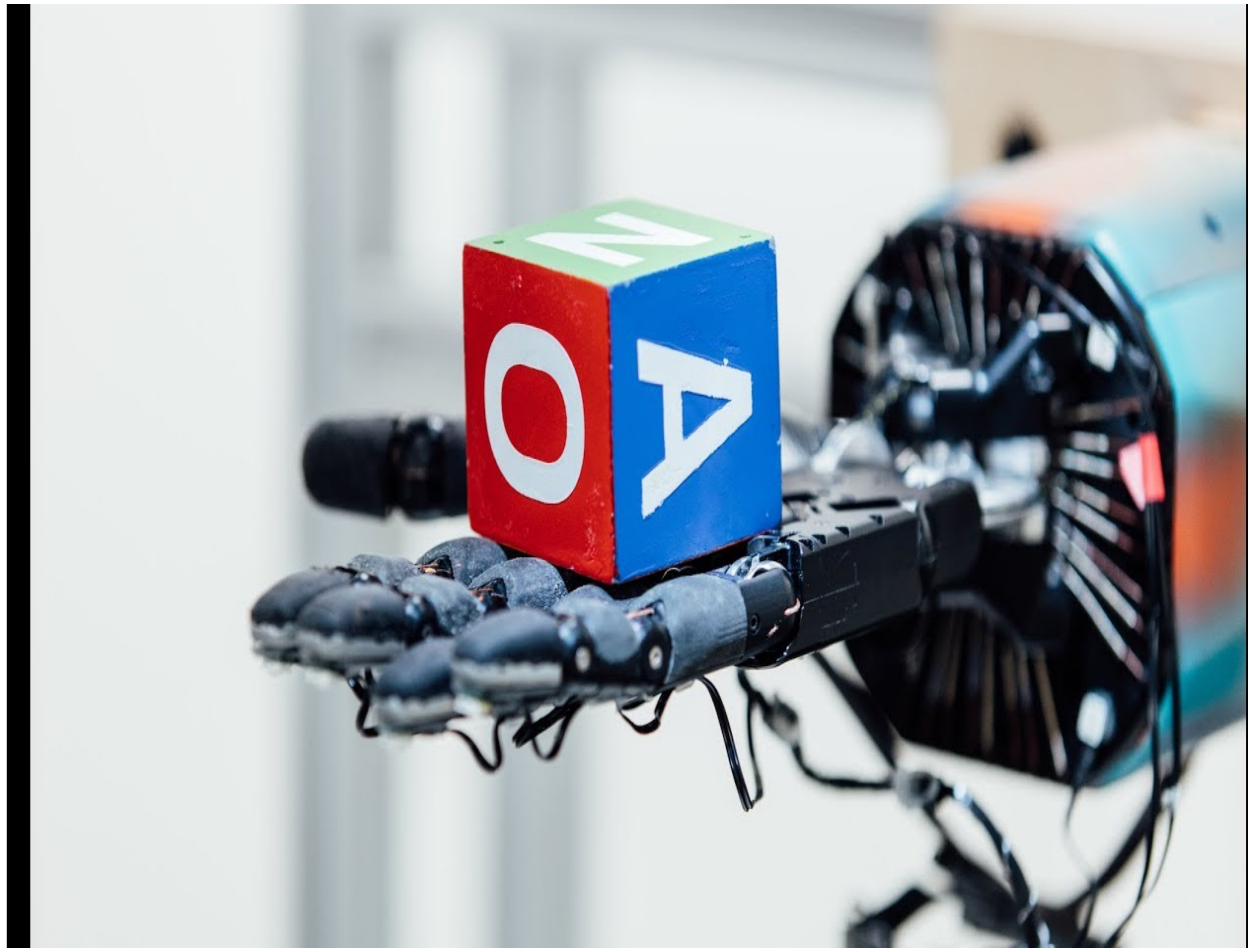
Policy Gradient:
REINFORCE, Variance Reduction, Convergence

Policy Optimization

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]



[OpenAI,19]

