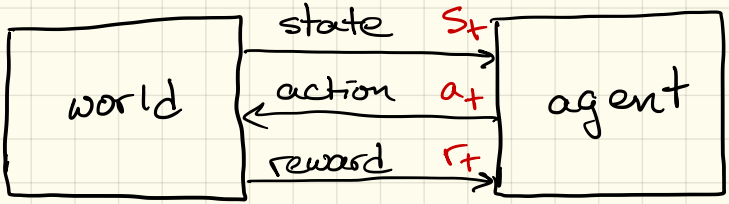
598 RL Fall 2018 T. Brecl

Policy Gradient (in class)

MODEL



$p(s_{t+1} \mid s_t, a_t)$

"policy"

$p(a_t \mid s_t ; \theta)$ ← $\pi_\theta(a_t \mid s_t)$

POLICY

$r_t = r(s_t, a_t)$

↑ sometimes called r_{t+1}

GOAL

$$\tau = (s_1, a_1, \ldots, s_T, a_T, s_{T+1}) \leftarrow \text{trajectory}$$

$$p(\tau; \theta) = p(s_1)\, p(a_1 | s_1; \theta)\, p(s_2 | s_1, a_1) \ldots$$

$$= p(s_1) \prod_{t=1}^{T} p(a_t | s_t; \theta)\, p(s_{t+1} | s_t, a_t)$$

PAYOFF

$$J(\theta) = \underbrace{E_{\tau \sim p(\tau;\theta)}}\left[\underbrace{\sum_{t=1}^{T} r(s_t, a_t)}_{r(\tau)\ \text{total reward}} \right]$$

MAXIMIZE

POLICY GRADIENT

we want to find

$$\theta^* = \arg\max_{\theta} J(\theta)$$

we can do that by gradient ascent

$$\theta_{k+1} = \theta_k + \alpha \boxed{\nabla_\theta J(\theta_k)}$$

$$J(\theta) = E_{\tau \sim p(\tau;\theta)}\left[r(\tau) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} r(\tau^i) \qquad \leftarrow \tau^1, \ldots, \tau^N \text{ are generated by } \theta$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} r(s_t^i, a_t^i)$$

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{N} \sum_i \sum_t \left(\frac{\partial r}{\partial s}(s_t^i, a_t^i) \frac{\partial s_t^i}{\partial \theta} + \frac{\partial r}{\partial a}(s_t^i, a_t^i) \frac{\partial a_t^i}{\partial \theta} \right)$$

$$r(s, a) = -(s - \mu)^2$$

$$J(\theta) = E_{\tau \sim p(\tau;\theta)}\left[r(\tau)\right]$$

$$= \int_{\tau} r(\tau)\, p(\tau;\theta)\, d\tau$$

$p(\tau;\theta)$: probability of sampling τ under θ

$r(\tau)$: total reward for τ

$\int_{\tau}$: integrate over the space of all trajectories

$$\nabla_\theta J(\theta) = \int_{\tau} r(\tau)\, \nabla_\theta p(\tau;\theta)\, d\tau$$

$$\frac{1}{12} \sum_{i=1}^{12} r(\tau^i)\, \nabla_\theta p(\tau^i;\theta)$$

$T = 1$

$\mathcal{A} = \{1, 2\}$

$S = \{1, 2, 3\}$

$\tau = (s_1, a_1, s_2)$

τ^1 1, 1, 1

τ^2 1, 1, 2

⋮ ⋮

12 trajectories

QUEUE

- infinite horizon?
- discounted reward?

$$\nabla_\theta J(\theta) = \int_\tau r(\tau)\, \nabla_\theta p(\tau;\theta)\, d\tau$$

This does not have the form

$$\int_\tau f(\tau)\, p(\tau;\theta)\, d\tau = E_{\tau \sim p(\tau;\theta)}\left[f(\tau)\right]$$

If it did, then we could approximate

$$E_{\tau \sim p(\tau;\theta)}\left[f(\tau)\right] = \frac{1}{N} \sum_{i=1}^{N} f(\tau^i)$$

$\tau^1, \ldots, \tau^N$ are generated by θ

$$\nabla_\theta J(\theta) = \int_\tau r(\tau) \, \nabla_\theta p(\tau;\theta) \, d\tau$$

$$p(\tau;\theta) = p(s_1, a_1, \dots, s_T, a_T, s_{T+1})$$
$$= p(s_1) \prod_{t=1}^{T} p(a_t|s_t;\theta) \, p(s_{t+1}|s_t, a_t)$$

$$\nabla_\theta p(\tau;\theta) = p(s_1) \sum_{t=1}^{T} \left(\prod_{\substack{k=1 \\ k \neq t}}^{T} p(a_k|s_k;\theta) \, p(s_{k+1}|s_k, a_k) \right.$$
$$\left. \cdot \left(\nabla_\theta p(a_t|s_t;\theta) \right) p(s_{t+1}|s_t, a_t) \right)$$

$$\log p(\tau;\theta) = \log\left(p(s_1)\prod_{t=1}^{T} p(a_t|s_t;\theta)\,p(s_{t+1}|s_t,a_t)\right)$$

$$= \cancel{\log p(s_1)} + \sum_{t=1}^{T}\left(\log p(a_t|s_t;\theta) + \cancel{\log p(s_{t+1}|s_t,a_t)}\right)$$

$$\nabla_\theta \log p(\tau;\theta) = \sum_{t=1}^{T} \nabla_\theta \log p(a_t|s_t;\theta)$$

$$\nabla_\theta \log p(\tau;\theta) = \frac{1}{p(\tau;\theta)}\nabla_\theta p(\tau;\theta)$$

$$\nabla_\theta p(\tau;\theta) = p(\tau;\theta)\,\nabla_\theta \log p(\tau;\theta)$$

