

Reinforcement Learning (Bellman Eqn)

Friday, March 27, 2020 3:05 PM

recall the policy value $V_{\pi}(s) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right]$

Def The State-action value function Q_{π} for a policy π is the expected return if

we're at state s and take action a and then follow policy π

$$\begin{aligned} Q_{\pi}(s, a) &= \mathbb{E}_{a_t \sim \pi(s_t)} [r(s, a) + \mathbb{E}_{s_t \sim P(s_t | s_0, a_0)} \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]] \\ &= \mathbb{E}_{a_t \sim \pi(s_t)} [r(s, a) + \gamma V_{\pi}(s_1) \mid s_0 = s, a_0 = a] \quad \text{dependence on } s_0 \text{ is implicit: } \\ &\qquad s_1 \sim P(s_1 | s_0, a_0) \\ &= \mathbb{E}[r(s, a)] + \gamma \cdot \sum_{s_1 \in S} V_{\pi}(s_1) \cdot P(s_1 | s, a) \quad \text{showing explicit dependence} \end{aligned}$$

like $V_{\pi}(s)$ except a is an input,

$$\text{i.e., } \mathbb{E}_{a \sim \pi(s)} Q(s, a) = V_{\pi}(s) \quad (*)$$

Thm 17.6: Policy-Improvement Thm

$$\left(\forall s \in S, \mathbb{E}_{a \sim \pi'(s)} Q_{\pi'}(s, a) \geq \mathbb{E}_{a \sim \pi(s)} Q_{\pi}(s, a) \right) \Rightarrow \left(\forall s \in S, V_{\pi'}(s) \geq V_{\pi}(s) \right)$$

← mismatch

(And if $\exists s$ s.t. left ineq. is strict $\Rightarrow \exists s$ s.t. right ineq. is strict)

Proof:

$$\begin{aligned} \text{Assume } (\dots), \text{ so } \underbrace{\mathbb{E}_{a \sim \pi(s)} Q_{\pi}(s, a)}_{V_{\pi}(s) \text{ by } (*)} &\leq \mathbb{E}_{a \sim \pi'(s)} Q_{\pi}(s, a) \quad \text{by our hypothesis} \\ &= \mathbb{E}_{a \sim \pi'(s)} [r(s, a) + \gamma V_{\pi}(s_1) \mid s_0 = s] \quad \text{by def'n of } Q \\ &= \mathbb{E}_{a \sim \pi'(s)} [r(s, a) + \gamma \mathbb{E}_{a_1 \sim \pi(s_1)} [Q_{\pi}(s_1, a_1)] \mid s_0 = s] \quad \text{by } (*) \\ &\leq \mathbb{E}_{a \sim \pi'(s)} [r(s, a) + \gamma \mathbb{E}_{a_1 \sim \pi'(s_1)} [Q_{\pi'}(s_1, a_1)] \mid s_0 = s] \quad \text{by our hypothesis (again)} \\ &= \mathbb{E}_{a \sim \pi'(s)} [r(s, a) + \gamma \cdot r(s_1, a_1) + \gamma^2 V_{\pi'}(s_2) \mid s_0 = s] \quad \text{by def'n of } Q \\ &\qquad \qquad \qquad \text{↑ } s_2 \text{ is a r.v., dependent on } (a, s, a_1, s_1) \\ &\cdots \leq \mathbb{E}_{a \sim \pi'(s)} \left[\sum_{t=0}^{T-1} \gamma^t \mathbb{E}[r(s_t, a_t)] + \gamma^T V_{\pi'}(s_T) \mid s_0 = s \right] \\ &\qquad \qquad \qquad \text{↑ } \text{if MDP is finite, or assume reward is bounded, then } V(s) \text{ is bounded} \\ &\qquad \qquad \qquad \text{as } T \rightarrow \infty \\ \text{so, } \lim_{T \rightarrow \infty} V_{\pi}(s) &\leq \lim_{T \rightarrow \infty} (\dots) = \mathbb{E}_{a \sim \pi'(s)} \left[\sum_{t=0}^{T-1} \gamma^t \mathbb{E}[r(s_t, a_t)] \mid s_0 = s \right] := V_{\pi'}(s) \quad \boxed{\text{□}} \quad \begin{aligned} &\text{* can show strict ineq. too,} \\ &\text{just correctly bookkeeping w/ } \mathbb{E}^{\pi'} \\ &\text{(needed since must be careful w/ limits!) } \\ &a_n < b_n \nRightarrow \lim a_n < \lim b_n \\ &\text{ex: } a_n = 0, b_n = \gamma_n \end{aligned} \end{aligned}$$

Thm 17.7 Bellman's Optimality Condition (^{# Bellman's Equation})

A policy π is optimal iff $\forall (s, a) \in S \times A$ with $\pi(s)(a) > 0$, the following holds:

$$a \in \arg\max_{a' \in A} Q_\pi(s, a') \quad \text{i.e. } \pi \text{ should recommend actions based on the value function } Q$$

