Policy Gradient: REINFORCE, Variance Reduction, Convergence

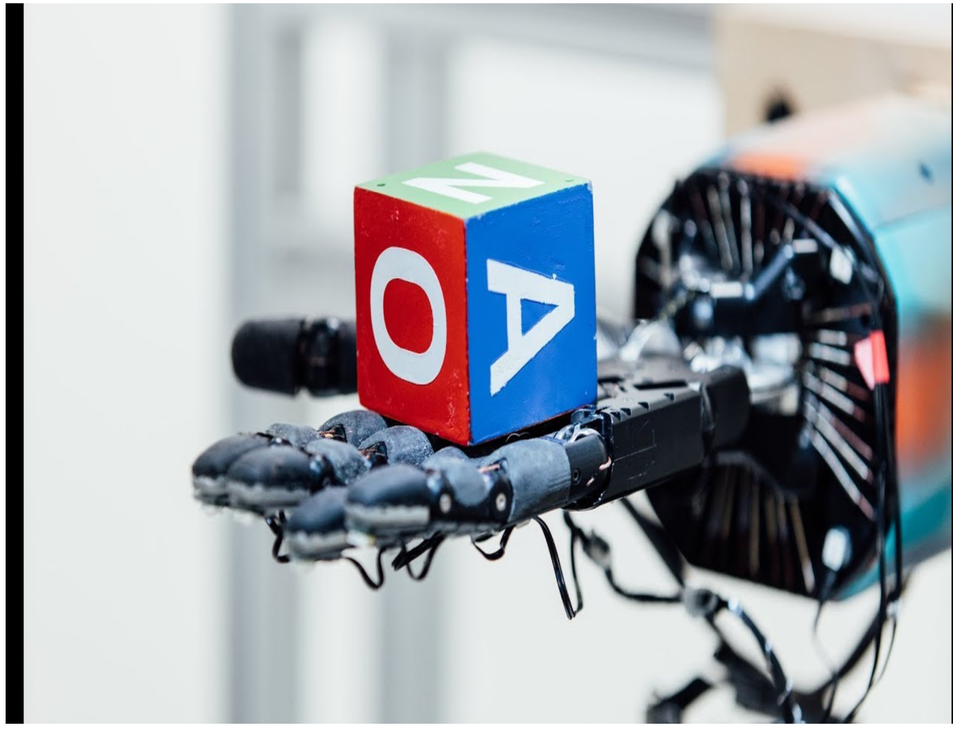
Sham Kakade and Kianté Brantley

CS 6789: Foundations of Reinforcement Learning

Policy Optimization

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]



[OpenAI,19]

Recap: Infinite Horizon Discounted MDPs

1. Two formulations of Policy Gradient

Recap: Infinite Horizon Discounted MDPs

1. Two formulations of Policy Gradient

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\nabla_\theta \ln \rho_\theta(\tau) R(\tau) \right]$$

Recap: Infinite Horizon Discounted MDPs

1. Two formulations of Policy Gradient

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\nabla_\theta \ln \rho_\theta(\tau) R(\tau)\right] \quad = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right) R(\tau)\right]$$

Recap: Infinite Horizon Discounted MDPs

1. Two formulations of Policy Gradient

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\nabla_\theta \ln \rho_\theta(\tau) R(\tau)\right] \quad = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right) R(\tau)\right]$$

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{s_0\sim\rho}\left[V^{\pi_\theta}(s_0)\right]$$

Recap: Infinite Horizon Discounted MDPs

1. Two formulations of Policy Gradient

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\nabla_\theta \ln \rho_\theta(\tau) R(\tau)\right] \quad = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right) R(\tau)\right]$$

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{s_0\sim\rho}\left[V^{\pi_\theta}(s_0)\right] \quad = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a)$$

Outline for today

1. Two formulations of Policy Gradient

2. Variance Reduction

3. Convergence of SGD

Derivation of unbiased Stochastic Policy Gradient

$$\nabla_\theta J(\theta) := \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s) Q^{\pi_\theta}(s,a)\right]$$

Derivation of unbiased Stochastic Policy Gradient

$$\nabla_\theta J(\theta) := \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s) Q^{\pi_\theta}(s,a)\right]$$

Draw $h \propto \gamma^h$, **roll-in** π_θ to generate $s_h, a_h \sim \mathbb{P}_h^{\pi_\theta}$

Derivation of unbiased Stochastic Policy Gradient

$$\nabla_\theta J(\theta) := \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s) Q^{\pi_\theta}(s,a)\right]$$

Draw $h \propto \gamma^h$, **roll-in** π_θ to generate $s_h, a_h \sim \mathbb{P}_h^{\pi_\theta}$

Roll-out π_θ from (s_h, a_h) : terminate with prob $1-\gamma$, $\widetilde{Q}^{\pi_\theta}(s_h, a_h) = \sum_{\tau=h}^{t\geq h} r_\tau$

