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~~

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

↳ For all times t.

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

all policy parameters θ_i

① $\forall i,\ \theta_i \leftarrow \theta_i + \alpha \dfrac{\partial \pi_\theta(S_t, A_t)}{\partial \theta_i}$

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

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.
- For each time t:
  - $\forall i,\ \theta_i \leftarrow \theta_i + \alpha \left(\delta_t \right) \dfrac{\partial \pi_\theta(S_t, A_t)}{\partial \theta_i}$

play the games

learn from (after) the game.

$$\delta_t = R_t + \gamma v^\pi(S_{t+1}) - v^\pi(S_t)$$

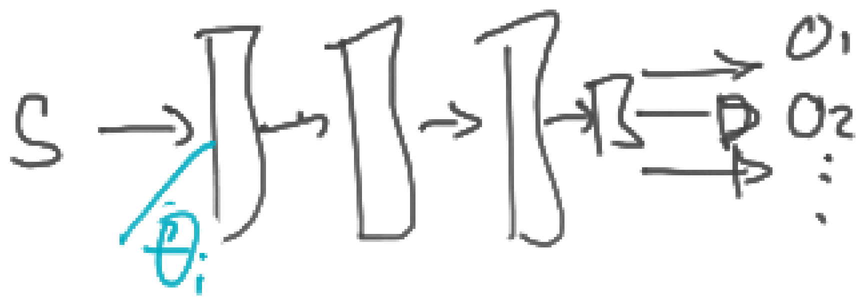
$$\forall i,\ \theta_i \leftarrow \theta_i + \alpha \underbrace{\left(\sum_{t'=0}^{\infty} \gamma^{t'} R_{t'} \right)}_{\text{How much more likely should we make } A_t?} \underbrace{\frac{\partial \pi_\theta(S_t, A_t)}{\partial \theta_i}}_{\text{How to change } \theta_i \text{ to make } A_t \text{ more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

↳ For all times t.

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

$t=5$

$$\pi(s,a) = \frac{e^{o_a}}{\sum_{a'} e^{o_{a'}}}$$



REINFORCE

unbiased issue

Policy update:

$$\forall i,\quad \theta_i \leftarrow \theta_i + \alpha \underbrace{\cancel{\left(\sum_{t'=0}^{\infty}\gamma^{t'}R_{t'}\right)}}_{\text{How "good" was the outcome?}} \underbrace{\frac{\partial \pi_\theta(S_t,A_t)}{\partial\theta_i}}_{\text{How to change } \theta_i \text{ to make } A_t \text{ more likely in state } S_t.}$$

$t=5$

$$\cancel{A_5 \xrightarrow{\text{influence?}} R_1}$$

$$\left(\sum_{k=0}^{\infty}\gamma^{k}R_{t+k}\right)$$

$$A_t \rightarrow R_t$$

$$\searrow R_{t+1}$$

$$\searrow R_{t+2}$$

$$\vdots$$

$$\theta_i \leftarrow \theta_i + \alpha\left(\sum_{k=0}^{\infty}\gamma^{k}R_{t+k}\right)\frac{\partial \pi_\theta(S_t,A_t)}{\partial\theta_i}$$

Note: Discounting starts at time t.

$$\gamma^0 R_t + \gamma^1 R_{t+1} + \gamma^2 R_{t+2}$$

How do we update before the end of the episode?

→ Not a good "outcome" that matters, but whether the outcome was better or worse than expected.

outcome → outcome relative to what ~~agent~~ the agent was expecting.

What is the agent expecting?

↳ Not what is S_{t+1} or S_{t+2}

↳ Want to know what $\sum_{k=0}^{\infty} \gamma^k R_{t+k}$ is going to be.

The value of state S_t under policy π is the amount of reward the agent expects to get if it uses policy π starting from state S_t.

$$\forall s, \quad v^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s; \pi\right]$$

= Does not depend on t.

v^{π} → value function

- Assume v^{π} is given (for now)

$$v^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s; \pi\right]$$

Better than expected:

(S_t, A_t, R_t, S_{t+1})

$$\sum_{k=0}^{\infty} \gamma^k R_{t+k} > v^{\pi}(S_t)$$

$$R_t + (\gamma R_{t+1} + \gamma^2 R_{t+2} \ldots) > v^{\pi}(S_t)$$

$$R_t + \gamma\underbrace{(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots)}_{v^{\pi}(S_{t+1})} > v^{\pi}(S_t)$$

$$0 + \gamma\left(\gamma^0(1) + \gamma^1(2)\right) > 1$$

$$\underbrace{R_t}_{0} + \gamma\, \underbrace{v^{\pi}(S_{t+1})}_{=2} > \boxed{v^{\pi}(S_t)}$$

$+\gamma 2$

$$= \gamma^0(0) + \gamma^1(1) + \gamma^2(2)$$

$$= 1(0) + .5(1) + .25(2) = 1$$

Worse than expected:

$$\sum_{k=0}^{\infty} \gamma^k R_{t+k} < v^{\pi}(S_t)$$

$$R_t + \gamma v^{\pi}(S_{t+1}) < v^{\pi}(S_t)$$

$R_t = 0$ | $R_{t+1} = 1$ $R_{t+2} = 2$

$R_t = 1$ $R_{t+1} = 2$

S_t = state at time t
Rand. Variable.

s = any state.
Not a R.V.

$v^{\pi}(S_t)$

$v^{\pi}(S_t)$ $v^{\pi}(s)$

$\gamma = 0.5$

Expecting 1 cookie in 1 min
2 cookies in 2 minutes
(3 total)

$$R_t + \gamma v^\pi(S_{t+1}) \quad \underline{\text{vs}} \quad v^\pi(S_t)$$

But not just R_t.

after one time step.

δ_t = "was it better or worse than expected"
= positive if better than expected
= negative if worse than expected.

δ_t positive → outcome $(R_t + \gamma v^\pi(S_{t+1}))$ was better than expected.

δ_t negative → outcome was worse than expected.

$$\delta_t = R_t + \gamma v^\pi(S_{t+1}) - v^\pi(S_t)$$

1) R_t ~~smaller~~ bigger than expected

2) S_{t+1} ~~worse~~ better. than expected.

t = get MM

$t+1$ = April 1.

δ_t = "temporal difference error"
TD error.