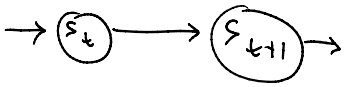
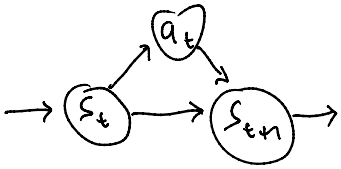
Tuesday, June 7, 2022 9:00 AM

$\mathcal{S} = \{1, 2\}$

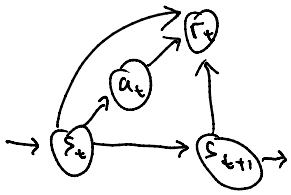
MP: $\mathbb{P}(S_{t+1} = j \mid S_t = i) =: A_{ij}$, MLE $\hat{A}_{ij} = \frac{n_{ij}}{\sum_{j'} n_{ij'}}$



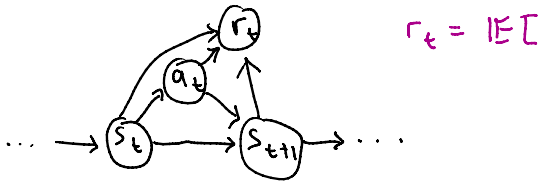
MDP:



$\mathbb{P}(S_{t+1} = j \mid S_t = i, a_t = a) =: A^a_{ij}$

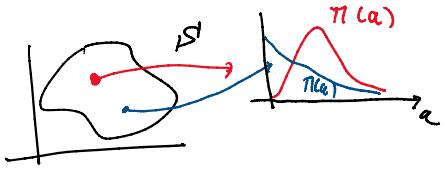


$\mathbb{P}(S_{t+1} = j, r_t = r \mid a_t = a, S_t = i) =: A^a_{ij,r}$



$r_t = \mathbb{E}[R_t \mid S_t = s]$

π policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$



→ $V^\pi(s) := \mathbb{E}[r_1 + \gamma^1 r_2 + \cdots \mid S_0 = s]$, $\gamma \in [0, 1)$

state value of the policy π, being in state s

$Q^\pi(s, a) := \mathbb{E}[r_1 + \gamma^1 r_2 + \cdots \mid S_0 = s, a_0 = a]$

state action value of the policy π, being in state s & taking action a.

Bellman principle

$$V^\pi(s) := \mathbb{E}[r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots \mid S_0 = s]$$
$$= \mathbb{E}[r_1 + \underbrace{\mathbb{E}[\gamma r_1 + \gamma^2 r_3 + \cdots \mid S_1]}_{\gamma\, \mathbb{E}[r_2 + \gamma^1 r_3 + \gamma^2 r_4 + \cdots \mid S_1]} \mid S_0 = s]$$

$$\gamma \underbrace{\mathbb{E}[r_2 + \gamma r_3 + \gamma^2 r_4 + \dots \mid s_1]}_{V^\pi(s_1)}$$

$$\boxed{V^\pi(s) = \mathbb{E}[r_1 + \gamma V^\pi(s_1) \mid s_0 = s]}$$

Bellman equation

$\mathcal{S} = \{1, 2, \dots, I\}$

$$\vec{V}^\pi = \begin{pmatrix} V^\pi(1) \\ V^\pi(2) \\ \vdots \\ V^\pi(I) \end{pmatrix}$$

$\mathbb{E}[r_1 \mid s_0 = s]$

$\hookrightarrow \begin{pmatrix} \mathbb{E}[r_1 \mid s_0 = 1] \\ \mathbb{E}[r_1 \mid s_0 = 2] \\ \vdots \\ \mathbb{E}[r_1 \mid s_0 = I] \end{pmatrix} = \int r(a,1)\pi(a)\,da$

$\vec{r} =: \begin{pmatrix} \mathbb{E}[r_1 \mid s_0 = 1] \\ \mathbb{E}[r_1 \mid s_0 = 2] \\ \vdots \\ \mathbb{E}[r_1 \mid s_0 = I] \end{pmatrix}$

$$\underbrace{\vec{V}^\pi_s}_{} = V^\pi(s) = \iint \big(r(a,s) + \gamma V^\pi(s')\big) \underbrace{\mathbb{P}(a, s' \mid s)}_{\mathbb{P}(s' \mid a,s)\,\mathbb{P}(a \mid s)} \, da\, ds'$$