Recap: Infinite Horizon Discounted MDPs

$$\mathcal{M} = \{P, r, \gamma, \rho, S, A\}$$

where $s_0 \sim \rho$

Objective: $J(\pi) := \mathbb{E}_\pi \left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 \sim \rho, s_{h+1} \sim P_{s_h, a_h}, a_h \sim \pi(\cdot \mid s_h) \right]$

Recap: Infinite Horizon Discounted MDPs

State-action distribution $\mathbb{P}_h^{\pi}(s, a)$: probability of π hitting (s, a) at h

Recap: Infinite Horizon Discounted MDPs

State-action distribution $\mathbb{P}_h^\pi(s, a)$: probability of π hitting (s, a) at h

State-distribution $\mathbb{P}_h^\pi(s)$: probability of π hitting (s) at h

Recap: Infinite Horizon Discounted MDPs

State-action distribution $\mathbb{P}_h^\pi(s, a)$: probability of π hitting (s, a) at h

State-distribution $\mathbb{P}_h^\pi(s)$: probability of π hitting (s) at h

Discounted visitation $d^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a)$

Recap: Infinite Horizon Discounted MDPs

State-action distribution $\mathbb{P}_h^\pi(s, a)$: probability of π hitting (s, a) at h

State-distribution $\mathbb{P}_h^\pi(s)$: probability of π hitting (s) at h

Discounted visitation $d^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a)$

Advantage function: $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$

Today: Policy Gradient Derivation

e.g., Reinforce, Natural Policy Gradient, TRPO, PPO:

(Williams 92, Kakade 02, Schulman et al 15, 17)

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta)$$

Today: Policy Gradient Deriviation

e.g., Reinforce, Natural Policy Gradient, TRPO, PPO:

(Williams 92, Kakade 02, Schulman et al 15, 17)

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \qquad J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

Today: Policy Gradient Derivation

e.g., Reinforce, Natural Policy Gradient, TRPO, PPO:

(Williams 92, Kakade 02, Schulman et al 15, 17)

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \qquad J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\pi_\theta)\,|_{\theta=\theta_t}$$

Today: Policy Gradient Deriviation

e.g., Reinforce, Natural Policy Gradient, TRPO, PPO:

(Williams 92, Kakade 02, Schulman et al 15, 17)

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \qquad J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\pi_\theta) \mid_{\theta=\theta_t}$$

Main question for today's lecture:
how to compute the gradient?

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

3. Neural Policy:

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

3. Neural Policy:

Neural network $f_\theta : S \times A \mapsto \mathbb{R}$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (e.g., for linear MDPs):

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

3. Neural Policy:

Neural network $f_\theta : S \times A \mapsto \mathbb{R}$

$$\pi_\theta(a \mid s) = \frac{\exp(f_\theta(s, a))}{\sum_{a'} \exp(f_\theta(s, a'))}$$

Warm Up

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Warm Up

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

Warm Up

$$J(\theta) = \mathbb{E}_{x \sim P_\theta} \left[f(x) \right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x)$$

Warm Up

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) \quad = \mathbb{E}_{x \sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x)$$

Warm Up

$$J(\theta) = \mathbb{E}_{x\sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x\sim\rho} \frac{P_\theta(x)}{\rho(x)} f(x) \;\; = \mathbb{E}_{x\sim\rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \;\; \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

Warm Up

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) = \mathbb{E}_{x \sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)\,|_{\theta=\theta_0}$$

Warm Up

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) = \mathbb{E}_{x \sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)\,|_{\theta=\theta_0}$$

We can set sampling distribution $\rho = P_{\theta_0}$

Warm Up

$$J(\theta) = \mathbb{E}_{x\sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x\sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) \;\; = \mathbb{E}_{x\sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \;\; \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

