Last Time

Last Time

- Bandits

Explore - Exploit

Greedy

→ ε- Greedy

softmax

→ UCB

Thompson

Interval

Guiding Questions

Guiding Questions

- What is Policy Gradient?

Guiding Questions

- What is Policy Gradient?
- What tricks are needed for it to work effectively?

Map

Map

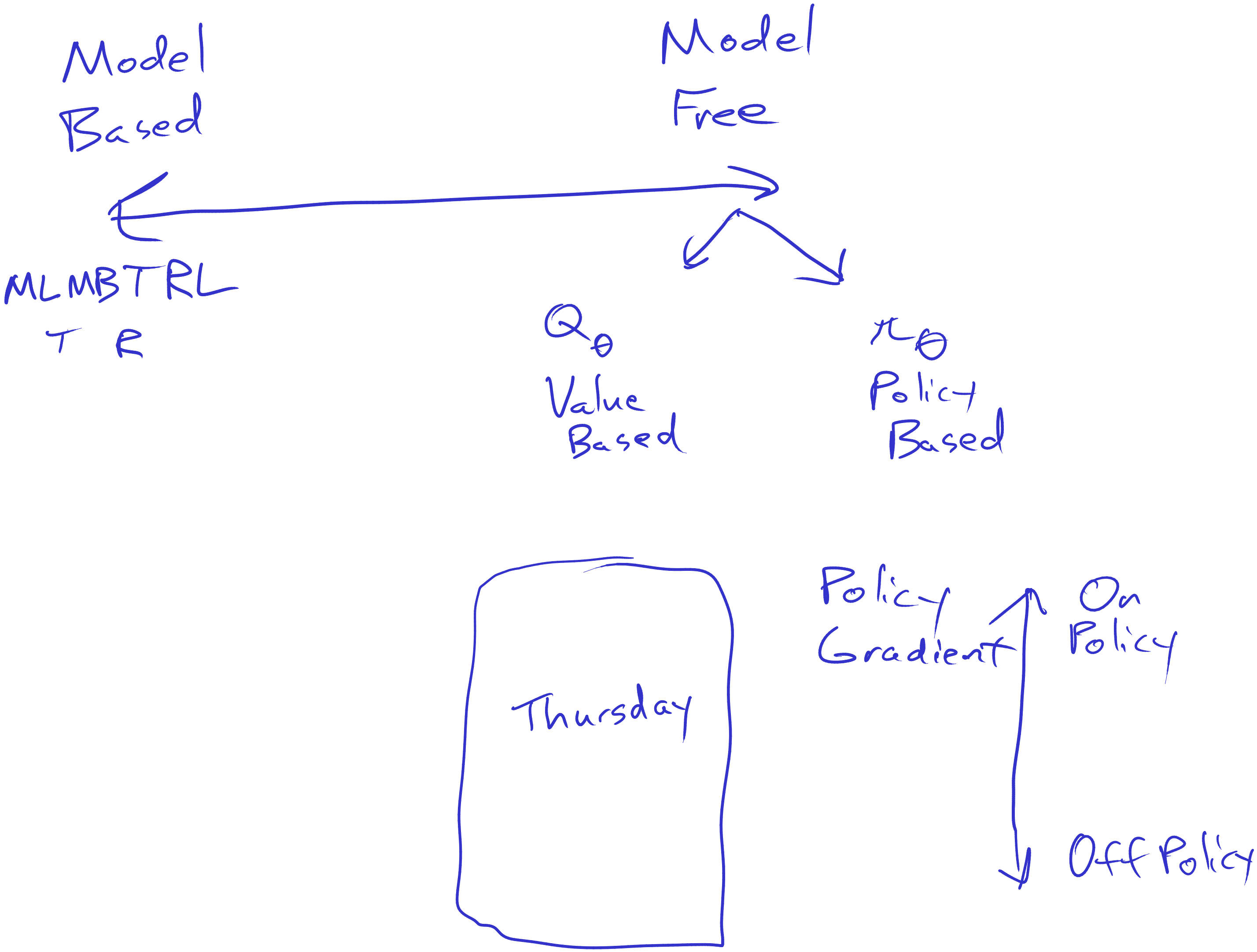
Challenges in RL

- Exploration and Exploitation
- Credit Assignment
- Generalization

Map

Challenges in RL

- Exploration and Exploitation
- Credit Assignment ←
- Generalization



Review: Gradient Ascent

Review: Gradient Ascent

$U(\theta)$

- Definition of Gradient

Review: Gradient Ascent

- Definition of Gradient
- Gradient Ascent

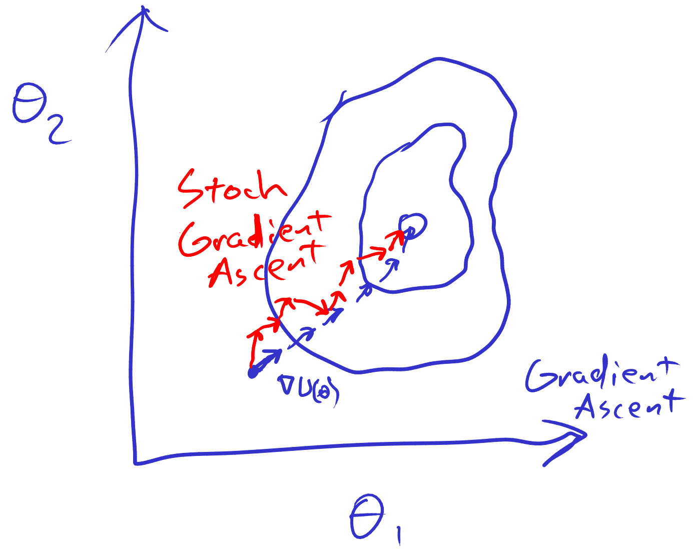
Review: Gradient Ascent

$U(\theta)$

$$\nabla U(\theta) = \left[\frac{\partial U}{\partial \theta_1}, \ldots \frac{\partial U}{\partial \theta_n}\right]$$

