HW 3 Due next Tuesday

Last Time

- Bandits
  - ε-greedy
  - UCB
  - Bayesian
    - Quantile
    - Thompson Sampling

$$Q(s,a) + c\sqrt{\frac{\log(N)}{N(a)}}$$

Guiding Questions

Guiding Questions

- What is Policy Optimization?
- What is Policy Gradient?

Guiding Questions

- What is Policy Optimization?
- 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

Policy Optimization

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s \sim b}{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t) \right]$$

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s \sim b}{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t) \right]$$

$$\underset{\pi}{\text{maximize}} \, U(\pi) = \underset{s \sim b}{E} \left[U^{\pi}(s) \right]$$

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s \sim b}{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t) \right]$$

$$\underset{\pi}{\text{maximize}} \, U(\pi) = \underset{s \sim b}{E} \left[U^\pi(s) \right]$$

Two approximations:

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s\sim b}{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t)\right]$$

$$\underset{\pi}{\text{maximize}}\, U(\pi) = \underset{s\sim b}{E}\left[U^\pi(s)\right]$$

Two approximations:

$\pi(s) = \theta s$

1. Parameterized stochastic policies

$$\underset{\theta}{\text{maximize}} \quad U(\pi_\theta) = U(\theta)$$

$$a \sim \pi_\theta(a \mid s)$$

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s \sim b}{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t) \right]$$

$$\underset{\pi}{\text{maximize}} \, U(\pi) = \underset{s \sim b}{E} \left[U^\pi(s) \right]$$

Two approximations:

1. Parameterized stochastic policies

$$\underset{\theta}{\text{maximize}} \quad U(\pi_\theta) = U(\theta)$$

$$a \sim \pi_\theta(a \mid s)$$

2. Monte Carlo Utility

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

trajectory:

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

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s \sim b}{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t) \right]$$

$$\underset{\pi}{\text{maximize}} \, U(\pi) = \underset{s \sim b}{E} \left[U^\pi(s) \right]$$

Two approximations:

1. Parameterized stochastic policies

$$\underset{\theta}{\text{maximize}} \quad U(\pi_\theta) = U(\theta)$$

$$a \sim \pi_\theta(a \mid s)$$

2. Monte Carlo Utility

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

trajectory:

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

Two classes of optimization algorithms:

Policy Optimization

$$\underset{\pi}{\text{maximize}} \underset{s\sim b}{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t)\right]$$

$$\underset{\pi}{\text{maximize}}\, U(\pi) = \underset{s\sim b}{E}\left[U^\pi(s)\right]$$

Two approximations:

1. Parameterized stochastic policies $\underset{\theta}{\text{maximize}} \quad U(\pi_\theta) = U(\theta)$ $\quad a \sim \pi_\theta(a \mid s)$

2. Monte Carlo Utility $U(\pi) \approx \frac{1}{m}\sum_{i=1}^{m} R(\tau^{(i)})$

trajectory:

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

Two classes of optimization algorithms:

1. Zeroth order (use only $U(\theta)$)

2. First order (use $U(\theta)$ and $\nabla_\theta U(\theta)$)

1. Zeroth-Order Optimization

1. Zeroth-Order Optimization

Common zeroth-order aproaches:

1. Genetic Algorithms
2. Pattern Search
3. Cross-Entropy

1. Zeroth-Order Optimization

Common zeroth-order aproaches:

1. Genetic Algorithms
2. Pattern Search
3. Cross-Entropy

Cross Entropy:

Initialize d

loop:

population $\leftarrow$ sample(d)

elite $\leftarrow m$ with highest $U(\theta)$

$d \leftarrow$ fit(elite)

1. Zeroth-Order Optimization

Common zeroth-order aproaches:

1. Genetic Algorithms
2. Pattern Search
3. Cross-Entropy

Cross Entropy:

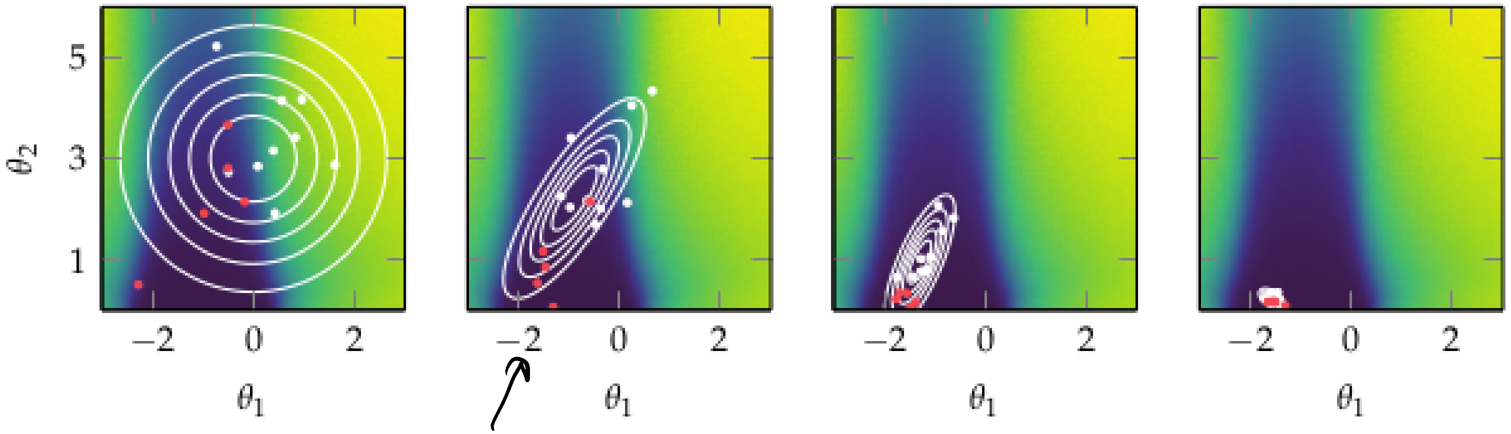
Initialize d

loop:

  population $\leftarrow$ sample(d)

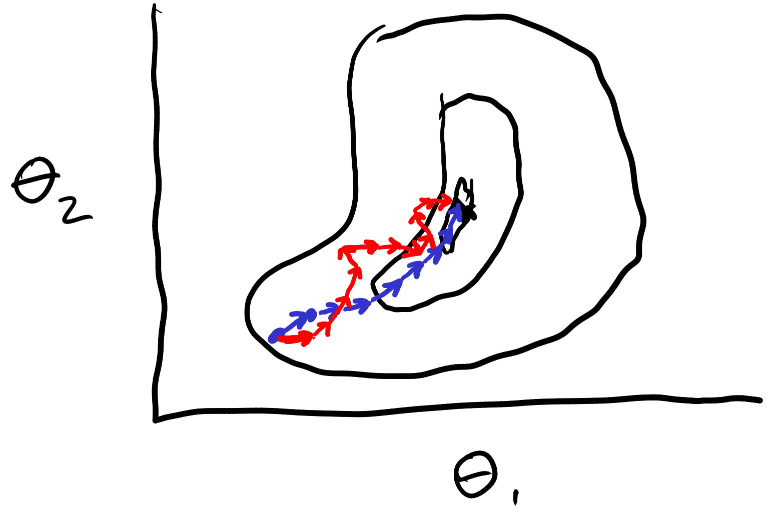
  elite $\leftarrow m$ with highest $U(\theta)$

  $d \leftarrow$ fit(elite)



