Policy Gradient with Baseline

Shusen Wang

Policy Gradient

- Use policy network, $\pi(a|s; \boldsymbol{\theta})$, for controlling the agent.

Policy Gradient

- Use policy network, $\pi(a|s;\mathbf{\theta})$, for controlling the agent.
- State-value function:

$$V_\pi(s) = \mathbb{E}_{A\sim\pi}[Q_\pi(s,A)]$$

$$= \sum_a \pi(a|s;\mathbf{\theta}) \cdot Q_\pi(s,a).$$

Policy Gradient

- Use policy network, $\pi(a|s;\boldsymbol{\theta})$, for controlling the agent.
- State-value function:

$$V_\pi(s) = \mathbb{E}_{A\sim\pi}[Q_\pi(s,A)]$$

$$= \sum_a \pi(a|s;\boldsymbol{\theta}) \cdot Q_\pi(s,a).$$

- Policy gradient:

$$\frac{\partial\, V_\pi(s)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot Q_\pi(s,A)\right].$$

Baseline

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \, \frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \right]$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \pi(a|s;\boldsymbol{\theta}) \cdot \frac{\partial \ln \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \pi(a|s;\boldsymbol{\theta}) \frac{\partial \ln \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$$= \underline{\frac{1}{\pi(a \mid s;\boldsymbol{\theta})}} \cdot \underline{\frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}}$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \pi(a|s;\boldsymbol{\theta}) \cdot \left[\frac{1}{\pi(a \mid s;\boldsymbol{\theta})} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \cancel{\pi(a|s;\boldsymbol{\theta})} \cdot \left[\frac{1}{\cancel{\pi(a \mid s;\boldsymbol{\theta})}} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \cancel{\pi(a|s;\boldsymbol{\theta})} \cdot \left[\frac{1}{\cancel{\pi(a \mid s;\boldsymbol{\theta})}} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

$$= b \cdot \sum_a \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \cancel{\pi(a|s;\boldsymbol{\theta})} \cdot \left[\frac{1}{\cancel{\pi(a \mid s;\boldsymbol{\theta})}} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

$$= b \sum_a \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \pi(a|s;\boldsymbol{\theta}) \cdot \left[\frac{1}{\pi(a \mid s;\boldsymbol{\theta})} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

$$= b \cdot \sum_a \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$$= b \cdot \frac{\partial \sum_a \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$= 1$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \pi(a|s;\boldsymbol{\theta}) \cdot \left[\frac{1}{\pi(a \mid s;\boldsymbol{\theta})} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

$$= b \cdot \sum_a \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$$= b \cdot \frac{\partial \sum_a \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$$= b \cdot \frac{\partial\, 1}{\partial\, \boldsymbol{\theta}}$$

Baseline

- Let the baseline, b, be anything independent of A.
- $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = b \cdot \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$

$$= b \cdot \sum_a \pi(a|s;\boldsymbol{\theta}) \cdot \left[\frac{1}{\pi(a \mid s;\boldsymbol{\theta})} \cdot \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right]$$

$$= b \cdot \sum_a \frac{\partial\, \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$$= b \cdot \frac{\partial \sum_a \pi(a \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}$$

$$= b \cdot \frac{\partial\, 1}{\partial\, \boldsymbol{\theta}} \quad = \quad 0.$$

Policy Gradient with Baseline

If b is independent of A, then $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = 0.$

Policy Gradient with Baseline

If b is independent of A, then $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = 0.$

- Policy gradient:

$$\frac{\partial V_\pi(s)}{\partial \boldsymbol{\theta}}$$

$$= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot Q_\pi(s, A)\right]$$

Policy Gradient with Baseline

If b is independent of A, then $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = 0.$

- Policy gradient:

$$\frac{\partial\, V_\pi(s)}{\partial\, \boldsymbol{\theta}}$$

$$= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot Q_\pi(s, A)\right] - \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot b\right]$$

Equal to zero

Policy Gradient with Baseline

If b is independent of A, then $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = 0.$

- Policy gradient:

$$\frac{\partial V_\pi(s)}{\partial \boldsymbol{\theta}}$$

$$= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot Q_\pi(s, A)\right] - \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot b\right]$$

Policy Gradient with Baseline

If b is independent of A, then $\mathbb{E}_{A\sim\pi}\left[b \cdot \frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}\right] = 0.$

- Policy gradient:

$$\frac{\partial\, V_\pi(s)}{\partial\, \boldsymbol{\theta}}$$

$$= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot Q_\pi(s,A)\right] - \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot b\right]$$

$$= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A \mid s;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s,A) - b\right)\right].$$

Policy Gradient with Baseline

Theorem. If b is independent of A_t, then policy gradient is:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - b\right)\right].$$

Monte Carlo Approximation

Monte Carlo Approximation

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\boxed{\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - b\right)}\right].$$

$$= \mathbf{g}(A_t)$$

Monte Carlo Approximation

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\boxed{\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - b\right)}\right].$$

$$= \mathbf{g}(A_t)$$

- Randomly sample $a_t \sim \pi(\cdot \mid s_t; \boldsymbol{\theta})$ and compute $\mathbf{g}(a_t)$.

Monte Carlo Approximation

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t\sim\pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \big(Q_\pi(s_t, A_t) - b\big)\right].$$

- Randomly sample $a_t \sim \pi(\cdot \mid s_t; \boldsymbol{\theta})$ and compute $\mathbf{g}(a_t)$.
- $\mathbf{g}(a_t)$ is an unbiased estimate of the policy gradient:

$$\mathbb{E}_{A_t\sim\pi}[\mathbf{g}(A_t)] = \frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}}.$$

Stochastic Policy Gradient

Stochastic policy gradient:

$$\mathbf{g}(a_t) = \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, a_t) - b\right).$$

- Stochastic policy gradient ascent:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta \cdot \mathbf{g}(a_t).$$

Stochastic Policy Gradient

Stochastic policy gradient:

$$\mathbf{g}(a_t) = \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, a_t) - b\right).$$

- Whatever b (independent of A_t) we use, the policy gradient $\mathbb{E}_{A_t \sim \pi}[\mathbf{g}(A_t)]$ remains the same.

Stochastic Policy Gradient

Stochastic policy gradient:

$$\mathbf{g}(a_t) = \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, a_t) - b\right).$$

- Whatever b (independent of A_t) we use, the policy gradient $\mathbb{E}_{A_t \sim \pi}[\mathbf{g}(A_t)]$ remains the same.
- However, b affects $\mathbf{g}(a_t)$.
- A good b leads to small variance and speeds up convergence.

Choices of Baselines

Choice 1: b=0

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t\sim\pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - b\right)\right].$$

- We can simply set $b = 0$.
- It becomes the standard policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t\sim\pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot Q_\pi(s_t, A_t)\right].$$

Choice 2: b is state-value

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - b\right)\right].$$

- Because s_t has been observed, $b = V_\pi(s_t)$ is independent of A_t.

Choice 2: b is state-value

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t\sim\pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - b\right)\right].$$

- Because s_t has been observed, $b = V_\pi(s_t)$ is independent of A_t.
- Why using such a baseline?
- $V_\pi(s_t)$ is close to $Q_\pi(s_t, A_t)$:

$$V_\pi(s_t) = \mathbb{E}_{A_t}[Q_\pi(s_t, A_t)].$$

Thank you!