indeed in a quadratic form, which completes the inductive step. □

2. (**PG: Alternative Expressions and Baselines**)

Proof.

1. We use Adam's Law, conditioning on the states and actions $s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t$. Note that $\sum_{k=0}^{t-1} r_k$ is a function of the above. Therefore, we have

$$\mathbb{E}_{\tau\sim\rho_\theta}\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{k=0}^{t-1} r_k\right) = \mathbb{E}\left(\sum_{k=0}^{t-1} r_k \cdot \mathbb{E}_{\tau\sim\rho_\theta}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \mid s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t\right]\right)$$

Consider the inner expectation. We see that

$$\mathbb{E}_{\tau\sim\rho_\theta}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \mid s_0, a_0, \ldots, s_t\right] = \sum_\tau \rho_\theta(\tau \mid s_0, a_0, \ldots, s_t)\nabla_\theta \log \pi_\theta(a_t|s_t).$$

But note that for each τ, the conditional probability $\rho_\theta(\tau \mid s_0, a_0, \ldots, s_t)$ is equal to the probability of taking the respective action a_t given state s_t according to the policy π_θ. Therefore, we have

$$\begin{aligned}
\sum_\tau \rho_\theta(\tau \mid s_0, a_0, \ldots, s_t)\nabla_\theta \log \pi_\theta(a_t|s_t) &= \sum_{a_t\in\mathcal{A}} \pi_\theta(a_t|s_t)\nabla_\theta \log \pi_\theta(a_t|s_t). \\
&= \sum_{a_t\in\mathcal{A}} \pi_\theta(a_t|s_t)\frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \\
&= \nabla_\theta \sum_{a_t\in\mathcal{A}} \pi_\theta(a_t|s_t) \\
&= \nabla_\theta 1 \\
&= 0.
\end{aligned}$$

Since the inner expectation evaluates to 0, we conclude that

$$\mathbb{E}_{\tau\sim\rho_\theta}\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{k=0}^{t-1} r_k\right) = 0.$$

2. We have

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\tau\sim\rho_\theta}\left(\sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot R(\tau)\right) \\
&= \mathbb{E}_{\tau\sim\rho_\theta}\left(\sum_{t=0}^{H-1}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \sum_{k=0}^{H-1} r_k\right)\right) \\
&= \mathbb{E}_{\tau\sim\rho_\theta}\left(\sum_{t=0}^{H-1}\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\left[\sum_{k=0}^{t-1} r_k + \sum_{k=t}^{H-1} r_k\right]\right)\right) \\
&= \mathbb{E}_{\tau\sim\rho_\theta}\left(\sum_{t=0}^{H-1}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \sum_{k=0}^{t-1} r_k\right)\right) + \mathbb{E}_{\tau\sim\rho_\theta}\left(\sum_{t=0}^{H-1}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \sum_{k=t}^{H-1} r_k\right)\right)
\end{aligned}$$

However, by part (a), we see that

$$\mathbb{E}_{\tau\sim\rho_\theta}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \sum_{k=0}^{t-1} r_k\right) = 0$$

for all $t = 0, \ldots, H-1$. Therefore, the first term vanishes, and we are left with

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta} \left(\sum_{t=0}^{H-1} \left(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot \sum_{k=t}^{H-1} r_k \right) \right).$$

3. We use Adam's Law again, conditioning on s_t and a_t. Starting from the result in part (b), we have

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\tau \sim \rho_\theta} \left(\sum_{t=0}^{H-1} \left(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot \sum_{k=t}^{H-1} r_k \right) \right) \\
&= \sum_{t=0}^{H-1} \mathbb{E}_{\tau \sim \rho_\theta} \left(\mathbb{E} \left(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot \sum_{k=t}^{H-1} r_k \,\middle|\, s_t, a_t \right) \right) \\
&= \sum_{t=0}^{H-1} \mathbb{E}_{\tau \sim \rho_\theta} \left(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot \mathbb{E} \left(\sum_{k=t}^{H-1} r_k \,\middle|\, s_t, a_t \right) \right).
\end{aligned}$$