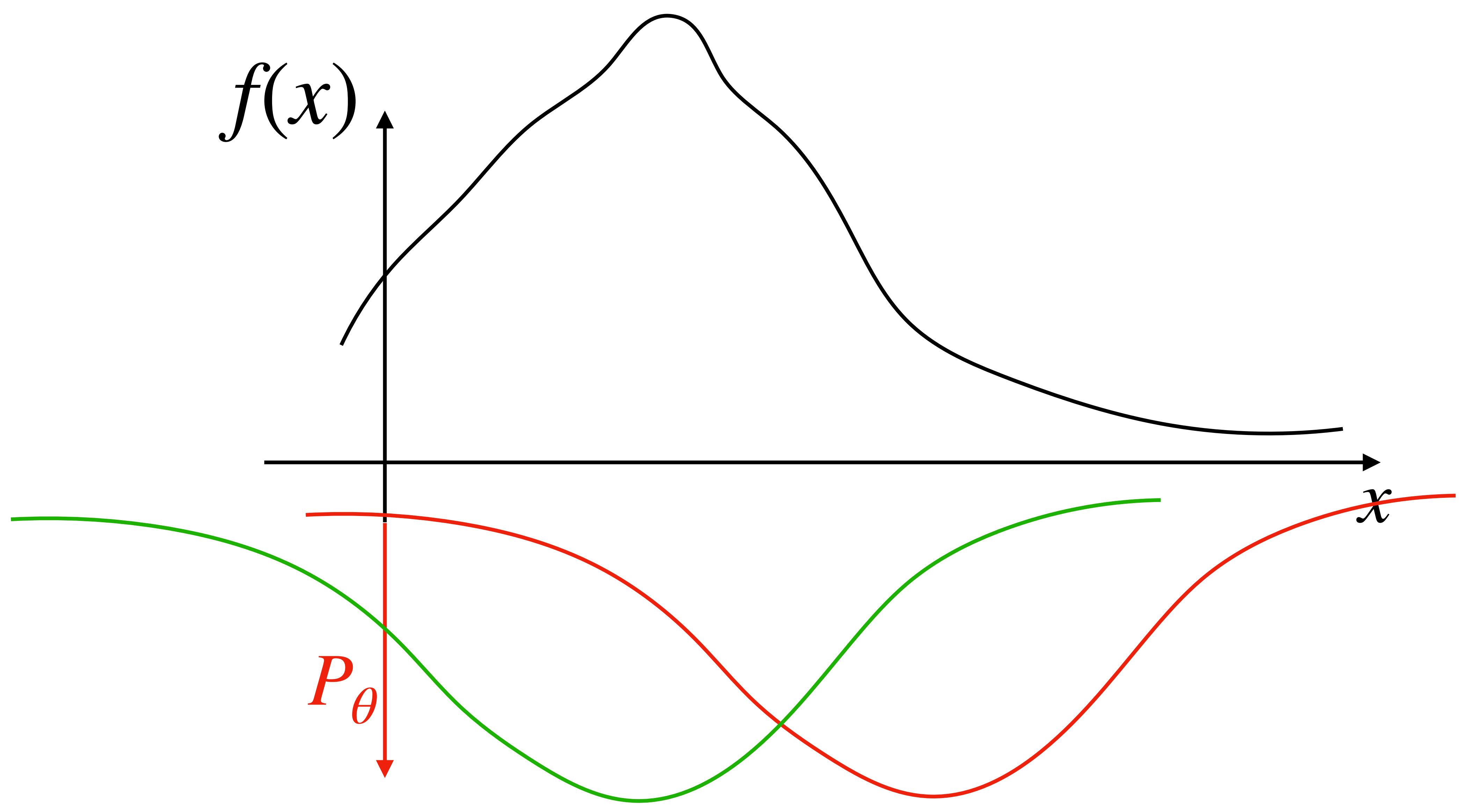
$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x)\,|_{\theta=\theta_0}$$

We can set sampling distribution $\rho = P_{\theta_0}$

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{x\sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x) f(x)$$

Warm Up

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{x\sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x) f(x)$$



Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \rho(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \rho(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \right]}_{R(\tau)}$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \rho(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

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

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \dots\}$$

$$\rho_\theta(\tau) = \rho(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\dots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

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

$$= \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\left[\nabla_\theta\big(\ln \rho(s_0) + \ln \pi_\theta(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \dots\big) R(\tau)\right]$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \rho(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

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

$$= \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \left[\nabla_\theta \left(\ln \rho(s_0) + \ln \pi_\theta(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \ldots\right) R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \left[\nabla_\theta \left(\ln \pi_\theta(a_0 \mid s_0) + \ln \pi_\theta(a_1 \mid s_1)\ldots\right) R(\tau)\right]$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \rho(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \right]}_{R(\tau)}$$

