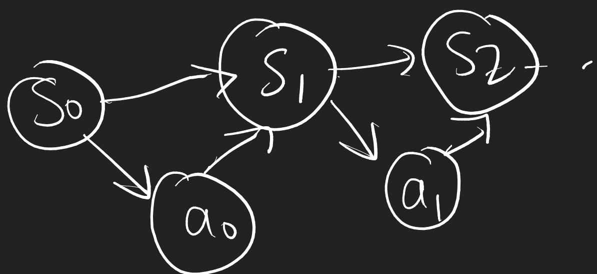
5) State-Action Distribution

Trajectory of MDP up to t,

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

Consider possible stochastic policies

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



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

Discounted Average State-Action Distribution

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

HW0: is this a valid distribution?

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

2) Optimal Policies & Bellman Optimality

A^S policies!

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

<u>Fact</u>: there always exist

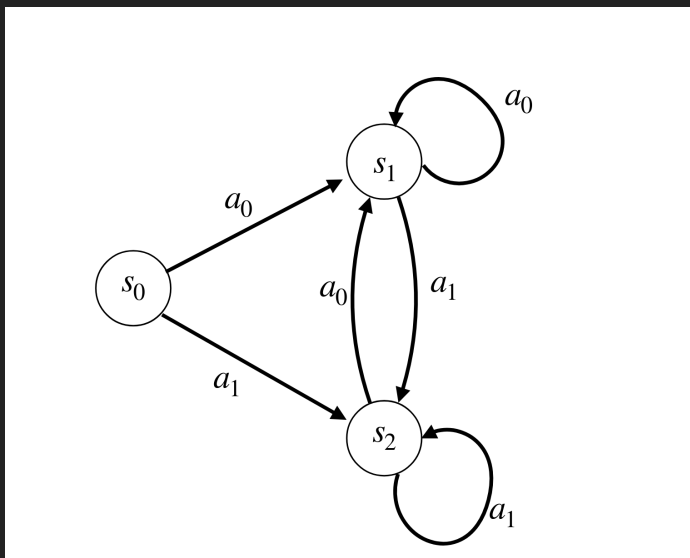
$\pi^* : \mathcal{S}\to\mathcal{A}$ s.t. $\underline{V^{\pi^*}(s) \geq V^{\pi}(s)}$

for all π and $s\in\mathcal{S}$

dominates

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

Example: deterministic MDP



$r(s_1, a_0) = 1$

0 otherwise

$$V^{\pi_0}(s) = \begin{cases} \frac{\gamma}{1-\gamma} & s_0, s_2 \\ \frac{1}{1-\gamma} & s_1 \end{cases}$$

$\pi_0(s) = a_0 \quad \forall s$

PollEV.com/sarahdean011

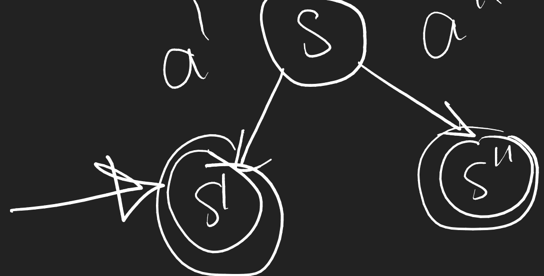
Bellman Optimality

Theorem 1:

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

$\forall s \in \mathcal{S}$

If $V^*(s')$ is known, can compute $V^*(s)$



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

Proof: We show
$$\hat{\pi}(s) = \underset{a \in \mathcal{A}}{\arg\max}\, Q^*(s,a)$$
that $V^{\hat{\pi}}(s) = V^*(s)$

a) by definition
$$V^*(s) \geq V^{\hat{\pi}}(s) \quad \forall s.$$

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

$$\begin{aligned}
V^*(s) &= r(s,\pi^*(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^*(s))}\left[V^*(s')\right] \\
&\leq \max_a r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[V^*(s')\right] \\
&= r(s,\hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\hat{\pi}(s))}\left[\boxed{V^*(s')}\right] \\
&\leq r(s,\hat{\pi}(s)) + \gamma \mathbb{E}\left[r(s',\hat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s,\hat{\pi})}\left[V^*(s'')\right]\right] \\
&\leq \mathbb{E}_{\substack{s' \sim P(s) \\ s'' \sim P(s')}}\Big(r(s,\hat{\pi}(s)) + \gamma r(s',\hat{\pi}(s')) \\
&\qquad + \gamma^2 r(s'',\hat{\pi}(s'')) + \dots\Big)
\end{aligned}$$

$= V^{\hat{\pi}}(s)$ $\square$

This means that $\hat{\pi}$ acheives optimal value, so

$$\pi^*(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^*(s,a)$$

is an optimal policy

Now, show Bellman optimality is sufficient to characterize $V^*(s)$.

