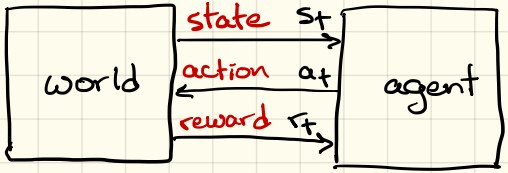
598 RL Fall 2018 T. Bretl

Policy Gradient

and M. West, personal correspondence

(see also Levine, CS294, Lecture 5, Fall 2018)

MODEL



$p(s_{t+1} \mid s_t, a_t)$ = probability that the next state is s_{t+1} given that the state is s_t and the agent takes action a_t

often called $\pi_\theta(a_t \mid s_t)$

$p(a_t \mid s_t ; \theta)$ = probability that the agent takes action a_t given that the state is s_t

POLICY

it is parameterized by θ

$r_t = r(s_t, a_t)$ = reward for taking action a_t at state s_t

GOAL

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

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

↑ probability of generating this trajectory with a given policy

PAYOFF

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

$r(\tau)$ total reward

WE WANT TO MAXIMIZE THIS

take expectation wrt the distribution $p(\tau;\theta)$ on the space of trajectories

METHOD - POLICY GRADIENT

we want to find

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

we can do so by gradient ascent

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

← NEED TO COMPUTE THIS

↑ learning rate (step size)

First try

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

$$\approx \frac{1}{N} \sum_{i=1}^{N} r(\tau^i)$$

$\tau^1, \ldots, \tau^N$ generated with θ

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

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \underbrace{\nabla_\theta r(s_t^i, a_t^i)}_{\frac{\partial r}{\partial s}(s_t^i, a_t^i) \frac{\partial s_t^i}{\partial \theta} + \frac{\partial r}{\partial a}(s_t^i, a_t^i) \frac{\partial a_t^i}{\partial \theta}}$$

↑ this makes no sense at all (and can't be computed)

SECOND TRY

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

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

probability of generating this trajectory

total reward for this trajectory

space of trajectories

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

how good is τ?

what change in θ would make the probability of τ increase the fastest?

PROBLEM #1 $$\nabla_\theta J(\theta) = \int_\tau r(\tau)\, \nabla_\theta p(\tau;\theta)\, d\tau$$

This does not have the form

$$\int_\tau f(\tau)\; p(\tau;\theta)\, d\tau$$

so it can't be approximated by sampling τ with θ as

$$\frac{1}{N} \sum_{i=1}^{N} f(\tau^i).$$

Instead, τ must be sampled uniformly, and so $r(\tau)$ is likely to be very small.

PROBLEM #2

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

$$p(\tau;\theta) = p(s_1, a_1, \ldots, s_T, a_T, s_{T+1}; \theta)$$

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

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

this is a long product of probabilities, which is also likely to be very small (and numerically bad)

if only we were working with log-probabilities ...

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

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

then ...

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

this is a sum and not a product and so will be much better behaved

it also doesn't need $p(s_{t+1}|s_t,a_t)$!!

BEHOLD!

$$\underbrace{\nabla_\theta \log p(\tau;\theta)}_{\text{what we want to find}} = \frac{1}{p(\tau;\theta)} \underbrace{\nabla_\theta p(\tau;\theta)}_{\text{what we are supposed to find}}$$

$$\Downarrow$$

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

$$\Downarrow$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \int_\tau r(\tau)\, \nabla_\theta p(\tau;\theta)\, d\tau \\
&= \int_\tau \left(r(\tau)\, \nabla_\theta \log p(\tau;\theta) \right) p(\tau;\theta)\, d\tau \\
&= \underset{\tau \sim p(\tau;\theta)}{E} \left[r(\tau)\, \nabla_\theta \log p(\tau;\theta) \right]
\end{aligned}$$

↑ can be computed by sampling τ with θ

THE PAYOFF GRADIENT

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

$$\approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \nabla_\theta \log p(a_t^i | s_t^i; \theta) \right) \left(\sum_{t=1}^{T} r(s_t^i, a_t^i) \right)$$

$\tau^i = (s_1^i, a_1^i, \ldots, s_T^i, a_T^i, s_{T+1}^i)$ is generated with θ

REINFORCE algorithm

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

repeat until convergence

how to compute this?