loop

$$\theta \leftarrow \theta + \alpha \underline{\widehat{\nabla U(\theta)}}$$



$$\nabla U(\theta) = E\left[\widehat{\nabla U(\theta)}\right]$$

π_θ

$$U(\theta) = E\left[\sum_{k=0}^{d} \gamma^k r_k\right]$$

- Definition of Gradient
- Gradient Ascent
- Stochastic Gradient Ascent

Additional Notation

Additional Notation

- Probabilistic parameterized policies

Additional Notation

- Probabilistic parameterized policies
- initial state distribution

Additional Notation

- Probabilistic parameterized policies
- initial state distribution
- trajectory: $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$

Additional Notation

$\pi_\theta(a|s)$

$(S, A, R, T, \gamma) \longrightarrow (S, A, R, T, \gamma, \rho(s_0))$

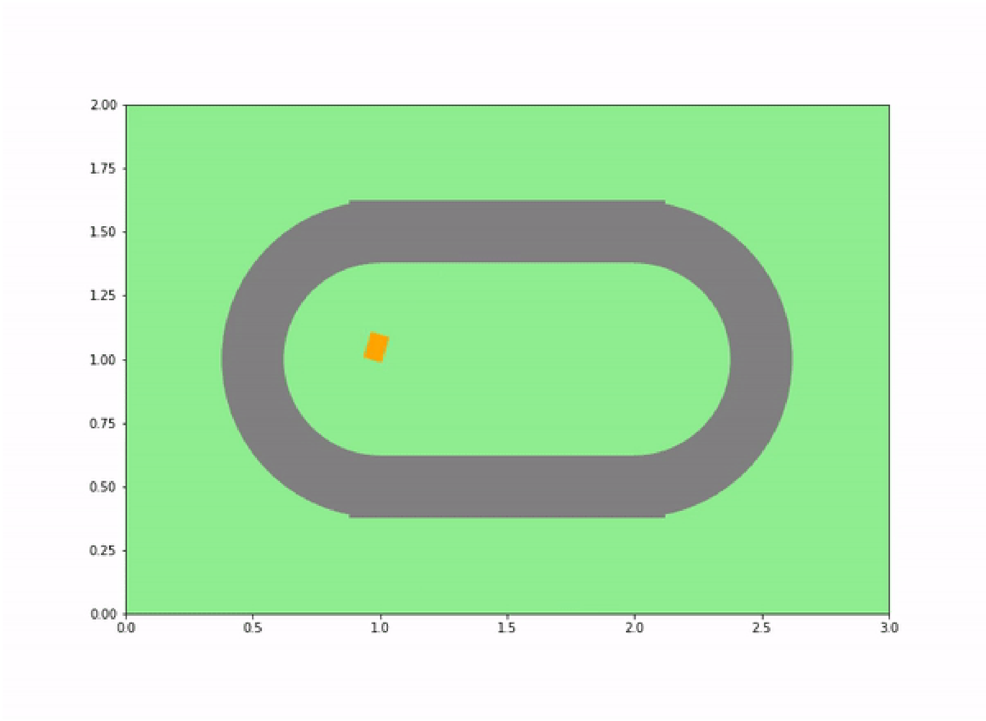
$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$

