# ECEN 743: Reinforcement Learning

## Policy Gradient Theorem

Dileep Kalathil
Assistant Professor
Department of Electrical and Computer Engineering
Texas A&M University

# References

- [SB, Chapter 13]
- [AJKS, Section 3]

# Reinforcement Learning Questions

- How do we learn the value of a policy $\pi$?
- How do we learn the optimal action-value function $Q^*$?
- How do we learn the optimal policy $\pi^*$?

... without the knowledge of the model $P$

- Need to learn from the observed sequence of states, actions, and rewards

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$
- Define $V_{\pi,\mu_0} = \mathbb{E}_{s\sim\mu_0(\cdot)}\left[V_\pi(s)\right]$

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$
- Define $V_{\pi,\mu_0} = \mathbb{E}_{s \sim \mu_0(\cdot)}\left[V_\pi(s)\right]$
  - We will often just use $V_\pi$, making the dependence on $\mu_0$ implicit

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$
- Define $V_{\pi,\mu_0} = \mathbb{E}_{s\sim\mu_0(\cdot)}\left[V_\pi(s)\right]$
  - We will often just use $V_\pi$, making the dependence on $\mu_0$ implicit

- Consider the parameterized policies $\pi_\theta$, where $\theta \in \Theta \subset \mathbb{R}^d$

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$
- Define $V_{\pi,\mu_0} = \mathbb{E}_{s\sim\mu_0(\cdot)}\left[V_\pi(s)\right]$
  - We will often just use $V_\pi$, making the dependence on $\mu_0$ implicit

- Consider the parameterized policies $\pi_\theta$, where $\theta \in \Theta \subset \mathbb{R}^d$
- $V_{\pi_\theta}$ is the value of the policy $\pi_\theta$

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$
- Define $V_{\pi,\mu_0} = \mathbb{E}_{s \sim \mu_0(\cdot)}\left[V_\pi(s)\right]$
  - We will often just use $V_\pi$, making the dependence on $\mu_0$ implicit

- Consider the parameterized policies $\pi_\theta$, where $\theta \in \Theta \subset \mathbb{R}^d$
- $V_{\pi_\theta}$ is the value of the policy $\pi_\theta$
- Policy optimization problem:

$$\max_{\theta \in \Theta} \; V_{\pi_\theta}$$

# Finding the Optimal Policy Parameter

- Value of a policy; $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$
- Assume that the initial state $s_0$ is sampled according to a distribution $\mu_0$
- Define $V_{\pi,\mu_0} = \mathbb{E}_{s\sim\mu_0(\cdot)}\left[V_\pi(s)\right]$
  - We will often just use $V_\pi$, making the dependence on $\mu_0$ implicit

- Consider the parameterized policies $\pi_\theta$, where $\theta \in \Theta \subset \mathbb{R}^d$
- $V_{\pi_\theta}$ is the value of the policy $\pi_\theta$
- Policy optimization problem:

$$\max_{\theta\in\Theta} \; V_{\pi_\theta}$$

- Can we develop (stochastic) gradient algorithms for solving the above problem?

# Policy Parameterization

- Tabular representation

$$\pi(s,a) = \theta_{s,a}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

# Policy Parameterization

- Tabular representation

$$\pi(s,a) = \theta_{s,a}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies for tabular representation

$$\pi(s,a) = \frac{\exp(\theta_{s,a})}{\sum_b \exp(\theta_{s,b})}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

# Policy Parameterization

- Tabular representation

$$\pi(s,a) = \theta_{s,a}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies for tabular representation

$$\pi(s,a) = \frac{\exp(\theta_{s,a})}{\sum_b \exp(\theta_{s,b})}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies / log-linear policies
  Consider the basis function $\phi(s,a) \in \mathbb{R}^d$, $d \ll |\mathcal{S}||\mathcal{A}|$. Let $\theta \in \mathbb{R}^d$. Then,

$$\pi(s,a) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_b \exp(\theta^\top \phi(s,b))}$$

# Policy Parameterization

- Tabular representation

$$\pi(s,a) = \theta_{s,a}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies for tabular representation

$$\pi(s,a) = \frac{\exp(\theta_{s,a})}{\sum_b \exp(\theta_{s,b})}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies / log-linear policies
  Consider the basis function $\phi(s,a) \in \mathbb{R}^d$, $d \ll |\mathcal{S}||\mathcal{A}|$. Let $\theta \in \mathbb{R}^d$. Then,

$$\pi(s,a) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_b \exp(\theta^\top \phi(s,b))}$$

- Neural softmax policies

$$\pi(s,a) = \frac{\exp(f_\theta(s,a))}{\sum_b \exp(f_\theta(s,b))}$$

# Policy Parameterization

- Tabular representation

$$\pi(s,a) = \theta_{s,a}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies for tabular representation

$$\pi(s,a) = \frac{\exp(\theta_{s,a})}{\sum_b \exp(\theta_{s,b})}, \quad \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

- Softmax policies / log-linear policies
  Consider the basis function $\phi(s,a) \in \mathbb{R}^d$, $d \ll |\mathcal{S}||\mathcal{A}|$. Let $\theta \in \mathbb{R}^d$. Then,

$$\pi(s,a) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_b \exp(\theta^\top \phi(s,b))}$$

- Neural softmax policies

$$\pi(s,a) = \frac{\exp(f_\theta(s,a))}{\sum_b \exp(f_\theta(s,b))}$$

- Gaussian policies

$$\pi(s,a) = \mathcal{N}(\mu_\theta(s), \Sigma_\theta(s))$$

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy
- We can directly compute $\nabla_\theta \pi_\theta$

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy
- We can directly compute $\nabla_\theta \pi_\theta$
- But, the gradient of the (stationary) distribution of the states induced by the policy cannot be computed easily

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy
- We can directly compute $\nabla_\theta \pi_\theta$
- But, the gradient of the (stationary) distribution of the states induced by the policy cannot be computed easily
  - Requires the knowledge of the transition probability, which is unknown

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy
- We can directly compute $\nabla_\theta \pi_\theta$
- But, the gradient of the (stationary) distribution of the states induced by the policy cannot be computed easily
  - Requires the knowledge of the transition probability, which is unknown
- How do we estimate the gradient?

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter

$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$

- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy
- We can directly compute $\nabla_\theta \pi_\theta$
- But, the gradient of the (stationary) distribution of the states induced by the policy cannot be computed easily
  - Requires the knowledge of the transition probability, which is unknown
- How do we estimate the gradient?

# Gradient of the Value Function

- Stochastic gradient ascent for learning the optimal policy parameter
$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta V_{\pi_\theta}}$$
- Computing $\nabla_\theta V_{\pi_\theta}$ is challenging
- Value of a policy depends on both action selections (policy) and the (stationary) distribution of the states resulting from the policy
- We can directly compute $\nabla_\theta \pi_\theta$
- But, the gradient of the (stationary) distribution of the states induced by the policy cannot be computed easily
    - Requires the knowledge of the transition probability, which is unknown
- How do we estimate the gradient?

Policy Gradient Theorem!

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T), \;\; a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot | s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T),\ \ a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\mathrm{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T),\ \ a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\mathrm{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T),\; a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\mathrm{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t)\; P(s_{t+1}|s_t, a_t)$$

$$P_{\mathrm{traj},\theta}(\tau) = \mathbb{P}(s_T, a_T, \tau[0:T-1]) = \mathbb{P}(s_T, a_T|\tau[0:T-1])\; P_{\mathrm{traj},\theta}(\tau[0:T-1])$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T),\ a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\text{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)$$

$$\begin{aligned} P_{\text{traj},\theta}(\tau) &= \mathbb{P}(s_T, a_T, \tau[0:T-1]) = \mathbb{P}(s_T, a_T|\tau[0:T-1])\ P_{\text{traj},\theta}(\tau[0:T-1]) \\ &= \mathbb{P}(a_T|s_T, \tau[0:T-1])\ \mathbb{P}(s_T|\tau[0:T-1])\ P_{\text{traj},\theta}(\tau[0:T-1]) \end{aligned}$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T), \;\; a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\mathrm{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t) \; P(s_{t+1}|s_t, a_t)$$

$$\begin{aligned}
&P_{\mathrm{traj},\theta}(\tau) = \mathbb{P}(s_T, a_T, \tau[0:T-1]) = \mathbb{P}(s_T, a_T|\tau[0:T-1]) \; P_{\mathrm{traj},\theta}(\tau[0:T-1]) \\
&= \mathbb{P}(a_T|s_T, \tau[0:T-1]) \; \mathbb{P}(s_T|\tau[0:T-1]) \; P_{\mathrm{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \; P(s_T|s_{T-1}, a_{T-1}) \; P_{\mathrm{traj},\theta}(\tau[0:T-1])
\end{aligned}$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T), \;\; a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\text{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t) \; P(s_{t+1}|s_t, a_t)$$

$$\begin{aligned}
P_{\text{traj},\theta}(\tau) &= \mathbb{P}(s_T, a_T, \tau[0:T-1]) = \mathbb{P}(s_T, a_T|\tau[0:T-1]) \; P_{\text{traj},\theta}(\tau[0:T-1]) \\
&= \mathbb{P}(a_T|s_T, \tau[0:T-1]) \; \mathbb{P}(s_T|\tau[0:T-1]) \; P_{\text{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \; P(s_T|s_{T-1}, a_{T-1}) \; P_{\text{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \; P(s_T|s_{T-1}, a_{T-1}) \; P_{\text{traj},\theta}(s_{T-1}, a_{T-1}, \tau[0:T-2])
\end{aligned}$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,
$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T), \; a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$
- What is the probability of the trajectory $\tau$?

