Policy Gradient
(continue)

Recap: the REINFORCE Algorithm

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$$

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

Recap: the REINFORCE Algorithm

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$$

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

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]}_{R(\tau)}$$

Recap: the REINFORCE Algorithm

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$$

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

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]}_{R(\tau)}$$

$$\nabla_\theta J(\pi_\theta)\mid_{\theta=\theta_0} := \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_{\theta_0}(a_h \mid s_h)\right) R(\tau)\right]$$

Recap: the REINFORCE Algorithm

Initialize θ

While True:

Recap: the REINFORCE Algorithm

Initialize θ

While True:

Generate n i.i.d trajectories $\tau^1, \ldots, \tau^n$ using π_θ

Recap: the REINFORCE Algorithm

Initialize θ

While True:

Generate n i.i.d trajectories $\tau^1, \ldots, \tau^n$ using π_θ

Compute gradient: $g = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{h=0}^{H-1}\nabla_\theta \ln \pi_\theta(a_h^i \mid s_h^i) \cdot R(\tau^i)\right)$

Recap: the REINFORCE Algorithm

Initialize θ

While True:

Generate n i.i.d trajectories $\tau^1, \ldots, \tau^n$ using π_θ

Compute gradient: $g = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h^i \mid s_h^i) \cdot R(\tau^i) \right)$

update: $\theta \Leftarrow \theta + \eta g$ (or adaptive methods like Adam)

REINFORCE can have high uncertainty

REINFORCE can have high uncertainty

Gradient: $g = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h^i \mid s_h^i) \cdot R(\tau^i)\right)$

REINFORCE can have high uncertainty

Gradient: $g = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h^i \,|\, s_h^i) \cdot R(\tau^i) \right)$

Often require large n to reduce the variance, especially when policy π_θ is quite random

Today's Question:

How to reduce Variance in Policy Gradient?

Outline:

1. A $Q(s, a)$ based Policy Gradient

2. Variance Reduction via A Baseline

Notations

$$\mathscr{M} = \{P, r, \gamma, \mu, S, A\} \quad \text{where } s_0 \sim \mu$$

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, s_0 = s, a_h \sim \pi\right]$$

$$\text{Objective: } J(\pi) := \mathbb{E}_{s_0 \sim \mu}\left[V^\pi(s_0)\right]$$

Notations

$$\mathscr{M} = \{P, r, \gamma, \mu, S, A\} \quad \text{where } s_0 \sim \mu$$

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, s_0 = s, a_h \sim \pi\right]$$

Objective: $J(\pi) := \mathbb{E}_{s_0 \sim \mu}\left[V^\pi(s_0)\right]$

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

Notations

$$\mathcal{M} = \{P, r, \gamma, \mu, S, A\} \quad \text{where } s_0 \sim \mu$$

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, s_0 = s, a_h \sim \pi\right]$$

$$\text{Objective: } J(\pi) := \mathbb{E}_{s_0 \sim \mu}\left[V^\pi(s_0)\right]$$

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

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

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 $V^{\pi_\theta}(s)$

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

Derivation of Policy Gradient w/ Q^π

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

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

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

Derivation of Policy Gradient w/ Q^π

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

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

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

Derivation of Policy Gradient w/ Q^π

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

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

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

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

Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

Derivation of Policy Gradient w/ Q^{π}

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

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

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

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

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

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

Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

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

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

Derivation of Policy Gradient w/ Q^π

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

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

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

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

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

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

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

Summary so far:

chain rule + Important weighting + Recursion:

Summary so far:

chain rule + Important weighting + Recursion:

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

Summary so far:

chain rule + Important weighting + Recursion:

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

for finite horizon MDP (try this out!)

$$\nabla_\theta J(\pi_\theta) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim\mathbb{P}_h^{\pi_\theta}} \left[\nabla \ln \pi_\theta(a_h | s_h) \cdot Q_h^{\pi_\theta}(s_h, a_h) \right]$$