$A(s,a) = Q(s,a) - V(s)$

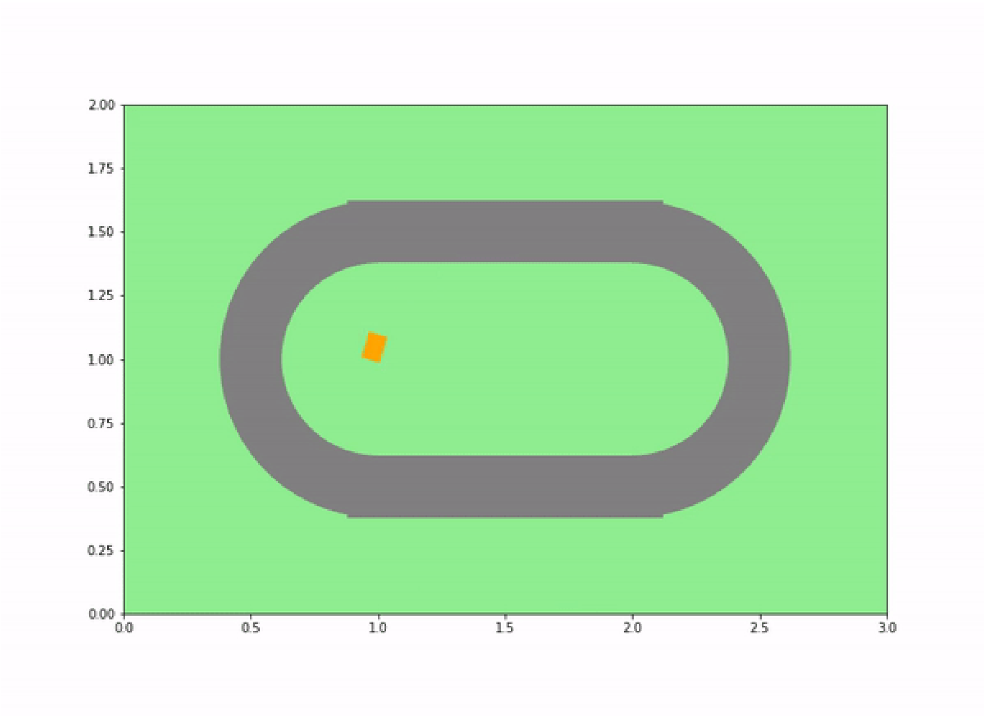
- Probabilistic parameterized policies
- initial state distribution
- trajectory: $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$
- advantage function

Tricks

Tricks



Tricks



$\nabla_\theta U(\theta)$

For policy gradient, 3 tricks

- Likelihood Ratio
- Reward to go
- Baseline Subtraction

Log Derivative

$\sum \gamma^t r_t = R(\tau)$

Log Derivative

$$U(\theta) = \mathbf{E}[R(\tau)]$$

Log Derivative

$$
\begin{aligned}
U(\theta) &= \mathrm{E}[R(\tau)] \\
&= \int p_\theta(\tau) R(\tau)\, d\tau
\end{aligned}
$$

Log Derivative

$$\begin{aligned} U(\theta) &= \mathbf{E}[R(\tau)] \\ &= \int p_\theta(\tau) R(\tau)\, d\tau \end{aligned}$$

$$\nabla U(\theta) = \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau$$

Log Derivative

$$
\begin{aligned}
U(\theta) &= \mathbf{E}[R(\tau)] \\
&= \int p_\theta(\tau) R(\tau)\, d\tau
\end{aligned}
$$

$\frac{\partial \log}{dx}$ $\quad$ $1/x$

$\nabla \log \theta =$

$$
\begin{aligned}
\nabla U(\theta) &= \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau \\
&= \int \underline{\nabla_\theta\, p_\theta(\tau)} R(\tau)\, d\tau
\end{aligned}
$$

Log Derivative

$$
\begin{aligned}
U(\theta) &= \mathrm{E}[R(\tau)] \\
&= \int p_\theta(\tau) R(\tau)\, d\tau
\end{aligned}
$$

$$
\begin{aligned}
\nabla U(\theta) &= \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau \\
&= \int \nabla_\theta \, p_\theta(\tau) R(\tau)\, d\tau
\end{aligned}
$$

