

CS 747, Autumn 2023: Lecture 21

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2023

Reinforcement Learning

1. Policy gradient methods
2. Variance reduction
3. Actor-critic methods

Reinforcement Learning

1. Policy gradient methods
2. Variance reduction
3. Actor-critic methods

A Variety of Applications

- **Learning to Trade via Direct Reinforcement**

Moody and Saffell (2001)

- **Reinforcement learning of motor skills with policy gradients**

Peters and Schaal (2008).

- **Mastering the game of Go with deep neural networks and tree search**

Silver et al. (2016)

- **Deep Reinforcement Learning for Autonomous Driving: A Survey**

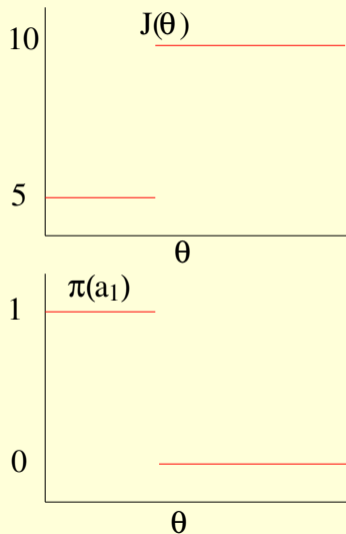
Ravi Kiran et al. (2021)

Stochastic Policies

- Single state; actions a_1, a_2 .
- $R(a_1) = 5$; $R(a_2) = 10$.
- Policy π ; parameter θ .

$$\pi(a_1) = \begin{cases} 1 & \text{if } \theta < 0.6, \\ 0 & \text{otherwise.} \end{cases}$$

$$J(\theta) = \pi(a_1) \cdot 5 + \pi(a_2) \cdot 10.$$



Stochastic Policies

- Single state; actions a_1, a_2 .
- $R(a_1) = 5$; $R(a_2) = 10$.
- Policy π ; parameter θ .

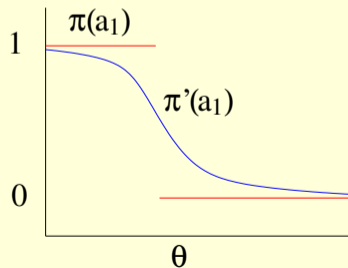
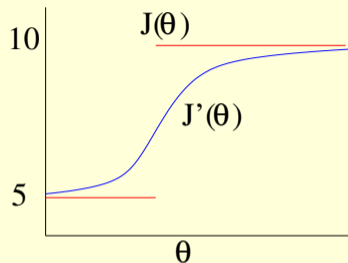
$$\pi(a_1) = \begin{cases} 1 & \text{if } \theta < 0.6, \\ 0 & \text{otherwise.} \end{cases}$$

$$J(\theta) = \pi(a_1) \cdot 5 + \pi(a_2) \cdot 10.$$

- Policy π' ; parameter θ .

$$\pi'(a_1) = \frac{1}{1 + e^{\theta - 0.6}}.$$

$$J'(\theta) = \pi'(a_1) \cdot 5 + \pi'(a_2) \cdot 10.$$



Idea

- If π is differentiable w.r.t. θ , so is (scalar) “policy value” J .

Idea

- If π is differentiable w.r.t. θ , so is (scalar) “policy value” J .
- We can “search” for “good” θ by iterating:

$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \alpha \nabla_{\theta} J(\theta_{\text{old}}).$$

Idea

- If π is differentiable w.r.t. θ , so is (scalar) “policy value” J .
- We can “search” for “good” θ by iterating:

$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \alpha \nabla_{\theta} J(\theta_{\text{old}}).$$

- **Example.** If we have features $x(\mathbf{s}, \mathbf{a}) \in \mathbb{R}^d$ for $\mathbf{s} \in \mathcal{S}$, $\mathbf{a} \in \mathcal{A}$, a common template for π is:

$$\pi(\mathbf{s}, \mathbf{a}) = \frac{e^{\theta \cdot x(\mathbf{s}, \mathbf{a})}}{\sum_{b \in \mathcal{A}} e^{\theta \cdot x(\mathbf{s}, b)}},$$

where $\theta \in \mathbb{R}^d$ is the vector of policy parameters.