But note that by definition, $\mathbb{E} \left(\sum_{k=t}^{H-1} r_k \,\middle|\, s_t, a_t \right) = Q_t^{\pi_\theta}(s_t, a_t)$. It follows that

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{H-1} \mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot Q_t^{\pi_\theta}(s_t, a_t) \big) \\
&= \mathbb{E}_{\tau \sim \rho_\theta} \left(\sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t | s_t) \cdot Q_t^{\pi_\theta}(s_t, a_t) \right),
\end{aligned}$$

as desired.

4. By part (c) and linearity of expectation, it suffices to show that

$$\mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot b_t(s_t) \big) = 0$$

for an arbitrary function $b_t : S \to \mathbb{R}$ and $t = 0, \ldots, H-1$. Note that

$$\begin{aligned}
\mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot b_t(s_t) \big) &= \mathbb{E} \big(\mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot b_t(s_t) \mid s_t \big) \big) \\
&= \mathbb{E} \big(b_t(s_t) \cdot \mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \mid s_t \big) \big).
\end{aligned}$$

For the same reason as in part (a), we see that $\mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \mid s_t \big) = 0$. Therefore, we have

$$\mathbb{E}_{\tau \sim \rho_\theta} \big(\nabla_\theta \log \pi_\theta(a_t | s_t) \cdot b_t(s_t) \big) = 0$$

for all functions b_t and $t = 0, \ldots, H-1$, which completes the proof.

□

3. (**Off-policy Policy Gradient Estimation**)

Proof. We want to show that the expectation of the expression

$$\hat{J} := \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{h=0}^{H-1}\frac{\pi(a_h^i|s_h^i)}{\pi^b(a_h^i|s_h^i)}\right)\sum_{h=0}^{H-1}\nabla\log\pi(a_h^i|s_h^i)\big(R(\tau^i)\big)$$

is equal to the original PG of π, which is $\nabla_\theta J(\theta)$. Here, we are taking the expectations over trajectories given by the off-policy π^b. Therefore, we have

$$\rho_{\pi^b}(\tau^i) = \mu_0(s_0^i)\prod_{h=0}^{H-1}\pi^b(a_h^i|s_h^i)\prod_{h=0}^{H-2}P(s_{k+1}^i|s_k^i,a_k^i).$$

Note that the transition probabilities $P(\cdot|s_k^i,a_k^i)$ are independent from the policy for all k, and therefore they are equal between the two policies. It follows that

$$\begin{aligned}
\mathbb{E}_{\tau^i\sim\rho_{\pi^b}}\big(\hat{J}\big) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\tau^i\sim\rho_{\pi^b}}\left[\left(\prod_{h=0}^{H-1}\frac{\pi(a_h^i|s_h^i)}{\pi^b(a_h^i|s_h^i)}\right)\sum_{h=0}^{H-1}\nabla\log\pi(a_h^i|s_h^i)\big(R(\tau^i)\big)\right]\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{\tau^i}\rho_{\pi^b}(\tau^i)\left(\prod_{h=0}^{H-1}\frac{\pi(a_h^i|s_h^i)}{\pi^b(a_h^i|s_h^i)}\right)\sum_{h=0}^{H-1}\nabla\log\pi(a_h^i|s_h^i)\big(R(\tau^i)\big)\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{\tau^i}\mu_0(s_0^i)\prod_{h=0}^{H-1}\pi(a_h^i|s_h^i)\prod_{h=0}^{H-2}P(s_{k+1}^i|s_k^i,a_k^i)\sum_{h=0}^{H-1}\nabla\log\pi(a_h^i|s_h^i)\big(R(\tau^i)\big)\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\tau^i\sim\rho_{\pi}}\left[\sum_{h=0}^{H-1}\nabla\log\pi(a_h^i|s_h^i)\big(R(\tau^i)\big)\right].
\end{aligned}$$