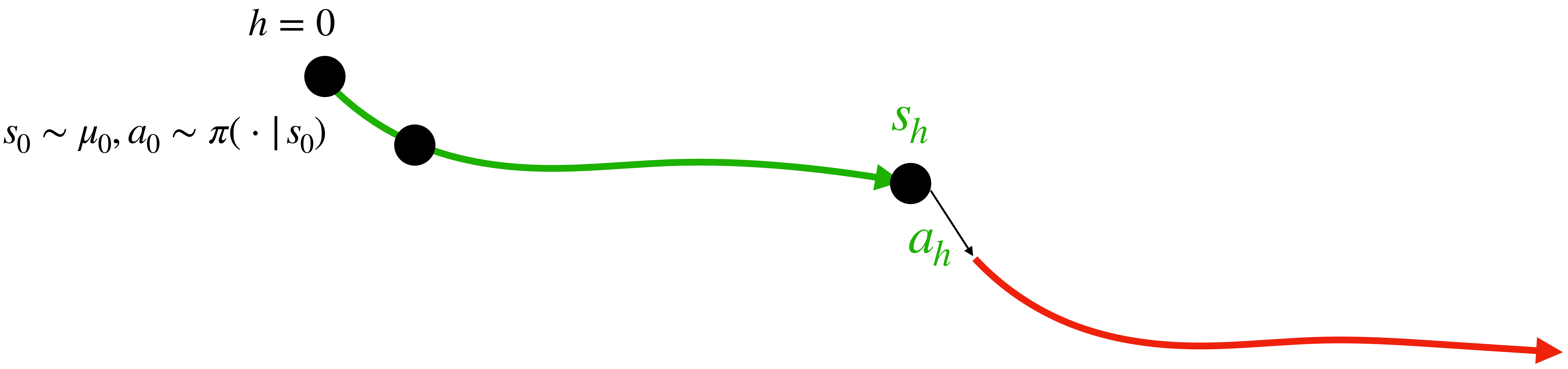
Summary so far:

chain rule + Important weighting + Recursion:

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

for finite horizon MDP (try this out!)

$$\nabla_\theta J(\pi_\theta) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim \mathbb{P}_h^{\pi_\theta}} \left[\nabla \ln \pi_\theta(a_h | s_h) \cdot Q_h^{\pi_\theta}(s_h, a_h) \right]$$



Outline:

✓ 1. A $Q(s, a)$ based Policy Gradient

2. Variance Reduction via A Baseline

Q-based PG with a baseline

Q-based PG with a baseline

$$\nabla_\theta J(\pi_\theta) = \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a\sim\mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot Q^{\pi_\theta}(s,a)$$

Q-based PG with a baseline

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

This still contains the entire future, could still have high variance

The Advantage-based PG:

The Advantage-based PG:

Denote $b(s)$ as a state-dependent ***baseline, turns out that***

The Advantage-based PG:

Denote $b(s)$ as a state-dependent ***baseline, turns out that***

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

The Advantage-based PG:

Denote $b(s)$ as a state-dependent ***baseline, turns out that***

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

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\nabla_\theta \ln \pi_\theta(a\,|\,s) b(s)$$

The Advantage-based PG:

Denote $b(s)$ as a state-dependent ***baseline, turns out that***

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

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\nabla_\theta \ln \pi_\theta(a\,|\,s) b(s)$$

$$= \sum_a \pi_\theta(a\,|\,s)\frac{\nabla \pi_\theta(a\,|\,s)}{\pi_\theta(a\,|\,s)} b(s)$$

The Advantage-based PG:

Denote $b(s)$ as a state-dependent ***baseline, turns out that***

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

$$\mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)} \nabla_\theta \ln \pi_\theta(a \mid s) b(s)$$

$$= \sum_a \pi_\theta(a \mid s) \frac{\nabla \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)} b(s) \quad = b(s) \sum_a \nabla \pi_\theta(a \mid s) = b(s) \nabla \left[\sum_a \pi_\theta(a \mid s) \right]$$

The Advantage-based PG:

Denote $b(s)$ as a state-dependent ***baseline, turns out that***

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

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\nabla_\theta \ln \pi_\theta(a\,|\,s) b(s)$$

$$= \sum_a \pi_\theta(a\,|\,s)\frac{\nabla \pi_\theta(a\,|\,s)}{\pi_\theta(a\,|\,s)} b(s) \quad = b(s)\sum_a \nabla \pi_\theta(a\,|\,s) = b(s)\nabla\left[\sum_a \pi_\theta(a\,|\,s)\right] \quad = b(s)\nabla 1 = 0$$