$$P_{\text{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t) \; P(s_{t+1}|s_t, a_t)$$

$$\begin{aligned}
P_{\text{traj},\theta}(\tau) &= \mathbb{P}(s_T, a_T, \tau[0:T-1]) = \mathbb{P}(s_T, a_T|\tau[0:T-1]) \; P_{\text{traj},\theta}(\tau[0:T-1]) \\
&= \mathbb{P}(a_T|s_T, \tau[0:T-1]) \; \mathbb{P}(s_T|\tau[0:T-1]) \; P_{\text{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \; P(s_T|s_{T-1}, a_{T-1}) \; P_{\text{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \; P(s_T|s_{T-1}, a_{T-1}) \; P_{\text{traj},\theta}(s_{T-1}, a_{T-1}, \tau[0:T-2]) \\
&= \pi_\theta(s_T, a_T) \; P(s_T|s_{T-1}, a_{T-1}) \; \pi_\theta(s_{T-1}, a_{T-1}) \; P(s_{T-1}|s_{T-2}, a_{T-2}) \; P_{\text{traj},\theta}(\tau[0:T-2])
\end{aligned}$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T), \ a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot | s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\mathrm{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t) \ P(s_{t+1} | s_t, a_t)$$

$$\begin{aligned}
&P_{\mathrm{traj},\theta}(\tau) = \mathbb{P}(s_T, a_T, \tau[0:T-1]) = \mathbb{P}(s_T, a_T | \tau[0:T-1]) \ P_{\mathrm{traj},\theta}(\tau[0:T-1]) \\
&= \mathbb{P}(a_T | s_T, \tau[0:T-1]) \ \mathbb{P}(s_T | \tau[0:T-1]) \ P_{\mathrm{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \ P(s_T | s_{T-1}, a_{T-1}) \ P_{\mathrm{traj},\theta}(\tau[0:T-1]) \\
&= \pi_\theta(s_T, a_T) \ P(s_T | s_{T-1}, a_{T-1}) \ P_{\mathrm{traj},\theta}(s_{T-1}, a_{T-1}, \tau[0:T-2]) \\
&= \pi_\theta(s_T, a_T) \ P(s_T | s_{T-1}, a_{T-1}) \ \pi_\theta(s_{T-1}, a_{T-1}) \ P(s_{T-1} | s_{T-2}, a_{T-2}) \ P_{\mathrm{traj},\theta}(\tau[0:T-2]) \\
&= \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t) \ P(s_{t+1} | s_t, a_t)
\end{aligned}$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T), \;\; a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot | s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\text{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t) \;\, P(s_{t+1} | s_t, a_t)$$

# Probability of a Trajectory

- Let $\tau$ be a trajectory generated using $\pi_\theta$,

$$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T),\ a_t \sim \pi_\theta(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 \sim \mu_0(\cdot)$$

- What is the probability of the trajectory $\tau$?

$$P_{\text{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t=0}^{T} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)$$

- We can extend this to arbitrary long trajectories, $\tau = (s_t, a_t)_{t\geq 0}$,

$$P_{\text{traj},\theta}(\tau) = \mu_0(s_0) \prod_{t\geq 0} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)$$

# Gradient of the Value Function: Form 1

- Let $\tau = (s_t, a_t)_{t \geq 0}$, $s_0 \sim \mu_0(\cdot)$. Denote the cumulative discounted reward of trajectory $\tau$ as

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

# Gradient of the Value Function: Form 1

- Let $\tau = (s_t, a_t)_{t\geq 0}$, $s_0 \sim \mu_0(\cdot)$. Denote the cumulative discounted reward of trajectory $\tau$ as
$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$
- It is straight forward to note that
$$V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\cdot)} [R(\tau)]$$

# Gradient of the Value Function: Form 1

- Let $\tau = (s_t, a_t)_{t \geq 0}$, $s_0 \sim \mu_0(\cdot)$. Denote the cumulative discounted reward of trajectory $\tau$ as
$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$
- It is straight forward to note that
$$V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} [R(\tau)]$$

**Theorem (Policy Gradient Theorem - Form 1)**

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(s_t, a_t) \right]$$

# Gradient of the Value Function: Form 1

- Let $\tau = (s_t, a_t)_{t \geq 0}$, $s_0 \sim \mu_0(\cdot)$. Denote the cumulative discounted reward of trajectory $\tau$ as
$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$
- It is straight forward to note that
$$V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} [R(\tau)]$$

**Theorem (Policy Gradient Theorem - Form 1)**

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- Gradient above does not depend on the gradient of the (stationary) distribution induced by the policy!

# Proof of Policy Gradient Theorem: Form 1

Proof:

# Proof of Policy Gradient Theorem: Form 1

Proof:

# Proof of Policy Gradient Theorem: Form 1

Proof:

$$\nabla V_{\pi_\theta} = \nabla \mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\cdot)}[R(\tau)] = \nabla \sum_{\tau} R(\tau) P_{\text{traj},\theta}(\tau)$$

# Proof of Policy Gradient Theorem: Form 1

Proof:

$$\begin{aligned}\nabla V_{\pi_\theta} &= \nabla \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)}\left[R(\tau)\right] = \nabla \sum_{\tau} R(\tau) P_{\mathrm{traj},\theta}(\tau) \\ &= \sum_{\tau} R(\tau)\ \nabla P_{\mathrm{traj},\theta}(\tau) = \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log P_{\mathrm{traj},\theta}(\tau)\end{aligned}$$

# Proof of Policy Gradient Theorem: Form 1

Proof:

$$\begin{aligned}\nabla V_{\pi_\theta} &= \nabla \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)}\left[R(\tau)\right] = \nabla \sum_{\tau} R(\tau) P_{\mathrm{traj},\theta}(\tau) \\ &= \sum_{\tau} R(\tau)\ \nabla P_{\mathrm{traj},\theta}(\tau) = \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log P_{\mathrm{traj},\theta}(\tau) \\ &= \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log\left(\mu_0(s_0)\ \prod_{t=0}^{\infty} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)\right)\end{aligned}$$

# Proof of Policy Gradient Theorem: Form 1

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \nabla \mathbb{E}_{\tau\sim P_{\mathrm{traj},\theta}(\cdot)}\left[R(\tau)\right] = \nabla \sum_{\tau} R(\tau) P_{\mathrm{traj},\theta}(\tau) \\
&= \sum_{\tau} R(\tau)\ \nabla P_{\mathrm{traj},\theta}(\tau) = \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log P_{\mathrm{traj},\theta}(\tau) \\
&= \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log \left(\mu_0(s_0)\ \prod_{t=0}^{\infty} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t)\right) \\
&= \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla(\sum_{t=0}^{\infty} \log \pi_\theta(s_t, a_t))
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 1

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \nabla \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)}\left[R(\tau)\right] = \nabla \sum_{\tau} R(\tau) P_{\mathrm{traj},\theta}(\tau) \\
&= \sum_{\tau} R(\tau)\ \nabla P_{\mathrm{traj},\theta}(\tau) = \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log P_{\mathrm{traj},\theta}(\tau) \\
&= \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla \log \left( \mu_0(s_0)\ \prod_{t=0}^{\infty} \pi_\theta(s_t, a_t)\ P(s_{t+1}|s_t, a_t) \right) \\
&= \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ \nabla (\sum_{t=0}^{\infty} \log \pi_\theta(s_t, a_t)) \\
&= \sum_{\tau} R(\tau)\ P_{\mathrm{traj},\theta}(\tau)\ (\sum_{t=0}^{\infty} \nabla \log \pi_\theta(s_t, a_t))
\end{aligned}
$$

$\square$

# Proof of Policy Gradient Theorem: Form 1

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \nabla \mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\cdot)}\left[R(\tau)\right] = \nabla \sum_{\tau} R(\tau) P_{\text{traj},\theta}(\tau) \\
&= \sum_{\tau} R(\tau)\ \nabla P_{\text{traj},\theta}(\tau) = \sum_{\tau} R(\tau)\ P_{\text{traj},\theta}(\tau)\ \nabla \log P_{\text{traj},\theta}(\tau) \\
&= \sum_{\tau} R(\tau)\ P_{\text{traj},\theta}(\tau)\ \nabla \log \left( \mu_0(s_0)\ \prod_{t=0}^{\infty} \pi_\theta(s_t, a_t)\ P(s_{t+1} | s_t, a_t) \right) \\
&= \sum_{\tau} R(\tau)\ P_{\text{traj},\theta}(\tau)\ \nabla (\sum_{t=0}^{\infty} \log \pi_\theta(s_t, a_t)) \\
&= \sum_{\tau} R(\tau)\ P_{\text{traj},\theta}(\tau)\ (\sum_{t=0}^{\infty} \nabla \log \pi_\theta(s_t, a_t)) \\
&= \mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(s_t, a_t) \right]
\end{aligned}
$$

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathtt{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?
  - Generate $n$ trajectories $(\tau_i)_{i=1}^{n}$

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?
  - Generate $n$ trajectories $(\tau_i)_{i=1}^n$
  - For each trajectory $\tau_i$, get $[R(\tau_i) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t^i, a_t^i)]$

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?
  - Generate $n$ trajectories $(\tau_i)_{i=1}^{n}$
  - For each trajectory $\tau_i$, get $[R(\tau_i) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t^i, a_t^i)]$
  - Average to get $\widehat{\nabla V_{\pi_\theta}}$