The equality in the last step holds because

$$\rho_{\pi}(\tau^i) = \mu_0(s_0^i)\prod_{h=0}^{H-1}\pi(a_h^i|s_h^i)\prod_{h=0}^{H-2}P(s_{k+1}^i|s_k^i,a_k^i)$$

is the probability of observing trajectory τ^i under the original policy π. But by definition, we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\tau^i\sim\rho_{\pi}}\left[\sum_{h=0}^{H-1}\nabla\log\pi(a_h^i|s_h^i)\big(R(\tau^i)\big)\right] = \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta J(\theta) = \nabla_\theta J(\theta),$$

and thus $\hat{J}$ is an unbiased estimate of $\nabla_\theta J(\theta)$. $\square$

4. (**Softmax Policy**)

Proof.

1. Define $\phi(s,a) := e_{s,a}$, where $e_{s,a} \in \mathbb{R}^d$ is the vector with 1 at the index corresponding to position (s,a) and zeroes everywhere else. Then, it is clear that $\theta^\top \phi(s,a) = \theta_{s,a}$, which then satisfies the property that $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$.

2. Note that if $\pi_\theta(a|s) \propto \exp(\theta^\top \phi(s,a))$, then we have

$$\pi_\theta(a|s) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))}$$

Then, the following expressions are equivalent:

$$\begin{aligned}
\pi_\theta(a|s) \geq \pi_\theta(a'|s) &\iff \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \geq \frac{\exp(\theta^\top \phi(s,a'))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \\
&\iff \exp(\theta^\top \phi(s,a)) \geq \exp(\theta^\top \phi(s,a')) \\
&\iff \theta^\top \phi(s,a) \geq \theta^\top \phi(s,a'),
\end{aligned}$$

where the last equivalence is because logarithms are monotonically increasing.

3. Note that $\nabla_\theta \log(\pi_\theta(a|s)) = \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}$. Consider the numerator,

$$\nabla_\theta \pi_\theta(a|s) = \nabla_\theta \left(\frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \right).$$

Via the quotient rule, we have

$$\nabla_\theta \left(\frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \right) = \frac{\nabla_\theta \exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} - \frac{\exp(\theta^\top \phi(s,a)) \cdot \nabla_\theta \sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))}{\left(\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))\right)^2}.$$

Note that $\nabla_\theta \exp(\theta^\top \phi(s,a)) = \phi(s,a) \cdot \exp(\theta^\top \phi(s,a))$. Therefore, the first term becomes

$$\begin{aligned}
\frac{\nabla_\theta \exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} &= \frac{\phi(s,a) \cdot \exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \\
&= \phi(s,a) \cdot \pi_\theta(a|s).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\nabla_\theta \sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a')) &= \sum_{a' \in \mathcal{A}} \nabla_\theta \exp(\theta^\top \phi(s,a')) \\
&= \sum_{a' \in \mathcal{A}} \phi(s,a') \cdot \exp(\theta^\top \phi(s,a'))
\end{aligned}$$

Therefore, the second term becomes

$$\begin{aligned}
\frac{\exp(\theta^\top \phi(s,a)) \cdot \nabla_\theta \sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))}{\left(\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))\right)^2} &= \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \cdot \frac{\sum_{a' \in \mathcal{A}} \phi(s,a') \cdot \exp(\theta^\top \phi(s,a'))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))} \\
&= \pi_\theta(a|s) \cdot \sum_{a' \in \mathcal{A}} \phi(s,a') \pi_\theta(a|s) \\
&= \pi_\theta(a|s) \cdot \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\phi(s,a')].
\end{aligned}$$

Therefore, the numerator is equal to $\phi(s,a)\cdot\pi_\theta(a|s)-\pi_\theta(a|s)\cdot\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}[\phi(s,a')]$, and so we conclude that

$$\nabla_\theta \log(\pi_\theta(a|s)) = \phi(s,a) - \mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}[\phi(s,a')],$$

as desired.

$\square$