Baseline can further reduce the variance

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

Baseline can further reduce the variance

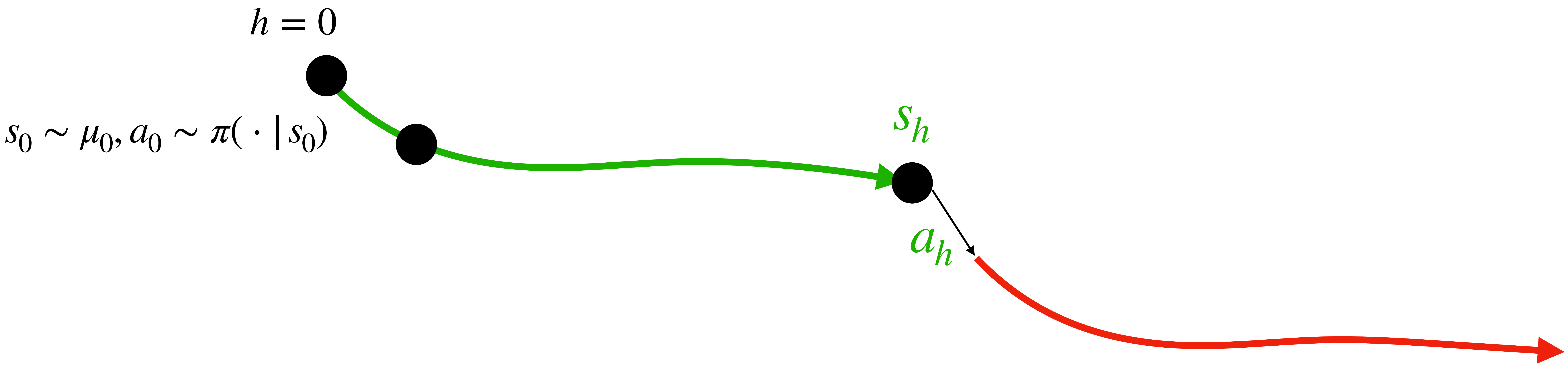
$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}}\left[\nabla_\theta \ln \pi_\theta(a\,|\,s)\cdot\left(Q^{\pi_\theta}(s,a) - b(s)\right)\right]$$

Baseline helps reduce variance (formal proof out of scope), and a common choice is $V^{\pi_\theta}(s)$:

Baseline can further reduce the variance

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

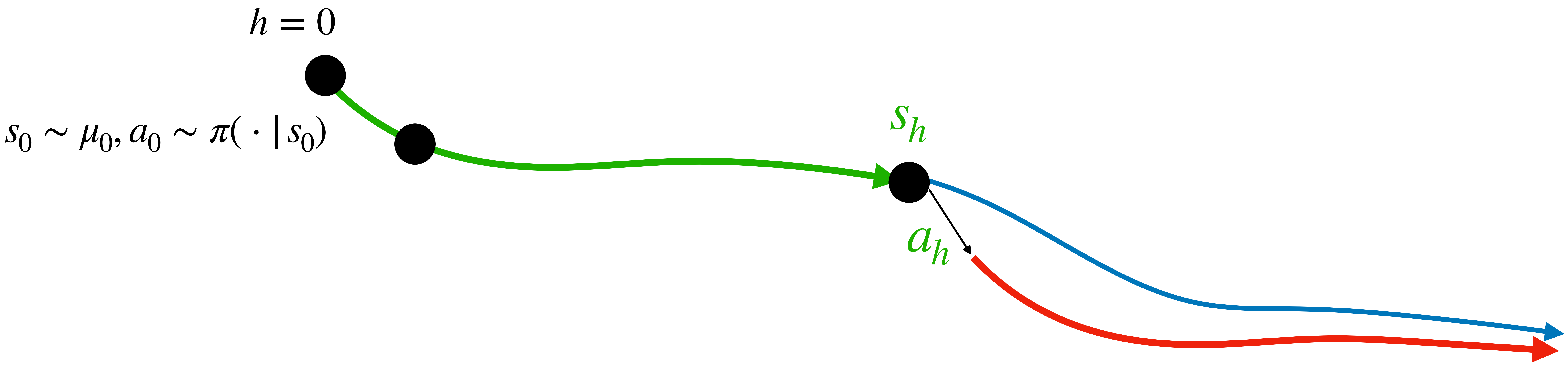
Baseline helps reduce variance (formal proof out of scope), and a common choice is $V^{\pi_\theta}(s)$:



Baseline can further reduce the variance

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

Baseline helps reduce variance (formal proof out of scope), and a common choice is $V^{\pi_\theta}(s)$:



Summary for PG:

Three common PG formulations:

REINFORCE

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

Summary for PG:

Three common PG formulations:

REINFORCE

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

PG w/ Q function

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

Summary for PG:

Three common PG formulations:

REINFORCE

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

PG w/ Q function

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

PG w/ A function (use $V^\pi(s)$ as a baseline)

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