Proof Recall optimality of π means $\forall s \in S, \forall \text{ policies } \pi', \text{ that } V_{\pi'}(s) \geq V_{\pi}(s)$

① optimal policy \Rightarrow eq'n holds (i.e., eq'n not true \Rightarrow not optimal)

Suppose $\exists s_0 \text{ s.t. } a_0 \notin \arg\max_a Q_\pi(s_0, a')$, let $\tilde{a} \in \arg\max_a Q_\pi(s_0, a')$

and $\pi(s_0) > 0$. Define $\pi'(s') = \begin{cases} \pi(s') & \text{if } s' \neq s_0 \\ \delta(\tilde{a}) & \text{if } s' = s_0, \end{cases}$ where $\delta(\cdot)$ is the probability distribution such that $P(a = \tilde{a}) = \begin{cases} 1 & a = \tilde{a} \\ 0 & \text{else} \end{cases}$

then

$$\mathbb{E}_{a \sim \pi'(s')} [Q_{\pi'}(s', a)] = \mathbb{E}_{a \sim \pi(s')} [Q_{\pi'}(s', a)] \quad \text{if } s' \neq s_0 \text{ since then } \pi'(s') = \pi(s)$$

$$\underbrace{\mathbb{E}_{a \sim \pi'(s')} [Q_{\pi'}(s', a)]}_{= \max_a Q_{\pi'}(s=s_0, a)} > \mathbb{E}_{a \sim \pi(s')} [Q_{\pi'}(s', a)] \quad \text{if } s' = s_0$$

since $\max_a f(a) \geq \mathbb{E}_a f(a)$ in general
and $\max_a f(a) > \mathbb{E}_a f(a)$ if "a" has nonzero support on submaximal elements

So.. apply the Policy Improvement Thm (Thm 17.6) we just proved...

$$\Rightarrow \exists s \text{ s.t. } V_{\pi'}(s) > V_{\pi}(s) \Leftrightarrow \pi \text{ isn't optimal.}$$

② eq'n holds \Rightarrow policy optimal (i.e. policy not-optimal \Rightarrow eq'n can't hold)

Let π' be a suboptimal policy $\Leftrightarrow \exists s, \pi' \text{ s.t. } V_{\pi'}(s) < V_{\pi}(s)$

so by contrapositive of Policy Improvement Thm, $\exists s \text{ s.t.}$

$$\mathbb{E}_{a \sim \pi(s)} [Q_{\pi'}(s, a)] < \mathbb{E}_{a \sim \pi(s)} [Q_{\pi}(s, a)]$$

if eq'n holds true,

$$= \max_a Q_{\pi'}(s, a) \dots \text{but } \max_a f(a) < \mathbb{E}_a f(a) \text{ is impossible. } \square$$

Thm 17.8 Existence of optimal policy: Any finite MDP admits an optimal deterministic policy

Proof Recall optimal means $V_{\pi^*}(s) \geq V_{\pi}(s) \forall s$

Let π^* be defined as the maximizer of $\sum_{s \in S} V_{\pi}(s)$ among all possible deterministic policies

Suppose π^* isn't optimal ...

then by previous Thm (17.7), $\exists s \text{ s.t. } \pi^*(s) \notin \arg\max_a Q_{\pi^*}(s, a')$

then we could improve π^* by redefining it, just for this s , to $\left. \begin{array}{l} \text{essentially Thm 17.6} \\ \text{be } \pi'(s) \in \arg\max_a Q_{\pi^*}(s, a') \quad (\text{and } \pi'(s') = \pi^*(s') \forall s' \neq s) \end{array} \right\}$

but then $V_{\pi^*}(s') = V_{\pi'}(s')$ if $s' \neq s$

$V_{\pi^*}(s') < V_{\pi'}(s')$ if $s' = s$ and π' is deterministic

$$\Rightarrow \sum_{s' \in S} V_{\pi^*}(s') < \sum_{s' \in S} V_{\pi'}(s') \text{ contradicting how we chose } \pi^* \quad \square$$

Therefore, for simplicity, we'll only consider deterministic policies
(so for finite MDP, this is WLOG)

π^* is a deterministic optimal policy, w/ corresponding policy value $V^*(s) \rightarrow = V_{\pi^*}(s)$

$$\text{State-action value function } Q^*(s, a) \rightarrow Q^*(s, a) = Q_{\pi^*}(s, a)$$

By Thm. 17.7, $(\forall s \in S) \quad \pi^*(s) \in \underset{a \in A}{\operatorname{argmax}} Q^*(s, a)$
i.e. only need to know $Q^*(s, a)$ (implicit, not explicit, dependence on $r(s, a)$ and $P(s'|s, a)$)

Now, $\underset{\text{on } \pi^*(s)}{\mathbb{E}[Q_{\pi^*}(s, a)]} = V_{\pi^*}(s)$. Since π^* is deterministic this means

$$Q_{\pi^*}(s, \pi^*(s)) = V_{\pi^*}(s).$$

$$= \underset{a \in A}{\operatorname{max}} Q^*(s, a), \text{ and recall } Q^*(s, a) = \mathbb{E}[r(s, a)] + \gamma \sum_{s' \in S} P(s'|s, a) \cdot V_{\pi^*}(s')$$

So...

