Policy Gradient (continue)
Recap: Policy Parameterization

Recall that we consider parameterized policy $\pi_\theta(\cdot | s) \in \Delta(A), \forall s$

1. Softmax linear Policy
(We will try this in HW2)

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

$$\pi_\theta(a | s) = \frac{\exp(\theta^T \phi(s, a))}{\sum_{a'} \exp(\theta^T \phi(s, a'))}$$

2. Neural Policy:

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

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

In high level, think about $\pi_\theta$ as a classifier which has its parameters to be optimized.
Recap: the REINFORCE Algorithm

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

\[ \rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 | s_0)P(s_1 | s_0, a_0)\pi_\theta(a_1 | s_1)\ldots \]
Recap: the REINFORCE Algorithm

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

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

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

\[ R(\tau) \]
Recap: the REINFORCE Algorithm

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

\[ \rho_\theta(\tau) = \mu(s_0) \pi_\theta(a_0 | s_0) P(s_1 | s_0, a_0) \pi_\theta(a_1 | s_1) \ldots \]

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

\[ \nabla_\theta J(\pi_\theta) \big|_{\theta = \theta_0} := \mathbb{E}_{\tau \sim \rho_{\theta_0}(\tau)} \left[ \left( \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_{\theta_0}(a_h | s_h) \right) R(\tau) \right] \]
Recap: the REINFORCE Algorithm

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

How to get an unbiased estimate of the PG?
Recap: the REINFORCE Algorithm

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

How to get an unbiased estimate of the PG?

\[ \tau \sim \rho_{\theta_0} \]
Recap: the REINFORCE Algorithm

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

How to get an unbiased estimate of the PG?

$$\tau \sim \rho_{\theta_0}$$

$$g := \sum_{h=0}^{H-1} \left[ \nabla \ln \pi_{\theta_0}(a_h | s_h) R(\tau) \right]$$
Recap: the REINFORCE Algorithm

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

How to get an unbiased estimate of the PG?

\[ \tau \sim \rho_{\theta_0} \]

\[ g := \sum_{h=0}^{H-1} \left[ \nabla \ln \pi_{\theta_0}(a_h | s_h) R(\tau) \right] \]

We have: \( \mathbb{E}[g] = \nabla_\theta J(\pi_{\theta_0}) \)
Recap: the REINFORCE Algorithm

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

How to get an unbiased estimate of the PG?

\[ \tau \sim \rho_{\theta_0} \]

\[ g := \sum_{h=0}^{H-1} \left[ \nabla \ln \pi_{\theta_0}(a_h | s_h) R(\tau) \right] \]

We have: \( \mathbb{E}[g] = \nabla_\theta J(\pi_{\theta_0}) \)

This formulation has large variance, i.e.,

\[ \mathbb{E} \left[ \| g - \nabla_\theta J(\pi_{\theta_0}) \|^2_2 \right] \]

could be as large as \( H^3 \)

(In practice, no one uses it)
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
   (i.e., an $A(s, a)$ based PG)

3. Algorithm: Put everything together
Notations

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

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

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

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

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

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

\[ d^\pi_\mu(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}^\pi_h(s, a; \mu) \]

\[ A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s, a) \]
Derivation of Policy Gradient w/ $Q^\pi$

Recall definition of value function $V^{\pi_\theta}(s)$
Derivation of Policy Gradient w/ $Q^\pi$

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/ $Q^\pi$

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/ $Q^\pi$

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

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

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

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

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

$$
= \mathbb{E}_{s_0 \sim \mu} \left[ \sum_{a_0 \in A} \pi_\theta(a_0 | s_0) \left( \frac{\nabla_\theta \pi_\theta(a_0 | s_0)}{\pi_\theta(a_0 | s_0)} \right) \cdot Q^\pi_\theta(s_0, a_0) + \sum_{a_0} \pi_\theta(a_0 | s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^\pi_\theta(s_1) \right]
$$
Derivation of Policy Gradient w/ $Q^\pi$