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?
  - Generate $n$ trajectories $(\tau_i)_{i=1}^{n}$
  - For each trajectory $\tau_i$, get $[R(\tau_i) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t^i, a_t^i)]$
  - Average to get $\widehat{\nabla V_{\pi_\theta}}$
- This approach, however, is unreliable due very high variance

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?
  - Generate $n$ trajectories $(\tau_i)_{i=1}^{n}$
  - For each trajectory $\tau_i$, get $[R(\tau_i) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t^i, a_t^i)]$
  - Average to get $\widehat{\nabla V_{\pi_\theta}}$
- This approach, however, is unreliable due very high variance
- This issue is avoided by exploiting the temporal structure and introducing a baseline

# REINFORCE: Monte Carlo Policy Gradient

- Consider a finite horizon (episodic) RL problem
- We have $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$, and

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t, a_t) \right]$$

- How do we get an estimate $\widehat{\nabla V_{\pi_\theta}}$?
  - Generate $n$ trajectories $(\tau_i)_{i=1}^n$
  - For each trajectory $\tau_i$, get $[R(\tau_i) \sum_{t=0}^{T} \nabla \log \pi_\theta(s_t^i, a_t^i)]$
  - Average to get $\widehat{\nabla V_{\pi_\theta}}$
- This approach, however, is unreliable due very high variance
- This issue is avoided by exploiting the temporal structure and introducing a baseline
  - Variance reduction for PG is an active area of research

# Visitation Distribution

- For a given policy $\pi$, (discounted) state-action visitation distribution is defined as
$$\rho_{\pi,\mu_0}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s, a_t = a),$$
where $(s_t, a_t)_{t\geq 0}$ is the trajectory generated according to $\pi$, and $s_0$ is sampled according to a given distribution $\mu_0$.

# Visitation Distribution

- For a given policy $\pi$, (discounted) state-action visitation distribution is defined as

$$\rho_{\pi,\mu_0}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s, a_t = a),$$

  where $(s_t, a_t)_{t\geq 0}$ is the trajectory generated according to $\pi$, and $s_0$ is sampled according to a given distribution $\mu_0$.
- We will often just use $\rho_\pi$, making the dependence on $\mu_0$ implicit.

# Visitation Distribution

- For a given policy $\pi$, (discounted) state-action visitation distribution is defined as
$$\rho_{\pi,\mu_0}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s, a_t = a),$$
where $(s_t, a_t)_{t\geq 0}$ is the trajectory generated according to $\pi$, and $s_0$ is sampled according to a given distribution $\mu_0$.
- We will often just use $\rho_\pi$, making the dependence on $\mu_0$ implicit.
- Similarly, for a given policy $\pi$, (discounted) state visitation distribution is defined as
$$\rho_{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s),$$
where $(s_t)_{t\geq 0}$ is the state-trajectory generated according to $\pi$ (we are *overloading* the notation)

# Visitation Distribution

- For a given policy $\pi$, (discounted) state-action visitation distribution is defined as
$$\rho_{\pi,\mu_0}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s, a_t = a),$$
where $(s_t, a_t)_{t\geq 0}$ is the trajectory generated according to $\pi$, and $s_0$ is sampled according to a given distribution $\mu_0$.
- We will often just use $\rho_\pi$, making the dependence on $\mu_0$ implicit.
- Similarly, for a given policy $\pi$, (discounted) state visitation distribution is defined as
$$\rho_{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s),$$
where $(s_t)_{t\geq 0}$ is the state-trajectory generated according to $\pi$ (we are *overloading* the notation)
- It is straight forward to show that
$$\rho_\pi(x,a) = \rho_\pi(x)\pi(x,a)$$

# Visitation Distribution and Value Function

**Lemma**

$$V_\pi = \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_\pi(\cdot,\cdot)}\left[r(s,a)\right] = \frac{1}{(1-\gamma)}\sum_{(s,a)}\rho_\pi(s,a)r(s,a)$$

# Visitation Distribution and Value Function

## Lemma

$$V_\pi = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\rho_\pi(\cdot,\cdot)}\left[r(s,a)\right] = \frac{1}{(1-\gamma)} \sum_{(s,a)} \rho_\pi(s,a) r(s,a)$$

Proof:

# Visitation Distribution and Value Function

## Lemma

$$V_\pi = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\rho_\pi(\cdot,\cdot)} \left[ r(s,a) \right] = \frac{1}{(1-\gamma)} \sum_{(s,a)} \rho_\pi(s,a) r(s,a)$$

*Proof:*

$$V_\pi = \mathbb{E}_{\tau\sim P_{\mathrm{traj},\pi}(\cdot)} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$$

□

# Visitation Distribution and Value Function

## Lemma

$$V_\pi = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\rho_\pi(\cdot,\cdot)}\left[r(s,a)\right] = \frac{1}{(1-\gamma)} \sum_{(s,a)} \rho_\pi(s,a) r(s,a)$$

**Proof:**

$$V_\pi = \mathbb{E}_{\tau\sim P_{\text{traj},\pi}(\cdot)}[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)]$$

$$= \sum_s \sum_a \sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\ \gamma^t r(s,a)$$

$\square$

# Visitation Distribution and Value Function

## Lemma

$$V_\pi = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\rho_\pi(\cdot,\cdot)} \left[r(s,a)\right] = \frac{1}{(1-\gamma)} \sum_{(s,a)} \rho_\pi(s,a) r(s,a)$$

**Proof:**

$$\begin{aligned}
V_\pi &= \mathbb{E}_{\tau\sim P_{\text{traj},\pi}(\cdot)}[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)] \\
&= \sum_s \sum_a \sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\ \gamma^t r(s,a) \\
&= \sum_s \sum_a r(s,a) (\sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\ \gamma^t)
\end{aligned}$$

□

# Visitation Distribution and Value Function

## Lemma

$$V_\pi = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\rho_\pi(\cdot,\cdot)} \left[ r(s,a) \right] = \frac{1}{(1-\gamma)} \sum_{(s,a)} \rho_\pi(s,a) r(s,a)$$

Proof:

$$\begin{aligned}
V_\pi &= \mathbb{E}_{\tau \sim P_{\text{traj},\pi}(\cdot)} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \\
&= \sum_s \sum_a \sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\ \gamma^t r(s,a) \\
&= \sum_s \sum_a r(s,a) (\sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\ \gamma^t) \\
&= \frac{1}{(1-\gamma)} \sum_x \sum_a r(s,a) \rho_\pi(s,a)
\end{aligned}$$

$\square$

# Gradient of the Value Function: Form 2

**Theorem (Policy Gradient Theorem - Form 2)**

$$\nabla V_{\pi_\theta} = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)} \left[ Q_{\pi_\theta}(s,a) \; \nabla \log \pi_\theta(s,a) \right]$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

# Proof of Policy Gradient Theorem: Form 2

Proof:

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$\nabla V_{\pi_\theta} = \sum_s \mathbb{P}(s_0 = s) \; \nabla V_{\pi_\theta}(s)$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$\nabla V_{\pi_\theta} = \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s)$$

$$= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)]$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s) \\
&= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \nabla \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)]
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s) \\
&= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \nabla \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)]
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s) \\
&= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \nabla \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla Q_{\pi_\theta}(s_0, a_0)]
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s) \\
&= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \nabla \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \\
&\qquad \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla (r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot | s_0, a_0)}[V_{\pi_\theta}(s_1)])
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s) \\
&= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \nabla \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \\
&\qquad \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla (r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot | s_0, a_0)}[V_{\pi_\theta}(s_1)])] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \\
&\qquad \gamma \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)} \mathbb{E}_{s_1 \sim P(\cdot | s_0, a_0)}[\nabla V_{\pi_\theta}(s_1)]
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

**Proof:**

$$
\begin{aligned}
\nabla V_{\pi_\theta} &= \sum_s \mathbb{P}(s_0 = s)\ \nabla V_{\pi_\theta}(s) \\
&= \nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] = \mathbb{E}_{s_0 \sim \mu_0}[\nabla \sum_{a_0} \pi_\theta(s_0, a_0) Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \nabla \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0}[\sum_{a_0} \pi_\theta(s_0, a_0)\ \nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla Q_{\pi_\theta}(s_0, a_0)] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \\
&\qquad \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla (r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot | s_0, a_0)}[V_{\pi_\theta}(s_1)])] \\
&= \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)}[\nabla \log \pi_\theta(s_0, a_0)\ Q_{\pi_\theta}(s_0, a_0)] + \\
&\qquad \gamma \mathbb{E}_{s_0 \sim \mu_0} \mathbb{E}_{a_0 \sim \pi_\theta(s_0, \cdot)} \mathbb{E}_{s_1 \sim P(\cdot | s_0, a_0)}[\nabla V_{\pi_\theta}(s_1)] \\
&= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s, a)\ Q_{\pi_\theta}(s, a)] + \gamma \sum_s \mathbb{P}(s_1 = s)\ \nabla V_{\pi_\theta}(s)
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

# Proof of Policy Gradient Theorem: Form 2

Proof:

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)]$$

$$+ \gamma \sum_{(s,a)} \mathbb{P}(s_1 = s, a_1 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] + \gamma^2 \sum_{s} \mathbb{P}(s_2 = s)\ \nabla V_{\pi_\theta}(s)$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
&= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&+ \gamma \sum_{(s,a)} \mathbb{P}(s_1 = s, a_1 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] + \gamma^2 \sum_{s} \mathbb{P}(s_2 = s)\ \nabla V_{\pi_\theta}(s) \\
&\quad \vdots
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
&= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&+ \gamma \sum_{(s,a)} \mathbb{P}(s_1 = s, a_1 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] + \gamma^2 \sum_{s} \mathbb{P}(s_2 = s)\ \nabla V_{\pi_\theta}(s) \\
&\quad \vdots \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{(s,a)} \mathbb{P}(s_t = s, a_t = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)]
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
&= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&+ \gamma \sum_{(s,a)} \mathbb{P}(s_1 = s, a_1 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] + \gamma^2 \sum_{s} \mathbb{P}(s_2 = s)\ \nabla V_{\pi_\theta}(s) \\
&\qquad \vdots \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{(s,a)} \mathbb{P}(s_t = s, a_t = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&= \sum_{(s,a)} [\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(x,a)] \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a)
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
&= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&+ \gamma \sum_{(s,a)} \mathbb{P}(s_1 = s, a_1 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] + \gamma^2 \sum_{s} \mathbb{P}(s_2 = s)\ \nabla V_{\pi_\theta}(s) \\
&\quad \vdots \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{(s,a)} \mathbb{P}(s_t = s, a_t = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&= \sum_{(s,a)} [\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(x,a)] \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a) \\
&= \frac{1}{(1-\gamma)} \sum_{(s,a)} [\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \rho_{\pi_\theta}(s,a)
\end{aligned}
$$

# Proof of Policy Gradient Theorem: Form 2

Proof:

$$
\begin{aligned}
&= \sum_{(s,a)} \mathbb{P}(s_0 = s, a_0 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&+ \gamma \sum_{(s,a)} \mathbb{P}(s_1 = s, a_1 = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] + \gamma^2 \sum_{s} \mathbb{P}(s_2 = s)\ \nabla V_{\pi_\theta}(s) \\
&\quad \vdots \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{(s,a)} \mathbb{P}(s_t = s, a_t = a)[\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \\
&= \sum_{(s,a)} [\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(x,a)] \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a) \\
&= \frac{1}{(1-\gamma)} \sum_{(s,a)} [\nabla \log \pi_\theta(s,a)\ Q_{\pi_\theta}(s,a)] \rho_{\pi_\theta}(s,a) \\
&= \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim \rho_{\pi_\theta}(\cdot)} [Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)]
\end{aligned}
$$

$\square$

# Estimating Policy Gradient

- We have $\nabla V_{\pi_\theta} = (1/(1-\gamma))\ \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we estimate $\widehat{\nabla V_{\pi_\theta}}$?

# Estimating Policy Gradient

- We have $\nabla V_{\pi_\theta} = (1/(1-\gamma))\ \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)\right]$
- How do we estimate $\widehat{\nabla V_{\pi_\theta}}$?
- We will use an equivalent form

# Estimating Policy Gradient

- We have $\nabla V_{\pi_\theta} = (1/(1-\gamma))\ \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we estimate $\widehat{\nabla V_{\pi_\theta}}$?
- We will use an equivalent form

## Proposition

$$\begin{aligned}\nabla V_{\pi_\theta} &= (1/(1-\gamma))\ \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right] \\ &= \mathbb{E}_{\tau\sim P_{\mathrm{traj},\theta}(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t\ Q_{\pi_\theta}(s_t,a_t)\ \nabla \log \pi_\theta(s_t,a_t)\right]\end{aligned}$$

# Estimating Policy Gradient

Proof:

$\square$

# Estimating Policy Gradient

Proof:

$$\mathbb{E}_{\tau\sim P_{\mathrm{traj},\theta}(\tau)}[\sum_{t=0}^{\infty}\gamma^t\ Q_{\pi_\theta}(s_t,a_t)\ \nabla\log\pi_\theta(s_t,a_t)]$$

$\square$

# Estimating Policy Gradient

Proof:

$$\mathbb{E}_{\tau\sim P_{\text{traj},\theta}(\tau)}[\sum_{t=0}^{\infty}\gamma^t\ Q_{\pi_\theta}(s_t,a_t)\ \nabla\log\pi_\theta(s_t,a_t)]$$

$$=\sum_x\sum_a\sum_{t=0}^{\infty}\mathbb{P}(s_t=s,a_t=a)\ \gamma^t\ Q_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)$$

$\square$

# Estimating Policy Gradient

Proof:

$$\mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\tau)}[\sum_{t=0}^{\infty} \gamma^t \; Q_{\pi_\theta}(s_t, a_t) \; \nabla \log \pi_\theta(s_t, a_t)]$$

$$= \sum_x \sum_a \sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a) \; \gamma^t \; Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a)$$

$$= \sum_s \sum_a Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a) \; (\sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\gamma^t)$$

$\square$

# Estimating Policy Gradient

Proof:

$$\mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\tau)}[\sum_{t=0}^{\infty} \gamma^t \; Q_{\pi_\theta}(s_t, a_t) \; \nabla \log \pi_\theta(s_t, a_t)]$$

$$= \sum_x \sum_a \sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a) \; \gamma^t \; Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a)$$

$$= \sum_s \sum_a Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a) \; (\sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a)\gamma^t)$$

$$= (1/(1-\gamma)) \sum_x \sum_a Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a) \rho_{\pi_\theta}(s, a)$$

$\square$

# Estimating Policy Gradient

Proof:

$$\mathbb{E}_{\tau \sim P_{\text{traj},\theta}(\tau)}[\sum_{t=0}^{\infty} \gamma^t \; Q_{\pi_\theta}(s_t, a_t) \; \nabla \log \pi_\theta(s_t, a_t)]$$

$$= \sum_x \sum_a \sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a) \; \gamma^t \; Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a)$$

$$= \sum_s \sum_a Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a) \; (\sum_{t=0}^{\infty} \mathbb{P}(s_t = s, a_t = a) \gamma^t)$$

$$= (1/(1-\gamma)) \sum_x \sum_a Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a) \rho_{\pi_\theta}(s, a)$$

$$= (1/(1-\gamma)) \mathbb{E}_{(s,a) \sim \rho_{\pi_\theta}(\cdot)} [Q_{\pi_\theta}(s, a) \; \nabla \log \pi_\theta(s, a)]$$

□

# Estimating Policy Gradient

Proof:

$$
\begin{aligned}
&\mathbb{E}_{\tau\sim P_{\text{traj},\theta}(\tau)}[\sum_{t=0}^{\infty}\gamma^t\ Q_{\pi_\theta}(s_t,a_t)\ \nabla\log\pi_\theta(s_t,a_t)]\\
&=\sum_{x}\sum_{a}\sum_{t=0}^{\infty}\mathbb{P}(s_t=s,a_t=a)\ \gamma^t\ Q_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)\\
&=\sum_{s}\sum_{a}Q_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)\ (\sum_{t=0}^{\infty}\mathbb{P}(s_t=s,a_t=a)\gamma^t)\\
&=(1/(1-\gamma))\sum_{x}\sum_{a}Q_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)\rho_{\pi_\theta}(s,a)\\
&=(1/(1-\gamma))\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)\right]\\
&=\nabla V_{\pi_\theta}
\end{aligned}
$$

□

# Estimating Policy Gradient

- We have $\nabla V_{\pi_\theta} = (1/(1-\gamma))\ \mathbb{E}_{(x,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(x,a)\ \nabla \log \pi_\theta(x,a)\right]$
- How do we estimate $\widehat{\nabla V_{\pi_\theta}}$?
- We will use an equivalent form

## Proposition

$$\begin{aligned}\nabla V_{\pi_\theta} &= (1/(1-\gamma))\ \mathbb{E}_{(x,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(x,a)\ \nabla \log \pi_\theta(x,a)\right] \\ &= \mathbb{E}_{\tau\sim P_{\mathrm{traj},\theta}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t\ Q_{\pi_\theta}(s_t,a_t)\ \nabla \log \pi_\theta(s_t,a_t)\right]\end{aligned}$$

# Estimating Policy Gradient

- We have $\nabla V_{\pi_\theta} = (1/(1-\gamma))\ \mathbb{E}_{(x,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(x,a)\ \nabla \log \pi_\theta(x,a)\right]$
- How do we estimate $\widehat{\nabla V_{\pi_\theta}}$?
- We will use an equivalent form

## Proposition

$$\begin{aligned}\nabla V_{\pi_\theta} &= (1/(1-\gamma))\ \mathbb{E}_{(x,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(x,a)\ \nabla \log \pi_\theta(x,a)\right] \\ &= \mathbb{E}_{\tau\sim P_{\mathrm{traj},\theta}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t\ Q_{\pi_\theta}(s_t,a_t)\ \nabla \log \pi_\theta(s_t,a_t)\right]\end{aligned}$$

- For a trajectory $\tau$, define:

$$\widehat{Q_{\pi_\theta}}(s_t,a_t) = \sum_{m\geq t}^{\infty} \gamma^{(m-t)} r(s_m,a_m)$$

$$\widehat{\nabla V_{\pi_\theta}} = \sum_{t=0}^{\infty} \gamma^t\ \widehat{Q_{\pi_\theta}}(s_t,a_t)\ \nabla \log \pi_\theta(s_t,a_t)$$

# Variance Reduction using Baseline

- Advantage function of a policy $\pi$ is defined as
$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$
- Note that $\mathbb{E}_{a\sim\pi(s,\cdot)}[A_\pi(s,a)] = 0$

# Variance Reduction using Baseline

- Advantage function of a policy $\pi$ is defined as

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

- Note that $\mathbb{E}_{a\sim\pi(s,\cdot)}[A_\pi(s,a)] = 0$

## Proposition

$$\nabla V_{\pi_\theta} = \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[A_{\pi_\theta}(s,a)\ \nabla\log\pi_\theta(s,a)\right]$$

# Variance Reduction using Baseline

**Proof:**

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0.$

# Variance Reduction using Baseline

Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0.$

# Variance Reduction using Baseline

Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0$.

$$\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right]$$

# Variance Reduction using Baseline

## Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0$.