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

$$= \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\nabla_\theta \left(\ln \rho(s_0) + \ln \pi_\theta(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \ldots \right) R(\tau) \right]$$

$$= \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\nabla_\theta \left(\ln \pi_\theta(a_0 \mid s_0) + \ln \pi_\theta(a_1 \mid s_1) \ldots \right) R(\tau) \right] = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right) R(\tau) \right]$$

Derivation of Policy Gradient w/ Q^{π}

Recall definition of value function $V^{\pi_\theta}(s)$

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right]$$

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0)\left[\frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}\right] \cdot Q^{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^{\pi_\theta}(s_1)\right]$$

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0)\left[\frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}\right] \cdot Q^{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^{\pi_\theta}(s_1)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) \cdot Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_1)$$

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0)\left[\frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}\right] \cdot Q^{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^{\pi_\theta}(s_1)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) \cdot Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_1)$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1 \mid s_1)} \nabla_\theta \ln \pi_\theta(a_1 \mid s_1) Q^{\pi_\theta}(s_1, a_1)\right] + \gamma^2 \mathbb{E}_{s_2 \sim \mathbb{P}_2^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_2)$$

Derivation of Policy Gradient w/ Q^{π}

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0)\left[\frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}\right] \cdot Q^{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^{\pi_\theta}(s_1)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) \cdot Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_1)$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1 \mid s_1)} \nabla_\theta \ln \pi_\theta(a_1 \mid s_1) Q^{\pi_\theta}(s_1, a_1)\right] + \gamma^2 \mathbb{E}_{s_2 \sim \mathbb{P}_2^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_2)$$

$$= \sum_{h=0}^{\infty} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) Q^{\pi_\theta}(s_h, a_h)$$

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0)\left[\frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}\right] \cdot Q^{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^{\pi_\theta}(s_1)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) \cdot Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_1)$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1 \mid s_1)} \nabla_\theta \ln \pi_\theta(a_1 \mid s_1) Q^{\pi_\theta}(s_1, a_1)\right] + \gamma^2 \mathbb{E}_{s_2 \sim \mathbb{P}_2^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_2)$$

$$= \sum_{h=0}^{\infty} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) Q^{\pi_\theta}(s_h, a_h) \quad = \frac{1}{1-\gamma} \mathbb{E}_{s, a \sim d^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a \mid s) Q^{\pi_\theta}(s, a)$$

Derivation of Policy Gradient w/ Q^π

Recall definition of value function $V^{\pi_\theta}(s)$

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

$$= \mathbb{E}_{s_0 \sim \rho}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^{\pi_\theta}(s_0, a_0)\right] = \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(a_0 \mid s_0) \nabla_\theta Q^{\pi_\theta}(s_0, a_0)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0)\left[\frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}\right] \cdot Q^{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^{\pi_\theta}(s_1)\right]$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0|s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) \cdot Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_1)$$

$$= \mathbb{E}_{s_0 \sim \rho}\left[\mathbb{E}_{a_0 \sim \pi_\theta(a_0|s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0)\right] + \gamma \mathbb{E}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1|s_1)} \nabla_\theta \ln \pi_\theta(a_1 \mid s_1) Q^{\pi_\theta}(s_1, a_1)\right] + \gamma^2 \mathbb{E}_{s_2 \sim \mathbb{P}_2^{\pi_\theta}} \nabla_\theta V^{\pi_\theta}(s_2)$$

$$= \sum_{h=0}^{\infty} \mathbb{E}_{s_h, a_h \sim \mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) Q^{\pi_\theta}(s_h, a_h) = \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a \mid s) Q^{\pi_\theta}(s, a)$$

Derivation of unbiased Stochastic Policy Gradient

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

Derivation of unbiased Stochastic Policy Gradient

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

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

Derivation of unbiased Stochastic Policy Gradient

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

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

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

Derivation of unbiased Stochastic Policy Gradient

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

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

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

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

Derivation of unbiased Stochastic Policy Gradient

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

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

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:

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

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

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

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:

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

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

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

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

Variance Reduction via Action-Independent Baseline

Unbiased Estimate:

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

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

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

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

Variance Reduction via Action-Independent Baseline

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

The best baseline:

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

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

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:

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

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

In practice:

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

Variance Reduction via Action-Independent Baseline

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

The best baseline:

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

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

In practice:

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

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

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 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?

