CS 4/5789 Prof Sarah Dean

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

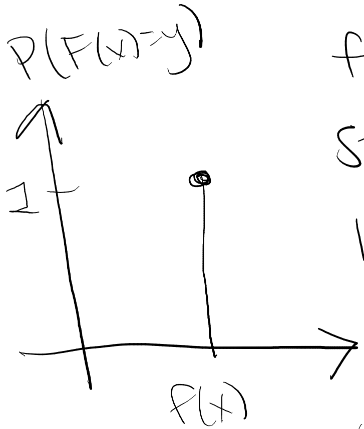
r : reward function $r: \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$

probability distributions

γ : discount factor $0 < \gamma < 1$

Aside: We can encode a derministic function $f: \mathcal{X} \to \mathcal{Y}$ as a stochastic one $F: \mathcal{X} \to \Delta(\mathcal{Y})$ by $F(x) = f(x)$ w.p. 1.

"with probability"



sometimes as shorthand we will overload notation and write e.g. $a = \pi(s)$ instead of $a \sim \pi(s)$ if the policy is deterministic.

Additionally, we will adopt the notation

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

(e.g $\pi(a|s)$, $P(s'|a,s)$)

In this notation we can write the goal:

finding a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes the (discounted) cumulative reward.

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

We will spend the semester learning how to solve this problem. In RL, we do not assume that $P(\cdot,\cdot)$ is known, and therefore we have to solve the optimization using data.
For now, we suppose that P is known.

2) Value and Q Function

allow us to reason about policy's long term effect.

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

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

Bellman Equations

Notice that

$$\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) = r(s_0, a_0) + \gamma \sum_{t=1}^{\infty} r(a_t, s_t)$$

$$(\text{let } t' = t+1) \qquad = r(s_0, a_0) + \gamma \sum_{t'=0}^{\infty} r(a_{t+1}, s_{t+1})$$

let's consider deterministic policies and reward fuctions.

This observation allows us to write

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

Poll EV: what property of expectations do we use?

How would the expressions change for stochastic reward functions and policies?

3) Policy Evaluation

How do we characterize how good a policy is? In terms of value function

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

function from $\mathcal{S} \to \mathbb{R}$

The Bellman equation:

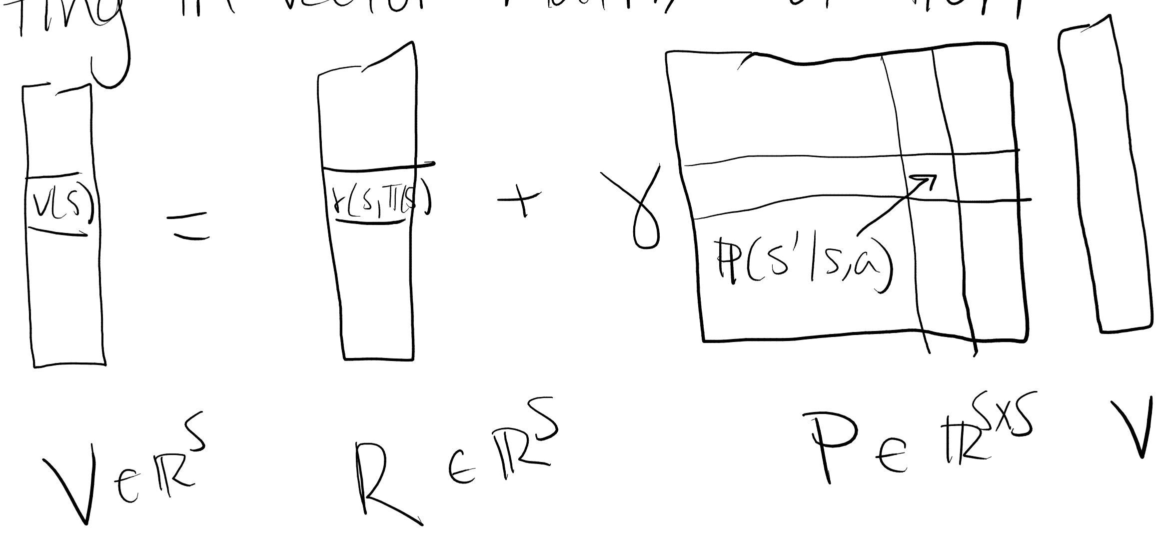
$$\forall s, \quad V^\pi(s) = r(s, \pi(s)) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s, \pi(s))}[V^\pi(s')]$$

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

denote $S = |\mathcal{S}|$ number of states

S linear constraints
S unknowns

Writing in vector-matrix notation



solving the linear equations

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

This is valid as long as $I - \gamma P$ is invertible (HW0)

Exact solution! But $O(S^3)$ for matrix inversion...

