Homework #15

Junwu Zhang
CME 241: Reinforcement Learning for Finance

March 04, 2020

Problem 1.
Write Proof (with precise notation) of the Policy Gradient Theorem

Solution.
For Policy Gradient (PG), we have the following crucial components to the problem:

- *States:* $s_t \in \mathcal{S}$
- *Actions:* $a_t \in \mathcal{A}$
- *Rewards:* $r_t \in \mathbb{R}, \forall t \in \{0, 1, 2, \cdots\}$
- State Transition Probabilities $\mathcal{P}^a_{s,s'} = \Pr\left(s_{t+1} = s' | s_t = s, a_t = a\right)$
- Initial State Probability Distribution $p_0 : \mathcal{S} \to [0, 1]$
- Expected Rewards $\mathcal{R}^a_s = E\left[r_t | s_t = s, a_t = a\right]$
- Policy Function Approximation $\pi(s, a; \theta) = \Pr\left(a_t = a | s_t = s, \theta\right), \theta \in \mathbb{R}^k$
- *Discount Factor:* γ

The Policy Gradient Theorem (PGT) can be written as:

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi(s, a; \theta) \cdot Q^\pi(s, a) \cdot da \cdot ds \tag{1}$$

where $\rho^\pi(s) = \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(s_0) \cdot p(s_0 \to s, t, \pi) \cdot ds_0$. We can it *Discounted-Aggregate State-Visitation Measure*, which is a key function for PG. $\rho^\pi(s)$ depends on θ. We also note that $\nabla_\theta \log \pi(s, a; \theta)$ is the Score Function.

To prove the PGT, we can first write:

$$J(\theta) = \int_{\mathcal{S}} p_0(s_0) \cdot V^\pi(s_0) \cdot ds_0 \tag{2}$$

$$= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi(s_0, a_0; \theta) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \tag{3}$$

Then, by calculating $\pi(s_0, a_0; \theta)$ and $Q^\pi(s_0, a_0)$, we can write $\nabla_\theta J(\theta)$ as:

$$\begin{aligned} \nabla_\theta J(\theta) = &\int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi(s_0, a_0; \theta) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\ &+ \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi(s_0, a_0; \theta) \cdot \nabla_\theta Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \end{aligned} \tag{4}$$

Using the Bellman operator, we can expand the action value function $Q^{\pi}(s_0, a_0)$ as:

$$Q^{\pi}(s_0, a_0) = \mathcal{R}_{s_0}^{a_0} + \int_{\mathcal{S}} \gamma \cdot \mathcal{P}_{s_0,s_1}^{a_0} \cdot V^{\pi}(s_1) \cdot ds_1 \tag{5}$$

Plugging Equation (5) into the original equation, we have:

$$\nabla_{\theta} J(\theta) = \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_{\theta} \pi(s_0, a_0; \theta) \cdot Q^{\pi}(s_0, a) \cdot da_0 \cdot ds_0 \tag{6}$$

$$+ \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi(s_0, a_0; \theta) \cdot \nabla_{\theta} \left(\int_{\mathcal{S}} \gamma \cdot \mathcal{P}_{s_0,s_1}^{a_0} \cdot V^{\pi}(s_1) \cdot ds_1 \right) \cdot da_0 \cdot ds_0 \tag{7}$$

since $\nabla_{\theta} \mathcal{R}_{s_0}^{a_0} = 0$.

Moving ∇_{θ} into $\int_{\mathcal{S}}$, and further moving the outside $\int_{\mathcal{A}}$, we have:

$$\begin{aligned}
\nabla_{\theta} J(\theta) &= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_{\theta} \pi(s_0, a_0; \theta) \cdot Q^{\pi}(s_0, a) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi(s_0, a_0; \theta) \int_{\mathcal{S}} \gamma \cdot \mathcal{P}_{s_0,s_1}^{a_0} \cdot \nabla_{\theta} V^{\pi}(s_1) \cdot ds_1 \cdot da_0 \cdot ds_0 \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_{\theta} \pi(s_0, a_0; \theta) \cdot Q^{\pi}(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \gamma \cdot p_0(s_0) \int_{\mathcal{A}} \pi(s_0, a_0; \theta) \cdot \mathcal{P}_{s_0,s_1}^{a_0} \cdot da_0 \cdot ds_0 \right) \cdot \nabla_{\theta} V^{\pi}(s_1) \cdot ds_1
\end{aligned} \tag{8}$$