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

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

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

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

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

Recall definition of value function $V^\pi(s)$

$$
\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{s_0 \sim \mu} \left[ V^\pi(s_0) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^\pi_\theta(s_0, a_0) \right] = \mathbb{E}_{s_0 \sim \mu} \left[ \sum_{a_0} \nabla_\theta \pi_\theta(a_0 \mid s_0) \cdot Q^\pi_\theta(s_0, a_0) + \sum_{a_0} \pi_\theta(a_0 \mid s_0) \cdot \nabla_\theta Q^\pi_\theta(s_0, a_0) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \sum_{a_0 \in A} \pi_\theta(a_0 \mid s_0) \left( \frac{\nabla_\theta \pi_\theta(a_0 \mid s_0)}{\pi_\theta(a_0 \mid s_0)} \right) \cdot Q^\pi_\theta(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 \mid s_0) \mathbb{E}_{s_1 \sim P_{s_0,a_0}} \nabla_\theta V^\pi_\theta(s_1) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) \cdot Q^\pi_\theta(s_0, a_0) \right] + \gamma \mathbb{E}_{s_1 \sim P_{s_0,a_0}} \nabla_\theta V^\pi_\theta(s_1) \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \mathbb{E}_{a_0 \sim \pi_\theta(a_0 \mid s_0)} \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) Q^\pi_\theta(s_0, a_0) \right] + \gamma \mathbb{E}_{s_1 \sim P_{s_0,a_0}} \left[ \mathbb{E}_{a_1 \sim \pi_\theta(a_1 \mid s_1)} \nabla_\theta \ln \pi_\theta(a_1 \mid s_1) Q^\pi_\theta(s_1, a_1) \right] + \gamma^2 \mathbb{E}_{s_2 \sim P_{s_1,a_1}} \nabla_\theta V^\pi_\theta(s_2) 
$$
**Derivation of Policy Gradient w/ \( Q^\pi \)**

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

\[
\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{s_0 \sim \mu} \left[ V^\pi_\theta(s_0) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)} Q^\pi_\theta(s_0, a_0) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \sum_{a_0} \nabla_\theta \pi_\theta(a_0 | s_0) \cdot Q^\pi_\theta(s_0, a_0) + \sum_{a_0} \pi_\theta(a_0 | s_0) \cdot \nabla_\theta Q^\pi_\theta(s_0, a_0) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \sum_{a_0 \in A} \pi_\theta(a_0 | s_0) \cdot \frac{\nabla_\theta \pi_\theta(a_0 | s_0)}{\pi_\theta(a_0 | s_0)} \cdot Q^\pi_\theta(s_0, a_0) + \gamma \sum_{a_0} \pi_\theta(a_0 | s_0) \mathbb{E}_{s_1 \sim P_{s_0, a_0}} \nabla_\theta V^\pi_\theta(s_1) \right] \\
= \mathbb{E}_{s_0 \sim \mu} \left[ \mathbb{E}_{a_0 \sim \pi_\theta(a_0 | s_0)} \nabla_\theta \ln \pi_\theta(a_0 | s_0) \cdot Q^\pi_\theta(s_0, a_0) \right] + \gamma \mathbb{E}_{s_1 \sim P^1_\pi} \nabla_\theta V^\pi_\theta(s_1) \\
= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s_h, a_h \sim P_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a_h | s_h) \cdot Q^\pi_\theta(s_h, a_h) 
\]
Derivation of Policy Gradient w/ $Q^\pi$

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

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

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

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

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

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

$$= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s_h \sim a_h \sim p_{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \cdot Q^\pi_\theta(s_h, a_h) = \frac{1}{1 - \gamma} \mathbb{E}_{s, a \sim d_{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a \mid s) \cdot Q^\pi_\theta(s, a)$$
Summary so far:

Product rule + Important weighting + Recursion:
Summary so far:

Product rule + Important weighting + Recursion:

\[
\nabla_{\theta} J(\pi_{\theta}) = \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a \sim \mathbb{D}_{\pi_{\theta}}^h} \nabla_{\theta} \ln \pi_{\theta}(a | s) \cdot Q_{\pi_{\theta}}(s, a)
\]

\[
= \frac{1}{1 - \gamma} \mathbb{E}_{s,a \sim \mathbb{d}_{\pi_{\theta}}^P} \left[ \nabla_{\theta} \ln \pi_{\theta}(a | s) \cdot Q_{\pi_{\theta}}(s, a) \right]
\]
Summary so far:

Product rule + Important weighting + Recursion:

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

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

For finite horizon setting, we have:

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

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

2. Variance Reduction via A Baseline
   (i.e., an $A(s, a)$ based PG)

3. Algorithm: Put everything together
Intuition behind Q-based PG:

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

We want to slowly adjust policy, such that \( \pi_\theta(a | s) \) is large at action \( a \) with large \( Q^\pi_\theta(s, a) \)
Intuition behind Q-based PG:

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

We want to slowly adjust policy, such that \( \pi_\theta(a | s) \) is large at action \( a \) with large \( Q^{\pi_\theta}(s, a) \)

Maybe we can slowly adjust policy, such that \( \pi_\theta(a | s) \) is large at action \( a \) with large \( A^{\pi_\theta}(s, a) \)?
Intuition behind Q-based PG:

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

We want to slowly adjust policy, such that \( \pi_\theta(a | s) \) is large at action \( a \) with large \( Q^{\pi_\theta}(s, a) \)

Maybe we can slowly adjust policy, such that \( \pi_\theta(a | s) \) is large at action \( a \) with large \( A^{\pi_\theta}(s, a) \)?

After all, recall PI, we know that \( \arg \max_a A^{\pi_\theta}(s, a) \) can work (subject to knowing \( A^{\pi_\theta} \) everywhere)
The Advantage-based PG:

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1 - \gamma} \mathbb{E}_{s,a \sim d_\mu} \left[ \nabla_\theta \ln \pi_\theta(a | s) \cdot A_\pi(s, a) \right]$$
The Advantage-based PG:

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

We will prove a more general version, denote $b(s)$ as a state-dependent baseline, we have:
The Advantage-based PG:

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

We will prove a more general version, denote \( b(s) \) as a state-dependent \textbf{baseline, we have:}

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

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

We will prove a more general version, denote \( b(s) \) as a state-dependent baseline, we have:

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

\[ \mathbb{E}_{a \sim \pi_\theta(a \mid s)} \nabla_\theta \ln \pi_\theta(a \mid s) b(s) \]
The Advantage-based PG:

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

We will prove a more general version, denote $b(s)$ as a state-dependent baseline, we have:

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

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

$$= \sum_a \pi_\theta(a | s) \frac{\nabla \pi_\theta(a | s)}{\pi_\theta(a | s)} b(s)$$
The Advantage-based PG:

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

We will prove a more general version, denote \( b(s) \) as a state-dependent \textbf{baseline}, \textbf{we have}:

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

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

\[
= \sum_a \pi_\theta(a \mid s) \frac{\nabla \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)} b(s) = b(s) \sum_a \nabla \pi_\theta(a \mid s) = b(s) \nabla \left[ \sum_a \pi_\theta(a \mid s) \right]
\]
The Advantage-based PG:

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

We will prove a more general version, denote \( b(s) \) as a state-dependent \textbf{baseline}, we have:

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

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

\[
= \sum_a \pi_\theta(a \mid s) \frac{\nabla \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)} b(s) = b(s) \sum_a \nabla \pi_\theta(a \mid s) = b(s) \nabla \left[ \sum_a \pi_\theta(a \mid s) \right] = b(s) \nabla 1 = 0
\]
Summary so far:

By a Baseline (proof undoes the importance weighting trick), we have:

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1 - \gamma} \mathbb{E}_{s,a \sim d_\mu^{\pi_\theta}} \left[ \nabla_\theta \ln \pi_\theta(a | s) \cdot (Q_\pi^{\pi_\theta}(s, a) - b(s)) \right]$$
By a Baseline (proof undoes the importance weighting trick), we have:

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

This holds for any baseline as long as it is action-independent (thus we can set \( b(s) = V^\pi_\theta(s) \)—the most common thing)
By a Baseline (proof undoes the importance weighting trick), we have:

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

This holds for any baseline as long as it is action-independent
(thus we can set $b(s) = V^{\pi_\theta}(s)$—the most common thing)

Baseline helps variance reduction (formal proof out of scope)
Outline:

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

2. Variance Reduction via A Baseline (i.e., an $A(s, a)$ based PG)

3. Algorithm: Put everything together
Algorithm that relies on Stochastic Gradient Ascent

Recall the PG:

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

Recall the PG:

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

To get unbiased estimate of gradient, recall we can roll-in \((s, a) \sim d^\pi_{\mu}\), and roll out to get \(y\) w/ \(\mathbb{E}[y] = Q^\pi_{\theta}(s, a)\)
Algorithm that relies on Stochastic Gradient Ascent

Recall the PG:

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

To get unbiased estimate of gradient, recall we can roll-in $(s, a) \sim d_\mu^\pi$, and roll out to get $y$ w/ $\mathbb{E}[y] = Q_\pi^\theta(s, a)$

