CS 4/5789 31 Jan 22 Prof Sarah Dean

1) State-Action Distribution

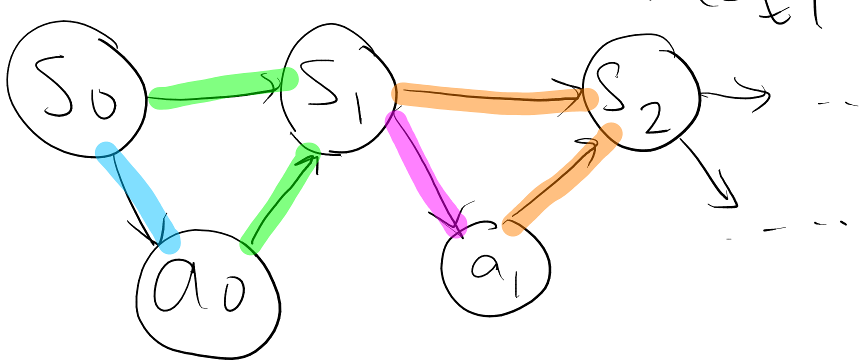
Trajectory of MDP up to step t:

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

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

considering possibly stocastic policies,

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



← This is a graphical model of transitions which illustrates condition indepence. (Markov Property)

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

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

Discounted Average State-Action Distribution

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

HWO: is this a valid distribution?

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

2) Optimal Policies & Bellman Optimality

we have A^S policies — which one is optimal?

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

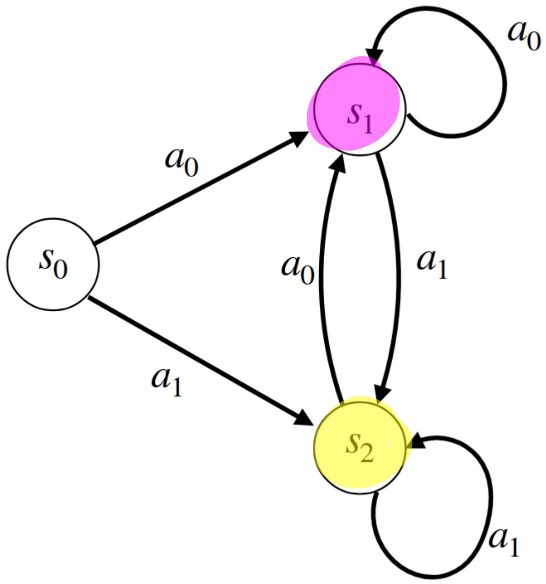
(deterministic policies & reward)

<u>Fact</u>: for infinite horizon discounted MDP, there always exists a deterministic $\pi^* : \mathcal{S} \to \mathcal{A}$ such that $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all $s \in \mathcal{S}$ and all π

i.e. π^* dominates all other π at all states!
This means it is the optimal policy

notation: $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$

Example: deterministic MDP with 2 actions & 3 states



Reward is always 0 except $r(s_1, a_0) = 1$

What is the optimal policy?

consider $\pi_1(s) = a_1 \ \forall s$

$V^{\pi_1}(s_0) = V^{\pi_1}(s_1) = V^{\pi_1}(s_2) = 0$

instead, $\pi_0(s) = a_0 \ \forall s$

$$V^{\pi_0}(s_0) = V^{\pi_0}(s_2) = 0 + \sum_{t=1}^{\infty} \gamma^t \cdot 1 = \frac{\gamma}{1-\gamma}$$

$$V^{\pi_0}(s_1) = \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$$

To be rigourous we would still have to argue about the other 6 possible policies...

Bellman Optimality

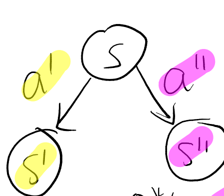
This is a key property of the optimal policy

Theorem 1 (Bellman Optimality)

$$V^*(s) = \max_{a \in \mathcal{A}} \overbrace{r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[V^*(s')\right]}^{Q^*(s,a)} \quad \text{for all } s \in \mathcal{S}$$

If we know the value of s', we can use this to compute the optimal action & value of s.

