ENGG6600 Reinforcement Learning Assignment-3

Fall 2022

Yaowen Mei
1177855

Yaowen Mei
1177855

Homework 3

Problem 1- Variance of return G_t

Prove that $Var(G_{t+1}) \geq Var(G_t)$ is true if we assume R_{t+1} is, on average, correlated with the previous rewards.

Given that:

- $G_t = \sum_{i=0}^{t} R_i$
- $\frac{1}{t+1} \sum_{i=0}^{t} Cov\left(R_i, R_{t+1}\right) > 0$

From statistic, we also have the following properties/definitions:

$$Cov\left(X, Y\right) = E\left[XY\right] - E\left[X\right] E\left[Y\right]$$

$$Var\left(X\right) = E\left[X^2\right] - E[X]^2 \geq 0 \tag{1}$$

$$\sum_{i=0}^{t} Cov\left(R_i, R_{t+1}\right) = Cov\left(\sum_{i=0}^{t} R_i, R_{t+1}\right) = Cov\left(G_t, R_{t+1}\right) > 0 \tag{2}$$

SOLUTION-1:

$$\begin{aligned} Var\left(G_{t+1}\right) =& Var\left(G_t + R_{t+1}\right) \\ =& E\left[\left(G_t + R_{t+1}\right)^2\right] - E[G_t + R_{t+1}]^2 \\ =& E\left[G_t^2 + R_{t+1}^2 + 2G_t R_{t+1}\right] - E[G_t]^2 - E[R_{t+1}]^2 - 2E\left[G_t\right] E\left[R_{t+1}\right] \\ =& E\left[G_t^2\right] - E[G_t]^2 + E\left[R_{t+1}^2\right] - E[R_{t+1}]^2 + 2E\left[G_t R_{t+1}\right] - 2E\left[G_t\right] E\left[R_{t+1}\right] \\ =& Var\left(G_t\right) + Var\left(R_{t+1}\right) + 2Cov\left(G_t, R_{t+1}\right) \end{aligned}$$

Therefore, by using Eq. 1 and Eq. 2, the equation above can be rewrite as:

$$Var\left(G_{t+1}\right) - Var\left(G_t\right) = \underbrace{Var\left(R_{t+1}\right)}_{\geq 0} + 2\underbrace{Cov\left(G_t, R_{t+1}\right)}_{>0} \geq 0 \qquad \square$$

Problem 2–Variance Reduce in Policy Gradient Method

Potentially, at the cost of increased bias, the variance in policy gradient methods could be reduced. Let us consider an infinite horizon MDP $< S, A, Pr, \gamma >$, let us define:
The advantage function:

$$A_{\pi_\theta}(S_t, A_t) = q_{\pi_\theta}(S_t, A_t) - v_{\pi_\theta}(S_t)$$

The policy gradient function:

$$\nabla_\theta J(\theta) = E_{\pi_\theta}\left[A_{\pi_\theta}(S_t, A_t)\nabla_\theta \log \pi(A_t|S_t;\theta)\right]$$

In practice, we do not have access to the true advantage function $A_{\pi_\theta}(S_t, A_t)$, so we would like to consider using the general form of an estimator $\hat{A}_{\pi_\theta}(S_t, A_t)$ that can be a function of the enitre trajectory.
Questions List:

1. ***Policy Gradient with Baseline:***

   Given that $\hat{q}_{\pi_\theta}(S_{t:\infty}, A_{t:\infty})$ is an unbiased estimator of the true $q_{\pi_\theta}(S_t, A_t)$, and b_t is an arbitrary function of the actions and states sampled before A_t, **prove** that by using this estimate of A_t, we obtain an unbiased estimate of the policy gradient.

2. ***TD error as unbiased estimator of the advantage function:***

   Let's look at another variants of $\hat{A}_{\pi_\theta}$.

   Recall that TD error:

   $$\delta_{\pi_\theta} = R_{t+1} + \gamma\hat{v}(S_{t+1}) - \hat{v}(S_t) \tag{3}$$

   Prove that $\delta_{\pi\theta}$ is an unbiased estimator of ${}_{\pi\theta}$ when $\hat{v} = v_{\pi\theta}$