Idea

- If π is differentiable w.r.t. θ , so is (scalar) “policy value” J .
- We can “search” for “good” θ by iterating:

$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \alpha \nabla_{\theta} J(\theta_{\text{old}}).$$

- **Example.** If we have features $x(\mathbf{s}, \mathbf{a}) \in \mathbb{R}^d$ for $\mathbf{s} \in \mathcal{S}$, $\mathbf{a} \in \mathcal{A}$, a common template for π is:

$$\pi(\mathbf{s}, \mathbf{a}) = \frac{e^{\theta \cdot x(\mathbf{s}, \mathbf{a})}}{\sum_{b \in \mathcal{A}} e^{\theta \cdot x(\mathbf{s}, b)}},$$

where $\theta \in \mathbb{R}^d$ is the vector of policy parameters. In this case, work out that

$$\nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) = \left(x(\mathbf{s}, \mathbf{a}) - \sum_{b \in \mathcal{B}} \pi(\mathbf{s}, b) x(\mathbf{s}, b) \right) \pi(\mathbf{s}, \mathbf{a}).$$

Idea

- If π is differentiable w.r.t. θ , so is (scalar) “policy value” J .
- We can “search” for “good” θ by iterating:

$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \alpha \nabla_{\theta} J(\theta_{\text{old}}).$$

- **Example.** If we have features $x(s, a) \in \mathbb{R}^d$ for $s \in \mathcal{S}$, $a \in \mathcal{A}$, a common template for π is:

$$\pi(s, a) = \frac{e^{\theta \cdot x(s, a)}}{\sum_{b \in \mathcal{A}} e^{\theta \cdot x(s, b)}},$$

where $\theta \in \mathbb{R}^d$ is the vector of policy parameters. In this case, work out that

$$\nabla_{\theta} \pi(s, a) = \left(x(s, a) - \sum_{b \in \mathcal{B}} \pi(s, b) x(s, b) \right) \pi(s, a).$$

- But what’s the connection between $\nabla_{\theta} J$ and $\nabla_{\theta} \pi(\cdot, \cdot)$?

Policy Gradient Theorem

- For simplicity assume episodic task with $\gamma = 1$.
- Assume there is a fixed start state s^0 .
- We leave it implicit that π is fixed by parameter vector θ .
- $J(\theta) = V^\pi(s^0)$.
- We shall derive the connection between $\nabla_\theta J$ and $\nabla_\theta \pi(\cdot, \cdot)$.

Policy Gradient Theorem

$$\text{For } \mathbf{s} \in \mathcal{S}, \nabla_{\theta} V^{\pi}(\mathbf{s}) = \nabla_{\theta} \sum_{a \in A} \pi(\mathbf{s}, a) Q^{\pi}(\mathbf{s}, a)$$

Policy Gradient Theorem

$$\begin{aligned}\text{For } \mathbf{s} \in \mathcal{S}, \nabla_{\theta} V^{\pi}(\mathbf{s}) &= \nabla_{\theta} \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &\quad + \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) \nabla_{\theta} \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') (R(\mathbf{s}, \mathbf{a}, \mathbf{s}') + V^{\pi}(\mathbf{s}'))\end{aligned}$$

Policy Gradient Theorem

$$\begin{aligned} \text{For } \mathbf{s} \in \mathcal{S}, \nabla_{\theta} V^{\pi}(\mathbf{s}) &= \nabla_{\theta} \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &\quad + \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) \nabla_{\theta} \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') (R(\mathbf{s}, \mathbf{a}, \mathbf{s}') + V^{\pi}(\mathbf{s}')) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \left[\nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) + \pi(\mathbf{s}, \mathbf{a}) \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') \nabla_{\theta} V^{\pi}(\mathbf{s}') \right] \end{aligned}$$

Policy Gradient Theorem

$$\begin{aligned} \text{For } \mathbf{s} \in \mathcal{S}, \nabla_{\theta} V^{\pi}(\mathbf{s}) &= \nabla_{\theta} \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &\quad + \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) \nabla_{\theta} \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') (R(\mathbf{s}, \mathbf{a}, \mathbf{s}') + V^{\pi}(\mathbf{s}')) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \left[\nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) + \pi(\mathbf{s}, \mathbf{a}) \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') \nabla_{\theta} V^{\pi}(\mathbf{s}') \right] \\ &= \dots = \sum_{\mathbf{x} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s} \rightarrow \mathbf{x}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{x}, \mathbf{a}) Q^{\pi}(\mathbf{x}, \mathbf{a}), \end{aligned}$$