← consider a simple example with two actions & deterministic transitions



$Q^*(s,a') = r(s,a') + \gamma V(s')$

$Q^*(s,a'') = r(s,a'') + \gamma V(s'')$

$$V^*(s) = \max\{Q^*(s,a'), Q^*(s,a'')\}$$

Proof of Bellman optimality

We show for $\hat{\pi}(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$,

that $V^{\hat{\pi}}(s) = V^*(s)$.

a) by definition of $V^*(s)$, $V^*(s) \geq V^{\pi}(s) \;\; \forall s$.

b) we now show that $V^*(s) \leq V^{\hat{\pi}}(s) \;\; \forall s$.

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

$$\left(f(a) \leq \max_u f(u)\right)$$

$$\star \quad \leq \max_a r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[V^*(s')\right]$$

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

$$(\star\star) \quad = r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s'}\left[r(s', \pi^*(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^*(s'))}\left[V^*(s'')\right]\right] \qquad \text{(defn. } V^*\text{)}$$

$$\leq r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi})}\left[r(s', \hat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \hat{\pi}(s'))}\left[V^*(s'')\right]\right] \qquad \text{(repeat } \star\text{)}$$

$$\leq r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))}\left[r(s', \hat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi(s'))}\left[r(s'', \hat{\pi}(s'')) + \gamma \mathbb{E}_{s''' \sim P(s'', \hat{\pi}(s''))}\left[V^*(s''')\right]\right]\right] \qquad \text{(repeat } \star\star\text{)}$$

$$\leq \mathbb{E}_{s, s', s''}\left[r(s, \hat{\pi}(s)) + \gamma r(s', \hat{\pi}(s')) + \gamma^2 r(s'', \hat{\pi}(s'')) \cdots\right] \qquad \text{(repeat)}$$

$$= V^{\hat{\pi}}(s) \qquad \text{(definition } V^{\hat{\pi}}\text{)}$$

$\square$

Therefore, $V^{\hat{\pi}}(s) = V^*(s) \quad \forall s$. This means that $\hat{\pi}$ acheives optimal value, so

$$\hat{\pi} = \arg\max_{a \in \mathcal{A}} Q^*(s,a) \quad \text{is an optimal policy.}$$

We now show that Bellman optimality is not only necessary, but also sufficient to characterize V^*.

Theorem 2:

For any $V : \mathcal{S} \to \mathbb{R}$, if $V(s) = \max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} [V(s')] \right]$ for all $s \in \mathcal{S}$, then $V(s) = V^*(s)$.

This means that finding optimal value function is equivalent to the Bellman optimality condition.

We can consider just one step between s and s' to check if $V = V^*$, we check if

$$\left| V(s) - \max_a \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} [V(s')] \right] \right| = 0 \quad \forall s$$

Proof:

$$|V(s) - V^*(s)| = \left| \max_a \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} (V(s')) \right] - \max_a \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} (V^*(s')) \right] \right|$$

basic inequalities (HW0)

$$\le \max_a \left| \cancel{r(s,a)} + \gamma \mathbb{E} V(s') - (\cancel{r(s,a)} - \gamma \mathbb{E} V^*(s')) \right|$$

$$\le \max_a \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} |V(s') - V^*(s')|$$

$$\le \max_a \gamma \mathop{\mathbb{E}}_{s'} \left[\max_{a'} \mathop{\mathbb{E}}_{s'' \sim P(s',a')} |V(s'') - V^*(s'')| \right]$$

(repeat)

$$\le \max_{a_1, a_2, \dots, a_{k-1}} \gamma^k \mathop{\mathbb{E}}_{s_k} |V(s_k) - V^*(s_k)| \longrightarrow 0 \text{ as } k \to \infty \quad \square$$

Example Recall the deterministic MDP.