SOLUTION-2.1:

$$\begin{aligned} LHS = \nabla_\theta J(\theta) &= E_{\pi_\theta}\left[A_{\pi_\theta}(S_t, A_t)\nabla_\theta \log \pi(A_t|S_t;\theta)\right] \\ &= E_{\pi_\theta}\left[\left(q_{\pi_\theta}(S_t, A_t) - v_{\pi\theta}(S_t)\right)\nabla_\theta \log \pi(A_t|S_t;\theta)\right] \\ &= \underbrace{E_{\pi_\theta}\left[q_{\pi_\theta}(S_t, A_t)\nabla_\theta \log \pi(A_t|S_t;\theta)\right]}_{(2.1)} - \underbrace{E_{\pi_\theta}\left[v_{\pi\theta}(S_t)\nabla_\theta \log \pi(A_t|S_t;\theta)\right]}_{(2.2)} \end{aligned} \tag{4}$$

$$\begin{aligned} RHS = E_{\pi_\theta}\left[\hat{A}_{\pi_\theta}(S_t, A_t)\nabla_\theta \log \pi(A_t|S_t;\theta)\right] &= E_{\pi_\theta}\left[\left(\hat{q}_{\pi_\theta}(S_{t:\infty}, A_{t:\infty}) - b_t(S_{0:t-1}, A_{0:t-1})\nabla_\theta \log \pi(A_t|S_t;\theta)\right)\right] \\ = \underbrace{E_{\pi_\theta}\left[\hat{q}_{\pi_\theta}(S_{t:\infty}, A_{t:\infty})\nabla_\theta \log \pi(A_t|S_t;\theta)\right]}_{(2.3)} &- \underbrace{E_{\pi_\theta}\left[b_t(S_{0:t-1}, A_{0:t-1})\nabla_\theta \log \pi(A_t|S_t;\theta)\right]}_{(2.4)} \end{aligned} \tag{5}$$

First, we will proof the (2.2) component in LHS equals the (2.4) componenet in RHS.

Theorem 1. *The (2.2) component in LHS equals the (2.4) componenet in RHS equals 0.*

$$E_{\pi_\theta}\left[v_{\pi\theta}\left(S_t\right)\nabla_\theta \log \pi\left(A_t | S_t;\theta\right)\right] = E_{\pi_\theta}\left[b_t\left(S_{0:t-1}, A_{0:t-1}\right)\nabla_\theta \log \pi\left(A_t | S_t;\theta\right)\right] = 0$$

Proof.

$$\begin{aligned}
E_{\pi_\theta}\left[v_{\pi\theta}\left(S_t\right)\nabla_\theta \log \pi\left(A_t | S_t;\theta\right)\right] &= E_{S_t\sim\pi_\theta}\left[\sum_a \pi\left(a | S_t;\theta\right) v_{\pi\theta}\left(S_t\right)\frac{\nabla_\theta \pi\left(a | S_t;\theta\right)}{\pi\left(a | S_t;\theta\right)}\right] \\
&= E_{S_t\sim\pi_\theta}\left[v_{\pi\theta}\left(S_t\right)\sum_a \pi\left(a | S_t;\theta\right)\frac{\nabla_\theta \pi\left(a | S_t;\theta\right)}{\pi\left(a | S_t;\theta\right)}\right] \\
&= E_{S_t\sim\pi_\theta}\left[v_{\pi\theta}\left(S_t\right)\sum_a \nabla_\theta \pi\left(a | S_t;\theta\right)\right] \\
&= E_{S_t\sim\pi_\theta}\left[v_{\pi\theta}\left(S_t\right)\sum_a \nabla_\theta 1\right] = 0
\end{aligned}$$

Also:

$$\begin{aligned}
&\mathrm{E}_{\pi_\theta}\left[b_t\left(S_{0:t-1}, A_{0:t-1}\right)\nabla_\theta \log \pi\left(A_t | S_t;\theta\right)\right] = E_{S_{t:\infty}\sim\pi_\theta}\left[\sum_a \pi\left(a | S_t;\theta\right) b_t\left(S_{0:t-1}, A_{0:t-1}\right)\tfrac{\nabla_\theta \pi\left(a | S_t;\theta\right)}{\pi\left(a | S_t;\theta\right)}\right] \\
&= E_{\pi_\theta}\left[E_{S_{t+1:\infty}, A_{t+1:\infty}\sim\pi_\theta}\left[\sum_a \pi\left(a | S_t;\theta\right) b_t\left(S_{0:t-1}, A_{0:t-1}\right)\tfrac{\nabla_\theta \pi\left(a | S_t;\theta\right)}{\pi\left(a | S_t;\theta\right)}\middle| S_t, A_t\right]\right] \\
&= E_{\pi_\theta}\left[E_{S_{t+1:\infty}, A_{t+1:\infty}\sim\pi_\theta}\left[\sum_a \pi\left(a | S_t;\theta\right)\tfrac{\nabla_\theta \pi\left(a | S_t;\theta\right)}{\pi\left(a | S_t;\theta\right)}\middle| S_t, A_t\right] b_t\left(S_{0:t-1}, A_{0:t-1}\right)\right] \\
&= E_{\pi_\theta}\left[E_{S_{t+1:\infty}, A_{t+1:\infty}\sim\pi_\theta}\left[\sum_a \nabla_\theta \pi\left(a | S_t;\theta\right)\middle| S_t, A_t\right] b_t\left(S_{0:t-1}, A_{0:t-1}\right)\right] \\
&= E_{\pi_\theta}\left[E_{S_{t+1:\infty}, A_{t+1:\infty}\sim\pi_\theta}\left[\sum_a \nabla_\theta 1\middle| S_t, A_t\right] b_t\left(S_{0:t-1}, A_{0:t-1}\right)\right] \\
&= 0
\end{aligned}$$

