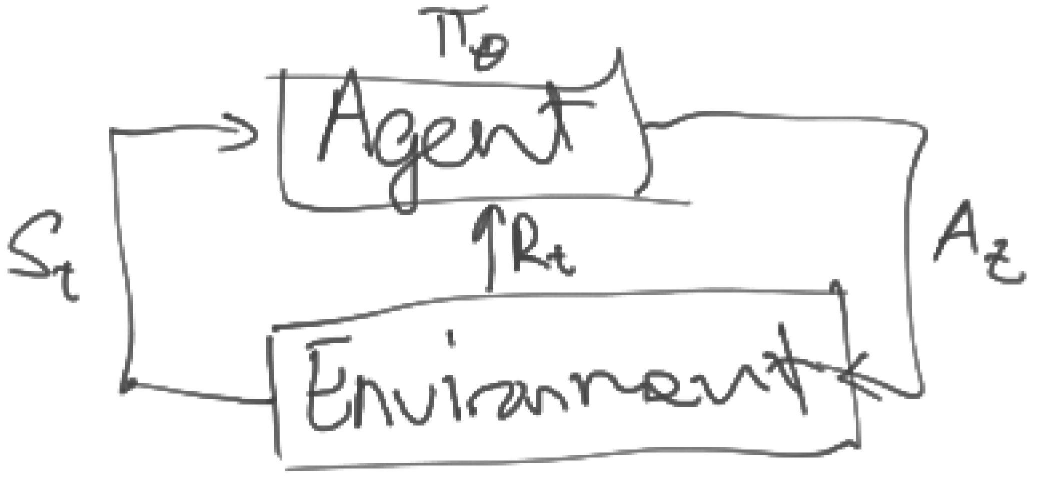
lecture 15
- Policy Gradient

- Midterm & course curve.



$$J(\theta) = \mathbb{E}\Big[\underbrace{\sum_{t=0}^{\infty} \gamma^t R_t}_{\text{Return}} \,;\, \theta\Big]$$

$$\pi(s,a) = \Pr(A_t = a \mid S_t = s)$$

$$\pi_\theta(s,a) = \Pr(A_t = a \mid S_t = s \,;\, \theta)$$

Simple RL Algorithm

Hyperparameter: Initial policy parameters θ → linear, $\theta = 0$. → ANN, use Weight init schemes (He)

For each episode:

For each time t:

Agent observes S_t

Agent selects A_t using π_θ

Environment responds by changing to state S_{t+1} and giving reward R_t.

~~If $\sum_{t=0}^{\infty} \gamma^t R_t$ is big~~

~~For all times t~~

~~Make A_t more likely in S_t~~

~~Else $\sum_{t=0}^{\infty} \gamma^t R_t$ is small~~

~~For all times t~~

~~Make A_t less likely in S_t~~

① $\forall i, \theta_i \leftarrow \theta_i + \alpha \frac{\partial \pi_\theta(S_t, A_t)}{\partial \theta}$ all policy parameters θ_i

② $\forall i, \theta_i \leftarrow \theta_i - \alpha \frac{\partial \pi_\theta(S_t, A_t)}{\partial \theta}$

$$\boxed{\forall i, \theta_i \leftarrow \theta_i + \alpha \underbrace{\left(\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'}\right)}_{\text{How much more likely should we make } A_t\text{?}} \underbrace{\frac{\partial \pi_\theta(S_t, A_t)}{\partial \theta_i}}_{\text{How to change } \theta_i \text{ to make } A_t \text{ more likely in } S_t}}$$

↳ For all times t.

Simple RL Algorithm

Hyperparameter: Initial policy parameters θ → linear, $\theta = 0$. → ANN, use Weight init schemes (He)

For each episode:

play the games

For each time t:

Agent observes S_t

Agent selects A_t using π_θ

Environment responds by changing to state S_{t+1} and giving reward R_t.

learn from (after) the game.

For each time t:

$$\forall i,\ \theta_i \leftarrow \theta_i + \alpha\left(\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'}\right) \frac{\partial \pi_{\theta_i}(S_t, A_t)}{\partial \theta_i}$$

$$\forall i,\ \theta_i \leftarrow \theta_i + \alpha\underbrace{\left(\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'}\right)}_{\text{How much more likely should we make } A_t?} \underbrace{\frac{\partial \pi_{\theta}(S_t, A_t)}{\partial \theta_i}}_{\text{How to change } \theta_i \text{ to make } A_t \text{ more likely in } S_t.}$$

↳ For all times t.

Only depends/uses most recent episode (game of Tic-Tac-Toe).

