CS234 Problem Session Solutions

Week 9: March 10

1) [CA Session] Importance Sampling [1]

Recall that importance sampling can be used to generate an estimate of the performance of one policy, called the evaluation policy, given a trajectory that was generated by a different policy, called the behavior policy. The importance sampling estimate for a trajectory τ is:

$$\text{IS}(\tau) = \prod_{t=1}^{H} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} \sum_{t=1}^{H} \gamma^{t-1} R_t,$$

where π_e is the evaluation policy, π_b is the behavior policy, γ is the discount factor, and H is the trajectory length. The product in the equation is called the importance weight:

$$\text{IW}(\tau) = \prod_{t=1}^{H} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}$$

and the sum is the return. If there are multiple trajectories, $\mathcal{D} = \{\tau_i\}_{i=1}^{n}$, then the mean IS estimator is:

$$\text{IS}(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \text{IS}(\tau_i).$$

(a) If τ is produced by π_b, show that the expected value of $\text{IS}(\tau)$ is the expected return of π_e.

Solution Recall from problem 2 of Feb 17's Problem Session that we can write $\frac{Pr(\tau|\pi_e)}{Pr(\tau|\pi_b)} = \prod_{t=1}^{H} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} = \text{IW}(\tau)$. Thus, we can write

[1] For those interested in learning more about this line of work, Section 3.8 of Phil Thomas's thesis is a great read: `https://people.cs.umass.edu/~pthomas/papers/Thomas2015c.pdf`

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi_b}[\text{IS}(\tau)] &= \mathbb{E}_{\tau\sim\pi_b}\left[\prod_{t=1}^{H}\frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}\sum_{t=1}^{H}\gamma^{t-1}R_t\right] \\
&= \mathbb{E}_{\tau\sim\pi_b}\left[\frac{Pr(\tau|\pi_e)}{Pr(\tau|\pi_b)}\sum_{t=1}^{H}\gamma^{t-1}R_t\right] \\
&= \mathbb{E}_{\tau\sim\pi_e}\left[\sum_{t=1}^{H}\gamma^{t-1}R_t\right]
\end{aligned}$$

(b) Show that $\mathbb{E}[IW(\tau)|\tau\sim\pi_b] = 1$, and therefore that $\mathbb{E}[\sum_{i=1}^{n}\text{IW}(\tau_i)] = n$.

Solution

$$\begin{aligned}
\mathbb{E}[IW(\tau)|\tau\sim\pi_b] &= \mathbb{E}\left[\prod_{t=1}^{H}\frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}|\tau\sim\pi_b\right] \\
&= \mathbb{E}\left[\frac{Pr(\tau|\pi_e)}{Pr(\tau|\pi_b)}|\tau\sim\pi_b\right] \\
&= \sum_{\tau}P(\tau|\pi_b)\frac{Pr(\tau|\pi_e)}{Pr(\tau|\pi_b)} \\
&= \sum_{\tau}Pr(\tau|\pi_e) \\
&= 1
\end{aligned}$$

(c) Due to this result, a researcher proposes using $\frac{1}{\sum_{i=1}^{n}\text{IW}(\tau_i)}$ rather than $\frac{1}{n}$ when averaging the importance sampling estimates from many trajectories. The researcher calls this new estimator approximate importance sampling and is defined as:

$$\text{AIS}(\mathcal{D}) = \frac{1}{\sum_{i=1}^{n}\text{IW}(\tau_i)}\sum_{i=1}^{n}\text{IS}(\tau_i)$$

Show that $\text{AIS}(\mathcal{D}) \in [0, HR_{max}]$ if the rewards are bounded by $R_t \in [0, R_{max}]$.