4) Approximate Policy Evaluation

can we trade accuracy for faster computation?

yes! Iterative Algorithm for fixed point.

> **Algorithm (Iterative PE)**
>
> initialize V^0
>
> for $t = 0, \ldots$:
>
> $\quad V^{t+1} \leftarrow R + \gamma P V^t$

Q: complexity per iteration?

A: matrix-vector multiply is $O(S^2)$

To show that this algorithm works, we will show a <u>contraction</u>, which is a general strategy for fixed point algorithms.

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

Proof:

$$\|V^{t+1} - V^{\pi}\|_\infty = \|R + \gamma P V^t - V^{\pi}\|_\infty \quad \text{(alg)}$$

$$= \|R + \gamma P V^t - (R + \gamma P V^{\pi})\|_\infty \quad \text{(Bellman)}$$

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

recall that each entry of this vector represents the expectation

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

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

so $\|P(V^t - V^{\pi})\|_\infty \leq \max\limits_{s} \mathbb{E}_{s' \sim P(s, \pi(s))} [|V^t(s') - V^{\pi}(s')|]$

(expectation upper bounded by max)

$$\leq \max_{s'} |V^t(s') - V^{\pi}(s')| = \|V^t - V^{\pi}\|_\infty$$

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

$\boxed{\begin{array}{l}\underline{\text{Theorem}}\text{: after } t \text{ iterations,} \\ \qquad \|V^t - V^\pi\|_\infty \leq \gamma^t \|V^0 - V^\pi\|_\infty\end{array}}$

i.e.,

$$\forall s, \quad |V^t(s) - V^\pi(s)| \leq \gamma^t \max_s |V^0(s) - V^\pi(s)|$$

follows by repeated application of Lemma.

How many iterations necessary for ε accurate solution?

$$\gamma^t \|V^0 - V^\pi\|_\infty \leq \varepsilon$$

$$\Rightarrow \quad t \geq \log\left(\frac{\|V^0 - V^\pi\|_\infty}{\varepsilon}\right) / \log(1/\gamma)$$

overall complexity

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

compare with $O(S^3)$ for exact.

5) State-Action Distribution

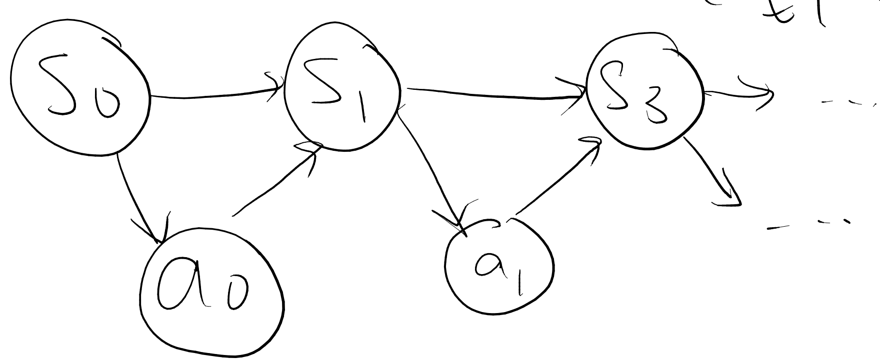
Trajectory of MDP up to step t:

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

What is the probability of a particular trajectory under policiy π?

considering possibly stocastic policies,

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



What is the probability of seeing (s, a) at timestep t, starting from s_0?

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

Discounted Average State-Action Distribution

$$d_{s_0}^{\pi}(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_h^{\pi}(s,a;s_0)$$

HW0: is this a valid distribution?

$$V^{\pi}(s_0) = \frac{1}{1-\gamma} \sum_{s,a} d_{s_0}^{\pi}(s,a)\, r(s,a) \,?$$