(Eq. 17.4) Bellman Equations or Optimality Equations

π^* is an optimal deterministic policy iff

$$\textcircled{A} \quad (\forall s \in S) \quad V_{\pi^*}(s) := V^*(s) = \underset{a \in A}{\operatorname{max}} \left(\mathbb{E}[r(s, a)] + \gamma \sum_{s' \in S} P(s'|s, a) \cdot V^*(s') \right)$$

For a finite MDP, $|S| < \infty$, so V^* is a vector: $\begin{bmatrix} V^*(s_1) \\ \vdots \\ V^*(s_n) \end{bmatrix}$

This is almost like a linear eqn for V^* ,

except for the $\max_{a \in A}$

Now, if π is any policy (not necessarily optimal), V_π is not immediately obvious how to calculate,
though we'll assume deterministic

but $\underset{\text{on } \pi(s)}{\mathbb{E}[r(s, a)]}$ is a simple calculation,

and $P(s'|s, \pi(s))$ is also known (i.e., a known matrix) ← "modern" learning algorithms
don't assume this info is known

Turns out we can compute V_π easily:

Thm 17.9 (Sometimes also called Bellman Eq, but not optimality conditions)

(∞ -horizon discounted MDP w/ deterministic policy π) \rightarrow Math means $\mathbb{E}[r(s, \pi(s))]$

$$(\forall s \in S) \quad V_\pi(s) = \underset{a, \sim \pi(s)}{\mathbb{E}[r(s, a)]} + \gamma \sum_{s' \in S} P(s'|s, \pi(s)) \cdot V_\pi(s')$$

i.e.

$$V = R + \gamma P V \quad \begin{array}{l} \text{row } s, \text{ column } s' \text{ of } P \\ \text{is } P(s'|s, \pi(s)) \end{array} \quad \begin{array}{l} \text{Rows of } V \text{ is } V_\pi(s) \\ \text{Row } s \text{ of } R \text{ is } \mathbb{E}[r(s, \pi(s))] \end{array}$$

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

Proof:

$$\begin{aligned} V_\pi(s) &:= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, \pi(s_t)) \mid s_0 = s\right] \\ &= \mathbb{E}[r(s, \pi(s))] + \gamma \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_{t+1}, \pi(s_{t+1})) \mid s_0 = s\right] \\ &= \mathbb{E}[r(s, \pi(s))] + \gamma \cdot \sum_{s_t \in S} \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, \pi(s_{t+1})) \mid s_0 = s\right]}_{V_\pi(s_t)} \cdot P(s_t \mid s, \pi(s)) \end{aligned}$$

□

Thm 17.10 For a finite MDP, the eq'n $V = R + \gamma PV$ has a unique sol'n

proof:

Prove $(I - \gamma P)$ is nonsingular, hence $\underbrace{(I - \gamma P)V = R}_{\text{i.e.}} \text{ has unique sol'n}$

$$\text{Define } \|P\|_{\infty} = \|P\|_{\infty \rightarrow \infty} := \sup_{\|v\|_{\infty}=1} \|Pv\|_{\infty} = \max_{\text{row } s} \sum_{s'} |P_{ss'}|$$

operator norm on the Banach space $(\mathbb{R}^n, \|\cdot\|_{\infty})$

proved in Numerical Analysis or Functional Analysis (use Hölder's Ineq.)

Our matrix $P_{s,s'} = P(s'|s, \pi(s))$ is a **stochastic** matrix: entries ≥ 0 and rows sum to 1

i.e. $\sum_{s'} P(s'|s, \pi(s)) = 1$ since it's a (conditional) probability

So $\|P\|_{\infty} = 1$, and hence $\|\gamma P\|_{\infty} = \gamma < 1$

Then at least two ways to finish proof:

① $(I - \gamma P)$ is strictly diagonally dominant: $1 - \gamma P_{s,s} > \sum_{s' \neq s} |\gamma P_{s,s'}|$, so use **Gershgorin's disc theorem**

for \forall MDP if S is bdd.

② $\|\gamma P\|_{\infty} < 1$, $\{\mathbb{R}^n, \|\cdot\|_{\infty}\}$ is Banach $\Rightarrow (I - \gamma P)^{-1}$ a bdd. linear operator

"Neumann Series"

□ generalizes the 1D fact
 $\frac{1}{1-p}$ exists and equals its Taylor Series $\sum_{k=0}^{\infty} p^k$ $\forall |p| < 1$