$A^a_{s,s'}$ $\quad \pi(a \mid s)$

$$= \sum_{s',a} \big(r(a,s) + \gamma V^\pi(s')\big) \underbrace{\mathbb{P}(s' \mid a,s)\,\mathbb{P}(a \mid s)}$$

$$= \vec{r}_s + \gamma \sum_{s'} A_{s,s'} \vec{V}^\pi_{s'}$$

$$\vec{V}^\pi = \vec{r} + \gamma A \vec{V}^\pi$$

$$\Rightarrow \quad (\mathbb{1} - \gamma A)\vec{V}^\pi = \vec{r}$$

$$\Rightarrow \quad \boxed{\vec{V}^\pi = (\mathbb{1} - \gamma A)^{-1} \vec{r}}$$

analogously we have

$$\boxed{Q^\pi(s,a) = \mathbb{E}[r_1 + \gamma Q^\pi(s_1, a_1) \mid s_0 = s, a_0 = a]}$$

$$Q^\pi(s,a) = \mathbb{E}[\, r_1 + \gamma\, Q^\pi(s_1, a_1) \mid s_0 = s,\ a_0 = a\,]$$

$$V^\pi(s) = \mathbb{E}^{\mathbb{P}_\pi}[\, r(s,a) + \gamma\, V^\pi(s') \mid s = s\,]$$

$$V^\pi(s) = \mathbb{E}^{\mathbb{P}_\pi}[\, Q^\pi(s,a) \mid s = s\,]$$

$$\mathbb{E}^{\mathbb{P}_\pi}[\, r(a,s) + \gamma\, r(a', s') + \gamma^2 r(a'', s'') + \dots \mid s = s, A = a\,]$$

suppose I take an arbitrary action at $t=0$, then follow the optimal policy π^*

call this $\tilde{\pi}$

$$V^{\tilde{\pi}}(s) = \mathbb{E}^{\mathbb{P}_{\tilde{\pi}}}[\, r(s,a) + \gamma\, V^{\tilde{\pi}}(s') \mid s = s\,]$$

$$= V^{\pi^*}(s')$$

$$V^{\tilde{\pi}}(s) = \mathbb{E}^{\mathbb{P}_{\tilde{\pi}}}[\, r(s,a) + \gamma\, V^{\pi^*}(s') \mid s = s\,]$$

$$V^{\pi^*}(s) = \max_{\tilde{\pi} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}_{\tilde{\pi}}}[\, r(s,a) + \gamma\, V^{\pi^*}(s') \mid s = s\,]$$

Bellman optimality equation

$$Q^{\pi^*}(s,a) = \mathbb{E}^{\mathbb{P}\pi^*}\left[r(s,a) + \gamma V^{\pi^*}(s') \mid s=s, A=a \right]$$

$$= \mathbb{E}^{\mathbb{P}\pi^*}\left[r(s,a) + \gamma \mathbb{E}^{\mathbb{P}\pi^*}\left[Q^{\pi^*}(s',a') \mid s' \right] \mid s=s, A=a \right]$$

Suppose $\pi(a|s)$ are "deterministic" policies

$$V^{\pi^*}(s) = \max_{a\in\mathcal{A}} \mathbb{E}^{\mathbb{P}\pi^*}\left[r(s,a) + \gamma V^{\pi^*}(s') \mid s=s \right]$$

$$Q^{\pi^*}(s,a) = \mathbb{E}^{\mathbb{P}\pi^*}\left[r(s,a) + \gamma \max_{a'\in\mathcal{A}} Q(s',a') \mid s=s, A=a \right]$$

$$V^{\pi^*}(s') = \max_{a'\in\mathcal{A}} Q^{\pi^*}(s',a')$$

$$\pi \longrightarrow V^{\pi}(s) \xrightarrow[\text{greedy}]{} \pi' = \arg\max_{a\in\mathcal{A}} \mathbb{E}^{\mathbb{P}\pi}[\dots]$$

$$\longrightarrow V^{\pi'}(s)$$

$$\Rightarrow V^{\pi'}(s) \geq V^{\pi}(s)$$

Define the Bellman operator $T^{\pi} : F_{\mathcal{S}} \mapsto F_{\mathcal{S}}$

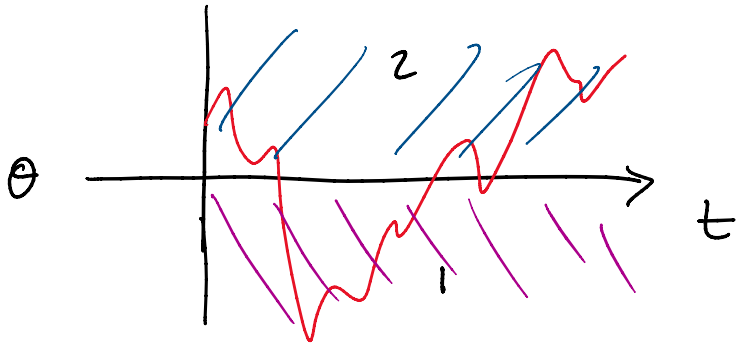
$f \in \mathcal{F}_{\mathcal{S}}$

$$(T^{\pi} f)(s) = \mathbb{E}^{\mathbb{P}\pi}\left[r(s,a) + \gamma f(s') \mid s=s \right]$$