Now we can verify that $\pi_0(s) = a_0 \ \forall s$ is the optimal policy!

$$s_0: \ \left| V(s_0) - \max\left(0 + \gamma V(s_1),\ 0 + \gamma V(s_2)\right)\right|$$
$$= \left| \frac{\gamma}{1-\gamma} - \max\left(\frac{\gamma}{1-\gamma},\ \frac{\gamma^2}{1-\gamma}\right)\right| = 0$$

$$s_1: \ \left| V(s_1) - \max\left(1 + \gamma V(s_1),\ 0 + \gamma V(s_2)\right)\right|$$
$$= \left| \frac{1}{1-\gamma} - \max\left(1 + \frac{\gamma}{1-\gamma},\ \frac{\gamma^2}{1-\gamma}\right)\right| = 0$$

$$s_2: \ \left| V(s_2) - \max\left(0 + \gamma V(s_1),\ 0 + \gamma V(s_2)\right)\right|$$
$$= \left| \frac{\gamma}{1-\gamma} - \frac{\gamma}{1-\gamma}\right| = 0.$$

3) Value Iteration

How to find the optimal policy?

Algorithm: Enumeration

for all $\pi: \mathcal{S} \to \mathcal{A}$:

  compute $V^\pi = \text{Exact-PE}(\pi)$

select $\hat{\pi}$ such that

  $V^{\hat{\pi}}(s) \geq V^{\pi} \ \forall s, \pi$

a naive approach

The computation time is $O(A^S \cdot S^3)$

Exponential complexity is a problem!

Define Bellman operator $\mathcal{T}$:

given function $Q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the Bellman operator $\mathcal{T}Q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines another fn.

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right]$$

Consider Tabular representation of Q,

$Q \in \mathbb{R}^{SA}$ $\quad$ $S = |\mathcal{S}|$ number of states

$A = |\mathcal{A}|$ number of actions

Then we also have

$\mathcal{T}Q \in \mathbb{R}^{SA}$ so we can think of $\mathcal{T}$ as a map (nonlinear) from $\mathbb{R}^{SA}$ to $\mathbb{R}^{SA}$

Fixed Point Motivation

By Bellman optimality,

$$Q^*(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s \sim P(s,a)} \max_{a'} Q^*(s',a')$$

Thus $Q^* = \mathcal{T}Q^*$ the optimal Q fn. is a fixed point solution to $Q = \mathcal{T}Q$

Algorithm: Value Iteration

Initialize Q^0

for $t = 0, 1, 2, \ldots$

$$Q^{t+1} \leftarrow \mathcal{T} Q^t$$

"fixed point iteration"

like inexact policy iteration.

$$Q^{t+1}(s,a) \leftarrow r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)}\left[\max_{a'} Q^t(s',a')\right]$$

$\forall\, s, a$

Convergence of Value Iteration

We will use a contraction argument.

Lemma: (contraction) for any Q, Q'

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \le \gamma \|Q - Q'\|_\infty$$

Proof:

$$\begin{aligned}
|\mathcal{T}Q(s,a) - \mathcal{T}Q'(s,a)| &= \Big|\cancel{r(s,a)} + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)}\left[\max_{a'} Q(s',a')\right] \\
&\qquad - \Big(\cancel{r(s,a)} + \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)}\left[\max_{a'} Q'(s',a')\right]\Big)\Big| \\
&\le \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)} \left|\max_{a'} Q(s',a') - \max_{a'} Q'(s',a')\right| && \text{(basic inequalities HW0)} \\
&\le \gamma \mathop{\mathbb{E}}_{s' \sim P(s,a)}\left[\max_{a'} |Q(s',a') - Q'(s',a')|\right] && \left(f(s') \le \max_{s} f(s)\right) \\
&\le \gamma \max_{s'} \max_{a'} |Q(s',a') - Q'(s',a')| \\
&= \gamma \|Q - Q'\|_\infty && \text{(definition of } \|\cdot\|_\infty\text{)}
\end{aligned}$$

$\square$

Lemma: (convergence) for any Q^0

$$\|Q^t - Q^*\|_\infty \le \gamma^t \|Q^0 - Q^*\|_\infty$$

Proof:

$$\|Q^t - Q^*\|_\infty = \|\mathcal{J}Q^{t-1} - \mathcal{J}Q^*\|_\infty \le \gamma\|Q^{t-1} - Q^*\|_\infty$$
$$\le \gamma^2 \|Q^{t-2} - Q^*\|_\infty$$
$$\cdots$$
$$\le \gamma^t \|Q^0 - Q^*\|_\infty \quad \square$$

From Q functions to policies

We know $\pi^*(s) = \arg\max_a Q^*(s,a)$

since $Q^t(s,a) \approx Q^*(s,a)$ during value iteration,

$\pi^t(s) = \arg\max_a Q^t(s,a)$

a good choice?