Solution

$$\begin{aligned}
\text{AIS}(\mathcal{D}) &= \frac{1}{\sum_{i=1}^{n}\text{IW}(\tau_i)}\sum_{i=1}^{n}\text{IS}(\tau_i) \\
&= \frac{1}{\sum_{i=1}^{n}\text{IW}(\tau_i)}\sum_{i=1}^{n}\left[\text{IW}(\tau_i)\sum_{t=1}^{H}\gamma^{t-1}R_t^i\right]
\end{aligned}$$

In order to find an upper bound, notice that $\sum_{t=1}^{H} \gamma^{t-1} R_t \leq H R_{max}$. Thus, we can see

$$\begin{aligned}
\text{AIS}(\mathcal{D}) &\leq \frac{1}{\sum_{i=1}^{n} \text{IW}(\tau_i)} \sum_{i=1}^{n} [IW(\tau_i) H R_{max}] \\
&= H R_{max} \frac{1}{\sum_{i=1}^{n} \text{IW}(\tau_i)} \sum_{i=1}^{n} \text{IW}(\tau_i) \\
&= H R_{max}.
\end{aligned}$$

Likewise, we can find a lower bound, by seeing that $\sum_{t=1}^{H} \gamma^{t-1} R_t \geq 0$. Thus, $\text{AIS}(\mathcal{D}) \geq \frac{1}{\sum_{i=1}^{n} \text{IW}(\tau_i)} \sum_{i=1}^{n} [IW(\tau_i) 0] = 0$.

(d) Why is the result in part c) important? Why does it suggest that approximate importance sampling might give better estimates than ordinary importance sampling?

Solution The expected value of AIS is bounded within the same range as the expected value of the returns under π_e. This can reduce variance, as the IW could dramatically vary in the ordinary IS estimate. The ordinary IS estimate can have very high variance, although being unbiased.

(e) What is $\text{AIS}(\mathcal{D})$ an unbiased estimator of if $\mathcal{D}$ contains only a single trajectory? Show this result mathematically.

Solution $\text{AIS}(\mathcal{D})$ is an unbiased estimator of the expected return of π_b. We can see this as follows:

$$\begin{aligned}
\mathbb{E}_{\tau_i \sim \pi_b}[\text{AIS}(\mathcal{D})] &= \mathbb{E}_{\tau \sim \pi_b} \left[\frac{1}{\text{IW}(\tau)} \text{IS}(\tau) \right] \\
&= \mathbb{E}_{\tau \sim \pi_b} \left[\frac{1}{\text{IW}(\tau)} \text{IW}(\tau) \sum_{t=1}^{H} \gamma^{t-1} R_t \right] \\
&= \mathbb{E}_{\tau \sim \pi_b} \left[\sum_{t=1}^{H} \gamma^{t-1} R_t \right]
\end{aligned}$$

2) [Breakout Rooms] True/False

For each of these problems, answer true or false and provide a short justification for your answer.

(a) We can apply REINFORCE when the policy is not differentiable.

Solution False, the policy must be differentiable in order to compute the policy gradient.

(b) Two instances of the Thompson Sampling algorithm will choose the same actions at every timestep on a bandit problem with deterministic rewards.

Solution False, due to random sampling of the parameters, the algorithm can choose different actions.

(c) Importance sampling does not require knowledge of the transition function and does not rely on the Markov assumption.

Solution True, see lecture slides on importance sampling.

(d) Suppose you are in a multi-armed bandit setting where your algorithm selects an arm, and then your algorithm must select another arm before observing the first arm's reward. If your algorithm is UCB, UCB will deterministically select the same arm twice.

Solution True, the optimal arm according to UCB will remain the same since no updates have been made in between selections.

(e) Suppose you are in a multi-armed bandit setting where your algorithm selects an arm, and then your algorithm must select another arm before observing the first arm's reward. If your algorithm is Thompson sampling, Thompson sampling will deterministically select the same arm twice.

Solution False, Thompson sampling can choose different arms due to randomness from sampling from the posterior.

(f) In a bandit problem with deterministic rewards, UCB will only visit each suboptimal arm once.

Solution False, as the counts for the optimal arm grows, UCB will begin to select the other suboptimal arms again.