2. First Order Optimization

$$\nabla_\theta U(\theta) = \left[\frac{\partial}{\partial\theta_1}U\Big|_\theta, \frac{\partial}{\partial\theta_2}U\Big|_\theta \dots \frac{\partial}{\partial\theta_2}U\Big|_\theta\right]$$



loop

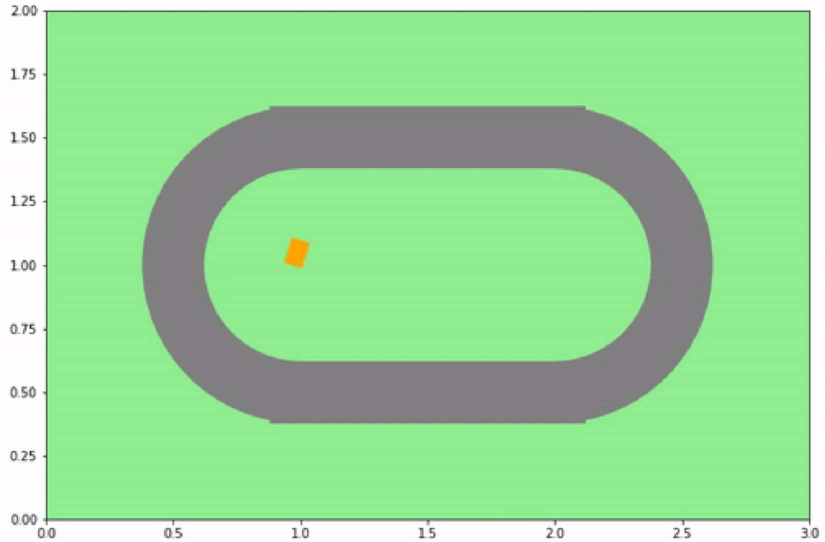
$$\theta \leftarrow \theta + \alpha \widehat{\nabla_\theta U(\theta)}$$

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

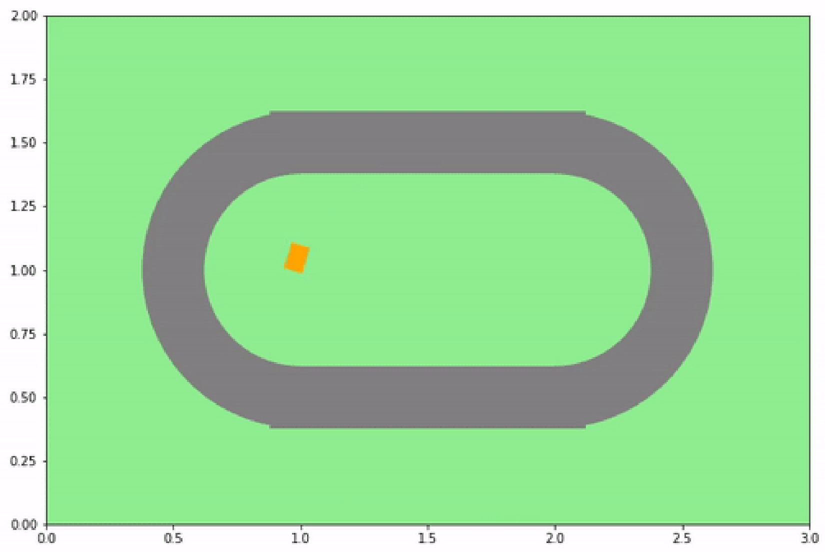
- Definition of Gradient
- Gradient Ascent
- Stochastic Gradient Ascent

Tricks

Tricks



Tricks



For policy gradient, 3 tricks

- Likelihood Ratio/Log Derivative
- Reward to go
- Baseline Subtraction

Log Derivative

$$U(\theta) = E[R(\tau)]$$
$$= \int p_\theta(\tau) R(\tau) d\tau$$

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

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

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

$$= E[\underbrace{\nabla_\theta \log p_\theta(\tau) R(\tau)}_{\nabla \hat{U}(\theta)}]$$