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \mathbb{E}_{a\sim\pi_\theta(s,\cdot)}[\nabla \log \pi_\theta(s,a)]]
\end{aligned}
$$

# Variance Reduction using Baseline

## Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0$.

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \mathbb{E}_{a\sim\pi_\theta(s,\cdot)}[\nabla \log \pi_\theta(s,a)]] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \pi_\theta(s,a)\nabla \log \pi_\theta(s,a)]
\end{aligned}$$

# Variance Reduction using Baseline

## Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0$.

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \mathbb{E}_{a\sim\pi_\theta(s,\cdot)}[\nabla \log \pi_\theta(s,a)]] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \pi_\theta(s,a)\nabla \log \pi_\theta(s,a)] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \nabla\pi_\theta(s,a)] = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \nabla \sum_a \pi_\theta(s,a)]
\end{aligned}
$$

# Variance Reduction using Baseline

## Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0$.

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \mathbb{E}_{a\sim\pi_\theta(s,\cdot)}[\nabla \log \pi_\theta(s,a)]] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \pi_\theta(s,a)\nabla \log \pi_\theta(s,a)] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \nabla\pi_\theta(s,a)] = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \nabla\sum_a \pi_\theta(s,a)] \\
&\qquad = \mathbb{E}_{x\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla(1)\right] = 0
\end{aligned}
$$

# Variance Reduction using Baseline

Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] = 0$.

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla \log \pi_\theta(s,a)\right] \\
&\quad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \mathbb{E}_{a\sim\pi_\theta(s,\cdot)}[\nabla \log \pi_\theta(s,a)]] \\
&\quad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \pi_\theta(s,a)\nabla \log \pi_\theta(s,a)] \\
&\quad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \nabla\pi_\theta(s,a)] = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \nabla\sum_a \pi_\theta(s,a)] \\
&\quad = \mathbb{E}_{x\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla(1)\right] = 0
\end{aligned}
$$

So, for any such baseline $b_\pi(s)$ , we get

$$
\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot)}\left[(Q_{\pi_\theta}(s,a) - b_{\pi_\theta}(s))\ \nabla \log \pi_\theta(s,a)\right] = \nabla V_{\pi_\theta}
$$

# Variance Reduction using Baseline

## Proof:

Let $b_{\pi_\theta}(s)$ be an action independent baseline.

We will first show that $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla\log\pi_\theta(s,a)\right] = 0$.

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla\log\pi_\theta(s,a)\right] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \mathbb{E}_{a\sim\pi_\theta(s,\cdot)}[\nabla\log\pi_\theta(s,a)]] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \pi_\theta(s,a)\nabla\log\pi_\theta(s,a)] \\
&\qquad = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(x)\ \sum_a \nabla\pi_\theta(s,a)] = \mathbb{E}_{s\sim\rho_{\pi_\theta}(\cdot)}[b_{\pi_\theta}(s)\ \nabla\sum_a \pi_\theta(s,a)] \\
&\qquad = \mathbb{E}_{x\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[b_{\pi_\theta}(s)\ \nabla(1)\right] = 0
\end{aligned}
$$

So, for any such baseline $b_\pi(s)$ , we get

$$
\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot)}\left[(Q_{\pi_\theta}(s,a) - b_{\pi_\theta}(s))\ \nabla\log\pi_\theta(s,a)\right] = \nabla V_{\pi_\theta}
$$

This is in particular true for the baseline $b_{\pi_\theta}(x) = V_{\pi_\theta}(x)$

# Policy Gradient Theorem(s)

## Theorem

*The following as expressions for* $\nabla V_{\pi_\theta}$*:*

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\cdot)} \left[ R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(s_t, a_t) \right] \tag{1}$$

$$\nabla V_{\pi_\theta} = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a) \sim \rho_{\pi_\theta}(\cdot,\cdot)} \left[ Q_{\pi_\theta}(s,a) \ \nabla \log \pi_\theta(s,a) \right] \tag{2}$$

$$\nabla V_{\pi_\theta} = \mathbb{E}_{\tau \sim P_{\mathrm{traj},\theta}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t \ Q_{\pi_\theta}(s_t, a_t) \ \nabla \log \pi_\theta(s_t, a_t) \right] \tag{3}$$

$$\nabla V_{\pi_\theta} = \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a) \sim \rho_{\pi_\theta}(\cdot,\cdot)} \left[ A_{\pi_\theta}(s,a) \ \nabla \log \pi_\theta(s,a) \right] \tag{4}$$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$
- Then, $\nabla V_{\pi_\theta} \approx \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_w(s,a)\ \nabla \log \pi_\theta(s,a)\right]$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$
- Then, $\nabla V_{\pi_\theta} \approx \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_w(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- Can we then do the update:

$$\theta \leftarrow \theta + \alpha\ Q_w(s,a)\ \nabla \log \pi_\theta(s,a)$$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$
- Then, $\nabla V_{\pi_\theta} \approx \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_w(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- Can we then do the update:

$$\theta \leftarrow \theta + \alpha\ Q_w(s,a)\ \nabla \log \pi_\theta(s,a)$$

- What is the $w$ to be used?

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$
- Then, $\nabla V_{\pi_\theta} \approx \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_w(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- Can we then do the update:

$$\theta \leftarrow \theta + \alpha\ Q_w(s,a)\ \nabla \log \pi_\theta(s,a)$$

- What is the $w$ to be used?
  - $w$ should give a good approximation of Q-value function for the current policy $\pi_\theta$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$
- Then, $\nabla V_{\pi_\theta} \approx \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_w(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- Can we then do the update:

$$\theta \leftarrow \theta + \alpha\ Q_w(s,a)\ \nabla \log \pi_\theta(s,a)$$

- What is the $w$ to be used?
  - $w$ should give a good approximation of Q-value function for the current policy $\pi_\theta$
  - $Q_{\pi_\theta}$ is unknown. So, we need to learn $w$

# Actor-Critic Algorithm

- We have, $\nabla V_{\pi_\theta} = \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_{\pi_\theta}(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- How do we get $Q_{\pi_\theta}(s,a)$?
- We can approximate $Q$-function: $Q_w(s,a) \approx Q_{\pi_\theta}(s,a)$
- Then, $\nabla V_{\pi_\theta} \approx \frac{1}{(1-\gamma)}\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(\cdot,\cdot)}\left[Q_w(s,a)\ \nabla \log \pi_\theta(s,a)\right]$
- Can we then do the update:

$$\theta \leftarrow \theta + \alpha\ Q_w(s,a)\ \nabla \log \pi_\theta(s,a)$$

- What is the $w$ to be used?
  - $w$ should give a good approximation of Q-value function for the current policy $\pi_\theta$
  - $Q_{\pi_\theta}$ is unknown. So, we need to learn $w$
  - $\theta$ is changing. So, $w$ should also change with respect to $\theta$

# Actor-Critic Algorithm

- Critic: Updates Q-value parameter $w$
  - (*Evaluate* the policy corresponding to $\theta$)
- Actor: Updates policy parameter $\theta$
  - (*Improve* the policy corresponding to based on the feedback from critic)

# Actor-Critic Algorithm

- Critic: Updates Q-value parameter $w$
  - (*Evaluate* the policy corresponding to $\theta$)
- Actor: Updates policy parameter $\theta$
  - (*Improve* the policy corresponding to based on the feedback from critic)

**for** each step **do**
  $\theta = \theta + \alpha_\theta \ Q_w(s, a) \ \nabla_\theta \log(\pi_\theta(s, a))$
  Sample the next state $s'$, sample the next action $a' \sim \pi_\theta(s', \cdot)$
  $\delta = r(s, a) + \gamma \ Q_w(s', a') - Q_w(x, a)$
  $w = w + \alpha_w \ \delta \ \nabla_w Q_w(s, a)$
**end for**

# Actor-Critic Algorithm

- Critic: Updates Q-value parameter $w$
  - (*Evaluate* the policy corresponding to $\theta$)
- Actor: Updates policy parameter $\theta$
  - (*Improve* the policy corresponding to based on the feedback from critic)

**for** each step **do**
 $\theta = \theta + \alpha_\theta \; Q_w(s,a) \; \nabla_\theta \log(\pi_\theta(s,a))$
 Sample the next state $s'$, sample the next action $a' \sim \pi_\theta(s', \cdot)$
 $\delta = r(s,a) + \gamma \; Q_w(s',a') - Q_w(x,a)$
 $w = w + \alpha_w \; \delta \; \nabla_w Q_w(s,a)$
**end for**

- Learning rate $\alpha_\theta$ and $\alpha_w$ should be selected appropriately for convergence

# Actor-Critic Algorithm

- Critic: Updates Q-value parameter $w$
  - (*Evaluate* the policy corresponding to $\theta$)
- Actor: Updates policy parameter $\theta$
  - (*Improve* the policy corresponding to based on the feedback from critic)

**for** each step **do**
  $\theta = \theta + \alpha_\theta \; Q_w(s,a) \; \nabla_\theta \log(\pi_\theta(s,a))$
  Sample the next state $s'$, sample the next action $a' \sim \pi_\theta(s', \cdot)$
  $\delta = r(s,a) + \gamma \; Q_w(s',a') - Q_w(x,a)$
  $w = w + \alpha_w \; \delta \; \nabla_w Q_w(s,a)$
**end for**