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta \, p_\theta(\tau) / p_\theta(\tau)$$

Log Derivative

$$U(\theta) = \mathbf{E}[R(\tau)]$$

$$= \int p_\theta(\tau) R(\tau)\, d\tau$$

$$\nabla U(\theta) = \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau$$

$$= \int \nabla_\theta\, p_\theta(\tau) R(\tau)\, d\tau$$

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta\, p_\theta(\tau) / p_\theta(\tau)$$

$$\therefore \quad \nabla_\theta\, p_\theta(\tau) = p_\theta(\tau)\, \nabla_\theta \log p_\theta(\tau)$$

Log Derivative

$$\begin{aligned} U(\theta) &= \mathbf{E}[R(\tau)] \\ &= \int p_\theta(\tau) R(\tau)\, d\tau \end{aligned}$$

$$\begin{aligned} \nabla U(\theta) &= \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau \\ &= \int \nabla_\theta\, p_\theta(\tau) R(\tau)\, d\tau \end{aligned}$$

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta\, p_\theta(\tau) / p_\theta(\tau)$$

$$\therefore \quad \boxed{\nabla_\theta\, p_\theta(\tau) = p_\theta(\tau)\, \nabla_\theta \log p_\theta(\tau)}$$

Log Derivative

$$U(\theta) = \mathrm{E}[R(\tau)]$$
$$= \int p_\theta(\tau) R(\tau)\, d\tau$$

$$\underline{\nabla U(\theta)} = \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau$$
$$= \int \nabla_\theta\, p_\theta(\tau) R(\tau)\, d\tau$$

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta\, p_\theta(\tau) / p_\theta(\tau)$$

$$\therefore \boxed{\nabla_\theta\, p_\theta(\tau) = p_\theta(\tau)\, \nabla_\theta \log p_\theta(\tau)}$$

$$= \int \underline{p_\theta(\tau)} \underline{\nabla_\theta \log p_\theta(\tau) R(\tau)}\, d\tau$$

$$\nabla U = E[\nabla U)$$

Log Derivative

$$U(\theta) = \mathrm{E}[R(\tau)]$$
$$= \int p_\theta(\tau) R(\tau)\, d\tau$$

$$\nabla U(\theta) = \nabla_\theta \int p_\theta(\tau) R(\tau)\, d\tau$$
$$= \int \nabla_\theta p_\theta(\tau) R(\tau)\, d\tau$$
$$= \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) R(\tau)\, d\tau$$
$$= \mathrm{E}\left[\underline{\nabla_\theta \log p_\theta(\tau)} R(\tau)\right]$$

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta p_\theta(\tau) / p_\theta(\tau)$$

$$\therefore \quad \boxed{\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)}$$

$\nabla_\theta \log \pi_\theta$

$\nabla U(\theta)$

Trajectory Probability Gradient

Trajectory Probability Gradient

$$\nabla_\theta \log p_\theta(\tau)$$

Trajectory Probability Gradient

$\nabla_\theta \log p_\theta(\tau)$

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$

Trajectory Probability Gradient

$$\nabla_\theta \log p_\theta(\tau) \qquad \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$$

$$p_\theta(\tau) = p(s_0) \prod_{k=0}^{d} \underline{T(s_{k+1} \mid s_k, a_k)}\, \underline{\pi_\theta(a_k \mid s_k)}$$

Trajectory Probability Gradient

$\nabla_\theta \log p_\theta(\tau)$

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$

$$p_\theta(\tau) = p(s_0) \prod_{k=0}^{d} T(s_{k+1} \mid s_k, a_k)\, \pi_\theta(a_k \mid s_k)$$

$\log p_\theta(\tau)$

Trajectory Probability Gradient

$$\underline{\nabla_\theta} \log p_\theta(\tau) \qquad \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$$

$\log(ab) = \log a + \log b$

$$p_\theta(\tau) = p(s_0) \prod_{k=0}^{d} T(s_{k+1} \mid s_k, a_k)\, \pi_\theta(a_k \mid s_k)$$

$$\nabla \log p_\theta(\tau) = \overset{0}{\cancel{\underline{\log p(s_0)}}} + \overset{0}{\cancel{\underline{\sum_{k=0}^{d} \log T(s_{k+1} \mid s_k, a_k)}}} + \sum_{k=0}^{d} \log \pi_\theta(a_k \mid s_k)$$

