Mathematics of Machine Learning Summer School 2019

Reinforcement Learning, Problem Session 3

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:

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

where π_e is the evaluation policy, π_b is the behavior policy, γ is the reward discount factor, H is the trajectory length, and S_t, A_t, and R_t are the state, action, and reward at time t. The product in this equation is called the *importance weight*:

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

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

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

A) Show that the expected value of $IS(\tau)$ if τ is produced by π_b is the expected return of π_e.

B) Show that

$$E\left[IW(\tau)|\tau\sim\pi_b\right] = 1,$$

and therefore that

$$E\left[\sum_{i=1}^{n} IW(\tau_i)\right] = n.$$

C) Due to this result, a researcher proposes using $\frac{1}{\sum_{i=1}^{n} IW(\tau_i)}$ 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:

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

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

D) Why is the result in C) important? Why does it suggest that approximate importance sampling might give better estimates than ordinary importance sampling?

E) Show that $AIS(D)$ is not always an unbiased estimator of the expected return of π_e.

F) What is $AIS(D)$ an unbiased estimator of if D contains only a single trajectory? Show this result mathematically.

G) Given what we know about approximate importance sampling, could we use it with the Chernoff-Hoeffding inequality to produce a confidence interval on the performance of the evaluation policy?

H) If you are given an evaluation policy, π_e, but can select the behavior policy, π_b, you might do so with the goal of minimizing the variance of *IS*. Show (e.g., by counter-example) that using $\pi_b = \pi_e$ is not necessarily optimal.