Reinforcement Learning (Bellman Eqn)

Friday, March 27, 2020 3:05 PM

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

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

$$Q_\pi(s,a) = \mathbb{E}\, r(s,a) + \mathbb{E}_{a_t \sim \pi(s_t)}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right]$$

(If in case reward is stochastic)

$$= \mathbb{E}\left[r(s,a) + \gamma V_\pi(s_1) \mid s_0 = s, a_0 = a\right]$$

(dependence on s_0 is implicit: $s_1 \sim \mathbb{P}(s_1 \mid s_0, a_0)$)

$$= \mathbb{E}[r(s,a)] + \gamma \cdot \sum_{s_1 \in S} V(s_1) \cdot \mathbb{P}(s_1 \mid s, a)$$

(showing explicit dependence)

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

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

Thm 17.6: Policy-Improvement Thm

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

($\mathbb{E}_{a \sim \pi'(s)}$: mismatch)

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

Proof:

Assume (...), so

$$\mathbb{E}_{a \sim \pi(s)} Q_\pi(s,a) \leq \mathbb{E}_{a \sim \pi'(s)} Q_\pi(s,a) \quad \text{by our hypothesis}$$

($\mathbb{E}_{a \sim \pi(s)} Q_\pi(s,a) = V_\pi(s)$ by (*))

$$= \mathbb{E}_{a \sim \pi'(s)}\left[r(s,a) + \gamma V_\pi(s_1) \mid s_0 = s\right] \quad \text{by def'n of } Q$$

(also includes other randomness over reward, Markov Chain, etc.)

$$= \mathbb{E}_{a \sim \pi'(s)}\left[r(s,a) + \gamma \mathbb{E}_{a_1 \sim \pi(s_1)}\left[Q_\pi(s_1, a_1)\right] \mid s_0 = s\right] \quad \text{by (*)}$$

$$\leq \mathbb{E}_{a \sim \pi'(s)}\left[r(s,a) + \gamma \mathbb{E}_{a_1 \sim \pi'(s_1)}\left[Q_\pi(s_1, a_1)\right] \mid s_0 = s\right] \quad \text{by our hypothesis (again)}$$

$$= \mathbb{E}_{\substack{a \sim \pi'(s) \\ a_1 \sim \pi'(s_1)}}\left[r(s,a) + \gamma \cdot r(s_1, a_1) + \gamma^2 V_\pi(s_2) \mid s_0 = s\right] \quad \text{by def'n of } Q$$

(s_2 is a r.v., dependent on (a, s, a_1, s_1))

$$\cdots$$

$$\leq \mathbb{E}_{a_t \sim \pi'(s_t)}\left[\sum_{t=0}^{T} \gamma^t \mathbb{E}[r(s_t, a_t)] + \gamma^{T+1} V_\pi(s_{T+1}) \mid s_0 = s\right]$$

($\gamma^{T+1} \to 0$ as $T \to \infty$; if MDP is finite, or assume reward is bounded, then $V(s)$ is bounded)

So, $\lim_{T \to \infty} V_\pi(s) \leq \lim_{T \to \infty} (\ldots) = \mathbb{E}_{a_t \sim \pi'(s_t)}\left[\sum_{t=0}^{T} \gamma^t \mathbb{E}[r(s_t, a_t)] \mid s_0 = s\right] := V_{\pi'}(s)$

(and strict ineq. follows since no limit on LHS)

i.e., $a_n < b_n \not\Rightarrow \lim a_n < \lim b_n$

but $a < b_n \not\Rightarrow a < \lim b_n$

Thm 17.7 Bellman's Optimality Condition (≠ "Bellman's Equation")

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

$$a \in \underset{a' \in A}{\operatorname{argmax}}\ Q_\pi(s,a')$$

i.e. π should recommend actions based on the value function Q

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

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

Suppose $\exists\, s_0$ s.t. $a_0 \notin \underset{a'}{\operatorname{argmax}}\ Q_\pi(s_0,a')$, let $\tilde{a} \in \underset{a'}{\operatorname{argmax}}\ Q_\pi(s_0,a')$

and $\pi(s_0)(a_0) > 0$

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

then

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

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

$$= \max_a Q_\pi(s'=s_0,a) \qquad \text{since } \max_a f(a) \geq \mathop{\mathbb{E}}_a f(a) \text{ in general}$$