Trajectory Probability Gradient

$$\nabla_\theta \log p_\theta(\tau) \qquad \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$$

$$p_\theta(\tau) = p(s_0) \prod_{k=0}^{d} T(s_{k+1} \mid s_k, a_k)\, \pi_\theta(a_k \mid s_k)$$

$$\log p_\theta(\tau) = \log p(s_0) + \sum_{k=0}^{d} \log T(s_{k+1} \mid s_k, a_k) + \sum_{k=0}^{d} \log \pi_\theta(a_k \mid s_k)$$

$$\nabla_\theta \log p_\theta(\tau)$$

Trajectory Probability Gradient

$$\nabla_\theta \log p_\theta(\tau) \qquad \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots s_d, a_d, r_d)$$

$$p_\theta(\tau) = p(s_0) \prod_{k=0}^{d} T(s_{k+1} \mid s_k, a_k)\, \pi_\theta(a_k \mid s_k)$$

$$\log p_\theta(\tau) = \log p(s_0) + \sum_{k=0}^{d} \log T(s_{k+1} \mid s_k, a_k) + \sum_{k=0}^{d} \log \pi_\theta(a_k \mid s_k)$$

$$\nabla_\theta \log p_\theta(\tau) = \sum_{k=0}^{d} \underline{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}$$

Grid World

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

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

Trajectory Probability Gradient

$$\underline{\nabla_\theta \log p_\theta(\tau)} \qquad \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots s_d, a_d, r_d)$$

$$p_\theta(\tau) = p(s_0) \prod_{k=0}^{d} T(s_{k+1} \mid s_k, a_k)\, \pi_\theta(a_k \mid s_k)$$

$$\log p_\theta(\tau) \;= \log p(s_0) + \sum_{k=0}^{d} \log T(s_{k+1} \mid s_k, a_k) + \sum_{k=0}^{d} \log \pi_\theta(a_k \mid s_k)$$

$$\nabla_\theta \log p_\theta(\tau) \;= \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)$$

$$\nabla U(\theta) = E\left[\nabla_\theta \log p_\theta(\tau)\, R(\tau) \right]$$

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau) \right]$$

Policy Gradient

Policy Gradient

```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \underline{\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)}$
```

Policy Gradient

On Policy - only use
data from current policy

Off Policy

```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \underline{\nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)}$
```

On Policy!

Causality

Causality

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right]$$

Causality

$$
\begin{aligned}
\nabla U(\theta) &= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right] \\
&= \mathrm{E}\left[\left(\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]
\end{aligned}
$$

Causality

$$\begin{aligned}
\nabla U(\theta) &= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right] \\
&= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]
\end{aligned}$$

Causality

$$\begin{aligned}
\nabla U(\theta) &= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right] \\
&= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right] \\
&= \mathrm{E}\left[\left(f_0 + \ldots + f_d\right)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right]
\end{aligned}$$

Causality

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right]$$

$$= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(\underline{a_k} \mid \underline{s_k})}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]$$

$$= \mathrm{E}\left[\left(f_0 + \ldots + f_d\right)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right]$$

$$= E\begin{bmatrix} f_0\gamma^0 r_0 + f_0\gamma^1 r_1 + f_0\gamma^2 r_2 + \ldots + f_0\gamma^d r_d \\ + f_1\gamma^0 r_0 + f_1\gamma^1 r_1 + f_1\gamma^2 r_2 + \ldots f_1\gamma^d r_d \\ \vdots \\ + f_d\gamma^0 r_0 + f_d\gamma^1 r_1 + f_d\gamma^2 r_2 + \ldots f_d\gamma^d r_d \end{bmatrix}$$

Causality

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right]$$

$$= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]$$

$$= \mathrm{E}\left[\left(f_0 + \ldots + f_d\right)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right]$$

$$= E\left[\begin{array}{l} f_0\gamma^0 r_0 + f_0\gamma^1 r_1 + f_0\gamma^2 r_2 + \ldots + f_0\gamma^d r_d \\ + \cancel{f_1\gamma^0} r_0 + f_1\gamma^1 r_1 + f_1\gamma^2 r_2 + \ldots f_1\gamma^d r_d \\ \vdots \\ + \cancel{f_d\gamma^0} r_0 + \cancel{f_d\gamma^1} r_1 + \cancel{f_d\gamma^2} r_2 + \ldots f_d\gamma^d r_d \end{array}\right]$$