- Learning rate $\alpha_\theta$ and $\alpha_w$ should be selected appropriately for convergence
- This is an example of two timescale stochastic approximation algorithm

# Natural Policy Gradient

# Steepest Ascent Direction

- Goal: solve $\max_\theta \; V_\theta$

# Steepest Ascent Direction

- Goal: solve $\max_\theta \ V_\theta$
- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

# Steepest Ascent Direction

- Goal: solve $\max_\theta \; V_\theta$
- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

## Steepest Ascent Direction

- Goal: solve $\max_\theta \; V_\theta$
- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

where $\Delta(\theta)$ is defined as

$$\arg\max_{\delta} \; \delta^\top (\nabla V), \quad \text{s.t. } \frac{1}{2}\|\delta\|^2 \leq \epsilon$$

# Steepest Ascent Direction

- Goal: solve $\max_\theta \ V_\theta$
- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

where $\Delta(\theta)$ is defined as

$$\arg\max_{\delta} \ \delta^\top(\nabla V), \quad \text{s.t. } \frac{1}{2}\|\delta\|^2 \le \epsilon$$

- In general, we can define $\|\delta\|^2 = \delta^\top G \delta$, where $G$ is a positive definite matrix

# Steepest Ascent Direction

- Goal: solve $\max_\theta \; V_\theta$
- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

  where $\Delta(\theta)$ is defined as

$$\arg\max_{\delta} \; \delta^\top (\nabla V), \quad \text{s.t. } \frac{1}{2}\|\delta\|^2 \leq \epsilon$$

- In general, we can define $\|\delta\|^2 = \delta^\top G \delta$, where $G$ is a positive definite matrix
  - If $G = I$, we recover $L^2$ norm

# Steepest Ascent Direction

- Goal: solve $\max_\theta \ V_\theta$
- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

where $\Delta(\theta)$ is defined as

$$\arg\max_{\delta} \ \delta^\top(\nabla V), \ \text{ s.t. } \frac{1}{2}\|\delta\|^2 \leq \epsilon$$

- In general, we can define $\|\delta\|^2 = \delta^\top G\delta$, where $G$ is a positive definite matrix
  - If $G = I$, we recover $L^2$ norm
- When $\|\delta\|^2 = \delta^\top G\delta$, we can show that

$$\Delta(\theta) = \sqrt{\frac{\epsilon}{(\nabla V)^\top G^{-1}(\nabla V)}} \ G^{-1}(\nabla V)$$

# Steepest Ascent Direction

- We need to solve: $\arg\max_\delta \;\delta^\top(\nabla V),$ s.t. $\frac{1}{2}\delta^\top G\delta \le \epsilon$

# Steepest Ascent Direction

- We need to solve: $\arg\max_\delta \; \delta^\top(\nabla V), \;$ s.t. $\frac{1}{2}\delta^\top G\delta \leq \epsilon$
- Taking Lagrangian, $L(\lambda, \delta) = \delta^\top(\nabla V) - \lambda(\frac{1}{2}\delta^\top G\delta - \epsilon)$.

# Steepest Ascent Direction

- We need to solve: $\arg\max_\delta \; \delta^\top(\nabla V), \quad$ s.t. $\frac{1}{2}\delta^\top G\delta \le \epsilon$
- Taking Lagrangian, $L(\lambda, \delta) = \delta^\top(\nabla V) - \lambda(\frac{1}{2}\delta^\top G\delta - \epsilon)$.
- To find $\delta(\lambda)$, we solve $\nabla_\delta L(\lambda, \delta) = 0$

# Steepest Ascent Direction

- We need to solve: $\arg\max_\delta \ \delta^\top(\nabla V), \quad$ s.t. $\frac{1}{2}\delta^\top G\delta \leq \epsilon$
- Taking Lagrangian, $L(\lambda, \delta) = \delta^\top(\nabla V) - \lambda(\frac{1}{2}\delta^\top G\delta - \epsilon)$.
- To find $\delta(\lambda)$, we solve $\nabla_\delta L(\lambda, \delta) = 0$
- we get $\delta(\lambda) = \frac{1}{\lambda}G^{-1}\nabla V$

# Steepest Ascent Direction

- We need to solve: $\arg\max_\delta \ \delta^\top(\nabla V)$, s.t. $\frac{1}{2}\delta^\top G\delta \leq \epsilon$
- Taking Lagrangian, $L(\lambda, \delta) = \delta^\top(\nabla V) - \lambda(\frac{1}{2}\delta^\top G\delta - \epsilon)$.
- To find $\delta(\lambda)$, we solve $\nabla_\delta L(\lambda, \delta) = 0$
- we get $\delta(\lambda) = \frac{1}{\lambda}G^{-1}\nabla V$
- To find $\lambda^\star$, we first get $D(\lambda) = \max_\delta L(\lambda, \delta) = \frac{1}{2\lambda}(\nabla V^\top G^{-1}\nabla V) + \lambda\epsilon$

# Steepest Ascent Direction

- We need to solve: $\arg\max_{\delta}\ \delta^\top(\nabla V),$ s.t. $\frac{1}{2}\delta^\top G\delta \leq \epsilon$
- Taking Lagrangian, $L(\lambda, \delta) = \delta^\top(\nabla V) - \lambda(\frac{1}{2}\delta^\top G\delta - \epsilon)$.
- To find $\delta(\lambda)$, we solve $\nabla_\delta L(\lambda, \delta) = 0$
- we get $\delta(\lambda) = \frac{1}{\lambda}G^{-1}\nabla V$
- To find $\lambda^\star$, we first get $D(\lambda) = \max_\delta L(\lambda, \delta) = \frac{1}{2\lambda}(\nabla V^\top G^{-1}\nabla V) + \lambda\epsilon$
- We solve $\nabla_\lambda D(\lambda) = 0$ to get $\lambda^\star = \sqrt{\frac{(\nabla V^\top G^{-1}\nabla V)}{2\epsilon}}$

# Steepest Ascent Direction

- We need to solve: $\arg\max_\delta \ \delta^\top(\nabla V), \quad$ s.t. $\frac{1}{2}\delta^\top G\delta \leq \epsilon$
- Taking Lagrangian, $L(\lambda, \delta) = \delta^\top(\nabla V) - \lambda(\frac{1}{2}\delta^\top G\delta - \epsilon)$.
- To find $\delta(\lambda)$, we solve $\nabla_\delta L(\lambda, \delta) = 0$
- we get $\delta(\lambda) = \frac{1}{\lambda}G^{-1}\nabla V$
- To find $\lambda^\star$, we first get $D(\lambda) = \max_\delta L(\lambda, \delta) = \frac{1}{2\lambda}(\nabla V^\top G^{-1}\nabla V) + \lambda\epsilon$
- We solve $\nabla_\lambda D(\lambda) = 0$ to get $\lambda^\star = \sqrt{\frac{(\nabla V^\top G^{-1}\nabla V)}{2\epsilon}}$
- Using $\lambda^\star$, we get $\delta^* = \delta(\lambda^\star)$ as

$$\delta^* = \sqrt{\frac{2\epsilon}{(\nabla V^\top G^{-1}\nabla V)}}\, G^{-1}\nabla V$$

# Steepest Ascent Direction

- Gradient ascent algorithms takes a small step in the "steepest direction"

  $$\theta \leftarrow \theta + \Delta(\theta),$$

  where $\delta\theta$ is defined as

  $$\arg\max_{\delta} \ \delta^\top(\nabla V), \quad \text{s.t. } \|\delta\|^2 \leq \epsilon$$

- In general, we can define $\|\delta\|^2 = \delta^\top G\delta$, where $G$ is a positive definite matrix
  - If $G = I$, we recover $L^2$ norm
- When $\|\delta\|^2 = \delta^\top G\delta$, we can show that

$$\Delta(\theta) = \sqrt{\frac{\epsilon}{(\nabla V)^\top G^{-1}(\nabla V)}}\, G^{-1}(\nabla V)$$

# Steepest Ascent Direction

- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

where $\delta\theta$ is defined as

$$\arg\max_{\delta} \ \delta^\top(\nabla V), \quad \text{s.t. } \|\delta\|^2 \leq \epsilon$$

- In general, we can define $\|\delta\|^2 = \delta^\top G \delta$, where $G$ is a positive definite matrix
  - If $G = I$, we recover $L^2$ norm
- When $\|\delta\|^2 = \delta^\top G \delta$, we can show that

$$\Delta(\theta) = \sqrt{\frac{\epsilon}{(\nabla V)^\top G^{-1}(\nabla V)}} \, G^{-1}(\nabla V)$$

- Vanilla policy gradient algorithm consider $G = I$, and $\theta \leftarrow \theta + \alpha(\nabla V)$

# Steepest Ascent Direction

- Gradient ascent algorithms takes a small step in the "steepest direction"

$$\theta \leftarrow \theta + \Delta(\theta),$$

where $\delta\theta$ is defined as

$$\arg\max_{\delta} \ \delta^\top(\nabla V), \quad \text{s.t. } \|\delta\|^2 \leq \epsilon$$

- In general, we can define $\|\delta\|^2 = \delta^\top G\delta$, where $G$ is a positive definite matrix
  - If $G = I$, we recover $L^2$ norm
- When $\|\delta\|^2 = \delta^\top G\delta$, we can show that

$$\Delta(\theta) = \sqrt{\frac{\epsilon}{(\nabla V)^\top G^{-1}(\nabla V)}}\, G^{-1}(\nabla V)$$

- Vanilla policy gradient algorithm consider $G = I$, and $\theta \leftarrow \theta + \alpha(\nabla V)$
- Is $(\nabla V)$ the best direction?