Trajectory Probability Gradient

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

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

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

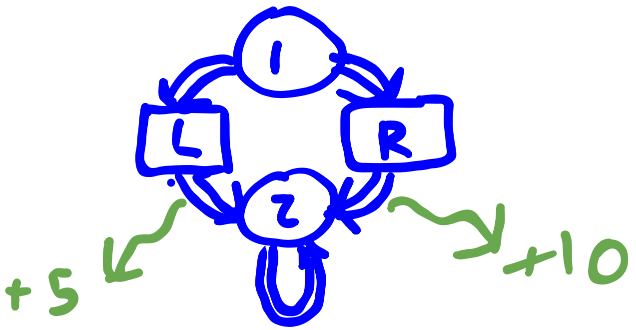
$$\log p_\theta(\tau) = \underline{\log b(s_0)} + \underline{\sum_{k=0}^{d} \log T(s_{k+1} \mid s_k, a_k)} + \sum_{k=0}^{d} \log \pi_\theta(a_k \mid s_k)$$

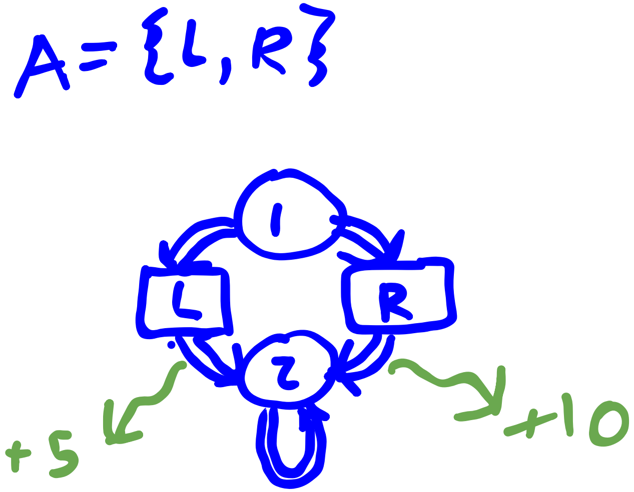
$$\nabla_\theta \log p_\theta(\tau) = \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k)$$

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

Discuss

$A = \{L, R\}$



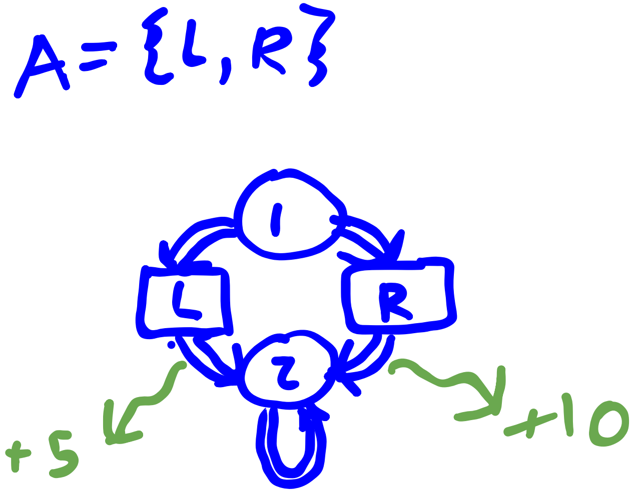


Discuss

$$\pi_\theta(a = L \mid s = 1) = \text{clamp}(\theta, 0, 1)$$

$$\pi_\theta(a = R \mid s = 1) = \text{clamp}(1 - \theta, 0, 1)$$

Discuss



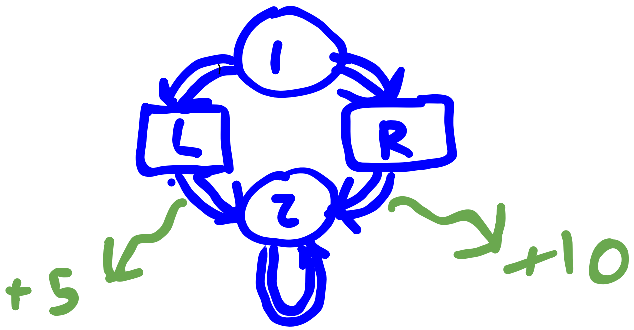
$$\pi_\theta(a = L \mid s = 1) = \text{clamp}(\theta, 0, 1)$$

