Class 10

Shikhar Saxena

February 10, 2023

Contents

Recap

- α-optimal policy
- V_α policy and associated Bellman Equation

Ross' Notation

- $B(\mathcal{S})$: Set of all bounded (real-valued) Functions on the state space
- $u \in B(\mathcal{S})$
- Map $T_f : B(\mathcal{S}) \to B(\mathcal{S})$ for policy vector f over the state space. Each $f(i)$ belongs to the action space.

$$(T_f u)(i) = r(i, f(i)) + \alpha \sum_{j=0}^{\infty} P_{ij}(f(i))u(j)$$

Intuition for the map is the policy evaluation equation:

$$V^\pi(s) = r(s, \pi(s)) + \alpha E_{s,\pi}\left[V^\pi(S')\right]$$

- **Terminal Reward**: Just the reward of terminating at the state (doesn't consider the immediate reward).

So, interpretation of $T_f u$ has the interpretation of following policy f for one step before terminating with terminal reward αu ($\alpha u(j)$ when final state is j).

Definition 1. *$T_f^n = T_f(T_f^{n-1})$ or $T_f(T_f(T_f(...)))$ (n times).*

Definition 2. *For any two functions $u, v, u_n \in B(\mathcal{S})$:*

1. *$u \leq v$ if $u(i) \leq v(i)$ for all i*
2. *$u = v$ if $u(i) = v(i)$ for all i*
3. *$u_n \to u$ if $u_n(i) \to u(i)$ for all i*

Lemma 1. *For $u, v \in B(\mathcal{S})$ and a stationary policy f*

1. *$u \leq v \implies T_f u \leq T_f v$*
2. *$T_f V^f = V^f$*
3. *$T_f^n u \to V^f, \ \forall u \in B(\mathcal{S})$*

Proof. 1. Easy to proof

2. Place revenue back in the conditioned expectation and it will be same as V_f
3. $T_f^n u$ is following f for n steps and obtaining a terminal reward of $\alpha^n u$ and then taking $n \to \infty$. Basically, $T_f^n u$ indiciates n-period value function and as $n \to \infty$, this will become the infinite horizon value function.

□

Policy Evaluation Algorithm

- For a policy f, keep applying T_f and you'll get V_f

Underlying MRP under f

Remark. *Markov Reward Process: No action*

The *Policy Evaluation* equation can be rewritten as

$$V^f = r^f + \alpha P^f V^f$$

Optimal stationary policy f_α

Let's say you have the optimal stationary policy f_α.

Theorem 1.

$$V^{f_\alpha}(i) = V_\alpha(i) \ \forall i \geq 0$$

and hence f_α is optimal

Proof. Apply T_{f_α} operator to V_α. This gives us $T_{f_\alpha} V_\alpha = V_\alpha$. Now repeatedly apply. This gives $\lim_{n\to\infty} T_{f_\alpha}^n = V_{f_\alpha} = V_\alpha$

□

Improving a policy f