If terminate (w/ p $1 - \gamma$), we return

$$y := \sum_{i=h}^{t} r_i$$
Algorithm that relies on Stochastic Gradient Ascent

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

To get unbiased estimate of gradient, recall we can
roll-in \((s, a) \sim d_\mu^\pi\), and roll out to get \(y\) w/ \(\mathbb{E}[y] = Q_{\pi_\theta}(s, a)\)

Repeat roll-in & roll-out \(N\) times, with the mini-batch \(\{s^i, a^i, y^i\}_{i=1}^N\),
Algorithm that relies on Stochastic Gradient Ascent

Recall the PG:

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

To get unbiased estimate of gradient, recall we can roll-in \((s, a) \sim d_\mu^{\pi_\theta}\), and roll out to get \(y \) w/ \(\mathbb{E}[y] = Q^{\pi_\theta}(s,a)\)

If terminate (w/ \(p \ 1 - \gamma\)), we return

\[ y := \sum_{i=h}^{t} r_i \]

Repeat roll-in & roll-out \(N\) times, with the mini-batch \(\{s^i, a^i, y^i\}_{i=1}^{N}\),

\[ g = \sum_{i=1}^{N} \frac{1}{N} \left[ \nabla_{\theta} \ln \pi_\theta(a^i \mid s^i) \cdot y^i \right] \]
Algorithm that relies on Stochastic Gradient Ascent

Initialization $\theta_0$

For $t = 0, \ldots$

Sample $\{s_i, a_i, y_i\}_{i=1}^N$, w/ $s_i, a_i \sim d_{\mu}^{\pi_{\theta_t}}$, $\mathbb{E}[y_i] = Q^{\pi_{0_t}}(s^i, a^i)$

Form gradient estimate: $g_t = \sum_{i=1}^N \nabla_{\theta} \ln \pi_{\theta_t}(a_i | s^i) \cdot y^i / N$

Stochastic GA: $\theta_{t+1} = \theta_t + \eta g_t$
In practice, we often use supervised learning to estimate $Q^{\pi_{\theta^*}}$

$$\nabla_{\theta} J(\theta) = \frac{1}{1 - \gamma} \mathbb{E}_{s, a \sim d_{\pi_{\theta}}} \left[ \nabla_{\theta} \ln \pi_{\theta}(a \mid s) \left( Q^{\pi_{\theta}}(s, a) - V^{\pi_{\theta}}(s) \right) \right]$$
In practice, we often use supervised learning to estimate $Q^{\pi_{\theta^*}}$:

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

$$\hat{f} = \arg\min_f \mathbb{E}_{s \sim d_{\mu}, a \sim U(A)} \left( f(s,a) - Q^{\pi_{\theta}}(s,a) \right)^2 \text{ (e.g., regression oracle!)}$$
In practice, we often use supervised learning to estimate $Q^{\pi_{\theta^*}}$:

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

$$\hat{f} = \arg \min_{f} \mathbb{E}_{s \sim d_{\mu}, a \sim U(A)} (f(s, a) - Q^{\pi_{\theta}}(s, a))^2 \text{ (e.g., regression oracle!)}$$

We can form an approximated Gradient (could be unbiased) using $\hat{f}$:

$$\nabla_{\theta} \ln \pi_{\theta}(a_h \mid s_h) \left( \hat{f}(s_h, a_h) - \mathbb{E}_{a' \sim \pi_{\theta}(a' \mid s_h)} \hat{f}(s_h, a') \right)$$
In practice, we often use supervised learning to estimate $Q^{\pi_\theta}$:

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

$$\hat{f} = \arg \min_f \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim U(A)} \left( f(s, a) - Q^{\pi_\theta}(s, a) \right)^2 \text{ (e.g., regression oracle!)}$$

We can form an approximated Gradient (could be unbiased) using $\hat{f}$:

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

Bisa-variance tradeoff

(our $\hat{f}$ is a function now, we no-longer rely on a roll-out)
Summary for PG:

Three common PG formulations:

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

REINFORCE
Summary for PG:

Three common PG formulations:

**REINFORCE**

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

**PG w/ Q function**

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

Three common PG formulations:

**REINFORCE**

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

**PG w/ \( Q \) function**

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

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

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

Trust-region policy optimization (Natural Policy Gradient)