Policy Gradient Theorem

$$\begin{aligned} \text{For } \mathbf{s} \in \mathcal{S}, \nabla_{\theta} V^{\pi}(\mathbf{s}) &= \nabla_{\theta} \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) \\ &\quad + \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) \nabla_{\theta} \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') (R(\mathbf{s}, \mathbf{a}, \mathbf{s}') + V^{\pi}(\mathbf{s}')) \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \left[\nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}) + \pi(\mathbf{s}, \mathbf{a}) \sum_{\mathbf{s}' \in \mathcal{S}} T(\mathbf{s}, \mathbf{a}, \mathbf{s}') \nabla_{\theta} V^{\pi}(\mathbf{s}') \right] \\ &= \dots = \sum_{\mathbf{x} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s} \rightarrow \mathbf{x}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{x}, \mathbf{a}) Q^{\pi}(\mathbf{x}, \mathbf{a}), \end{aligned}$$

where $\mathbb{P}\{\mathbf{s} \rightarrow \mathbf{x}, t, \pi\}$ is the probability of reaching \mathbf{x} from \mathbf{s} in t steps following π . 19

Policy Gradient Theorem

- Recall that $J(\theta) = V^\pi(\mathbf{s}^0)$.

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}).$$

Policy Gradient Theorem

- Recall that $J(\theta) = V^\pi(\mathbf{s}^0)$.

$$\nabla_\theta J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{a \in \mathcal{A}} \nabla_\theta \pi(\mathbf{s}, a) Q^\pi(\mathbf{s}, a).$$

- But how to do gradient ascent? We don't know $\mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\}$, $Q^\pi(\mathbf{s}, a)$!

Policy Gradient Theorem

- Recall that $J(\theta) = V^\pi(\mathbf{s}^0)$.

$$\nabla_\theta J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{a \in \mathcal{A}} \nabla_\theta \pi(\mathbf{s}, a) Q^\pi(\mathbf{s}, a).$$

- But how to do gradient ascent? We don't know $\mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\}$, $Q^\pi(\mathbf{s}, a)$!
- We perform **stochastic** gradient ascent.
- We use the following fact. For any discrete, real-valued random variable X with pmf $p : X \rightarrow [0, 1]$,

$$\sum_{x \in X} p(x)x = \mathbb{E}[X].$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_{\top}$ by acting according to π , parameterised by θ .

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_{\top}$ by acting according to π , parameterised by θ . Now consider:

$$\nabla_{\theta} J(\theta) = \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a)$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_\top$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a) \\ &= \sum_{t=0}^{\infty} \sum_{s \in S} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a)\end{aligned}$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_\top$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a) \\ &= \sum_{t=0}^{\infty} \sum_{s \in \mathcal{S}} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a) \\ &= \sum_{t=0}^{\infty} \mathbb{E}_{\pi} \left[\sum_{a \in A} \nabla_{\theta} \pi(s^t, a) Q^{\pi}(s^t, a) \right]\end{aligned}$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_T$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a) \\ &= \sum_{t=0}^{\infty} \sum_{s \in S} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a) \\ &= \sum_{t=0}^{\infty} \mathbb{E}_{\pi} \left[\sum_{a \in A} \nabla_{\theta} \pi(s^t, a) Q^{\pi}(s^t, a) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \nabla_{\theta} \pi(s^t, a) Q^{\pi}(s^t, a) \right].\end{aligned}$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_\top$ by acting according to π , parameterised by θ . Now consider:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \nabla_{\theta} \pi(\mathbf{s}^t, a) Q^{\pi}(\mathbf{s}^t, a) \right]$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_\top$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a}) Q^{\pi}(\mathbf{s}^t, \mathbf{a}) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \pi(\mathbf{s}^t, \mathbf{a}) \frac{\nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a})}{\pi(\mathbf{s}^t, \mathbf{a})} Q^{\pi}(\mathbf{s}^t, \mathbf{a}) \right]\end{aligned}$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_\top$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a}) Q^{\pi}(\mathbf{s}^t, \mathbf{a}) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \pi(\mathbf{s}^t, \mathbf{a}) \frac{\nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a})}{\pi(\mathbf{s}^t, \mathbf{a})} Q^{\pi}(\mathbf{s}^t, \mathbf{a}) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \frac{\nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a}^t)}{\pi(\mathbf{s}^t, \mathbf{a}^t)} Q^{\pi}(\mathbf{s}^t, \mathbf{a}^t) \right]\end{aligned}$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_T$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a}) Q^{\pi}(\mathbf{s}^t, \mathbf{a}) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \pi(\mathbf{s}^t, \mathbf{a}) \frac{\nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a})}{\pi(\mathbf{s}^t, \mathbf{a})} Q^{\pi}(\mathbf{s}^t, \mathbf{a}) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \frac{\nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a}^t)}{\pi(\mathbf{s}^t, \mathbf{a}^t)} Q^{\pi}(\mathbf{s}^t, \mathbf{a}^t) \right] = \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \frac{\nabla_{\theta} \pi(\mathbf{s}^t, \mathbf{a}^t)}{\pi(\mathbf{s}^t, \mathbf{a}^t)} G_{t:T} \right]\end{aligned}$$