We can further expand the above equations to:

$$\begin{aligned}
\nabla_{\theta} J(\theta) &= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \cdot \nabla_{\theta} \pi(s_0, a_0; \theta) \cdot Q^{\pi}(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \gamma \cdot p_0(s_0) \cdot p(s_0 \to s_1, 1, \pi) \cdot ds_0 \right) \cdot \nabla_{\theta} V^{\pi}(s_1) \cdot ds_1
\end{aligned} \tag{9}$$

based on the fact that:

$$\int_{\mathcal{A}} \pi(s_0, a_0; \theta) \cdot \mathcal{P}_{s_0,s_1}^{a_0} \cdot da_0 = p(s_0 \to s_1, 1, \pi) \tag{10}$$

Next, expand the above equations to:

$$\begin{aligned}
\nabla_{\theta} J(\theta) &= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \cdot \nabla_{\theta} \pi(s_0, a_0; \theta) \cdot Q^{\pi}(s_0, a_0) \cdot da \cdot ds_0 \\
&+ \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \gamma \cdot p_0(s_0) \, p(s_0 \to s_1, 1, \pi) \, ds_0 \right) \cdot \nabla_{\theta} \left(\int_{\mathcal{A}} \pi(s_1, a_1; \theta) \, Q^{\pi}(s_1, a_1) \, da_1 \right) ds_1
\end{aligned} \tag{11}$$

based on the fact that:

$$V^{\pi}(s_1) = \int_{\mathcal{A}} \pi(s_1, a_1; \theta) \cdot Q^{\pi}(s_1, a_1) \cdot da_1 \tag{12}$$

We can then split πQ^π and calculate the gradient of Q^π using Bellman expansion, we have:

$$\begin{aligned}\nabla_\theta J(\theta) &= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \cdot \nabla_\theta \pi(s_0, a_0; \theta) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\ &+ \int_{\mathcal{S}} \int_{\mathcal{S}} \gamma p_0(s_0)\, p(s_0 \to s_1, 1, \pi)\, ds_0 \left(\int_{\mathcal{A}} \nabla_\theta \pi(s_1, a_1; \theta)\, Q^\pi(s_1, a_1)\, da_1 + \ldots \right) ds_1 \\ &= \sum_{t=0}^{\infty} \int_{\mathcal{S}} \int_{\mathcal{S}} \gamma^t \cdot p_0(s_0) \cdot p(s_0 \to s_t, t, \pi) \cdot ds_0 \int_{\mathcal{A}} \nabla_\theta \pi(s_t, a_t; \theta) \cdot Q^\pi(s_t, a_t) \cdot da_t \cdot ds_t \end{aligned} \tag{13}$$

Since we know $\int_{\mathcal{A}} \nabla_\theta \pi(s_t, a_t; \theta) \cdot Q^\pi(s_t, a_t) \cdot da_t$ is independent of t, we can move $\sum_{t=0}^{\infty}$ inside the integrals and write:

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(s_0) \cdot p(s_0 \to s, t, \pi) \cdot ds_0 \int_{\mathcal{A}} \nabla_\theta \pi(s, a; \theta) \cdot Q^\pi(s, a) \cdot da \cdot ds \tag{14}$$

Since we also have:

$$\int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0(s_0) \cdot p(s_0 \to s, t, \pi) \cdot ds_0 \overset{\text{def}}{=} \rho^\pi(s) \tag{15}$$

We can finally write the overall equation as:

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi(s, a; \theta) \cdot Q^\pi(s, a) \cdot da \cdot ds \tag{16}$$

which is the same as Equation (1). ■

Problem 2.
Derive the score function for softmax policy (for finite set of actions)

Solution. Since we know that θ is n-vector and features vector is:

$$\phi(s, a) = \big(\phi_1(s, a), \ldots, \phi_n(s, a)\big) \text{ for all } s \in \mathcal{S}, a \in \mathcal{A} \tag{17}$$

we can weight the actions by doing a linear combination of all the features, namely $\theta^T \cdot \phi(s, a)$. Then, we can see that action probabilites are proportional to weights exponentiated, which can be written as:

$$\pi(s, a; \theta) = \frac{e^{\theta^\top \cdot \phi(s,a)}}{\sum_b e^{\theta^T \cdot \phi(s,b)}} \text{ for all } s \in \mathcal{S}, a \in \mathcal{A} \tag{18}$$