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

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

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

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

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

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

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

if eq'n holds true,

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

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

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

Let π^* be defined as the maximizer of $\sum_{s \in S} V_\pi(s)$ among all possible deterministic policies (idea: can choose $\pi(s)$ independently of $\pi(s')$; finite set, so max well-defined)

Suppose π^* isn't optimal...

then by previous Thm (17.7), $\exists\, s$ s.t. $\pi^*(s) \notin \underset{a'}{\operatorname{argmax}}\ Q_{\pi^*}(s,a')$ (our notation for deterministic policies)

then we could improve π^* by redefining it, just for this s, to be $\pi'(s) \in \underset{a'}{\operatorname{argmax}}\ Q_{\pi^*}(s,a')$ (and $\pi'(s') = \pi^*(s')\ \forall s' \neq s$) } essentially Thm 17.6

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

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

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 $V^*(s)$ ($= V_{\pi^*}(s)$) and state-action value function $Q^*(s,a)$. ($Q^*(s,a) = Q_{\pi^*}(s,a)$)

By Thm. 17.7, $(\forall s \in S)$ $\pi^*(s) \in \underset{a \in A}{\operatorname{argmax}}\, Q^*(s,a)$

i.e. only need to know $Q^*(s,a)$ (implicit, not explicit, dependence on reward $r(s,a)$ and transition prob. $P(s'|s,a)$)

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

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

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

So...

(Eq. 17.4) **Bellman Equations** or **Optimality Equations**

π^* is an optimal deterministic policy iff

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

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 eq'n for V^*, except for the $\max_{a \in A}$

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

but $\underset{a \sim \pi(s)}{\mathbb{E}} r(s,a)$ 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)

$$(\forall s \in S)\quad V_\pi(s) = \underset{a_1 \sim \pi(s)}{\mathbb{E}}[r(s,a_1)] + \gamma \sum_{s' \in S} P(s'|s,\pi(s)) \cdot V_\pi(s')$$

(Note: $\mathbb{E}[r(s,\pi(s))]$ means $\mathbb{E}[r(s,a_1)]$ where $a_1 \sim \pi(s)$)

i.e. $V = R + \gamma P V$ (P: matrix; V: vector (if finite MDP))

row s, column s' of P is $P(s'|s,\pi(s))$

Row s of V is $V_\pi(s)$

Row s of R is $\mathbb{E}\, r[s,\pi(s)]$

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

Proof:

$$V_\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, \pi(s_t)) \;\middle|\; s_0 = s \right]$$

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

(s_0 dependence only implicit)

$$= \mathbb{E}[r(s,\pi(s))] + \gamma \cdot \sum_{s_1 \in S} \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, \pi(s_{t+1}))\right]}_{V_\pi(s_1)} \cdot P(s_1 | s, \pi(s))$$

□

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 $\underbrace{(I-\gamma P)V = R}_{\text{ie } V = R + \gamma PV}$ has unique sol'n

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

- $\|P\|_{\infty \to \infty}$: operator norm on the Banach space $(\mathbb{R}^n, \|\cdot\|_\infty)$
- $\sup_{\|v\|_\infty = 1} \|Pv\|_\infty$: proved in Numerical Analysis or Functional Analysis (use Hölder's Ineq.)

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

i.e. $\sum_{s'} \mathbb{P}(s' \mid s, \pi(s)) = 1$ 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 Gerschgorin's disc theorem

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

- works for ∞ MDP if S is bdd.
- not the same norm
- this is induced operator norm, induced by
- "Neumann Series"
- generalizes the 1D fact $\frac{1}{1-p}$ exists and equals its Taylor Series $\sum_{k=0}^{\infty} p^k$ $\forall |p| < 1$

□