1) Infinite Horizon Discounted MDP

$$\mathcal{M} = \{ \mathcal{S}, \mathcal{A}, P, r, \gamma \}$$

$\mathcal{S}$: space of possible states $s \in \mathcal{S}$

$\mathcal{A}$: space of possible actions $a \in \mathcal{A}$

P : transition function

$$P : \mathcal{S} \times \mathcal{A} \longrightarrow \Delta(\mathcal{S})$$

probability distributi

r : reward function

$$r : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathbb{R})$$

γ : discount factor $[0, 1]$

$$0 < \gamma < 1$$

cumulative reward:

$$\sum_{t=0}^{\infty} \underline{\gamma^t} \, r_t$$

Aside: deterministic $f : \mathcal{X} \rightarrow \mathcal{Y}$ $\quad f(x) = y$

encode as stochastic $F : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$

$$\rightarrow F(x) = \{ f(x) \text{ w.p. } 1$$

sometimes overload notation for deterministic policies, reward fns.

ex. $a = \pi(s)$ $\quad r_t = r(s_t, a_t)$

Adopt notation

$$F(y|x) = \mathbb{P}\{F(x) = y\}$$

e.g. $\pi(a|s)$

Goal of RL:

find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
maximize the expected discounted
cumulative reward.

$$\pi^* \quad \boxed{\underset{\pi}{\text{maximize}} \ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \begin{array}{l} a_t \sim \pi(s_t) \\ s_{t+1} \sim P(s_t, a_t) \\ s_0 \text{ given} \end{array} \right]}$$

2) <u>Value Function & Q function</u>

allow us to reason about
long term effects of policies

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \begin{array}{l} s_0 = s \\ s_{t+1} \sim P(s_t, a_t) \\ a_t \sim \pi(s_t) \end{array} \right]$$

$$Q^\pi(s,a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \Big| \begin{array}{l} s_0 = s \\ a_0 = a \\ s_{t+1} \sim P(s_t, a_t) \\ a_t \sim \pi(s_t) \end{array}\Big]$$

Bellman Equation

Notice that

$$\underbrace{\sum_{t=0}^{\infty} \gamma^t r_t}_{} = r_0 + \gamma \underbrace{\sum_{t=1}^{\infty} \gamma^{t-1} r_t}_{}$$

replace $t = t'+1$

$$\underbrace{\sum_{t'=0}^{\infty} \gamma^{t'} r_{t'+1}}_{}$$

assume deterministic π & r

we can write

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

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

$$\mathbb{E}[V^\pi(s') \mid s' \sim P(s,a)]$$

What property

$$\sum_{t=0}^{\infty} \gamma^t r_t = r_0 + \gamma r_1 + \gamma^2 r_t + \dots$$

$$= r_0 + \gamma \underbrace{\left(r_1 + \gamma r_2 + \dots \right)}_{\sum_{t=0}^{\infty} \gamma^t r_{t+1}}$$

3) Policy Evaluation

How good is a policy?
In terms of the value function.

Given MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, \gamma, r\}$
and a policy π, what is V^π?

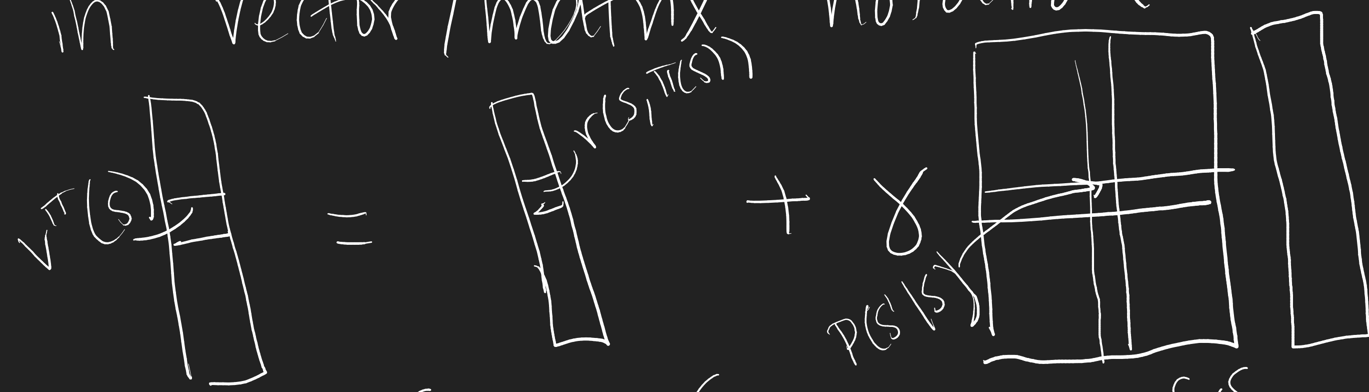
Bellman equation:

$$\forall s \in \mathcal{S}, \quad V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, \pi(s)) V(s')$$

