Policy Gradient

Last Time

- Bandits

Exploration vs. Exploitation

ε-greedy

→ UCB $\quad Q + c\sqrt{}$

Thompson Sampling

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 ← Bandits
- Credit Assignment ← Today
- 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:

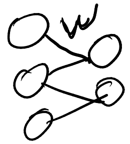
Policy Optimization

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

$\theta =$

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

$\pi_\theta(s) =$



Two approximations:

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 $U(\theta)$)

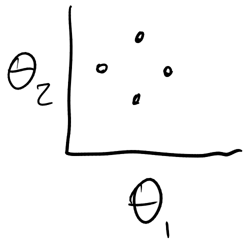
2. First order (use $U(\theta)$ and $\nabla_\theta U(\theta)$)

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:

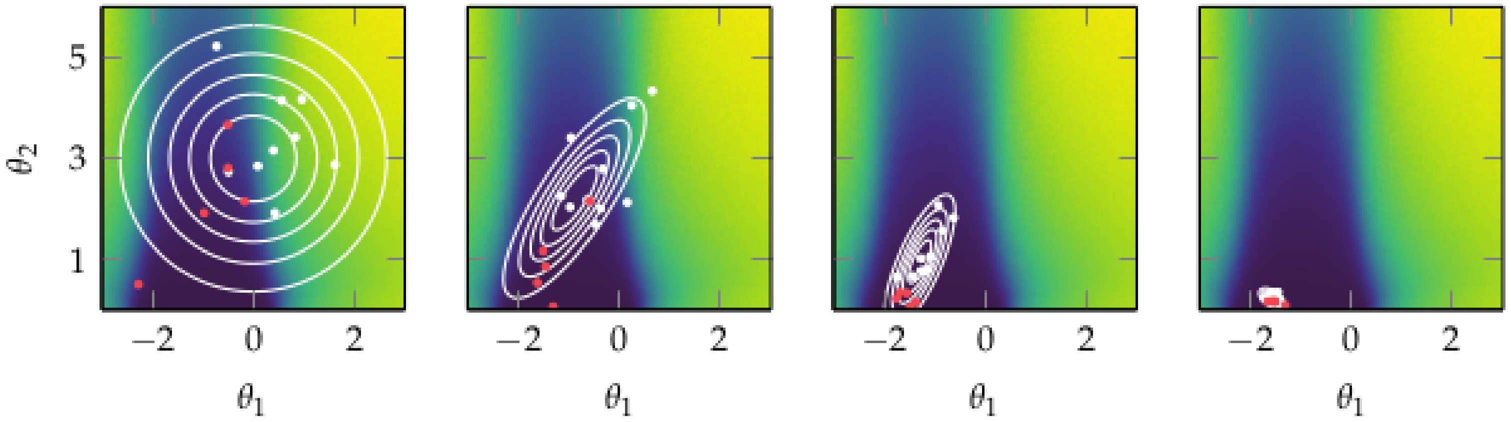
Initialize d

loop:

  population $\leftarrow$ sample(d)

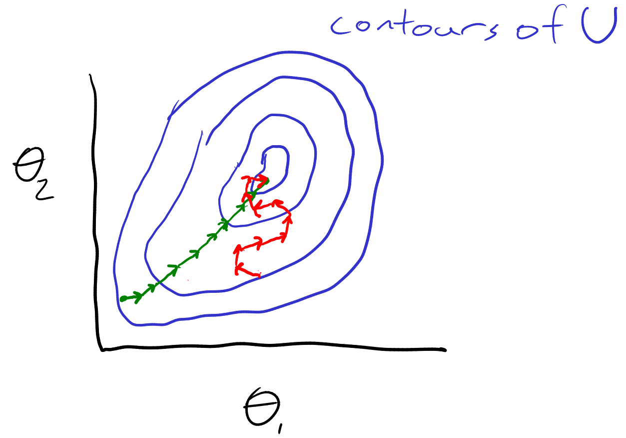
  elite $\leftarrow m$ with highest $U(\theta)$

  $d \leftarrow$ fit(elite)