Towards Gradient Ascent

- Generate episode $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \dots, s^T = s_T$ by acting according to π , parameterised by θ . Now consider:

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \nabla_{\theta} \pi(s^t, a) Q^{\pi}(s^t, a) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \sum_{a \in A} \pi(s^t, a) \frac{\nabla_{\theta} \pi(s^t, a)}{\pi(s^t, a)} Q^{\pi}(s^t, a) \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \frac{\nabla_{\theta} \pi(s^t, a^t)}{\pi(s^t, a^t)} Q^{\pi}(s^t, a^t) \right] = \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \frac{\nabla_{\theta} \pi(s^t, a^t)}{\pi(s^t, a^t)} G_{t:T} \right] \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} (\nabla_{\theta} \ln \pi(s^t, a^t)) G_{t:T} \right].\end{aligned}$$

REINFORCE Algorithm

- Reference: Williams (1992).
- For clarity we show explicit dependence of π on parameter vector $\theta \in \mathbb{R}^d$.
- Assume θ is initialised arbitrarily.

Repeat for ever:

$$\theta_{\text{new}} \leftarrow \theta.$$

Generate episode $s^0, a^0, r^0, s^1, \dots, s^T = s_T$, following π_θ .

For $t = 0, 1, \dots, T - 1$:

$$G \leftarrow \sum_{k=t}^{T-1} r^k. \text{ // This is } G_{t:T}.$$

$$\theta_{\text{new}} \leftarrow \theta_{\text{new}} + \alpha G \nabla_{\theta} \ln \pi_{\theta}(s^t, a^t).$$

$$\theta \leftarrow \theta_{\text{new}}.$$

REINFORCE Algorithm

- Reference: Williams (1992).
- For clarity we show explicit dependence of π on parameter vector $\theta \in \mathbb{R}^d$.
- Assume θ is initialised arbitrarily.

Repeat for ever:

$\theta_{\text{new}} \leftarrow \theta$.

Generate episode $s^0, a^0, r^0, s^1, \dots, s^T = s_T$, following π_θ .

For $t = 0, 1, \dots, T - 1$:

$G \leftarrow \sum_{k=t}^{T-1} r^k$. //This is $G_{t:T}$.

$\theta_{\text{new}} \leftarrow \theta_{\text{new}} + \alpha G \nabla_{\theta} \ln \pi_{\theta}(s^t, a^t)$.

//REward Increment = Nonnegative Factor \times

//Offset Reinforcement \times Characteristic Eligibility.

$\theta \leftarrow \theta_{\text{new}}$.

Reinforcement Learning

1. Policy gradient methods
2. Variance reduction
3. Actor-critic methods

Baseline Subtraction

- Policy Gradient Theorem

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}).$$

Baseline Subtraction

- Policy Gradient Theorem

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}).$$

- Let $B : \mathcal{S} \rightarrow \mathbb{R}$ be an *arbitrary* function of state.

Baseline Subtraction

- Policy Gradient Theorem

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(\mathbf{s}, a) Q^{\pi}(\mathbf{s}, a).$$

- Let $B : \mathcal{S} \rightarrow \mathbb{R}$ be an *arbitrary* function of state. We claim

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(\mathbf{s}, a) (Q^{\pi}(\mathbf{s}, a) - B(\mathbf{s})).$$

Baseline Subtraction

- Policy Gradient Theorem

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) Q^{\pi}(\mathbf{s}, \mathbf{a}).$$

- Let $B : \mathcal{S} \rightarrow \mathbb{R}$ be an *arbitrary* function of state. We claim

$$\nabla_{\theta} J(\theta) = \sum_{\mathbf{s} \in \mathcal{S}} \sum_{t=0}^{\infty} \mathbb{P}\{\mathbf{s}^0 \rightarrow \mathbf{s}, t, \pi\} \sum_{\mathbf{a} \in \mathcal{A}} \nabla_{\theta} \pi(\mathbf{s}, \mathbf{a}) (Q^{\pi}(\mathbf{s}, \mathbf{a}) - B(\mathbf{s})).$$

- How come?

