On a Connection between Importance Sampling and the Likelihood Ratio Policy Gradient

Jie Tang, Pieter Abbeel

Speaker: Jie Tang

UC Berkeley

Likelihood Ratio Policy Gradient

- Likelihood ratio policy gradients are some of the most successful reinforcement learning algorithms.
- Consider a class of stochastic policies parameterized by θ; let $\pi_\theta : S \times A \to [0,1]$ denote a policy in this class.
- Directly optimize expected reward over θ:

$$U(\theta) = \mathrm{E}\left[\sum_{t=1}^{H} R(s_t, a_t) \;\middle|\; \pi_\theta\right]$$

- Can compute gradient of U from sample trajectories:

$$\nabla_\theta U(\theta) \approx \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{t=1}^{H} R(s_t, a_t)$$

Importance Sampling

- Importance sampling:

$$\widehat{U}(\theta_1) = \mathrm{E}_{\{(s_t,a_t)\}\sim\theta_2}\left[\prod_{t=1}^{H}\frac{\pi_{\theta_1}(a_t \mid s_t)}{\pi_{\theta_2}(a_t \mid s_t)}\sum_{t=1}^{H} R(s_t, a_t)\right]$$

Proposition (Importance Sampling and Policy Gradients)

The sample estimate of the gradient of $\widehat{U}(\theta)$ evaluated using only sample trajectories drawn under π_θ is equal to the likelihood ratio based sample estimate of the gradient of $U(\theta)$.

- Implication: likelihood ratio policy gradient methods are not making full use of the data.
- However, importance sampling has not been widely adopted.

Optimal Baselines

- Optimal baselines for likelihood ratio PG methods:

$$\nabla_\theta U(\theta) = \mathrm{E}_{\{(s_t,a_t)\}\sim\theta}\left[\sum_{t=1}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\left(\sum_{t=1}^{H} R(s_t, a_t) - b\right)\right]$$

Proposition (Unbiased Baselines)

For any distribution $P_\theta(X)$, any scalar valued function $f(X)$, and any fixed vector b:

$$E_{P_\theta(X)}[f(X)] = E_{P_\theta(X)}\left[f(X) - b^T \nabla_\theta \log P_\theta(X)\right]$$

- We can set b to minimize the variance of the estimator.
- We use this generalized baseline to estimate $\widehat{U}(\theta)$.