AIL 722: Reinforcement Learning

Lecture 35: Variance Reduction

Raunak Bhattacharyya

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap & Today's Outline

- Policy gradients
- Bias and variance
- Baseline
- Optimal baseline
- Causality
- Actor-critic

Baseline

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right]$$

$$\text{Let } Y = \nabla_\theta \log p_\theta(\tau) \cdot \Big(r(\tau) - b\Big)$$

$$\mathbb{E}\Big[Y\Big] = \mathbb{E}\Big[\nabla_\theta \log p_\theta(\tau) \cdot r(\tau)\Big] - b \cdot \mathbb{E}\Big[\nabla_\theta \log p_\theta(\tau)\Big]$$

$$E[\nabla_\theta \log p_\theta(\tau)] = \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)\, d\tau$$

$$= \int \nabla_\theta p_\theta(\tau)\, d\tau$$

$$= \nabla_\theta \int p_\theta(\tau)\, d\tau \qquad = \nabla_\theta 1$$

$$E\Big[Y\Big] = E\Big[\nabla_\theta \log p_\theta(\tau) \cdot r(\tau)\Big]$$

Our revised estimator Y does not introduce bias

Optimal Baseline

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right]$$

$$Y = \nabla_\theta \log p_\theta(\tau) \cdot \left(r(\tau) - b\right)$$

$$\mathrm{Var}(\mathbb{E}[f(x)]) = \frac{\mathrm{Var}(f(x))}{N}$$

Thus, sufficient to analyse variance of Y. Then divide by N

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$$

Can we ignore the second term?

$$E\Big[Y\Big] = E\Big[\nabla_\theta \log p_\theta(\tau) \cdot r(\tau)\Big]$$

Independent of b

Therefore, we are interested in $\mathbb{E}[Y^2]$

Optimal Baseline

$$Y = \nabla_\theta \log p_\theta(\tau) \cdot \Big(r(\tau) - b\Big)$$

Goal: Find b such that the variance is lowest

Therefore, we are interested in $\mathbb{E}[Y^2]$

$$\mathbb{E}\Big[g(\tau)^2 (r(\tau) - b)^2\Big]$$

$$\begin{aligned}
\frac{d\text{Var}}{db} &= \frac{d}{db}\mathbb{E}\left[g(\tau)^2 (r(\tau) - b)^2\right] \\
&= \frac{d}{db}\mathbb{E}\left[g(\tau)^2 r(\tau)^2\right] - 2\mathbb{E}\left[g(\tau)^2 r(\tau) b\right] + b^2 \mathbb{E}\left[g(\tau)^2\right] \\
&= -2\mathbb{E}\left[g(\tau)^2 r(\tau)\right] + 2b\mathbb{E}\left[g(\tau)^2\right] \\
&= 0
\end{aligned}$$

$$b^{opt} = \frac{\mathbb{E}\left[g(\tau)^2 r(\tau)\right]}{\mathbb{E}\left[g(\tau)^2\right]}$$

In practice, we use the average reward as the baseline

From Sutton and Barto

REINFORCE with Baseline (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
    Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
    Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
        $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
        $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$

Variance Reduction: Causality

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t} \mid s_{i,t}) \right) \left(\sum_{t=1}^{T} r(s_{i,t}, a_{i,t}) \right)$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \left(\sum_{t'=1}^{T} r(s_{i,t'}, a_{i,t'}) \right)$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \left(\sum_{t'=t}^{T} r(s_{i,t'}, a_{i,t'}) \right)$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \hat{Q}_{i,t}$$

Reward-to-go

Reduced variance since we reduced the sum to be a smaller number