Causality

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right]$$

$$= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]$$

$$= \mathrm{E}\left[(f_0 + \ldots + f_d)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right]$$

$$= E\left[\begin{array}{l} f_0\gamma^0 r_0 + f_0\gamma^1 r_1 + f_0\gamma^2 r_2 + \ldots + f_0\gamma^d r_d \\ + \cancel{f_1\gamma^0 r_0} + f_1\gamma^1 r_1 + f_1\gamma^2 r_2 + \ldots f_1\gamma^d r_d \\ \vdots \\ + \cancel{f_d\gamma^0 r_0} + \cancel{f_d\gamma^1 r_1} + \cancel{f_d\gamma^2 r_2} + \ldots f_d\gamma^d r_d \end{array}\right]$$

$$= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\left(\sum_{l=k}^{d} \gamma^l r_l\right)\right]$$

Causality

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \underline{R(\tau)}\right]$$

$$= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]$$

$$= \mathrm{E}\left[(f_0 + \ldots + f_d)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right]$$

$$= E\left[\begin{array}{l} f_0\gamma^0 r_0 + f_0\gamma^1 r_1 + f_0\gamma^2 r_2 + \ldots + f_0\gamma^d r_d \\ + \cancel{f_1\gamma^0 r_0} + f_1\gamma^1 r_1 + f_1\gamma^2 r_2 + \ldots f_1\gamma^d r_d \\ \vdots \\ + \cancel{f_d\gamma^0 r_0} + \cancel{f_d\gamma^1 r_1} + \cancel{f_d\gamma^2 r_2} + \ldots f_d\gamma^d r_d \end{array}\right]$$

$$= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\left(\sum_{l=k}^{d} \gamma^l r_l\right)\right] \qquad = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k \underline{r_{k,\text{to-go}}}\right]$$

Causality

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right]$$

$$= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right]$$

$$= \mathrm{E}\left[(f_0 + \ldots + f_d)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right]$$

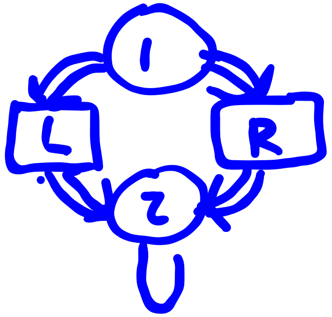
$$= E\begin{bmatrix} f_0\gamma^0 r_0 + f_0\gamma^1 r_1 + f_0\gamma^2 r_2 + \ldots + f_0\gamma^d r_d \\ + \cancel{f_1\gamma^0 r_0} + f_1\gamma^1 r_1 + f_1\gamma^2 r_2 + \ldots f_1\gamma^d r_d \\ \vdots \\ + \cancel{f_d\gamma^0 r_0} + \cancel{f_d\gamma^1 r_1} + \cancel{f_d\gamma^2 r_2} + \ldots f_d\gamma^d r_d \end{bmatrix}$$

$$= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\left(\sum_{l=k}^{d} \gamma^l r_l\right)\right] \qquad = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k \underline{r_{k,\text{to-go}}}\right] \; Q^\theta(s_k, a_k)$$

