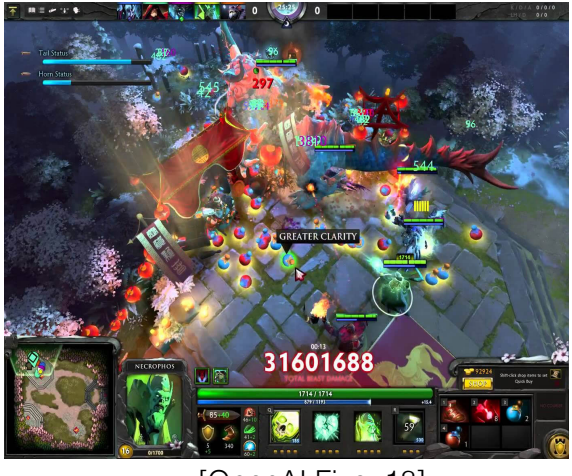
Policy Gradient: REINFORCE, Variance Reduction, Convergence

Sham Kakade and Kianté Brantley

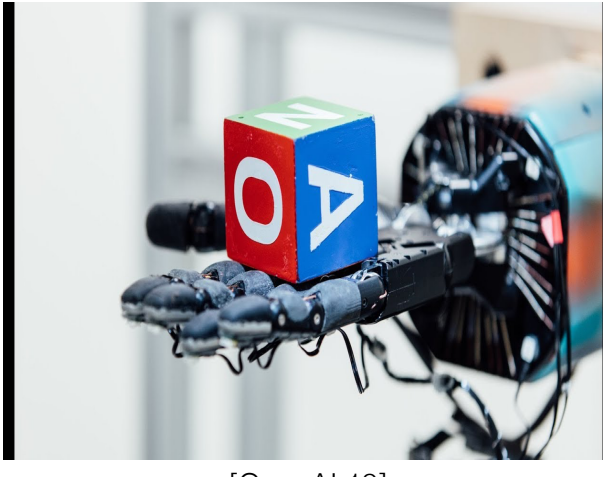
CS 6789: Foundations of Reinforcement Learning

Policy Optimization

[AlphaZero, Silver et.al, 17]



[OpenAI Five, 18]



[OpenAI,19]

Recap: Infinite Horizon Discounted MDPs

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

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

$$\mathbb{P}_h^\pi(s) = \sum_a \mathbb{P}_h^\pi(s,a)$$

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

$\sum_{s,a} d^\pi(s,a) = 1$

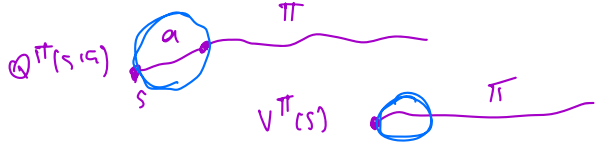
Recap: Infinite Horizon Discounted MDPs

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 Deriviation

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

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

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

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 Deriviation

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

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

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

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

$s_0 \sim \rho$

$a_0 \sim \pi_\theta$

$s_1 \sim P(\cdot \mid s_0, a_0)$

Today: Policy Gradient Deriviation

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

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

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

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

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

Outline for today

1. Two formulations of Policy Gradient

2. Variance Reduction

3. Convergence of SGD

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:

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

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

$\theta \in \mathbb{R}^{|S| \times |A|}$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

1. Softmax Policy for Tabular MDPs:

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

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

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

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

$\sum_a \pi_\theta(a|s) = 1$

Warm Up

$$\max_{\theta} \quad J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right] \qquad f : \mathcal{X} \to \mathbb{R}$$

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

$$= \int_{\mathcal{X}} p_\theta(x) f(x) \, dx$$

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., $\max_x P_\theta(x)/\rho(x) < \infty$

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

$$= \int_x \frac{P(x)}{\rho(x)} \cdot P_\theta(x) f(x)$$

$$= \int_x P(x) \frac{P_\theta(x)}{P(x)} f(x)$$

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., $\max_x P_\theta(x)/\rho(x) < \infty$

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

$x_i \sim \rho$

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., $\max_x P_\theta(x)/\rho(x) < \infty$

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

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

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., $\max_x P_\theta(x)/\rho(x) < \infty$

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

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

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., $\max_x P_\theta(x)/\rho(x) < \infty$