# Distance in Policy Space

- Distance in the parameter space need not be the same as the distance in the policy space

# Distance in Policy Space

- Distance in the parameter space need not be the same as the distance in the policy space
- We want a $\delta\theta$ such that $\pi_{(\theta+\delta\theta)}$ is better than $\pi_\theta$

# Distance in Policy Space

- Distance in the parameter space need not be the same as the distance in the policy space
- We want a $\delta\theta$ such that $\pi_{(\theta+\delta\theta)}$ is better than $\pi_\theta$
- Intuitively, we want to take a small step in the policy space, not necessarily in the parameter space

# Distance in Policy Space

- Distance in the parameter space need not be the same as the distance in the policy space
- We want a $\delta\theta$ such that $\pi_{(\theta+\delta\theta)}$ is better than $\pi_\theta$
- Intuitively, we want to take a small step in the policy space, not necessarily in the parameter space
- How do we define the distance in the policy space?

# Distance in Policy Space

- Distance in the parameter space need not be the same as the distance in the policy space
- We want a $\delta\theta$ such that $\pi_{(\theta+\delta\theta)}$ is better than $\pi_\theta$
- Intuitively, we want to take a small step in the policy space, not necessarily in the parameter space
- How do we define the distance in the policy space?
- We will use Kullback–Leibler divergence as a measure of distance between two policies

# Kullback–Leibler Divergence

- KL divergence is a measure of how one probability distribution is different from another

# Kullback–Leibler Divergence

- KL divergence is a measure of how one probability distribution is different from another
- Let $p$ and $q$ be two p.m.f. defined over a discrete probability space

$$D_{KL}(p, q) = \sum_x p(x) \ \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p}\left[\log \frac{p(x)}{q(x)}\right]$$

# Kullback–Leibler Divergence

- KL divergence is a measure of how one probability distribution is different from another
- Let $p$ and $q$ be two p.m.f. defined over a discrete probability space

$$D_{KL}(p,q) = \sum_x p(x) \ \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p}\left[\log \frac{p(x)}{q(x)}\right]$$

- A policy is a family of distribution

# Kullback–Leibler Divergence

- KL divergence is a measure of how one probability distribution is different from another
- Let $p$ and $q$ be two p.m.f. defined over a discrete probability space

$$D_{KL}(p, q) = \sum_x p(x) \ \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p}\left[\log \frac{p(x)}{q(x)}\right]$$

- A policy is a family of distribution
  - $\pi(s, \cdot)$ is a distribution over $\mathcal{A}$. For each $s \in \mathcal{S}$, we have one distribution

# Kullback–Leibler Divergence

- KL divergence is a measure of how one probability distribution is different from another
- Let $p$ and $q$ be two p.m.f. defined over a discrete probability space

$$D_{KL}(p, q) = \sum_x p(x) \; \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p} \left[ \log \frac{p(x)}{q(x)} \right]$$

- A policy is a family of distribution
  - $\pi(s, \cdot)$ is a distribution over $\mathcal{A}$. For each $s \in \mathcal{S}$, we have one distribution
- Define the expected KL divergence between two polices $\pi_1$ and $\pi_2$ as

$$D_{KL}^{\pi_1}(\pi_1, \pi_2) = \mathbb{E}_{s \sim \rho_{\pi_1}(\cdot)}[D_{KL}(\pi_1(s, \cdot), \pi_2(s, \cdot))]$$

# Kullback–Leibler Divergence

- KL divergence is a measure of how one probability distribution is different from another
- Let $p$ and $q$ be two p.m.f. defined over a discrete probability space

$$D_{KL}(p,q) = \sum_x p(x)\ \log \frac{p(x)}{q(x)} = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right]$$

- A policy is a family of distribution
  - $\pi(s,\cdot)$ is a distribution over $\mathcal{A}$. For each $s \in \mathcal{S}$, we have one distribution
- Define the expected KL divergence between two polices $\pi_1$ and $\pi_2$ as

$$D_{KL}^{\pi_1}(\pi_1,\pi_2) = \mathbb{E}_{s\sim\rho_{\pi_1}(\cdot)}[D_{KL}(\pi_1(s,\cdot),\pi_2(s,\cdot))]$$

- We we want to find a direction $\Delta(\theta)$, where

$$\arg\max_{\delta}\ \delta^\top(\nabla V_{\pi_\theta}), \quad \text{s.t. } D_{KL}^{\pi_\theta}(\pi_\theta,\pi_{(\theta+\delta\theta)}) \leq \epsilon$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of a function $f(y)$ around $y_0$

$$f(y) \approx f(y_0) + (y - y_0)^\top \nabla_y f(y)|_{y_0} + (y - y_0)^\top \nabla_y^2 f(y)|_{y_0} (y - y_0) + \cdots$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of a function $f(y)$ around $y_0$

$$f(y) \approx f(y_0) + (y - y_0)^\top \nabla_y f(y)|_{y_0} + (y - y_0)^\top \nabla_y^2 f(y)|_{y_0} (y - y_0) + \cdots$$

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$\begin{aligned} D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\ &+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} (\theta - \theta_0) + \cdots \end{aligned}$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$\begin{aligned} D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\ &+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots \end{aligned}$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} = \nabla_\theta \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s, \cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$\begin{aligned} D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\ &+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} (\theta - \theta_0) + \cdots \end{aligned}$$

$$\begin{aligned} \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\ &= -\mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla \pi_\theta(s,a)|_{\theta_0}}{\pi_{\theta_0}(s,a)} \right] \end{aligned}$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} (\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla \pi_\theta(s,a)|_{\theta_0}}{\pi_{\theta_0}(s,a)} \right] \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} [\sum_a \nabla \pi_\theta(s,a)|_{\theta_0}]
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla \pi_\theta(s,a)|_{\theta_0}}{\pi_{\theta_0}(s,a)} \right] \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} [\sum_a \nabla \pi_\theta(s,a)|_{\theta_0}] \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} [\nabla \sum_a \pi_\theta(s,a)|_{\theta_0}] = 0
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$\begin{aligned} D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\ &+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots \end{aligned}$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} = \nabla_\theta^2 \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s, \cdot)} \left[ \log \frac{\pi_{\theta_0}(s, a)}{\pi_\theta(s, a)} \right] |_{\theta_0}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta^2 \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s, \cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s, \cdot)} \nabla \left[ \frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0}
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta^2 \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[\log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)}\right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \nabla \left[\frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)}\right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[\frac{\nabla^2 \pi_\theta(s,a)}{\pi_\theta(s,a)} - \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2}\right] |_{\theta_0}
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta^2 \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \nabla \left[ \frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla^2 \pi_\theta(s,a)}{\pi_\theta(s,a)} - \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2} \right] |_{\theta_0} \\
&= \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2} \right] |_{\theta_0}
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta^2 \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)}\right] |_{\theta_0} \\
&= -\mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \nabla \left[\frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)}\right] |_{\theta_0} \\
&= -\mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\frac{\nabla^2 \pi_\theta(s,a)}{\pi_\theta(s,a)} - \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2}\right] |_{\theta_0} \\
&= \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2}\right] |_{\theta_0} \\
&= \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\nabla \log \pi_\theta(s,a)\ \nabla \log \pi_\theta(s,a)^\top\right] |_{\theta_0}
\end{aligned}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$
\begin{aligned}
D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) &+ (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\
&+ (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta^2 \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \nabla \left[ \frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)} \right] |_{\theta_0} \\
&= -\mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla^2 \pi_\theta(s,a)}{\pi_\theta(s,a)} - \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2} \right] |_{\theta_0} \\
&= \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2} \right] |_{\theta_0} \\
&= \mathbb{E}_{s \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(s,\cdot)} \left[ \nabla \log \pi_\theta(s,a)\ \nabla \log \pi_\theta(s,a)^\top \right] |_{\theta_0}
\end{aligned}
$$

- Fisher information matrix

$$
F(\theta_0) = \mathbb{E}_{x \sim \rho_{\pi_{\theta_0}}} \mathbb{E}_{a \sim \pi_{\theta_0}(x,\cdot)} \left[ \nabla \log \pi_\theta(x,a)\ \nabla \log \pi_\theta(x,a)^\top \right] |_{\theta_0}
$$

# Taylor Series Expansion of KL Divergence

- Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$\begin{aligned} D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) = D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_{\theta_0}) + (\theta - \theta_0)^\top \nabla_\theta D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} \\ + (\theta - \theta_0)^\top \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0}(\theta - \theta_0) + \cdots \end{aligned}$$

$$\begin{aligned} \nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} &= \nabla_\theta^2 \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\log \frac{\pi_{\theta_0}(s,a)}{\pi_\theta(s,a)}\right]|_{\theta_0} \\ &= -\mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \nabla \left[\frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)}\right]|_{\theta_0} \\ &= -\mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\frac{\nabla^2 \pi_\theta(s,a)}{\pi_\theta(s,a)} - \frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2}\right]|_{\theta_0} \\ &= \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\frac{\nabla \pi_\theta(s,a)\ \nabla \pi_\theta(s,a)^\top}{(\pi_\theta(s,a))^2}\right]|_{\theta_0} \\ &= \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(s,\cdot)} \left[\nabla \log \pi_\theta(s,a)\ \nabla \log \pi_\theta(s,a)^\top\right]|_{\theta_0} \end{aligned}$$