Example 1. Finite state and action space (tabular policy)

θ_{as} = weight of action a in state s

$$p(a|s;\theta) = \underbrace{\frac{e^{\theta_{as}}}{\sum_{a'=1}^{n_a} e^{\theta_{a's}}}}$$

softmax – exponentiate to make sure it is positive, then normalize

$$\log p(a|s;\theta) = \theta_{as} - \log\left(\sum_{a'=1}^{n_a} e^{\theta_{a's}}\right)$$

$$\frac{\partial}{\partial \theta_{\bar{a}\bar{s}}} \log p(a|s;\theta) = \delta_{\bar{a}a}\delta_{\bar{s}s} - \left(\sum_{a'=1}^{n_a} e^{\theta_{a's}}\right)^{-1} e^{\theta_{\bar{a}s}}\,\delta_{\bar{s}s}$$

$$= \delta_{\bar{s}s}\left(\delta_{\bar{a}a} - p(\bar{a}|s;\theta)\right)$$

Example 2. Continuous action space (Gaussian policy)

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

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

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

$$\downarrow$$

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

remember, it is easy to compute this by backpropagation if f is a neural net

VARIANCE (Lall, E207)

if x is a real-valued random variable, then

$$P\left(|x - E[x]| \geq a\right) \leq \frac{1}{a^2} \operatorname{var}[x]$$ } Chebyshev inequality

if $x_1, \ldots, x_N$ are real-valued and IID with mean μ and variance σ^2, then

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right| \leq \epsilon\right) \geq 1 - \frac{\sigma^2}{N\epsilon^2}$$ } law of large numbers

the confidence width that can be achieved at probability p_{conf} is

$$\epsilon = \sqrt{\frac{\sigma^2}{N(1-p_{conf})}} \quad \leftarrow \quad N = \frac{\sigma^2}{\epsilon^2(1-p_{conf})}$$

CAUSALITY

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

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

↓ see next page

$$= \sum_{t=1}^{T} \mathbb{E}_{\underbrace{(s_1, a_1, \dots, s_t, a_t, s_{t+1})}_{\tau_t} \sim p(\tau_t;\theta)} \left[r(s_t, a_t) \right]$$

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

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

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

$$\mathbb{E}_{(x,y)\sim p(x,y)}\left[f(x)\right]$$

$$= \sum_x \sum_y f(x)\, p(x,y)$$

$$= \sum_x \sum_y f(x)\, p(y|x)\, p(x)$$

$$= \sum_x f(x)\, p(x) \underbrace{\sum_y p(y|x)}_{1}$$

$$= \sum_x f(x)\, p(x)$$

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

$$= \mathbb{E} \Big[\nabla_\theta \log p(a_1|s_1;\theta) \left(r(s_1,a_1) + r(s_2,a_2) \right) + \nabla_\theta \log p(a_2|s_2;\theta)\, r(s_2,a_2) \Big]$$

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

$$= \mathbb{E} \Big[\left(\nabla_\theta \log p(a_1|s_1;\theta) + \nabla_\theta \log p(a_2|s_2;\theta) \right) \left(r(s_1,a_1) + r(s_2,a_2) \right) \Big]$$

$$= \mathbb{E} \Big[\nabla_\theta \log p(a_1|s_1;\theta) \left(r(s_1,a_1) + r(s_2,a_2) \right) + \nabla_\theta \log p(a_2|s_2;\theta) \left(r(s_1,a_1) + r(s_2,a_2) \right) \Big]$$

$$\mathbb{E}_{\tau \sim p(\tau;\theta)} \left[\nabla_\theta \log p(a_2|s_2;\theta)\, r(s_1,a_1) \right] = 0$$

$$E_{\tau \sim p(\tau;\theta)}\left[\nabla_\theta \log p(a_2|s_2;\theta)\, r(s_1,a_1)\right] \overset{?}{=} 0$$

$$= \sum_{s_1,a_1} \sum_{s_2,a_2} \nabla_\theta \log p(a_2|s_2;\theta)\, r(s_1,a_1)\, p(s_1,a_1,s_2,a_2;\theta)$$

$$= \sum_{s_1,a_1} r(s_1,a_1)\, p(s_1,a_1;\theta) \underbrace{\sum_{s_2,a_2} \nabla_\theta \log p(a_2|s_2;\theta)\, p(s_2,a_2|s_1,a_1;\theta)}$$