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

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

$\mathbb{E}_{x \sim P_{\theta_0}} \frac{\nabla_\theta P_\theta(x)}{P_{\theta_0}} f(x)$

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

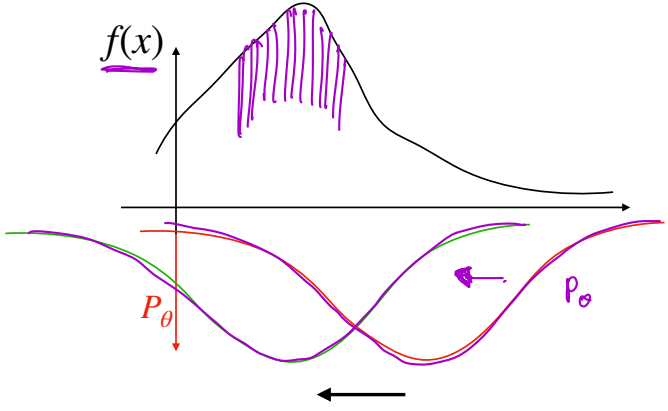
$\nabla_\theta \ln P_{\theta_0}$

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

Warm Up

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

$f: x \to \mathbb{R}$



1. Derivation of Policy Gradient: REINFORCE

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

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

1. Derivation of Policy Gradient: REINFORCE

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

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

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

$R(\tau) \in [0, \frac{1}{1-\gamma}]$

$\mathbb{E}_{x\sim p_\theta}[f(x)]$

$\mathbb{E}_{\tau\sim p_\theta}[R(\tau)]$

1. Derivation of Policy Gradient: REINFORCE

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

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

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

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

$\rho_\theta(\tau) \cdot \nabla \rho_\theta(\tau)$

$\rho_\theta(\tau)$

$\rho_\theta(\tau) \cdot \nabla_\theta \ln \rho_\theta(\tau)$

1. Derivation of Policy Gradient: REINFORCE

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

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

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

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

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

$$\nabla_\theta \ln \rho(s_0) + \nabla_\theta \ln \pi_\theta(a_0|s_0) + \nabla_\theta \ln P(s_1|s_0,a_0) + \dots$$

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

$$\nabla_\theta\left(\sum \pi_\theta(\tau) R(\tau)\right)$$

$$\nabla_\theta\left(\sum \frac{\pi_\theta(\tau)}{\pi_\theta(\tau)} \pi_\theta(\tau) R(\tau)\right)$$

$$\left(\sum \frac{\nabla_\theta \pi_\theta(\tau)}{\pi_\theta(\tau)} \pi_\theta(\tau) R(\tau)\right)$$

1. Derivation of Policy Gradient: REINFORCE

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

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

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

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

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

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

1. Derivation of Policy Gradient: REINFORCE

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

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

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

repeat
① $\tau \sim \rho_\theta(\tau)$
② $\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h)$
③ $\left[\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h)\right][R(\tau)]$
④ gradient ascent

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

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

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

2. Derivation of Policy Gradient w/ Q^π

$Q^\pi(s,a) = \mathbb{E}[\sum \gamma^h r_h]$
$(s_1, a_1 = (s,a)]$

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

2. Derivation of Policy Gradient w/ Q^{π}

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

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

2. Derivation of Policy Gradient w/ Q^π

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

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

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

$Q^\pi(s_0, a_0) = r(s_0, a_0) + \mathbb{E}_{s' \sim \rho}[V^\pi(s_1)]$

$\mathbb{E}_{a \sim \pi}[Q^\pi(s_0, a)] = \mathbb{E}_{a \sim \pi}[r(s_0, a) + \mathbb{E}_{s' \sim \rho}[V^\pi(s_1)]] = V^\pi(s_0)$

2. Derivation of Policy Gradient w/ Q^π

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

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

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

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

$\nabla_\theta \ln \pi_\theta(a_0 \mid s_0) = \frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)}$

$Q^\pi(s_0, a_0) = r(s_0, a_0) + \mathbb{E}_{s_1 \sim \rho}\{V^{\pi_\theta}(s_1)\}$

2. Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

2. Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

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

2. Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

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

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

2. Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

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

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

Derivation of unbiased Stochastic Policy Gradient

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