# Lecture 2

1. Recap of definitions
   MDPs, $\mathbb{P}^{\pi}_{\mu}$, $\mathbb{E}^{\pi}_{\mu}$, $v^{\pi}$, $v^*$
2. Fundamental theorem:
   ML policies & $v^*$ & BOE & DP
3. Value it. Complexity $\varepsilon$-opt
4. Policy it. $\tilde{O}\left(\frac{SA}{1-\gamma}\right)$ complexity
5. Lower bounds
6. Conclusions

**MDP**

$M = (S, A, P, r, \gamma)$

$P = (P_a(s))_{s,a}$

$r = (r_a(s))_{s,a}$

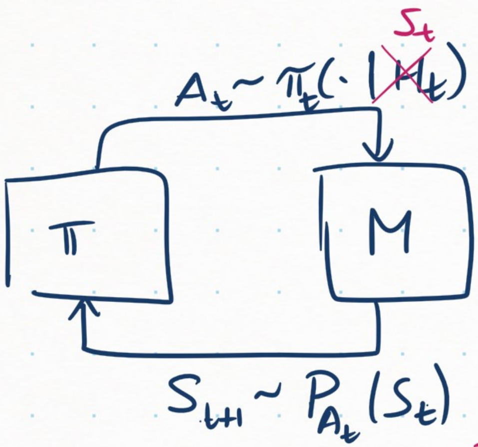
$0 \leq \gamma < 1$

**Policy** $\pi = (\pi_t)_{t \geq 0}$

$\pi_t : (S \times A)^{t-1} \times S \to \mathcal{M}_1(A)$

(discrete $S, A$)

$|S|, |A| < \infty$




$H_t = (S_0 \, A_0$
$S_1 \, A_1$
$\vdots$
$S_{t-1} A_{t-1}$
$S_t)$

Fix $\pi, M, \mu \in M_1(S)$

**Prop:** $\exists \, (\Omega, \mathcal{F}, \mathbb{P})$

$\exists \, S_0, A_0, S_1, A_1, \ldots\ldots$

random elements over $(\Omega, \mathcal{F}, \mathbb{P})$

Markov property

s.t.

(1) $\mathbb{P}(S_0 = s) = \mu(s) \quad \forall s \in S$

(b) $\mathbb{P}(A_t = a \mid H_t) = \pi_t(a \mid H_t) \quad \forall a \; \forall t \geq 0$

(c) $\mathbb{P}(S_{t+1} = s \mid \underline{H_t, A_t}) = P_{A_t}(\underline{S_t, s}) \quad \forall t \geq 0 \; \forall s \in S$

# Ionescu-Tulcea theorem

$\mathbb{P}^{\pi}_{\mu}$ ← policy ← initial state distr.

$\mathbb{E}^{\pi}_{\mu}$

$$R = \sum_{t \geq 0} \gamma^t r_{A_t}(S_t)$$

$$\mu = \delta_s \; : \; \mathbb{P}^{\pi}_{\delta_s} \doteq \mathbb{P}^{\pi}_s \; / \; \mathbb{E}^{\pi}_s$$

$$v^{\pi}(s) = \mathbb{E}^{\pi}_s [R]$$

$$\boxed{v^*(s) = \overset{\sup}{\cancel{\max_{\pi}}} \, \underbrace{v^{\pi}(s)}_{[0,1]}}$$

optimal value function

Goal: $\underline{\underline{\pi}}$ ? $v^\pi = v^*$

$v^\pi \le v^*$ $\qquad$ $v^\pi \ge v^*$ ?

$\Rightarrow v^\pi = v^*$

Find $\pi$ s.t. $v^\pi \ge \underline{v^* - \varepsilon 1}$

$1 : S \to \mathbb{R}$

$s \mapsto 1$

$\varepsilon$- optimal policy

memoryless policies (ML)

$\pi : S \to \mathcal{M}_1(\mathcal{A})$

$|S|, |A| < \infty$ $\boxed{\text{FT}}$

**Theorem:** (a) $\forall \pi$ greedy policy w.r.t. $v^* \Longrightarrow v^\pi = v^*$.

(b) $v^* = \underline{T} v^*$ Bellman optimality equation

Def: $\pi$ ML policy is greedy wrt $v : S \to \mathbb{R}$ if the following holds: $\forall s \in S$

$$\longrightarrow \sum_a \overbrace{\pi(a|s)\,\{r_a(s)}^{r^\pi(s)} + \gamma \langle \overbrace{P_a(s)}^{P^\pi}, v \rangle \}$$

$$= \max_a \{ r_a(s) + \gamma \langle P_a(s), v \rangle \}$$

$r^\pi(s) = \sum \pi(a|s) r_a(s)$

$r^\pi \in \mathbb{R}^S \qquad \mathcal{S} = \{1, \dots, S\}$

$P^\pi(s,s') = \sum_a \pi(a|s) P_a(s,s')$

$(P^\pi_{ss'})_{s,s'} \in [0,1]^{S\times S}$

$T^\pi : \mathbb{R}^S \to \mathbb{R}^S$

$T^\pi(v) = r^\pi + \gamma P^\pi v$

$T : \mathbb{R}^S \to \mathbb{R}^S$

$(T(v))(s) = \max_a r_a(s) + \gamma \langle P_a(s), v\rangle$

$T^\pi(v) = T^\pi v$

$T(v) = T v$ Operators

$\underbrace{Tv}(s) = (T(v))(s)$

$(Tv)_s$

$\boxed{\pi \text{ is greedy wrt } v: \quad T^\pi v = T v}$

$T$: Bellman operator

$T^\pi$: Policy eval. op. of $\pi$

$$\mathbb{I}(\mathrm{Log}) = \begin{cases} 1, \mathrm{Log} = \mathrm{True} \\ 0, \mathrm{Log} = \mathrm{False} \end{cases}$$

