CS234 Problem Session

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.

[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`

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

(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 $R_t \in [0, R_{max}]$.

(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?

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

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.

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

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

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

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

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