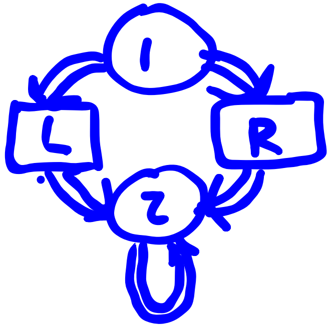
Discuss

U

Discuss

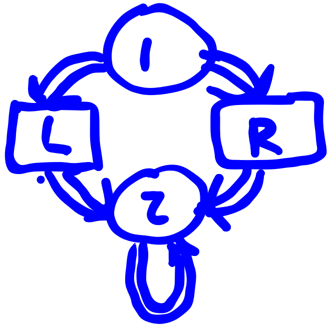


Discuss



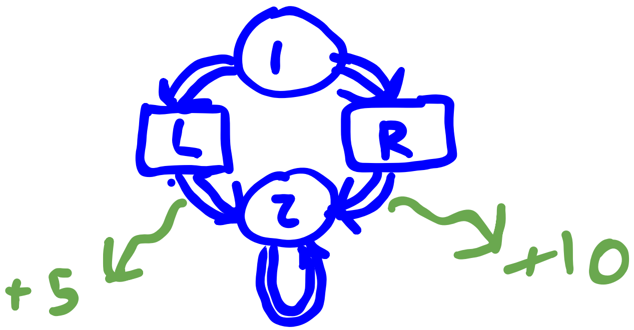
Discuss

$A = \{L, R\}$



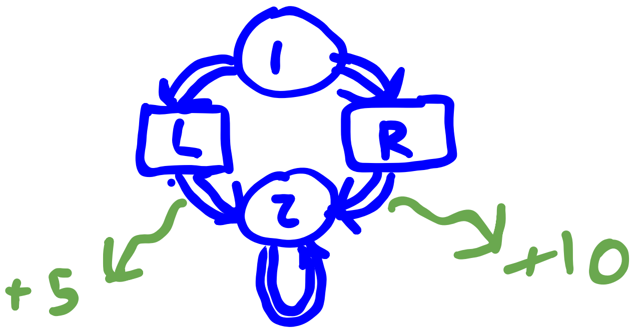
Discuss

$A = \{L, R\}$



Discuss

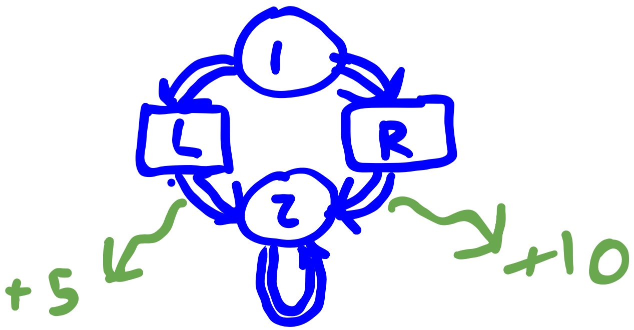
A={L,R}



```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta \underline{(a_k \mid s_k)} \gamma^k r_{k,\text{to go}}$
```

Discuss

$A = \{L, R\}$

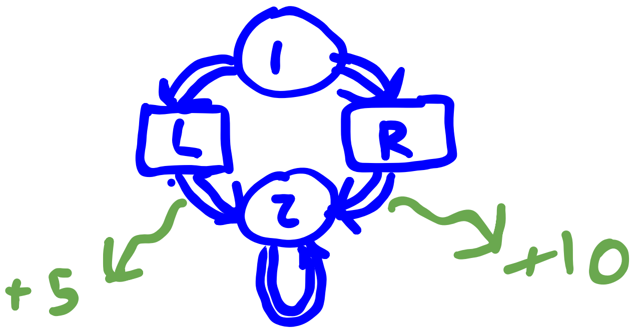


```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k r_{k,\text{to go}}$
```

1. 1. Given $\theta = (0.2, 0.8)$ calculate $\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k r_{k,\text{to go}}$ for two cases, (a) where $a_0 = R$ and (b) where $a_0 = L$

Discuss

$A = \{L, R\}$

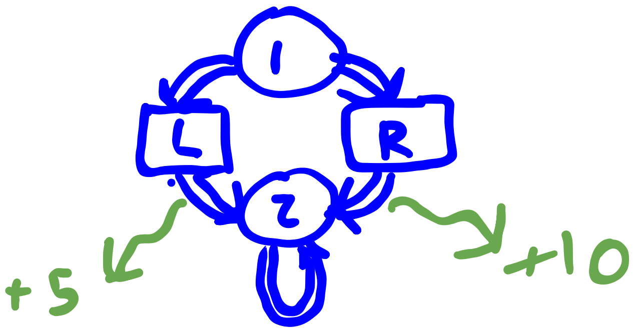


```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k r_{k,\text{to go}}$
```

1. 1. Given $\theta = (0.2, 0.8)$ calculate $\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k r_{k,\text{to go}}$ for two cases, (a) where $a_0 = R$ and (b) where $a_0 = L$
2. What happens if $\theta_1 \to 0$

Discuss

$A = \{L, R\}$



$$\pi_\theta(a = L \mid s) = \frac{\theta_1}{\theta_1 + \theta_2}$$

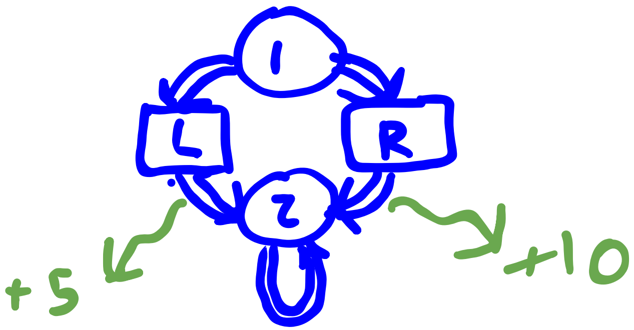
```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\gamma^k r_{k,\text{to go}}$
```