## Proof of the FT

$$\tilde{v}^*(s) = \sup_{\pi\ ML} v^\pi(s)$$

$\pi$ is greedy wrt $\tilde{v}^*$

$$\Rightarrow v^\pi = \tilde{v}^*$$

$$\Rightarrow \tilde{v}^* = T\tilde{v}^* \quad \text{fixed-point}$$

$$\tilde{v}^* = v^*$$

$$\tilde{v}^* \le v^* \quad \checkmark$$

$$v^* \le \tilde{v}^* \ ??$$

(Discounted) occupancy measures

$\mu \in M_1(S)$, $\pi$ policy

$$\to \nu_\mu^\pi \in M_1(S \times A):$$

$$\nu_\mu^\pi(s,a) \doteq \sum_{t\ge 0} \gamma^t \, \mathbb{P}_\mu^\pi(S_t = s, A_t = a)$$

$$v^\pi(s) = \sum_{s,a} \nu_\mu^\pi(s,a)\, r_a(s)$$

$$\doteq \langle \nu_\mu^\pi, r \rangle$$

$$= \mathbb{E}_\mu^\pi\Big[\sum_{t\ge 0} \gamma^t r_{A_t}(S_t)\Big] =$$

$$\sum_{s,a} \sum_{t\ge 0} \gamma^t\, \mathbb{E}_\mu^\pi\big[r_{A_t}(S_t)\, \mathbb{I}(S_t = s, A_t = a)\big]$$

$r_a(s)$

$$r_{A_t}(S_t)\,\mathbb{I}(S_t=s,\ A_t=a)$$

$$= r_a(s)\,\mathbb{I}(S_t=s,\ A_t=a)$$

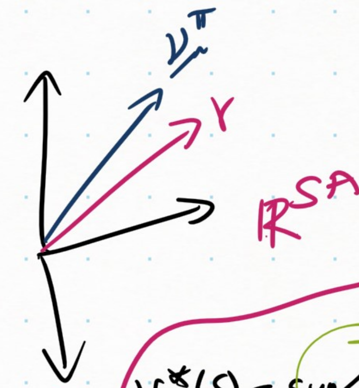
$$v^\pi(s) = \sum_{s,a} \sum_{t\ge 0} \gamma^t\, r_a(s)\, \mathbb{P}^\pi_\mu(S_t=s, A_t=a)$$

$$= \sum_{s,a} r_a(s) \underbrace{\sum_{t\ge 0} \gamma^t\, \mathbb{P}^\pi_\mu(S_t=s, A_t=a)}_{\nu^\pi_\mu(s,a)}$$

$\delta_s \to \nu^\pi_s$

$$= \langle \nu^\pi_\mu, r \rangle$$

$$\tilde{\nu}^\pi_\mu(s) = \sum_a \nu^\pi_\mu(s,a)$$




$\pi$ arb. $\to ML(\pi)$

$$v^\pi(s) = \langle \nu^\pi_s, r\rangle$$

$$= \langle \nu_s^{ML(\pi)}, r\rangle \Big/ \sup_\pi$$

$$v^*(s) = \sup_\pi \langle \nu_s^{ML(\pi)}, r\rangle \le \sup_{\pi\, ML} \langle \nu^\pi_s, r\rangle =: \tilde{v}^*(s)$$

**Theorem:** $\forall \pi,\ \forall \mu \in M_1(S)$

$\exists \pi'$ ML s.t.

$$\nu^\pi_\mu = \nu^{\pi'}_\mu .$$

**Proof (hint):**

$$\pi'(a|s) = \begin{cases} \dfrac{\nu^\pi_\mu(s,a)}{\tilde{\nu}^\pi_\mu(s)}, & \tilde{\nu}^\pi_\mu(s) \neq 0 \\[2ex] \pi_0(a), & \text{otherwise} \end{cases}$$

$\pi_0 \in M_1(\mathcal{A})$ arbitrary.

$\mathbb{R}^S$ (value fns)

$\Theta(A^S)$

FT : $v^*$ known

$T_\pi v^* = T v^*$

$\operatorname{argmax}_a \{ r_a(s) + \gamma \overbrace{\langle P_a(s), v^* \rangle}^{O(1)} \}$

$\downarrow$

$O(A)$ $\qquad\qquad$ $O(S)$

① Compute $v^*$

② Find greedy wrt $v^*$

$O(S^2 A)$

value iteration

$T v^* = v^*$

$v_0 = 0$

$v_{k+1} = T v_k$

$\downarrow$

$v^* \qquad k \to \infty$

Prop: $\|Tv - Tu\|_\infty \leq \gamma \|v - u\|_\infty$ ($\gamma$-contraction)

Banach's fixed point thm

Thm: $\|v_k - v^*\|_\infty \leq \gamma^k \|v^*\|_\infty \leq \varepsilon$

$\to 0, \varepsilon \to 0$

$\|v^*\|_\infty \leq \frac{1}{1-\gamma}$

$r_a(s) \in [0,1]$