REINFORCE

unbiased issue

Make A_t more likely in S_t

$$\forall i \quad \theta_i \leftarrow \theta_i + \alpha \boxed{\frac{\partial \pi_\theta(s,a)}{\partial \theta_i}}$$

all weights in policy.

"how much to increase action probability."

How to change θ_i to increase $\pi_\theta(s,a) = Pr(A_t = a \mid S_t = s)$

Make A_t less likely in S_t

$$\forall i \quad \theta_i \leftarrow \theta_i - \alpha \frac{\partial \pi_\theta(s,a)}{\partial \theta_i}$$

$\theta^{i\text{-action}}$ = weights for ith node in output layer.

$\theta^i = (\theta^i_1, \theta^i_2, \dots)$ ↳ $\mathbb{R}$

$\theta = (\theta^1, \theta^2, \theta^3, \dots)$

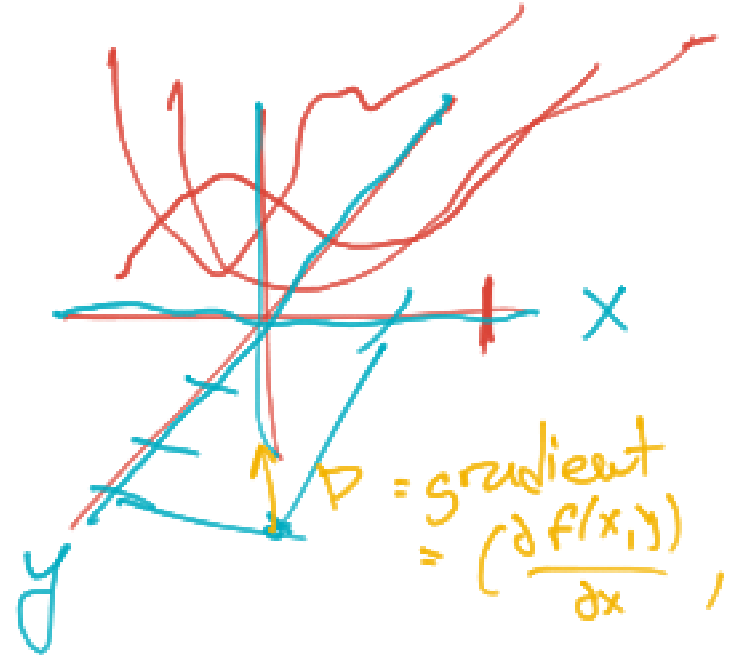
$f(x,y) = x^2 + y^2 + 7\sin(y)$

$(x,y) = (1,3)$

$2 \cdot 3 + 7\cos(3)$

$\frac{\partial f(x,y)}{\partial y} = 2y + 7\cos(y)$

$= -.93$



$\triangleright$ = gradient $= \left(\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}\right)$

Hint:

$\frac{\partial f(x,y)}{\partial y}$ = How to change y to increase f(x,y) as quickly as possible.

$\pi_\theta(s,a) \approx \pi(s,a,\theta)$

Is $\sum_{t=0}^{\infty} \gamma^t R_t$ big or small?

Idea: Scale how much we increase or decrease action probabilities by $\sum_{t=0}^{\infty} \gamma^t R_t$.

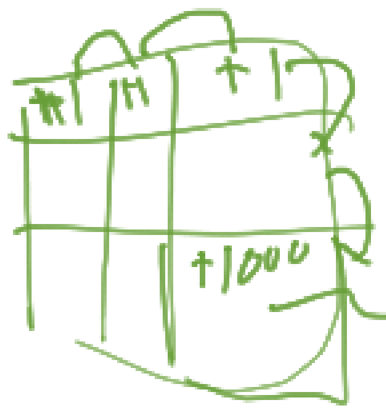
$$\forall i, \quad \theta_i \leftarrow \theta_i + \alpha \underbrace{\left(\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'} \right)} \frac{\partial \pi_\theta(S_t, A_t)}{\partial \theta}$$

step size to control learning speed. "learning rate"

like a step size, saying how much of a change to make to $\Pr(A_t \mid S_t)$

handles both good/bad outcomes – (positive and negative $\sum_t \gamma^t R_t$)

$\gamma = 0.5$



end.

× J(circles)

$1 + .5(1) + .5^2(1) + \ldots$

$= 1 + \frac{1}{2} + \frac{1}{4} + \ldots$

$= 2$

J(go to end)

$= 1 + .5(1)$

$+ .5^2(1) +$

$+ .5^3(1000)$

$\gamma^t \qquad R_t$

A	B	C
1/3	1/3	1/3

Softmax

	A	B	C
change in network output	3	3	3
	↓	↓	↓
	10	4	4
Change in probability	increase	decrease	decrease