1. 1. Given $\theta = (0.2, 0.8)$ calculate $\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\gamma^k r_{k,\text{to go}}$ for two cases, (a) where $a_0 = R$ and (b) where $a_0 = L$
2. What happens if $\theta_1 \to 0$

Discuss

$A = \{L, R\}$

$\theta = (\theta_1, \theta_2)$



$$\pi_\theta(a = L \mid s) = \frac{\theta_1}{\theta_1 + \theta_2}$$

$$\pi_\theta(a = R \mid s) = \frac{\theta_2}{\theta_1 + \theta_2}$$

$\log \pi_\theta(a=L|s) = \log \theta_1 - \log(\theta_1+\theta_2)$

$\frac{\partial \log \pi_\theta(a=L|s)}{\partial \theta_1} = \frac{1}{\theta_1} - \frac{1}{\theta_1+\theta_2}$

$\frac{\partial \log \pi_\theta(a=L|s)}{\partial \theta_2} = -\frac{1}{\theta_1+\theta_2}$

$\nabla_\theta \log \pi_\theta(a_0=L|s_0=1) = \left[\frac{1}{0.2} - \frac{1}{1}\right] = \begin{bmatrix}4\\-1\end{bmatrix}$

$\nabla_\theta \log \pi_\theta(a_0=L|s_0=1)\, r_{togo} = \begin{bmatrix}4\\-1\end{bmatrix}(5 - 7.5)$

$\nabla_\theta \log \pi_\theta(a_0=R|s_0=1)\, r_{togo} = \begin{bmatrix}-1\\0.25\end{bmatrix}(10 - 7.5)$

```
loop
    $\tau \leftarrow \text{simulate}(\pi_\theta)$
    $\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \underline{\nabla_\theta \log \pi_\theta(a_k \mid s_k)\cancel{\gamma^k} r_{k,\text{to go}}}$
```

1. 1. Given $\theta = (0.2, 0.8)$ calculate $\sum_{k=0}^{d} \underline{\nabla_\theta \log \pi_\theta(a_k \mid s_k)\gamma^k r_{k,\text{to go}}}$ for two cases, (a) where $\underline{a_0 = R}$ and (b) where $\underline{a_0 = L}$
2. What happens if $\theta_1 \to 0$

Baseline Subtraction

Baseline Subtraction

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k r_{k,\text{to-go}}\right]$$

Baseline Subtraction

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k r_{k,\text{to-go}}\right]$$

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k \left(r_{k,\text{to-go}} - r_{\text{base}}(s_k)\right)\right]$$

Baseline Subtraction

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k r_{k,\text{to-go}}\right]$$

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k \left(r_{k,\text{to-go}} - \underline{r_{\text{base}}(s_k)}\right)\right]$$

does not bias
(proof in book)

Baseline Subtraction

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k r_{k,\text{to-go}}\right]$$

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k \left(r_{k,\text{to-go}} - \underline{r_{\text{base}}(s_k)}\right)\right]$$

does not bias
(proof in book)

$$r_{\text{base},i} = \frac{\mathbb{E}_{a,s,r_{\text{to-go}},k}\left[\ell_i(a,s,k)^2 r_{\text{to-go}}\right]}{\mathbb{E}_{a,s,k}\left[\ell_i(a,s,k)^2\right]}$$

Baseline Subtraction

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k r_{k,\text{to-go}}\right]$$

$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\, \gamma^k \left(r_{k,\text{to-go}} - \underline{r_{\text{base}}(s_k)}\right)\right]$$

does not bias
(proof in book)

Optimal

$$r_{\text{base},i} = \frac{\mathbb{E}_{a,s,r_{\text{to-go}},k}\left[\ell_i(a,s,k)^2 r_{\text{to-go}}\right]}{\mathbb{E}_{a,s,k}\left[\ell_i(a,s,k)^2\right]}$$

$$\ell_i(a,s,k) = \gamma^{k-1} \frac{\partial}{\partial \theta_i} \log \pi_{\boldsymbol{\theta}}(a \mid s)$$

practical

$\hat{V}(s)$

Guiding Questions

π_θ

S.G.D.

$\nabla U(\theta)$

- What is Policy Gradient?
- What tricks are needed for it to work effectively?

- Log Derivative
- Causality
- Baseline subtraction