2. First Order Optimization

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



Gradient Ascent

loop

$\theta \leftarrow \theta + \alpha \nabla_\theta U(\theta)$

Stochastic Gradient Ascent

loop

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

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

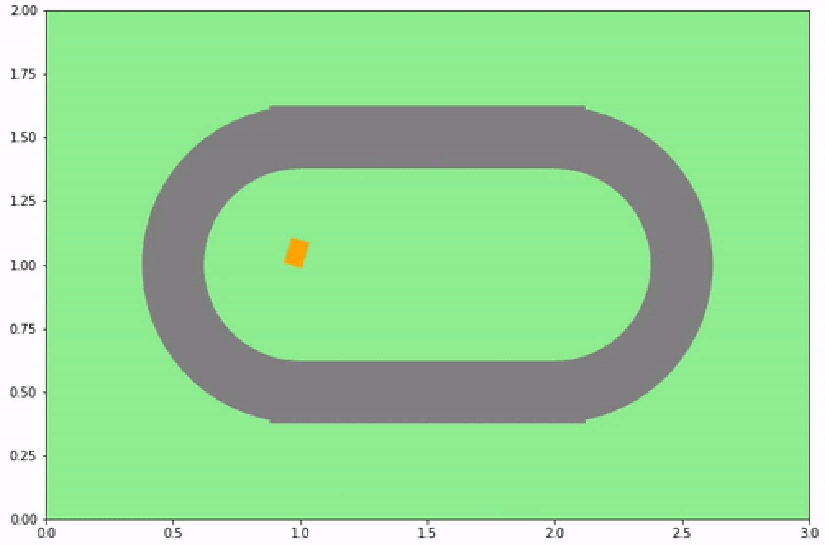
Roughly

Convergence to local optimum $\iff \sum_{k=1}^{\infty} \alpha^{(k)} = \infty$ and $\sum_{k=1}^{\infty} \left(\alpha^{(k)}\right)^2 < \infty$

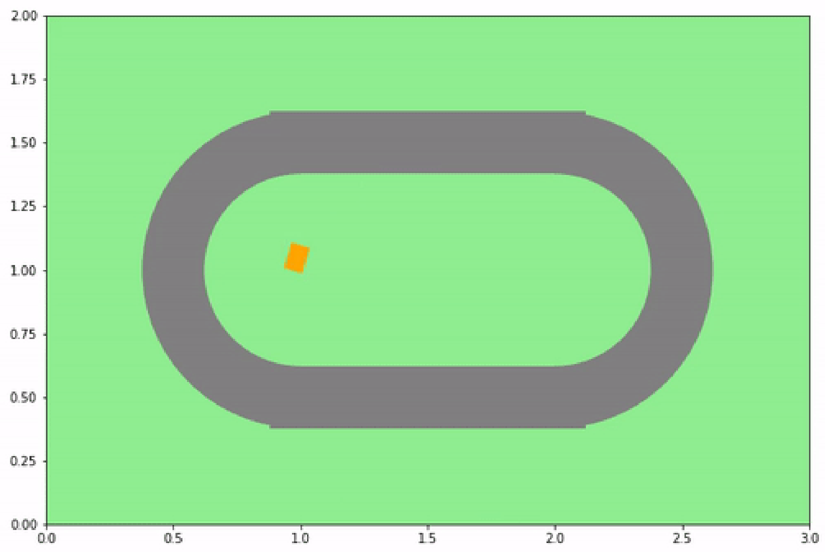
- Definition of Gradient
- Gradient Ascent
- Stochastic Gradient Ascent

Tricks

Tricks



Tricks



For policy gradient, 3 tricks

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

Log Derivative

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

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

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

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

Trajectory Probability Gradient

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

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

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

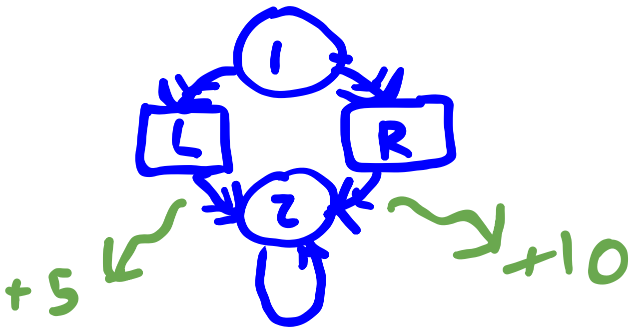
$$\log(p_\theta(\tau)) = \log(b(s_0)) + \sum_{k=0}^{d} \log T(s_{k+1} \mid s_k, a_k) + \sum_{k=0}^{d} \log \pi_\theta(a_k \mid s_k)$$