- Fisher information matrix

$$F(\theta_0) = \mathbb{E}_{x\sim\rho_{\pi_{\theta_0}}} \mathbb{E}_{a\sim\pi_{\theta_0}(x,\cdot)} \left[\nabla \log \pi_\theta(x,a)\ \nabla \log \pi_\theta(x,a)^\top\right]|_{\theta_0}$$

- $\nabla_\theta^2 D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)|_{\theta_0} = F(\theta_0)$

# Natural Policy Gradient

- Using Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) \approx (\theta - \theta_0)^\top F(\theta_0)(\theta - \theta_0)$$

# Natural Policy Gradient

- Using Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) \approx (\theta - \theta_0)^\top F(\theta_0)(\theta - \theta_0)$$

- We we want to find a direction $\delta\theta$, where

$$\arg\max_{\delta} \ \delta^\top(\nabla V_{\pi_\theta}), \quad \text{s.t. } D_{KL}^{\pi_\theta}\big(\pi_\theta, \pi_{(\theta+\delta\theta)}\big) \leq \epsilon$$

# Natural Policy Gradient

- Using Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) \approx (\theta - \theta_0)^\top F(\theta_0)(\theta - \theta_0)$$

- We we want to find a direction $\delta\theta$, where

$$\arg\max_{\delta} \; \delta^\top(\nabla V_{\pi_\theta}), \quad \text{s.t. } D_{KL}^{\pi_\theta}\left(\pi_\theta, \pi_{(\theta+\delta\theta)}\right) \leq \epsilon$$

- This is equivalent to

$$\arg\max_{\delta} \; \delta^\top(\nabla V_{\pi_\theta}), \quad \text{s.t. } \delta^\top F(\theta)\delta \leq \epsilon$$

# Natural Policy Gradient

- Using Taylor series expansion of $D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta)$

$$D_{KL}^{\pi_{\theta_0}}(\pi_{\theta_0}, \pi_\theta) \approx (\theta - \theta_0)^\top F(\theta_0)(\theta - \theta_0)$$

- We we want to find a direction $\delta\theta$, where

$$\arg\max_{\delta} \ \delta^\top(\nabla V_{\pi_\theta}), \quad \text{s.t. } D_{KL}^{\pi_\theta}(\pi_\theta, \pi_{(\theta+\delta\theta)}) \leq \epsilon$$

- This is equivalent to

$$\arg\max_{\delta} \ \delta^\top(\nabla V_{\pi_\theta}), \quad \text{s.t. } \delta^\top F(\theta)\delta \leq \epsilon$$

- Natural Policy Gradient

$$\theta \leftarrow \theta + \delta\theta$$

$$\delta\theta = \sqrt{\frac{\epsilon}{(\nabla V)^\top F(\theta)^{-1}(\nabla V)}} \ F(\theta)^\dagger(\nabla V)$$

# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t,\ w_t \sim \mathcal{N}(0, \sigma^2),\ R_t = -Qx_t^2 - Pu_t^2$$

# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t, \ w_t \sim \mathcal{N}(0, \sigma^2), \ R_t = -Qx_t^2 - Pu_t^2$$

- What is the optimal control policy?:

# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t, \; w_t \sim \mathcal{N}(0, \sigma^2), \; R_t = -Qx_t^2 - Pu_t^2$$

- What is the optimal control policy?:
  - Optimal control policy is linear: $u_t^* = \theta^* x_t$

# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t, \ w_t \sim \mathcal{N}(0, \sigma^2), \ R_t = -Qx_t^2 - Pu_t^2$$

- What is the optimal control policy?:
  - Optimal control policy is linear: $u_t^* = \theta^* x_t$
- Consider this as an RL problem, where $A$ and $B$ are unknown

# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t,\ w_t \sim \mathcal{N}(0, \sigma^2),\ R_t = -Qx_t^2 - Pu_t^2$$

- What is the optimal control policy?:
  - Optimal control policy is linear: $u_t^* = \theta^* x_t$
- Consider this as an RL problem, where $A$ and $B$ are unknown
- We parameterize the policy as $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$

# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t,\ w_t \sim \mathcal{N}(0, \sigma^2),\ R_t = -Qx_t^2 - Pu_t^2$$

- What is the optimal control policy?:
  - Optimal control policy is linear: $u_t^* = \theta^* x_t$
- Consider this as an RL problem, where $A$ and $B$ are unknown
- We parameterize the policy as $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$
- Perform vanilla PG: $\theta \leftarrow \theta + \alpha \widehat{\nabla V_{\pi_\theta}}$

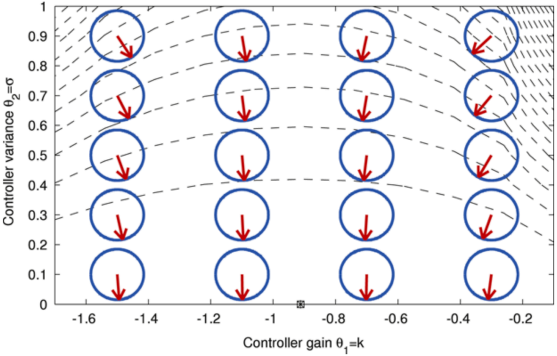
# Example: Policy Gradient for LQG

- Consider a scalar Linear Quadratic Gaussian (LQG) problem

$$x_{t+1} = Ax_t + Bu_t + w_t,\ w_t \sim \mathcal{N}(0, \sigma^2),\ R_t = -Qx_t^2 - Pu_t^2$$

- What is the optimal control policy?:
  - Optimal control policy is linear: $u_t^* = \theta^* x_t$
- Consider this as an RL problem, where $A$ and $B$ are unknown
- We parameterize the policy as $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$
- Perform vanilla PG: $\theta \leftarrow \theta + \alpha \widehat{\nabla V_{\pi_\theta}}$
- How will the vanilla PG perform?

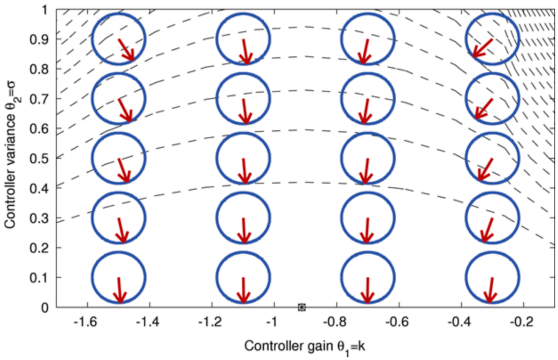
# Example: Vanilla Policy Gradient for LQG




- Control policy: $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$, $\theta \leftarrow \theta + \alpha \widehat{\nabla V_{\pi_\theta}}$

Figure from Peters, Jan, and Stefan Schaal. "Reinforcement learning of motor skills with policy gradients", *Neural*

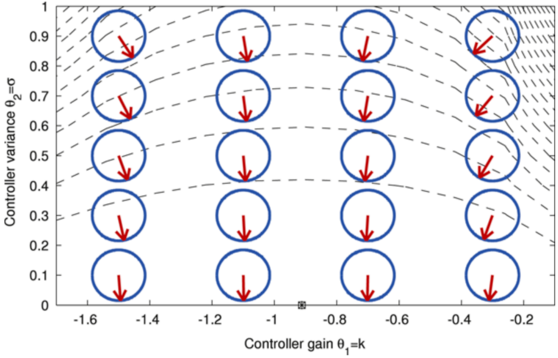
# Example: Vanilla Policy Gradient for LQG




- Control policy: $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$, $\theta \leftarrow \theta + \alpha \widehat{\nabla V_{\pi_\theta}}$
- Decrease of exploration has a stronger immediate effect on the expected return. So, the gradient mainly points in that direction.

Figure from Peters, Jan, and Stefan Schaal. "Reinforcement learning of motor skills with policy gradients", *Neural*

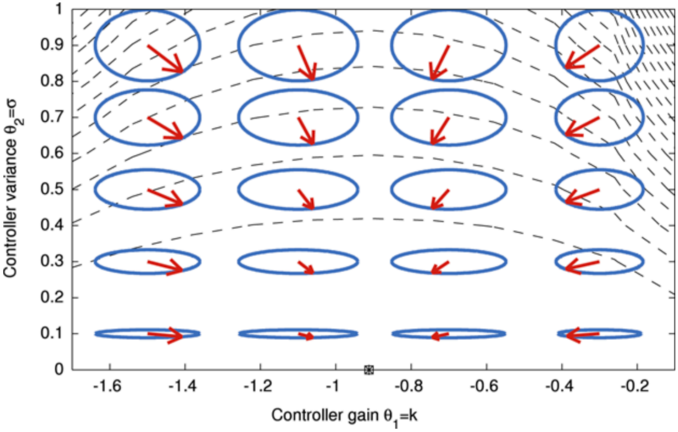
# Example: Vanilla Policy Gradient for LQG




- Control policy: $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$, $\theta \leftarrow \theta + \alpha \widehat{\nabla V_{\pi_\theta}}$
- Decrease of exploration has a stronger immediate effect on the expected return. So, the gradient mainly points in that direction.
- This will quickly reaches the 'plateau of zero exploration' and results in a very slow convergence

Figure from Peters, Jan, and Stefan Schaal. "Reinforcement learning of motor skills with policy gradients", *Neural*

# Example: Natural Policy Gradient for LQG




- Control policy: $u_t \sim \mathcal{N}(\theta_1 x_t, \theta_2)$, $\theta \leftarrow \theta + \alpha \widehat{F}(\theta)^{-1} \widehat{\nabla V_{\pi_\theta}}$

Figure from Peters, Jan, and Stefan Schaal. "Reinforcement learning of motor skills with policy gradients", *Neural networks*, 2008.