Lecture-13

- Today MC → TD → TD(λ)

* MC estimation of Action-Values

$\hat{q}(s,a)$ = average return from the first time action 'a' taken in state 's' in each episode

Problem - What if π never selects action 'a' in state s

- Need for estimation of value of all actions at each state, not only action that we currently favor.

- Solution 1: Exploring starts
  - Randomize s_0 & A_0 such that every (s,a) has a non-zero probability.
  - Sometimes not possible

- Solution 2: stochastic policy.
  - Non-zero prob. to every 'a' in every state

$$\forall s, a \quad \pi(s,a) > 0$$

* MC control with exploring starts

Property – avoid infinite episodes in evaluation step by improving after each episode
  - Accumulate Return over an episode

Pseudocode –

Initialize: for all $s \in S$, $a \in A$

$q(s,a) \leftarrow$ arbitrary

$\pi(s) \leftarrow$ arbitrary

Returns$(s,a) \leftarrow$ empty list

Repeat forever:

① Generate an episode using exploring starts & π

② for each (s,a) appearing in episode,

$G \leftarrow$ Returns following first occurrence

$q(s,a) \leftarrow$ mean(Returns(s,a))

③ for each $s \leftarrow S$ in the episode

$$\pi(s) \in \arg\max_a q(s,a)$$

* **ε-Greedy MC control**

Pseudo code:

$q(s,a) \leftarrow$ arbitrary
$\pi(s,a) \leftarrow$ arbitrary
Returns$(s,a) \leftarrow$ empty list

Repeat forever:

① Generate episode using π
② for each (s,a)

$G \leftarrow$

③ for each $s \leftarrow S$, in episode

$$a^* \leftarrow \arg\max_a q(s,a)$$

$$\pi(s,a) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A|} & \text{if } a = a^* \\ \frac{\varepsilon}{|A|} & \text{if } a \neq a^* \end{cases}$$

Properties: —

- By variations in policy imp. Thm
(sdB: —)

* TD (Temporal difference) learning

Properties — Like MC it learns directly from experiences, i.e. w/o knowledge of P & R

— Like DP update estimates based on other estimates

- Another MC algorithm

$$f(x) \leftarrow f(x) + \alpha\left(\underbrace{(Y - f(x))}_{\text{R.V. (Target)}}\right)$$

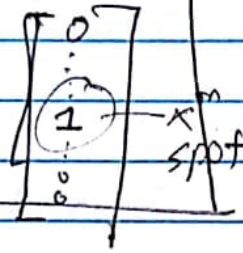
α : step size, or learning rate

$f_w(x)$,

$$L(w) = \mathbb{E}\left[\frac{1}{2}\left(Y - f_w(x)\right)^2\right]$$

$$w \leftarrow w + \mathbb{E}\left[(y - f_w(x)) \cdot \frac{\partial f_w(x)}{\partial w}\right] \times \alpha$$

$(y - f_w(x)) \frac{\partial f_w(x)}{\partial w}$

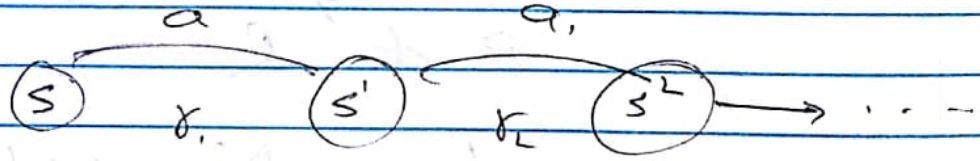
$(y - f_w(x))$ 

Similarly,

$$V(S_t) \leftarrow V(S_t) + \alpha \cdot (G_t - V(S_t))$$

... where, $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$

- Problem - wait till end of episode
- Que - how can we update before the end of the episode?



Bellman eqⁿ:

$$V(s) = \sum_a \pi(s,a) \sum_{s'} P(s,a,s') \left[R(s,a,s') + \gamma \cdot V^{\pi}(s') \right]$$

$$= \mathbb{E}\left[R_t + \gamma \cdot V^{\pi}(s_{t+1}) \mid s_t = s \right]$$

TD - target

$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid s_t = s, \pi \right]$$

MC - Target

TD - update (given transition (s, a, s'))

$$V(s) \leftarrow V(s) + \alpha \cdot (r + \gamma \cdot V(s') - V(s'))$$

Equivalently

$$V(S_t) \leftarrow V(S_t) + \alpha \left(R_t + \gamma \cdot V(S_{t+1}) - V(S_{t+1}) \right)$$

$R_t + \gamma \cdot V(S_{t+1})$ → actually what happened?

$V(S_{t+1})$ = Expectation

reward prediction error

$$\text{TD-error} = \left(R_t + \gamma \cdot V(S_{t+1}) - V(S_t) \right)$$

$$= \delta_t$$

$$\therefore V(s) \leftarrow V(s) + \alpha \delta_t$$

* How much better did things turn out than I expected?

- Is TD method SGD?

$$f(v) = \mathbb{E}\left[\frac{1}{2} \left(R_t + \gamma \cdot V(S_{t+1}) - V(S_t) \right)^2 \right]$$

$$= \mathbb{E}\left[\frac{1}{2} \delta_t^2 \right]$$

$$\therefore D\hat{f}(v) = \delta_t \cdot \frac{\partial \delta_t}{\partial v}$$

$$= \left(R_t + \gamma V(S_{t+1}) - V(S_t)\right) \left(\frac{\partial V(S_{t+1})}{\partial v} - \frac{\partial V(S_t)}{\partial v}\right)$$

$$= \left(R_t + \gamma \cdot V(S_{t+1}) - V(S_t)\right) \cdot \left[S_{t+1} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} S_t \right]$$

$$V(S_t) \leftarrow V(S_t) + \alpha \delta_t$$

drop this → $V(S_{t+1}) \leftarrow V(S_{t+1}) - \alpha \cdot \gamma \cdot \delta_t$