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

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

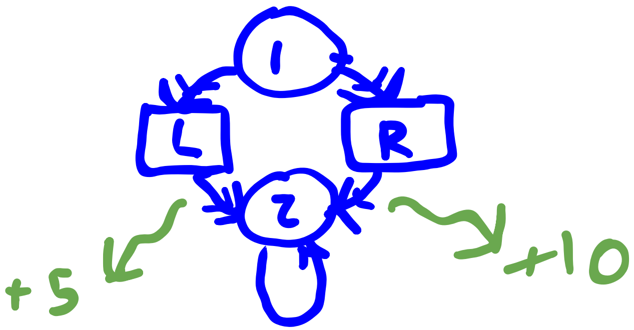
Example

$A = \{L, R\}$



Example

$A = \{L, R\}$

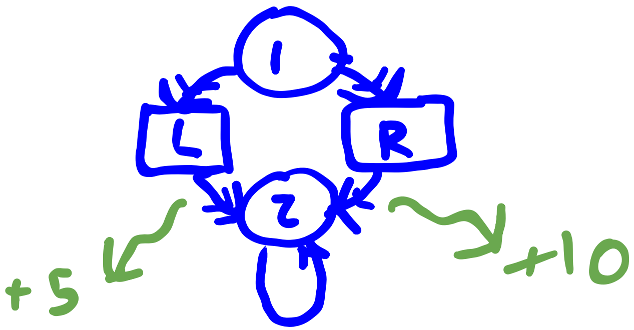


$$\pi_\theta(a = L \mid s = 1) = \text{clamp}(\theta, 0, 1)$$

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

Example

$A = \{L, R\}$



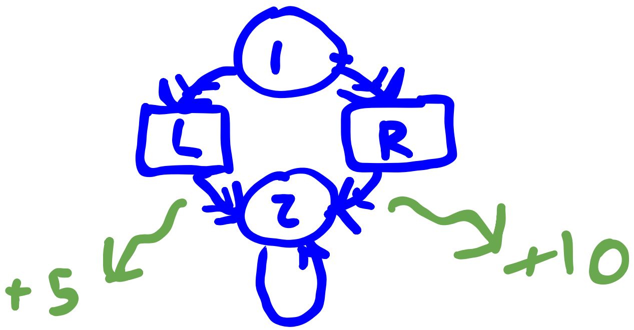
$$\pi_\theta(a = L \mid s = 1) = \text{clamp}(\theta, 0, 1)$$

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

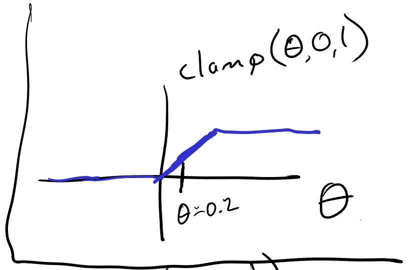
$$\nabla U(\theta) = \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right]$$

Example

$A = \{L, R\}$



$\theta = \pi_\theta(a=L \mid s=1)$



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

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

a) $\tau_{(a)} = (s_0 = 1, a_0 = L, r_0 = 5, s_1 = 2)$

$\nabla_\theta \sum_{k=0}^{0} \log \pi_\theta(a_k \mid s_k) = \frac{\partial}{\partial\theta} \log \text{clamp}(\theta, 0, 1)\Big|_{\theta=0.2} = \frac{\partial}{\partial\theta} \log\theta\Big|_{\theta=0.2} = \frac{1}{\theta} = \frac{1}{0.2}$

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

b) $\tau_{(b)} = (s_0 = 1, a_0 = R, r_0 = 10, s_1 = 2)$

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

$= \frac{1}{1-\theta}(-1) = -\frac{1}{0.8}$

$\widehat{\nabla_\theta U(\theta)} = -\frac{1}{0.8} \cdot 10 = \underline{-12.5}$

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

$E = p_\theta(\tau_{(a)})\, 25 + p_\theta(\tau_{(b)})(-12.5)$

$0.2 \cdot 25 + 0.8(-12.5) = -5.0$

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

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

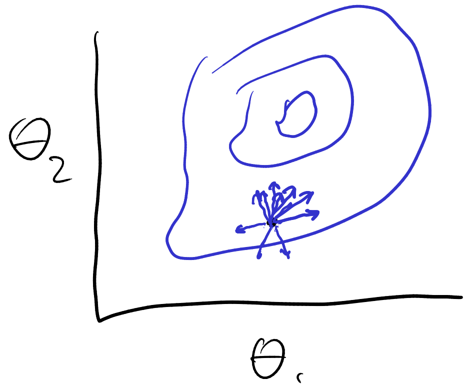
$$\theta \leftarrow \theta + \alpha \sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)$$