□

Now, the next task is to proof (2.1) component in LHS equals the (2.3) component in RHS:

Theorem 2.

$$E_{\pi_\theta}\left[q_{\pi_\theta}\left(S_t, A_t\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right] = E_{\pi_\theta}\left[\hat{q}_{\pi_\theta}\left(S_{t:\infty}, A_{t:\infty}\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right]$$

Proof.

$$\begin{aligned}
&E_{\pi_\theta}\left[q_{\pi_\theta}\left(S_t, A_t\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right] = E_{\pi_\theta}\left[\hat{q}_{\pi_\theta}\left(S_{t:\infty}, A_{t:\infty}\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right]\\
&\quad= E_{\pi_\theta}\left[E_{S_{t+1:\infty},A_{t+1:\infty},\pi_\theta}\left[\left.\hat{q}_{\pi_\theta}\left(S_{t:\infty}, A_{t:\infty}\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right|S_t,A_t\right]\right]\\
&\quad= E_{\pi_\theta}\left[\underbrace{E_{S_{t+1:\infty},A_{t+1:\infty},\pi_\theta}\left[\left.\hat{q}_{\pi_\theta}\left(S_{t:\infty}, A_{t:\infty}\right)\right|S_t,A_t\right]}_{q_{\pi_\theta}(S_t,A_t)}\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right]\\
&\quad= E_{\pi_\theta}\left[q_{\pi_\theta}\left(S_t, A_t\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right] = LHS
\end{aligned}$$

□

Since we have proofed component (2.1) equals component (2.3), and component (2.2) equals component (2.4), we can claim that Eq. 4 equals Eq. 5. That is:

$$\nabla_\theta J\left(\theta\right) = (2.1) + (2.2) = (2.3) + (2.4) = E_{\pi_\theta}\left[\hat{A}_{\pi_\theta}\left(S_t, A_t\right)\nabla_\theta \log\pi\left(\left.A_t\right|S_t;\theta\right)\right] \qquad \square$$

SOLUTION-2.2:
From the TD error equation, Eq. 3, and $\hat{v} = v_{\pi\theta}$, we have:

$$\delta_{\pi_\theta} = R_{t+1} + \gamma v_{\pi_\theta}\left(S_{t+1}\right) - v_{\pi_\theta}\left(S_t\right)$$

$$\begin{aligned}
E_{\pi_\theta}\left[\left.\delta_{\pi_\theta}\right|S_t,A_t\right] &= E_{\pi_\theta}\left[\left.\delta_{\pi_\theta}\right|S_t,A_t\right]\\
&= E_{\pi_\theta}\left[\left.R_{t+1} + \gamma v_{\pi_\theta}\left(S_{t+1}\right) - v_{\pi_\theta}\left(S_t\right)\right|S_t,A_t\right]\\
&= E_{\pi_\theta}\left[\left.R_{t+1} + \gamma v_{\pi_\theta}\left(S_{t+1}\right)\right|S_t,A_t\right] - v_{\pi_\theta}\left(S_t\right)
\end{aligned}$$

By definition of q:

$$E_{\pi_\theta}\left[\left.R_{t+1} + \gamma v_{\pi_\theta}\left(S_{t+1}\right)\right|S_t,A_t\right] = E_{S_{t+1:\infty},A_{t+1}:\infty,\pi_\theta}\left[\hat{q}_{\pi_\theta}\left(S_{t:\infty}, A_{t:\infty}\right)\right] = q_{\pi_\theta}\left(S_t, A_t\right)$$

Finally, we have proofed that:

$$E_{\pi_\theta}\left[\left.\delta_{\pi_\theta}\right|S_t,A_t\right] = \underbrace{E_{\pi_\theta}\left[\left.R_{t+1} + \gamma v_{\pi_\theta}\left(S_{t+1}\right)\right|S_t,A_t\right]}_{q_{\pi_\theta}(S_t,A_t)} - v_{\pi_\theta}\left(S_t\right) = A_{\pi_\theta}\left(S_t, A_t\right) \qquad \square$$