$$\sum_{s_2} \sum_{a_2} \nabla_\theta \log p(a_2|s_2;\theta)\, p(a_2|s_1,a_1,s_2;\theta)\, p(s_2|s_1,a_1;\theta)$$

$$= \sum_{s_2} p(s_2|s_1,a_1;\theta) \sum_{a_2} \nabla_\theta \log p(a_2|s_2;\theta)\, p(a_2|\underbrace{s_1,a_1}_{\text{redundant (cond. ind.)}},s_2;\theta)$$

$$= \sum_{s_2} p(s_2|s_1,a_1;\theta) \sum_{a_2} \nabla_\theta p(a_2|s_2;\theta)$$

$$= \sum_{s_2} p(s_2|s_1,a_1;\theta)\, \nabla_\theta \underbrace{\sum_{a_2} p(a_2|s_2;\theta)}_{1}$$

$$\underbrace{\qquad\qquad}_{0}$$

BASELINE SHIFT

variance of sampled $\nabla_\theta J(\theta)$ is high, so we have to use many samples

subtract a baseline from $r(\tau)$ to give an estimator with same mean and lower variance

$$\nabla_\theta J(\theta) = \nabla_\theta \left(J(\theta) - b \right)$$

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

← since $E_{\tau \sim p(\tau;\theta)}[b] = b$

$$= E_{\tau \sim p(\tau;\theta)} \left[\underbrace{\nabla_\theta \log p(\tau;\theta) \; (r(\tau) - b)}_{\text{call this } X} \right]$$

call this $f(\tau)$

$$\text{Var}_{\tau \sim p(\tau;\theta)}[x] = E_{\tau \sim p(\tau;\theta)}[x^2] - E_{\tau \sim p(\tau;\theta)}[x]^2$$

$$= E_{\tau \sim p(\tau;\theta)} \left[f(\tau)^2 (r(\tau) - b)^2 \right] - \left(\nabla_\theta J(\theta) \right)^2$$

↑ this is bogus unless applied to each element of ∇_θ separately

BASELINE SHIFT

assume θ is a scalar

$$\text{var}_{\tau \sim p(\tau;\theta)}[x] = E_{\tau \sim p(\tau;\theta)}[x^2] - E_{\tau \sim p(\tau;\theta)}[x]^2$$

$$= E_{\tau \sim p(\tau;\theta)}\left[f(\tau)^2 (r(\tau) - b)^2 \right] - \left(\nabla_\theta J(\theta)\right)^2$$

minimize over b:

$$\frac{\partial}{\partial b} \text{var}[\cdot] = E_{\tau \sim p(\tau;\theta)}\left[-2 f(\tau)^2 (r(\tau) - b) \right] = 0$$

$$\rightarrow \quad b = \frac{E\left[f(\tau)^2 r(\tau) \right]}{E\left[f(\tau)^2 \right]}$$

also look at 2nd derivative ...

$$E[xx^T] = \text{cov}(x) + E[x]\,E[x]^T$$

$$E\left[\nabla_\theta \log p(\tau)\,(r(\tau) - b)\right]$$

$$\text{tr}\,\text{cov}[x] = E[\|x\|^2] - \|E[x]\|^2$$

$$= E\left[\nabla_\theta \log p(\tau)^T \nabla_\theta \log p(\tau)\,(r(\tau) - b)^2\right] - \|\nabla_\theta J(\theta)\|^2$$

$$\frac{\partial}{\partial b}\left(\text{tr}\,\text{cov}[x]\right) = E\left[-2\,\nabla_\theta \log p(\tau)^T \nabla_\theta \log p(\tau)\,(r(\tau) - b)\right] = 0$$

$$\Rightarrow \quad b = \frac{E\left[\nabla_\theta \log p(\tau)^T \nabla_\theta \log p(\tau)\; r(\tau)\right]}{E\left[\nabla_\theta \log p(\tau)^T \nabla_\theta \log p(\tau)\right]}$$

↑ this at least makes sense but maybe isn't the best choice of baseline?

IMPORTANCE SAMPLING (for off-policy learning)

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

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

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

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

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

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

?

LIKELIHOOD RATIO

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

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