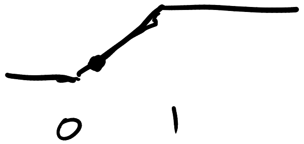
$$\pi_\theta(a = R \mid s = 1) = \text{clamp}(1 - \theta, 0, 1)$$

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

Discuss

$A=\{L,R\}$





$$\pi_\theta(a = L \mid s = 1) = \text{clamp}(\theta, 0, 1)$$

$$\pi_\theta(a = R \mid s = 1) = \text{clamp}(1-\theta, 0, 1)$$

(a) $a_0 = L$ $\quad R(\tau) = +5$

$$\nabla_\theta \log \pi_\theta(a \mid s) = \frac{\partial}{\partial \theta} \log \text{clamp}(\theta, 0, 1)\Big|_{\theta=0.2} = \frac{1}{\theta} = \frac{1}{0.2}$$

$$\widehat{\nabla U(\theta)} = \frac{1}{0.2} \cdot 5 = \underline{25}$$

(b) $a_0 = R$ $\quad R(\tau) = 10$

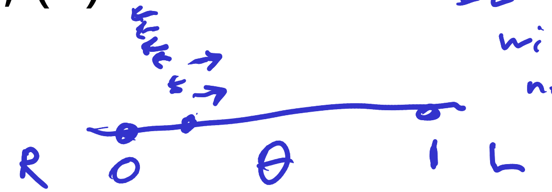
$$\nabla_\theta \log \pi_\theta(a \mid s) = \frac{\partial}{\partial \theta} \log \text{clamp}(1-\theta, 0, 1)\Big|_{\theta=0.2} = \frac{1}{1-\theta}(-1) = -\frac{1}{0.8}$$

$$\widehat{\nabla U(\theta)} = -12.5$$

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

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

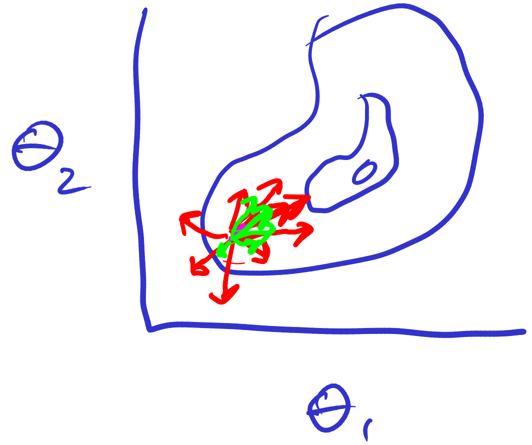
$\mathrm{E}[\quad]$ will be negative



Given $\theta = 0.2$ calculate $\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)$ for two cases, (a) where $a_0 = L$ and (b) where $a_0 = R$

Policy Gradient

Policy Gradient



loop

$\tau \leftarrow \text{simulate}(\pi_\theta)$

$\theta \leftarrow \theta + \alpha \underbrace{\sum_{m=1}^{M} \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_{k,m} \mid s_{k,m}) R(\tau_m)}_{\text{On-Policy}}$

Policy Gradient

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

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

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

Causality

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

Causality

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

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

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

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

Causality

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

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

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

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

Causality

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

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

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

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

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

Causality

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

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

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

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

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

Causality

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

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

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

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

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

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)

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

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

← Optimal Variance Reducing

In practice

$\hat{V}(s_k)$

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

$U(\theta)$

Guiding Questions

$\pi_\theta(a|s)$

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

log likelihood trick

Causality

Baseline subtraction