Simple RL Algorithm

Hyperparameter: Initial policy parameters

For each episode:
  For each time $t$:
    Agent observes $S_t$
    Agent selects $A_t$ using $\pi_{\theta}$
    Environment responds by changing to state $S'_{t+1}$ and giving reward $R_t$

If $\sum_{t=0}^{\infty} R_t$ is big for all times $t$
  Make $A_t$ more likely in $S_t$
Else $\sum_{t=0}^{\infty} R_t$ is small
  For all times $t$
    Make $A_t$ less likely in $S_t$

How likely should we make $A_t$?

For all times $t$.

$\forall i, \theta_i \leftarrow \theta_i + \alpha \left( \sum_{t=0}^{\infty} R_t \right) \frac{d\pi_{\theta}(S_t, A_t)}{d\theta_i}$

$\forall i, \theta_i \leftarrow \theta_i - \alpha \frac{d\pi_{\theta}(S_t, A_t)}{d\theta_i}$
For each episode:
Agent observes $S_t$ using its environment.

For each time $t$:
- Agent selects $A_t$ using its policy.
- Environment responds by changing to state $S_{t+1}$ and giving reward $R_{t+1}$.

For all times $t$:
$V_t(A_t) \leftarrow Q_t(A_t) + \alpha (R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a) - Q_t(A_t))$

For each time $t$:
- Learn new game:
- Play the game:
- Only depends lives
- Most recent episode
- Game Tic Tac Toe

Tic Tac Toe

How to change $Q_t$ to make $A_t$ more likely in the future?
Policy update:

\[ A_t, \theta_t \leftarrow \theta_t + \alpha \left( \frac{\sum_{t'=0}^{\infty} \delta^{t'} R_{t'}}{\delta^t} \right) \cdot \nabla \mathbb{E} (S_t, A_t) \frac{\partial \pi_\theta (S_t, A_t)}{\partial \theta_t} \]

How "good" was the outcome?

How to change \( \theta_t \) to make \( A_t \) more likely in state \( S_t \).

\[ \sum_{k=0}^{\infty} \delta^k R_{t+k} \]

\[ \theta_t \leftarrow \theta_t + \alpha \left( \sum_{k=0}^{\infty} \delta^k R_{t+k} \right) \cdot \nabla \mathbb{E} (S_t, A_t) \frac{\partial \pi_\theta (S_t, A_t)}{\partial \theta_t} \]

Note: Discounting starts at time \( t \).

\[ \delta^0 R_t + \delta^1 R_{t+1} + \delta^2 R_{t+2} \]
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 the agent was expecting.

What is the agent expecting?

/→ Not what is $s_{t+1}$ or $s_{t+2}$

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

The value of state $s_t$ under policy $\pi$ is the amount of reward the agent expects to get if it uses policy $\pi$ starting from state $s_t$.

$$V^\pi_t(s_t) = E\left[ \sum_{k=0}^{\infty} R_{t+k} \mid s_{t}=s_t, \pi \right] = \text{Does not depend on $\pi$.} \quad V^\pi_t \text{-value function}$$
- Assume $v^\pi$ is given (for now)

Better than expected:

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

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

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

Worse than expected:

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

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

$$v^\pi(S_t) = E[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = S]$$(1)

$$0 + 0.5(0.1 + 0.5(0.2)) > 1$$

$$R_t + 0.5 v^\pi(S_{t+1}) > v^\pi(S_t)$$

$$= v(0) + v'(1) + v^2(2)$$

$$= 1 + 0 + 0.5 + 0.25(2) = 1$$

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

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

$$R_t < 1 \quad | \quad R_{t+1} = 2$$

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

Expecting 1 cookie in 1 min

2 cookies in 2 min

$$(1, 0, 0, 2, 0, 0)$$

$$(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$$

$$(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$$
$R_t + \gamma V(S_{t+1}) \quad \text{vs} \quad V(S_t)$

$s_t = "\text{way it better or worse than expected}"

= positive if better than expected
= negative if worse than expected.

$s_t = R_t + \gamma V(S_{t+1}) - V(S_t)$

1) $R_t$ smaller than expected
2) $S_{t+1}$ worse than expected.

$s_t = "\text{temporal difference error}"$

$TD error.$