$|\mathcal{S}| = S$
\# states

S equations
S unknowns $(V^\pi(s))$

in vector/matrix notation



$V \in \mathbb{R}^S \quad R \in \mathbb{R}^S \quad P \in \mathbb{R}^{S \times S} \quad V$

solving linear equation

$$V = R + \gamma P V$$

$$V - \gamma P V = R$$

$$(I - \gamma P) V = R$$

$$V = (I - \gamma P)^{-1} R$$

HW0

Exact solution! $O(S^3)$

4) Approximate Policy Evaluation

$V = R + \gamma P V$

Algorithm

initialize V^0

for $t = 0, 1, \ldots T$

$V^{t+1} \leftarrow R + \gamma \underbrace{P V^t}$

$O(T S^2)$

Lemma: $\|V^{t+1} - V^{\pi}\|_\infty \leq \gamma \|V^t - V^\pi\|_\infty$

$\uparrow$ true

Proof

$$\|V^{t+1} - V^{\pi}\|_\infty = \| R + \gamma P V^t - \underbrace{V^\pi}\|_\infty$$

$$V^\pi = R + \gamma P V^\pi$$

$$= \gamma \| P(V^t - V^\pi)\|_\infty$$

@ index s, $\left| \mathbb{E}_{s' \sim P(s, \pi(s))} \left[V^t(s') - V^\pi(s') \right] \right|$

$$\leq \mathbb{E}_{s' \sim P(s,\pi(s))} |V^t(s') - V^\pi(s')|$$

$$\|P(V^t - V^\pi)\|_\infty \leq \max_s \mathbb{E}_{s' \sim P} \left[|V^t(s') - V^\pi(s')| \right]$$

$$\leq \cancel{\max_s} \max_{s'} \left[|V^t(s') - V^\pi(s')| \right]$$

$$\leq \|V^t - V^\pi\|_\infty$$

Theorem (Convergence) after T iterations,

$$\boxed{\|V^T - V^\pi\|_\infty \le \gamma^T \|V^0 - V^\pi\|_\infty}$$

What T give ε-approximation?

$$T \ge \frac{\log\left(\frac{\|V^0 - V^\pi\|_\infty}{\varepsilon}\right)}{\log(1/\gamma)}$$

$$O(S^2 \log(1/\varepsilon))$$

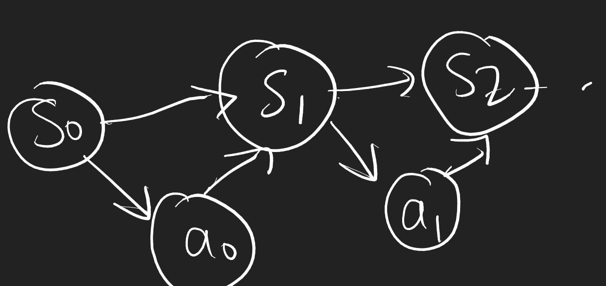
5) State-Action Distribution

Trajectory of MDP up to t,

$$(s_0, a_0, s_1, a_1, \dots, s_t, a_t)$$

Consider possible stochastic policies

$$\mathbb{P}^\pi(s_0, a_0, \dots, s_t, a_t) = \pi(a_0|s_0)P(s_1|s_0,a_0)$$
$$\times \pi(a_1|s_1)P(s_2|s_1,a_1)$$
$$\dots$$
$$\times P(s_t|s_{t-1},a_{t-1})$$
$$\times \pi(a_t|s_t)$$



$$\mathbb{P}_t^{\pi}(s, a; s_0) = \sum_{\substack{a_{0:t-1} \in \mathcal{A}^t \\ s_{1:t-1} \in \mathcal{S}^{t-1}}} \mathbb{P}^{\pi}(s_0, a_0, \dots, s_t, a_t)$$