Next: Stochastic Gradient Ascent Converges to Stationary Point

Convergence to Stationary Point

$J(\pi_\theta)$ is non-convex (see example in the monograph)

Convergence to Stationary Point

$J(\pi_\theta)$ is non-convex (see example in the monograph)

Def of β-smooth:

$$\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

$$\left| J(\theta) - J(\theta_0) - \nabla_\theta J(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2, \forall \theta, \theta_0$$

Convergence to Stationary Point

$J(\pi_\theta)$ is non-convex (see example in the monograph)

Def of β-smooth:

$$\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

$$\left| J(\theta) - J(\theta_0) - \nabla_\theta J(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2, \forall \theta, \theta_0$$

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t)$, $\mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2$,

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \le \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \le \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \le J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t) \le \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta}{2}\eta^2\sigma^2$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t) \leq \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta \sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \leq \sum_t \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta T}{2}\eta^2\sigma^2$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t) \leq \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta \sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \leq \sum_t \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta T}{2}\eta^2\sigma^2 \quad \Rightarrow \frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2 \leq \frac{1}{\eta T}M + \frac{\beta}{2}\eta\sigma^2$$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \le \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \le \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \le J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t) \le \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta \sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \le \sum_t \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta T}{2}\eta^2\sigma^2 \quad \Rightarrow \frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2 \le \frac{1}{\eta T}M + \frac{\beta}{2}\eta\sigma^2$$

Set $\eta = \sqrt{M/(\beta\sigma^2 T)}$

Convergence to Stationary Point

[Theorem] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta\nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta\nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta\nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t) \leq \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta\sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \leq \sum_t \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta T}{2}\eta^2\sigma^2 \quad \Rightarrow \frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2 \leq \frac{1}{\eta T}M + \frac{\beta}{2}\eta\sigma^2$$

$$\mathbb{E}\left[\left\| \ln \pi_\theta(a\,|\,s)\,\widetilde{Q}^{\pi_\theta}(s,a) \right\|_2^2\right]$$

Set $\eta = \sqrt{M/(\beta\sigma^2 T)}$

Convergence to Stationary Point

[**Theorem**] If $J(\theta)$ is β-smooth, and we run SGA: $\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t)$, $\quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2$,

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \le \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) - \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \le \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \le J(\theta_{t+1}) - J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t) \le \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta \sum_t \|\nabla_\theta J(\theta_t)\|_2^2 \le \sum_t \mathbb{E}\left[J(\theta_{t+1}) - J(\theta_t)\right] + \frac{\beta T}{2}\eta^2\sigma^2 \quad \Rightarrow \frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2 \le \frac{1}{\eta T}M + \frac{\beta}{2}\eta\sigma^2$$

$$\mathbb{E}\left[\left\| \ln \pi_\theta(a\,|\,s)\,\widetilde{Q}^{\pi_\theta}(s,a) \right\|_2^2\right]$$

$$\le \frac{1}{(1-\gamma)^2}\sup_{s,a}\left\| \nabla_\theta \ln \pi_\theta(a\,|\,s)\right\|_2^2$$

Set $\eta = \sqrt{M/(\beta\sigma^2 T)}$

Actor-Critic Algorithm

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

Critic $Q_w(s,a)$: approximately minimizes $\mathbb{E}_{s,a\sim\mu}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$

Actor-Critic Algorithm

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

Critic $Q_w(s,a)$: approximately minimizes $\mathbb{E}_{s,a\sim\mu}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$

Actor-Critic Gradient:

$$\nabla_\theta \ln \pi_\theta(a_h\,|\,s_h) Q_w(s_h, a_h)$$

Actor-Critic Algorithm

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

Critic $Q_w(s,a)$: approximately minimizes $\mathbb{E}_{s,a\sim\mu}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$