<u>Theorem 2</u>: for any $V : \mathcal{S} \to \mathbb{R}$

if $V(s) = \max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[V(s')]$

for all $s \in \mathcal{S}$, then

$V(s) = V^*(s)$.

This means we can confirm that $V = V^*$ by checking

$$\left| V(s) - \max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right| = 0 \quad \forall s.$$

Proof:

$$\left| \underline{V(s)} - \underline{V^*(s)} \right| =$$

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

(basic inequality HW0)

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

(Jensens)

$$\leq \max_a \gamma \mathbb{E}_{s'} \left| V(s') - V^*(s') \right|$$

$$|V(s) - V^*(s)| \leq \max_{a, a'} \gamma^2 \mathbb{E}_{\substack{s' \sim P(s,a) \\ s'' \sim P(s',a')}} \left| V(s'') - V^*(s'') \right|$$

iterate k times

limit to infinity

$|V(s) - V^*(s)| \longrightarrow 0$

$V(s) = V^*(s)$ $\square$

<u>Example</u>: check $\pi(s) = a_0 \ \forall s$ is optimal using Bellman optimality.

3) Value Iteration

How to find optimal policy? (π^*, V^*, Q^*)

enumeration: $O(A^S \, S^3)$

policy evaluation

<u>Define</u> Bellman Operator $\mathcal{T}$

given $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, Bellman operator $\mathcal{T}Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, defined as

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

Think of representing tabular

$$Q \in \mathbb{R}^{SA} \quad \& \quad \mathcal{T} : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$$

Fixed Point Iteration

Bellman optimality

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s \sim P(s,a)} \Big[\overbrace{\max_{a'} Q^*(s',a')}^{V^*(s')} \Big]$$

$$\mathcal{T} Q^* = Q^* \quad \text{(fixed point)}$$

Alg: Value Iteration (VI)

Initialize Q^0

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

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

Convergence

Lemma (contraction) for any Q, Q'

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

remember $\| Q \|_\infty = \max_{s,a} Q(s,a)$

Proof:

$$|JQ(s,a) - JQ'(s,a)|$$

$$= \left| \cancel{r(s,a)} + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q(s',a')\right] - \left(\cancel{r(s,a)} + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q'(s',a')\right]\right) \right|$$

(basic inequalities HW0)

$$\leq \gamma \mathbb{E}_{s'} \left| \max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right|$$

$$\leq \gamma \mathbb{E}_{s'} \left[\max_{a'} \left| \underbrace{Q(s',a') - Q'(s',a')}_{} \right| \right]$$

$f(s') \leq \max_s f(s)$

$$\leq \gamma \max_s \max_{a'} |Q(s,a') - Q'(s,a')|$$

$$= \gamma \|Q - Q'\|_\infty$$

Lemma (convergence)

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

Recall: $\pi^*(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^*(s,a)$

misc notes from office hours:

$$\mathbb{E}_{x\sim\Delta(\mathcal{X})}(f(x)) \le \max_{x\in\mathcal{X}} f(x)$$

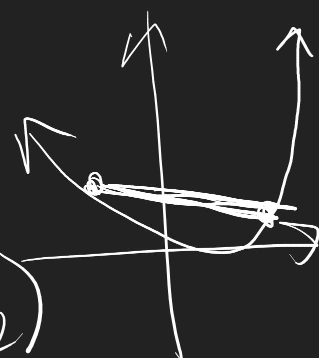
$$\Uparrow$$

$$f(x) \le \max_{y\in\mathcal{X}} f(y)$$

Jensen's inequality

$$f\left(\mathbb{E}_{x\sim D}(x)\right) \le \mathbb{E}_{x\sim D}[f(x)]$$

f is convex.



$$f(tx_1+(1-t)x_2) \ge t\,f(x_1) + (1-t)f(x_1)$$