Derivation of unbiased Stochastic Policy Gradient

$$\nabla_\theta J(\theta) := \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s) Q^{\pi_\theta}(s,a)\right]$$

Draw $h \propto \gamma^h$, **roll-in** π_θ to generate $s_h, a_h \sim \mathbb{P}_h^{\pi_\theta}$

Roll-out π_θ from (s_h, a_h) : terminate with prob $1-\gamma$, $\widetilde{Q}^{\pi_\theta}(s_h, a_h) = \sum_{\tau=h}^{t\geq h} r_\tau$

Unbiased estimate: $\nabla_\theta \ln \pi_\theta(a_h\,|\,s_h)\,\widetilde{Q}^{\pi_\theta}(s_h, a_h)$

Outline for today

1. Two formulations of Policy Gradient

2. Variance Reduction

3. Convergence of SGD

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h) \left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h) \right)$$

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h) \left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h) \right)$$

$$\mathbb{E}_{s,a \sim d^{\pi_\theta}} \left[\nabla_\theta \ln \pi_\theta(a \mid s) \right] b(s)$$

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)$$

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)\right] b(s)$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\mathbb{E}_{a\sim\pi_\theta(\cdot\mid s)} \nabla_\theta \ln \pi_\theta(a \mid s)\right]$$

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)$$

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)\right] b(s)$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\mathbb{E}_{a\sim\pi_\theta(\cdot\mid s)}\nabla_\theta \ln \pi_\theta(a \mid s)\right] \qquad = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \pi_\theta(a \mid s)\frac{\nabla_\theta \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)}\right]$$

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)$$

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)\right] b(s)$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \nabla_\theta \ln \pi_\theta(a \mid s)\right] \qquad = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \pi_\theta(a \mid s)\frac{\nabla_\theta \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)}\right]$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \nabla_\theta \pi_\theta(a \mid s)\right]$$

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)$$

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)\right] b(s)$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\nabla_\theta \ln \pi_\theta(a \mid s)\right] \qquad = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \pi_\theta(a \mid s)\frac{\nabla_\theta \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)}\right]$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \nabla_\theta \pi_\theta(a \mid s)\right] \quad = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\,\nabla_\theta\left(\sum_a \pi_\theta(a \mid s)\right)\right]$$

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)$$

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)\right] b(s)$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\nabla_\theta \ln \pi_\theta(a \mid s)\right] \qquad = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \pi_\theta(a \mid s)\frac{\nabla_\theta \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)}\right]$$

$$= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\sum_a \nabla_\theta \pi_\theta(a \mid s)\right] \quad = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[b(s)\,\nabla_\theta\left(\sum_a \pi_\theta(a \mid s)\right)\right] \quad = 0$$

Variance Reduction via Action-Independent Baseline

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)$$

The best baseline that minimizes variance:

$$\min_b \ \mathbb{E}\left[\Big(\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)\Big)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)\right]$$

Variance Reduction via Action-Independent Baseline

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)$$

The best baseline that minimizes variance:

$$\min_b \ \mathbb{E}\left[\left(\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)\right)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\left(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\right)\right]$$

$$b(s_h) = \frac{\mathbb{E}\left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\widetilde{Q}^{\theta}(s_h, a_h)\right]}{\mathbb{E}\left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right]}$$

Variance Reduction via Action-Independent Baseline

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)$$

The best baseline that minimizes variance:

$$\min_b \ \mathbb{E}\left[\left(\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)\right)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)\right]$$

$$b(s_h) = \frac{\mathbb{E}\left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\widetilde{Q}^{\theta}(s_h, a_h)\right]}{\mathbb{E}\left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right]}$$

In practice:

$$b(s_h) = V^{\pi_\theta}(s)$$

Variance Reduction via Action-Independent Baseline

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)$$

The best baseline that minimizes variance:

$$\min_b \ \mathbb{E}\left[\left(\nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)\right)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\Big(\widetilde{Q}^{\pi_\theta}(s_h, a_h) - b(s_h)\Big)\right]$$

$$b(s_h) = \frac{\mathbb{E}\left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\widetilde{Q}^{\theta}(s_h, a_h)\right]}{\mathbb{E}\left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h)^\top \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right]}$$

In practice:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s))\right] = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(\nabla_\theta \ln \pi_\theta(a \mid s)A^{\pi_\theta}(s,a)\right)$$

$$b(s_h) = V^{\pi_\theta}(s)$$

Summary so far:

The most commonly used formulation:
Policy Gradient with V^{π_θ} as a baseline:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)\right]$$

Summary so far:

The most commonly used formulation:
Policy Gradient with V^{π_θ} as a baseline:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s) A^{\pi_\theta}(s,a)\right]$$

Q: can you think about a way to get an unbiased estimate of $A^{\pi_\theta}(s,a)$ via one roll-out?

Summary

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a \mid s)\left(Q^{\pi_\theta}(s,a) - V_\theta^{\pi}(s)\right)\right]$$

Use an unbiased estimate of $\nabla_\theta J(\theta)$