Actor-Critic Gradient:

$$\nabla_\theta \ln \pi_\theta(a_h\,|\,s_h) Q_w(s_h, a_h)$$

Actor-Critic Gradient:

$$\nabla_\theta \ln \pi_\theta(a_h\,|\,s_h)\left(Q_w(s_h,a_h) - \mathbb{E}_{a'\sim\pi_\theta(a'|s_h)} Q_w(s_h, a')\right)$$

Actor-Critic Algorithm

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

Critic $Q_w(s,a)$: approximately minimizes $\mathbb{E}_{s,a\sim\mu} \left(Q_w(s,a) - Q^{\pi_\theta}(s,a) \right)^2$

Actor-Critic Gradient:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h) Q_w(s_h, a_h)$$

Actor-Critic Gradient:

$$\nabla_\theta \ln \pi_\theta(a_h \mid s_h) \left(Q_w(s_h, a_h) - \mathbb{E}_{a' \sim \pi_\theta(a' \mid s_h)} Q_w(s_h, a') \right)$$

$A_w(s_h, a_h)$

Compatible Function Assumption

Compatible Function Assumption

Assume we minimize $\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$ via SGD on w, and get to stationary point:

Compatible Function Assumption

Assume we minimize $\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$ via SGD on w, and get to stationary point:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q^{\pi_\theta}(s,a) - Q_w(s,a)\right)\nabla_w Q_w(s,a) = 0$$

Compatible Function Assumption

Assume we minimize $\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$ via SGD on w, and get to stationary point:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q^{\pi_\theta}(s,a) - Q_w(s,a)\right)\nabla_w Q_w(s,a) = 0$$

If compatible, i.e., $\nabla_w \widehat{Q}_w(s,a) = \nabla_\theta \ln \pi_\theta(a\,|\,s)$, then, we get unbiased gradient estimate:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\nabla_\theta \ln \pi_\theta(a\,|\,s)Q^{\pi_\theta}(s,a) = \mathbb{E}_{s,a\sim d^{\pi_\theta}}\nabla_\theta \ln \pi_\theta(a\,|\,s)Q_w(s,a)$$

Compatible Function Assumption

Assume we minimize $\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$ via SGD on w, and get to stationary point:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q^{\pi_\theta}(s,a) - Q_w(s,a)\right)\nabla_w Q_w(s,a) = 0$$

If compatible, i.e., $\nabla_w \widehat{Q}_w(s,a) = \nabla_\theta \ln \pi_\theta(a \mid s)$, then, we get unbiased gradient estimate:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\nabla_\theta \ln \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a) = \mathbb{E}_{s,a\sim d^{\pi_\theta}}\nabla_\theta \ln \pi_\theta(a \mid s) Q_w(s,a)$$

One possible parameterization for $Q_w(s,a) := w^\top \nabla_\theta \ln \pi_\theta(a \mid s)$ (Natural PG)

Compatible Function Assumption

Assume we minimize $\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q_w(s,a) - Q^{\pi_\theta}(s,a)\right)^2$ via SGD on w, and get to stationary point:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}}\left(Q^{\pi_\theta}(s,a) - Q_w(s,a)\right)\nabla_w Q_w(s,a) = 0$$

If compatible, i.e., $\nabla_w \widehat{Q}_w(s,a) = \nabla_\theta \ln \pi_\theta(a \mid s)$, then, we get unbiased gradient estimate:

$$\mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a) = \mathbb{E}_{s,a\sim d^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a \mid s) Q_w(s,a)$$

One possible parameterization for $Q_w(s,a) := w^\top \nabla_\theta \ln \pi_\theta(a \mid s)$ (Natural PG)

Another one: $Q_w(s,a) = w^\top \phi(s,a), \quad \pi_\theta(a \mid s) \propto \exp(\theta^\top \phi(s,a))$

Summary

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

Use unbiased estimate of $\nabla_\theta J(\theta)$, SG ascent converges to stationary point

Actor-Critic with Compatible function (warm up for Natural Policy Gradient)