Theorem: The quality of π^t is bounded below:

$$V^{\pi^t}(s) \ge V^*(s) - \frac{2\gamma^t}{1-\gamma}\|Q^0 - Q^*\|_\infty \quad \forall s \in \mathcal{S}$$

Proof:

Assume the following claim is true:

$$V^{\pi^t}(s) - V^*(s) \ge \gamma \mathop{\mathbb{E}}_{s' \sim P(s,\pi^t(s))}\left[V^{\pi^t}(s') - V^*(s')\right] - 2\gamma^t \|Q^0 - Q^*\|_\infty$$

Then recursing k times,

$$V^{\pi^t}(s) - V^*(s) \ge \gamma^k \mathop{\mathbb{E}}_{s \sim P(s,\pi^t(s))}\left[V^{\pi^t}(s') - V^*(s)\right] - 2\sum_{l=0}^{k}\gamma^{l+t}\|Q^0 - Q^*\|_\infty$$

letting $k \to \infty$,

$$V^{\pi^t}(s) - V^*(s) \geq -2\gamma^t \sum_{l=0}^{\infty} \gamma^l \|Q^0 - Q^*\|_\infty$$

$$= \frac{-2\gamma^t}{1-\gamma} \|Q^0 - Q^*\|_\infty$$

Proof of claim:

$$V^{\pi^t}(s) - V^*(s) = Q^{\pi^t}(s, \pi^t(s)) - Q^*(s, \pi^*(s)) \quad \text{(definition)}$$

$$\underbrace{- Q^*(s, \pi^t(s)) + Q^*(s, \pi^t(s))}_{=0}$$

$$= \gamma \mathbb{E}_{s' \sim p(s, \pi^t(s))}\left[V^{\pi^t}(s') - V^*(s')\right] + Q^*(s, \pi^t(s)) - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim p(s, \pi^t(s))}\left[V^{\pi^t}(s') - V^*(s')\right] + Q^*(s, \pi^t(s)) \underbrace{- Q^t(s, \pi^t(s)) + Q^t(s, \pi^*(s))}_{\leq 0 \text{ by } \pi^* \text{ optimality}} - Q^*(s, \pi^*(s))$$

$$\geq \gamma \mathbb{E}_{s' \sim p(s, \pi^t(s))}\left[V^{\pi^t}(s'), V^*(s')\right] - \|Q^t - Q^*\|_\infty - \|Q^t - Q^*\|_\infty$$

by definition of $\|\cdot\|_\infty$

$$\geq \gamma \mathbb{E}_{s' \sim p(s, \pi^t(s))}\left[V^{\pi^t}(s'), V^*(s')\right] - 2\gamma^t \|Q^0 - Q^*\|_\infty \quad \text{(convergence Lemma)}$$

$\square$

Summary of Value Iteration (VI)

1) VI (fixed point)

$$Q^{t+1} \leftarrow \mathcal{T} Q^t$$

contraction

2) VI convergence

$$\|Q^t - Q^*\|_\infty \leq \gamma^t \|Q^0 - Q^*\|_\infty$$

exponentially fast
"geometric rate"

$$\pi^t(s) = \arg\max_a Q^t(s,a)$$

3) policy performance

$$V^{\pi^t}(s) \geq V^*(s) - \frac{2\gamma^t}{1-\gamma} \|Q^0 - Q^*\|_\infty$$

Convergence argument is similar to Iterative Policy Eval (PE)

Bellman Eq: $V^\pi = R + \gamma P V^\pi$	Bellman Optimality $Q^* = \mathcal{T} Q^*$
Iterative PE $V^{t+1} \leftarrow R + P V^t$	VI $Q^{t+1} \leftarrow \mathcal{T} Q^t$
by contraction, $\|V^t - V^\pi\|_\infty \leq \gamma^t \|V^0 - V^\pi\|_\infty$	$\|Q^t - Q^*\|_\infty \leq \|Q^0 - Q^*\|_\infty$