By the definition of score function, we can take the logarithm of the above equation, and have:

$$\nabla_\theta \log \pi(s, a; \theta) = \phi(s, a) - \sum_b \pi(s, b; \theta) \cdot \phi(s, b) = \phi(s, a) - \mathbb{E}_\pi[\phi(s, \cdot)] \tag{19}$$

■

Problem 3.
Derive the score function for gaussian policy (for continuous actions)

Solution. Slightly differs from the last question, since the action space is now continuous, we have the feature vector as:

$$\phi(s) = \big(\phi_1(s), \ldots, \phi_n(s)\big) \text{ for all } s \in \mathcal{S} \tag{20}$$

Similar to the last question, we have Gaussian Mean as $\theta^T \cdot \phi(s)$ and with the Gaussian policy $a \sim \mathcal{N}\left(\theta^T \cdot \phi(s), \sigma^2\right)$ for all $s \in \mathcal{S}$, we have the score function for continuous action space as:

$$\nabla_\theta \log \pi(s, a; \theta) = \frac{\big(a - \theta^T \cdot \phi(s)\big) \cdot \phi(s)}{\sigma^2} \tag{21}$$

■

Problem 4.
Write code for the REINFORCE Algoithm (Monte-Carlo Policy Gradient Algorithm, i.e., no Critic)

Solution. Code is attached separately. ■

Problem 5.
Write Proof (with proper notation) of the Compatible Function Approximation Theorem

Solution. The *Compatible Function Approximation Theorem* states: If the following two conditions are satisfied:

- Critic gradient is compatible with the Actor score function

$$\nabla_w Q(s, a; w) = \nabla_\theta \log \pi(s, a; \theta) \tag{22}$$

- Critic parameters w minimize the following mean-squared error:

$$\epsilon = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s, a; \theta) \left(Q^\pi(s, a) - Q(s, a; w)\right)^2 \cdot da \cdot ds \tag{23}$$

Then the Policy Gradient using critic $Q(s, a; w)$ is exactly:

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi(s, a; \theta) \cdot Q(s, a; w) \cdot da \cdot ds \tag{24}$$

To prove this, we can first use the second point in the theorem and know that:

$$\begin{aligned} &\int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s, a; \theta) \cdot \big(Q^\pi(s, a) - Q(s, a; w)\big) \cdot \nabla_w Q(s, a; w) \cdot da \cdot ds = 0 \\ \forall w \text{ that minimizes } &\epsilon = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s, a; \theta) \cdot \big(Q^\pi(s, a) - Q(s, a; w)\big)^2 \cdot da \cdot ds \end{aligned} \tag{25}$$

Using the first bullet point of the theorem which is:

$$\nabla_w Q(s,a;w) = \nabla_\theta \log \pi(s,a;\theta), \tag{26}$$

we can replace the subscript and have:

$$\int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s,a;\theta) \cdot \big(Q^\pi(s,a) - Q(s,a;w)\big) \cdot \nabla_\theta \log \pi(s,a;\theta) \cdot da \cdot ds = 0 \tag{27}$$

Therefore, we can see that Equation (25) and Equation (27) are equal:

$$\begin{aligned} &\int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s,a;\theta) \cdot Q^\pi(s,a) \cdot \nabla_\theta \log \pi(s,a;\theta) \cdot da \cdot ds \\ &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s,a;\theta) \cdot Q(s,a;w) \cdot \nabla_\theta \log \pi(s,a;\theta) \cdot da \cdot ds \end{aligned} \tag{28}$$

Since we also know:

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s,a;\theta) \cdot Q^\pi(s,a) \cdot \nabla_\theta \log \pi(s,a;\theta) \cdot da \cdot ds \tag{29}$$

Combining Equation (28) and Equation (29), we have:

$$\begin{aligned} \nabla_\theta J(\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi(s,a;\theta) \cdot Q(s,a;w) \cdot \nabla_\theta \log \pi(s,a;\theta) \cdot da \cdot ds \\ &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi(s,a;\theta) \cdot Q(s,a;w) \cdot da \cdot ds \end{aligned} \tag{30}$$

and we can see that this is the same as Equation (24), therefore the theorem is proved. ■