Baseline Subtraction

- Policy Gradient Theorem

$$\nabla_{\theta} J(\theta) = \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) Q^{\pi}(s, a).$$

- Let $B : S \rightarrow \mathbb{R}$ be an *arbitrary* function of state. We claim

$$\nabla_{\theta} J(\theta) = \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) (Q^{\pi}(s, a) - B(s)).$$

- How come? Observe that

$$\begin{aligned} & \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} \sum_{a \in A} \nabla_{\theta} \pi(s, a) B(s) \\ &= \sum_{s \in S} \sum_{t=0}^{\infty} \mathbb{P}\{s^0 \rightarrow s, t, \pi\} B(s) \nabla_{\theta} \sum_{a \in A} \pi(s, a) = 0. \end{aligned}$$

Baseline Subtraction

- The policy gradient estimate can have high variance.

s	$Q^\pi(s, a_1)$	$Q^\pi(s, a_2)$	$Q^\pi(s, a_3)$	$V^\pi(s)$
s_1	105	79	100	90
s_2	10	6	13	12
s_3	-50	-60	-50	-55

Baseline Subtraction

- The policy gradient estimate can have high variance.

s	$Q^\pi(s, a_1)$	$Q^\pi(s, a_2)$	$Q^\pi(s, a_3)$	$V^\pi(s)$
s_1	105	79	100	90
s_2	10	6	13	12
s_3	-50	-60	-50	-55

- Common to subtract out $V^\pi(s)$ —approximated independently as $\hat{V}(s)$.
- REINFORCE with baseline: revise pseudocode to

$$\theta_{\text{new}} \leftarrow \theta_{\text{new}} + \alpha \sum_{t=0}^{T-1} (G_{t:T} - \hat{V}(s^t)) \nabla_{\theta} \ln \pi_{\theta}(s^t, a^t).$$

Reinforcement Learning

1. Policy gradient methods
2. Variance reduction
3. Actor-critic methods

Actor-critic Methods

- Even for fixed (s^t, a^t) , can have high variance in $G_{t:T}$.

Actor-critic Methods

- Even for fixed (s^t, a^t) , can have high variance in $G_{t:T}$.
- One approach is to do gradient ascent after averaging the gradient from a few episodes.

Actor-critic Methods

- Even for fixed (s^t, a^t) , can have high variance in $G_{t:T}$.
- One approach is to do gradient ascent after averaging the gradient from a few episodes.
- Another approach is to **bootstrap**: to use $r^t + \hat{V}(s^{t+1})$ in place of $G_{t:T}$, where $\hat{V}(s^{t+1})$ is estimated independently.

Actor-critic Methods

- Even for fixed (s^t, a^t) , can have high variance in $G_{t:T}$.
- One approach is to do gradient ascent after averaging the gradient from a few episodes.
- Another approach is to **bootstrap**: to use $r^t + \hat{V}(s^{t+1})$ in place of $G_{t:T}$, where $\hat{V}(s^{t+1})$ is estimated independently.
- Called the **Actor-Critic** architecture.
 - **Actor** updates θ and hence π_θ .
 - **Critic** evaluates π_θ (say using TD(0)) and provides input for the gradient ascent update.

$$\theta_{\text{new}} \leftarrow \theta_{\text{new}} + \alpha \sum_{t=0}^{T-1} (r^t + \hat{V}(s^{t+1}) - \hat{V}(s^t)) \nabla_{\theta} \ln \pi_{\theta}(s^t, a^t).$$

Actor-critic Methods

- Even for fixed (s^t, a^t) , can have high variance in $G_{t:T}$.
- One approach is to do gradient ascent after averaging the gradient from a few episodes.
- Another approach is to **bootstrap**: to use $r^t + \hat{V}(s^{t+1})$ in place of $G_{t:T}$, where $\hat{V}(s^{t+1})$ is estimated independently.
- Called the **Actor-Critic** architecture.
 - **Actor** updates θ and hence π_θ .
 - **Critic** evaluates π_θ (say using TD(0)) and provides input for the gradient ascent update.

$$\theta_{\text{new}} \leftarrow \theta_{\text{new}} + \alpha \sum_{t=0}^{T-1} (r^t + \hat{V}(s^{t+1}) - \hat{V}(s^t)) \nabla_{\theta} \ln \pi_{\theta}(s^t, a^t).$$

- Not always provably convergent, but widely used in practice.

Reinforcement Learning

1. Policy gradient methods
2. Variance reduction
3. Actor-critic methods

Reinforcement Learning

1. Policy gradient methods
2. Variance reduction
3. Actor-critic methods

Next class: Batch reinforcement learning