$$\nabla_\theta J(\theta) = \int_\tau r(\tau)\,\nabla_\theta p(\tau;\theta)\,d\tau$$

$$= \int_\tau \left(r(\tau)\,\nabla_\theta \log p(\tau;\theta)\right) p(\tau;\theta)\,d\tau$$

$$J(\theta) = E_{\tau \sim p(\tau;\theta)}\left[r(\tau)\right]$$

$$= \int_{\tau} r(\tau)\, p(\tau;\theta)\, d\tau$$

$$\nabla_\theta J(\theta) = \int_{\tau} r(\tau)\, \nabla_\theta p(\tau;\theta)\, d\tau$$

$$= \int_{\tau} r(\tau) \left(\nabla_\theta \log p(\tau;\theta)\right) p(\tau;\theta)\, d\tau$$

$$\sum_{t=1}^{T} \nabla_\theta \log p(a_t \mid s_t;\theta)$$

$$= E_{\tau \sim p(\tau;\theta)}\left[r(\tau)\, \nabla_\theta \log p(\tau;\theta)\right]$$

PAYOFF GRADIENT

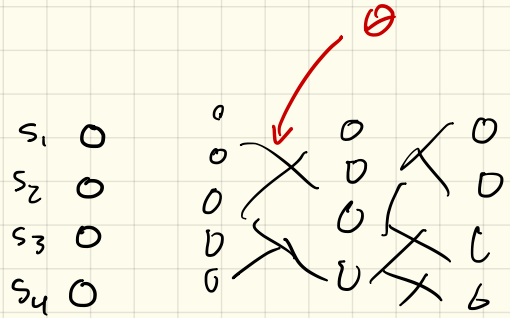
$$\nabla_\theta J(\theta) = \mathop{E}_{\tau \sim p(\tau;\theta)} \left[\left(\sum_{t=1}^{T} r(s_t, a_t) \right) \left(\sum_{t=1}^{T} \nabla_\theta \log p(a_t | s_t; \theta) \right) \right]$$

REINFORCE algorithm

0. initialize θ_0
1. sample $\tau^1, \ldots, \tau^N$ by running the policy $p(a_t | s_t; \theta_k)$
2. estimate $\nabla_\theta J(\theta)_k \approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} r(s_t^i, a_t^i) \right) \left(\sum_{t=1}^{T} \nabla_\theta \log p(a_t^i | s_t^i; \theta_k) \right)$
3. update $\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\theta_k)$

repeat until convergence

$p(a_t | s_t ; \theta)$ (e.g., for the acrobot)



$$p(a | s ; \theta) = \begin{cases} y_1 & \text{if } a = a_1 \\ y_2 & \text{if } a = a_2 \\ y_3 & \text{if } a = a_3 \end{cases}$$

$$p(a | s ; \theta) = \begin{cases} 1 & a = a_i \\ 0 & a \neq a_i \end{cases}$$

$$p(a | s ; \theta) = \begin{cases} e^{y_1} / \sum_{i=1}^{3} e^{y_i} & \text{if } a = a_1 \\ e^{y_2} / \cdot & \text{if } a = a_2 \\ e^{y_3} / \cdot & \text{if } a = a_3 \end{cases}$$

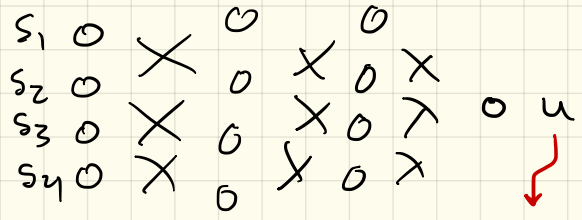
$$\log\left(e^{y_1} \Big/ \sum_{i=1}^{3} e^{y_i}\right)$$

$$= y_1 - \log\left(\sum_{i=1}^{3} e^{y_i}\right)$$

↑

cross-entropy (basically)

Example. Gaussian policy



$$p(a|s;\theta) = \mathcal{N}\, e^{-\frac{1}{2}(a-f(s;\theta))^T \Sigma^{-1} (a-f(s;\theta))}$$

$$\log p(a|s;\theta) = \log \mathcal{N} - \frac{1}{2}(a-f(s;\theta))^T \Sigma^{-1}(a-f(s;\theta))$$

$$\nabla_\theta \log p(a|s;\theta) = -\frac{\partial f}{\partial \theta}^T \Sigma^{-1}(a-f(s;\theta))$$

OFF-POLICY

importance sampling

$$E_{x\sim p(x)}[f(x)] = \int_x p(x) f(x)\, dx$$

$$= \int_x q(x) \left(\frac{p(x)}{q(x)}\right) f(x)\, dx$$

$$= E_{x\sim q(x)}\left[\left(\frac{p(x)}{q(x)}\right) f(x)\right]$$

$$E_{\tau\sim p(\tau;\theta')}\left[\nabla_\theta \log p(a_t|s_t;\theta')\, r(\tau)\right]$$

$$= E_{\tau\sim p(\tau;\theta)}\left[\frac{p(\tau;\theta')}{p(\tau;\theta)} \nabla_\theta \log p(a_t|s_t;\theta')\, r(\tau)\right]$$

this allows us to reuse old data and, maybe, to get better stability

$$\frac{p(\tau;\theta')}{p(\tau;\theta)} = \frac{p(s_1)\prod_{t=1}^{T} p(a_t|s_t;\theta')\,p(s_{t+1}|s_t,a_t)}{p(s_1)\prod_{t=1}^{T} p(a_t|s_t;\theta)\,p(s_{t+1}|s_t,a_t)}$$

$$= \prod_{t=1}^{T} \frac{p(a_t|s_t;\theta')}{p(a_t|s_t;\theta)}$$