Middle & course curve.

\[ J(\theta) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_t \right] \]

\[ \pi(s,a) = P(\pi = a | S_t = s) \]

\[ \pi_\theta(s,a) = P(\pi = a | S_t = s \; \theta) \]
**Simple RL Algorithm**

**Hyperparameter:** Initial policy parameters

For each episode:

For each time t:
- Agent observes $S_t$
- Agent selects $A_t$ using $\theta$
- 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 $S_t$

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

1. $\forall i, \theta_i \leftarrow \theta_i + \alpha \frac{\partial \log p(S_t, A_t)}{\partial \theta_i}$

2. $\forall i, \theta_i \leftarrow \theta_i - \alpha \frac{\partial \log p(S_t, A_t)}{\partial \theta_i}$

How much more likely should we make $A_t$?
Simple RL Algorithm

Hyperparameter: Initial policy parameters $\Theta$ for ANN, use Weight Init Schemes (He)

For each episode:
- Play the game
- Learn from after the game

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

For each time $t$:
- $A_t$, $\Theta_t \leftarrow \Theta_t + \alpha \left( \sum_{t'=0}^{\infty} g^{t'} R_{t'} \right) \frac{\partial T(S_t, A_t)}{\partial \Theta_t}$

For all times $t$. Only depends/use most recent episode (game of Tic-Tac-Toe).

How much more likely should we make $A_t$? How to change $\Theta_t$ make $A_t$ more likely in $S_t$. Overgeneral issues.
Make $A_t$ more likely in $S_t$:

$Q_i \leftarrow Q_i + \alpha \left( \frac{\pi_B(s,a) \cdot f(x,y)}{\pi_i(s)} \right)$

"how much to increase action probability."

Make $A_b$ less likely in $S_b$:

$Q_i \leftarrow Q_i - \alpha \frac{\pi_B(s,a)}{\pi_i(s)}$

$\hat{\theta} = \text{weights for i-th node in output layer}$

$b = (b_1, b_2, \ldots)$

$\theta = (\theta_1, \theta_2, \theta_3, \ldots)$

$\pi_B(s,a) = P(A_t = a | S_t = s)$

$\frac{\partial f(x,y)}{\partial y} = -0.93$

$\nabla f(x,y) = (\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y})$
Is $\sum_{t=0}^{\infty} \alpha^t R_t$ big or small?

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

$$H_t, \quad \theta_t \leftarrow \theta_t + \alpha \left( \sum_{t=0}^{\infty} \alpha^t R_t \right) \frac{\partial P(A_t | S_t)}{\partial \theta}$$

A handles both good/bad outcomes (positive and negative $\sum_{t=0}^{\infty} \alpha^t R_t$)

Step size to control learning speed. "Learning rate" like a step size, saying how much of a change to make to $P(A_t | S_t)$.

$$J = 0.5 \times J(circles)$$

$$J(g0 \rightarrow end) = 1 + 5(1) + 5^2(1) + \cdots = 1 + \frac{1}{2} + \frac{1}{4} + \cdots = 2$$

End.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Softmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
</tr>
</tbody>
</table>

Change in network output:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>\downarrow</td>
<td>\downarrow</td>
<td>\downarrow</td>
<td></td>
</tr>
</tbody>
</table>

Change in probability:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>\uparrow</td>
<td>\downarrow</td>
<td>\downarrow</td>
<td></td>
</tr>
</tbody>
</table>

increase decrease decrease