Last Time

- Bandits

Guiding Questions

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

Map

Challenges in RL

- Exploration and Exploitation
- Credit Assignment
- Generalization

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 $\quad \underset{\theta}{\text{maximize}} \quad U(\pi_\theta) = U(\theta) \qquad a \sim \pi_\theta(a \mid s)$

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

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

Common zeroth-order aproaches:

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

Cross Entropy:

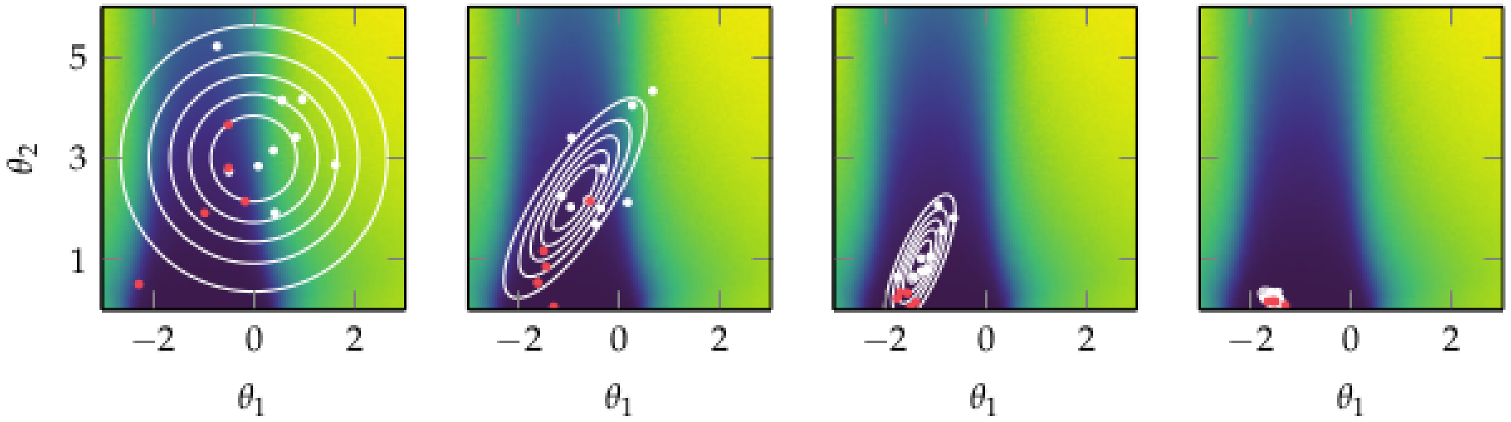
Initialize d

loop:

population $\leftarrow$ sample(d)

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

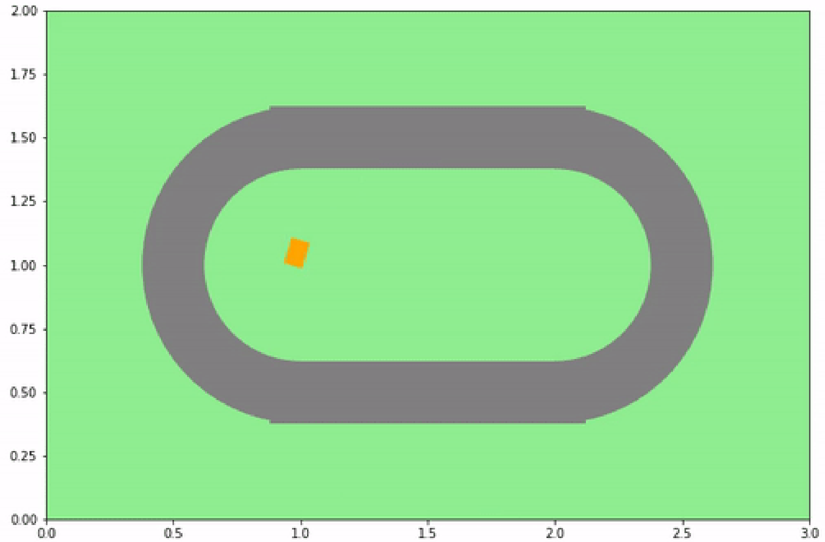
$d \leftarrow$ fit(elite)



2. First Order Optimization

- Definition of Gradient
- Gradient Ascent
- Stochastic Gradient Ascent

Tricks



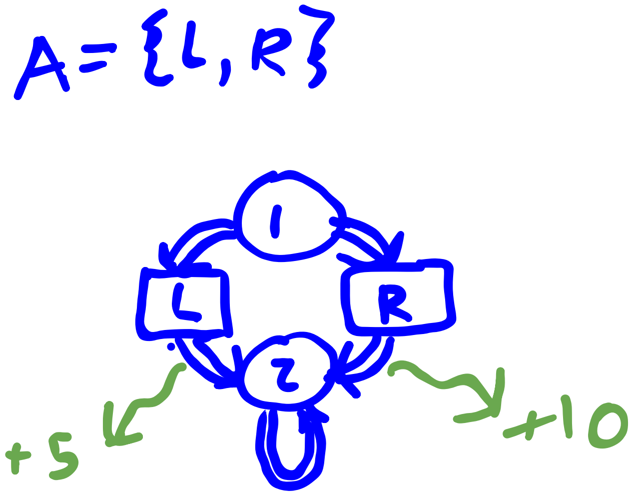
For policy gradient, 3 tricks

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

Log Derivative

Trajectory Probability Gradient

Discuss



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

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

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

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

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

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

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

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

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

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

Baseline Subtraction

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

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

does not bias (proof in book)

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

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

Guiding Questions

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