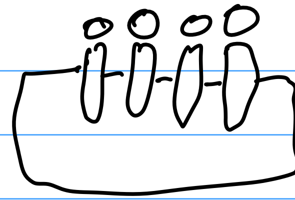
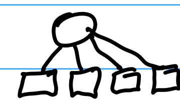
Last Time

Bandit

- ε greedy
- softmax
- Thompson Sampling
- Interval
- UCB
- Optimal Dynamic



Relationship to MCTS



Q

This Time

Policy Gradient

Model Based $\longleftrightarrow$ **Model Free**

Model Based: estimate T, R; solve with T, R

Model Free: directly optimize π or Q w/o T, R

$$\nabla_x f(x) = \left[\frac{\partial f}{\partial x_1}(x), \ldots \frac{\partial f}{\partial x_n}(x)\right]$$

Gradient Ascent

optimize $U(\theta)$

loop

$\theta' \leftarrow \theta + \alpha \nabla U(\theta)$

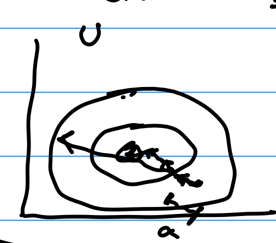
$\theta \leftarrow \theta'$

step size

decaying

ADAM

Stochastic Gradient Descent

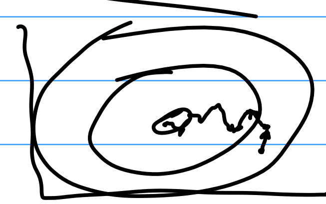


Probabilistic / Stochastic Parameterized Policies

$\pi_\theta(a|s)$

$\theta = |S| \times |A|$ matrix

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



initial state . 1st

$(S, A, T, R, \gamma, \rho_0)$

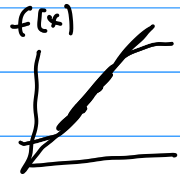
step

Episode / Trajectory $\quad \tau = (s^{(\,)}, a^{(0)}, r^{(0)}, \dots, s^{(d)}, a^{(d)}, r^{(d)})$

Advantage $\quad A(s,a) = Q(s,a) - V(s)$

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

$e^{(2)} = [0,1,0,0..]$ $\quad f(x)$



$$\nabla U(\theta) =$$

$[1,0,0,0]$

Finite Differencing

$$\nabla U(\theta) \approx \left[\frac{U(\theta + \delta e^{(1)}) - U(\theta)}{\delta}, \dots \frac{U(\theta + \delta e^{(n)}) - U(\theta)}{\delta} \right]$$

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

Leverage $\nabla \pi_\theta$

$\frac{\partial}{\partial x} \log x = \frac{1}{x}$

Likelihood Ratio Trick

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

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

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

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

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

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

$E[f(x)] = \int p(x) f(x)\, dx$

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

$$p_\theta(\tau) = p(s^{(1)}) \prod_{k=1}^{d} T(s^{(k+1)} \mid s^{(k)}, a^{(k)})\, \pi_\theta(a^{(k)} \mid s^{(k)})$$

p_0

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

$$\log p_\theta(\tau) = \log p(s^{(1)}) + \sum_{k=1}^{d} \log T(s^{k+1} \mid s^k, a^k) + \sum_{k=1}^{d} \log \pi_\theta(a^k \mid s^k)$$

$$\boxed{\nabla \log p_\theta(\tau) = \sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^k \mid s^k)}$$

Policy Gradient

$\theta \leftarrow \text{rand}()$

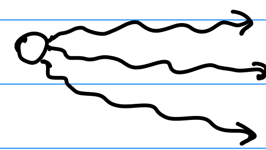
loop

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

Variance

$$\theta \leftarrow \theta + \alpha \sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^{(k)} \mid s^{(k)})\, R(\tau)$$

Break ~~at Rooms~~



Causality

$$\nabla U(\theta) = E\Big[\Big(\underbrace{\sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^{(k)} \mid s^{(k)})}_{f^k}\Big)\Big(\underbrace{\sum_{k=1}^{d} r^{(k)} \gamma^{k-1}}_{R(\tau)}\Big)\Big]$$

$$= E\Big[(f^1 + f^2 \ldots f^d)(r^{(1)} + r^{(2)}\gamma \ldots r^d \gamma^{d-1})\Big]$$

$$= E\begin{bmatrix} f^1 r^1 + f^1 r^2 \gamma + \ldots & f^{(1)} r^d \gamma^{d-1} \\ + f^2 r^1 + f^2 r^2 \gamma & f^2 r^d \gamma^{d-1} \\ \vdots & \\ f^d r^1 + & + f^d r^d \gamma^{d-1} \end{bmatrix}$$

$$\nabla U(\theta) = E\Big[\sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^k \mid s^k)\Big(\sum_{l=k}^{d} r^l \gamma^{l-k}\Big)\Big]$$

$$= E\Big[\sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^k \mid s^k)\, \gamma^{k-1}\, r^{(k)}_{\text{to go}}\Big]$$

$\hat{Q}(s^k, a^k)$

Baseline Subtraction

$$\nabla U(\theta) \approx E\left[\sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^k \mid s^k)\, \gamma^{k-1} \underbrace{\left(r_{\text{to-go}}^k - r_{\text{base}}(s^k)\right)}\right]$$

Does not bias grad est
(proof in book)

→ Good: $r_{\text{base}}(s^k) = \hat{V}(s^k) = \frac{1}{m}\sum_{i=1}^{m} r_{\text{to-go},i}^k$

↖ previous simulations

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

reduces variance

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

$r_{\text{to-go}}^k - r_{\text{base}}(s^k)$

→ $$\boxed{\nabla U(\theta) = E\left[\sum_{k=1}^{d} \nabla_\theta \log \pi_\theta(a^k \mid s^k)\, \gamma^{k-1}\, \hat{A}_\cdot(s^k, a^k)\right]}$$

Recap

Policy Gradient

Running a bunch of simulations
increasing $\pi_\theta(a \mid s)$ for a that resulted in high reward

$\nabla U(\theta)$ $\quad \nabla \pi_\theta$

- Likelihood Ratio
- Causality
- Baseline Subtraction

~~$R(s,a)$~~

$s', r \leftarrow$ step!(env, a)

$r \leftarrow$ act!(env, a)