Policy Gradient

loop

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

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



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

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

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

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

Causality

$$
\begin{aligned}
\nabla U(\theta) &= \mathrm{E}\left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) R(\tau)\right] \\
&= \mathrm{E}\left[\left(\sum_{k=0}^{d} \underbrace{\nabla_\theta \log \pi_\theta(a_k \mid s_k)}_{f_k}\right)\left(\sum_{k=0}^{d} \gamma^k r_k\right)\right] \\
&= \mathrm{E}\left[\left(f_0 + \ldots + f_d\right)\left(\gamma^0 r_0 + \ldots \gamma^d r_d\right)\right] \\
&= E\left[\begin{array}{l} f_0\gamma^0 r_0 + f_0\gamma^1 r_1 + f_0\gamma^2 r_2 + \ldots + f_0\gamma^d r_d \\ + f_1\gamma^0 r_0 + f_1\gamma^1 r_1 + f_1\gamma^2 r_2 + \ldots f_1\gamma^d r_d \\ \vdots \\ + f_d\gamma^0 r_0 + f_d\gamma^1 r_1 + f_d\gamma^2 r_2 + \ldots f_d\gamma^d r_d \end{array}\right]
\end{aligned}
$$

Causality

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

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

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

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

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

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

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

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

Causality

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

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

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

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

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

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)

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

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

In practice $\hat{V}^{\pi_\theta}(s_k) = r_{base}(s_k)$

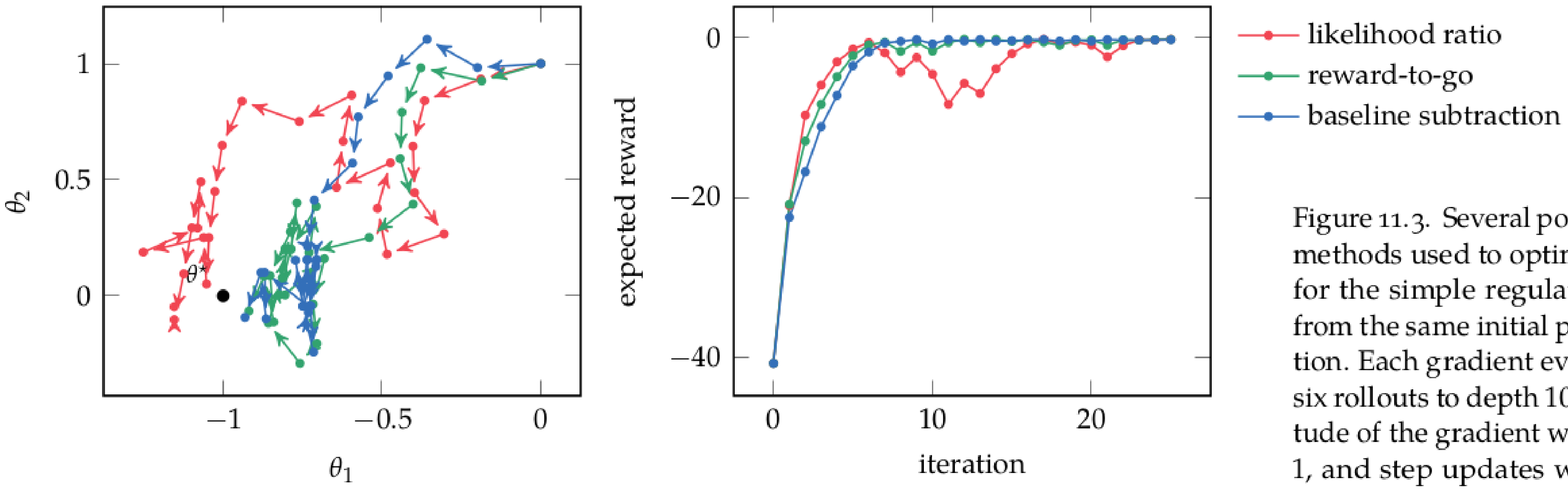


Figure 11.3. Several policy gradient methods used to optimize policies for the simple regulator problem from the same initial parameterization. Each gradient evaluation ran six rollouts to depth 10. The magnitude of the gradient was limited to 1, and step updates were applied with step size 0.2. The optimal policy parameterization is shown in black.

Guiding Questions

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