(for deterministic policies $a(s)$)

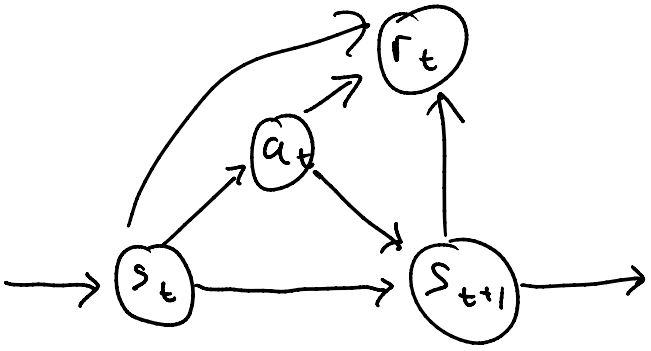
claim: $\lim_{k} (T^{\pi})^k f = V^{\pi}$

claim: $\lim_{k \to \infty} (T^{\pi})^{k} f = V^{\pi}$



$A = \{$ do nothing, 1
long position, 2
short position $\}$ 3

$$A = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}$$



rewards

s	s'	A 1	2	3
1	1	0	0	0
1	2	0	+1	-1
2	1	0	-1	+1
2	2	0	0	0

initial policy

s	π
1	1
2	1

Tuesday, June 7, 2022 12:58 PM

$$V^\pi(s) = \mathbb{E}^{\mathbb{P}_\pi}[\, r(s,a) + \gamma V^\pi(s') \mid s = s]$$

$$V^{\pi(k+1)}(s) = (T^\pi V^{\pi(k)})(s)$$

$$= \mathbb{E}^{\mathbb{P}_\pi}[\, r(s,a) + \gamma V^{\pi(k)}(s') \mid s = s]$$

sample from $\mathbb{P}_\pi$ to get $(s^{(1)}, a^{(1)}, r^{(1)}, s'^{(1)})$

$$V^\pi_{t+1}(s_t) = \mathbb{E}^{\mathbb{P}_\pi}[\, r(s_t, a_t) + \gamma V^\pi_t(s_{t+1}) \mid s_t]$$

Temporal - Difference (TD)

$$\underbrace{V^\pi_{t+1}(s_t)}_{u^{(t+1)}} = \underbrace{V^\pi_t(s_t)}_{u^{(t)}} + \alpha_t \Big(\underbrace{\underbrace{r(s_t, a_t) + \gamma V^\pi_t(s_{t+1})}_{y^{(t+1)}} - \underbrace{V^\pi_t(s_t)}_{u^{(t)}}}_{\text{TD error}} \Big)$$

$\alpha_t = 1/(t+1)$

target

$$\alpha_t > 0, \quad \lim_{t \to \infty} \alpha_t = 0, \quad \sum_t \alpha_t = +\infty, \quad \sum_t \alpha_t^2 < +\infty$$

$$Q^\pi(s,a) = \mathbb{E}^{\mathbb{P}_\pi}[\, r(s,a) + \gamma Q^\pi(s', a') \mid s = s, A = a]$$

(s, a, r, s', a') SARSA

$$Q^\pi_{t+1}(s_t, a_t) = Q^\pi_t(s_t, a_t) + \alpha_t \Big(\underbrace{r(s_t, a_t) + \gamma Q^\pi_t(s_{t+1}, a_{t+1})}_{} - Q^\pi_t(s_t, a_t) \Big)$$

target

TD error

$$Q^{\pi}_{t+1}(s_t, a_t) = Q^{\pi}_t(s_t, a_t) + \alpha_t [\, r(s_t, a_t) + \gamma \; Q^{\pi}_t(s_{t+1}, \underset{a}{\operatorname{argmax}}\, Q^{\pi}_t(s_{t+1}, a')) - Q^{\pi}_t(s_t, a_t) \,]$$

$$Q^{\pi}_t(s_{t+1}, \underset{a'}{\operatorname{argmax}}\, Q^{\pi}_t(s_{t+1}, a'))$$

$$= \max_{a'} Q^{\pi}_t(s_{t+1}, a')$$

$\uparrow \quad \pi(s_{t+1}) \neq \underset{a'}{\operatorname{argmax}}\, Q^{\pi}_t(s_{t+1}, a')$

Q-learning

$$Q(s, a) = \beta' g(s)$$

$$F_\theta(s, a)$$