EE6106 LECTURE 19 5th APRIL 2024

Markov Decision Processes

In bandits, we were considering a static environment. MDPs generalize this setting.

- State space S, assumed to be countable
- Action space A, assumed to be finite
- Say $X_t = s$ (state), $A_t = a$
- cost $C(s,a)$, may be stochastic or deterministic. In former case, use expected values
- $X_{t+1} = s'$ w.p. $P_{ss'}(a)$ $\left\{\sum_{s'} P_{ss'}(a) = 1\right\}$

$$P(X_{t+1} = s' \mid X_0 = x_0, A_0 = a_0, \ldots, X_t = s, A_t = a) = P_{ss'}(a)$$

- Policies are mappings from histories to actions
- An MAB is a trivial example of MDPs with only one state

Stationary policy : A_t is a deterministic function of S_t. $\pi : S \to A$

Two cost metrics :

- Discounted average cost $\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \alpha^t C(X_t, A_t)\right]$

  $\{\alpha \in (0,1)$ is a discount factor$\}$

  The summation may diverge when $\alpha = 1$

  We also need $|C(s,a)| \leq M \ \forall\ s,a$

- Average cost $\lim_{t\to\infty} \mathbb{E}_\pi\left[\frac{1}{t+1}\sum_{0}^{t} C(X_t, A_t)\right]$

Discounted Average Cost MDP

(S, A, P, C, α)

P : transition kernel

For any policy π,

$$V_\pi(i) = \mathbb{E}_\pi\left[\sum_0^\infty \alpha^t C(X_t, A_t) \,\middle|\, X_0 = i\right] \longrightarrow \text{Value function for policy } \pi$$

Define $V^*(i) = \inf_\pi V_\pi(i)$ $\longrightarrow$ Optimal value function

- This is a pointwise infimum. We first have to fix the starting state

Def: Policy π is optimal if $V_\pi(i) = V^*(i)$ $\forall i$

Note that there is no adversary in this setting, this is a generalization of the stochastic MAB setting

Theorem: Optimal value function V^* satisfies

$$V^*(i) = \min_a \left[C(i,a) + \alpha \sum P_{ij}(a) V^*(j)\right]$$

Bellman Equations

Note that this is just a necessity

Proof: (LB)

Consider π

$$V_\pi(i) = \sum_a P_{\pi,a}\left[C(i,a) + \alpha \sum P_{ij}(a) W_\pi(j)\right]$$

$P_{\pi,a}$: starting at $X_0 = i$, $\mathbb{P}$(action a is played by π)

Claim: $W_\pi(j) \geq V^*(j)$

Q: Is $W_\pi(j) = V_\pi(j)$?

A: No, unless we have a Markovian policy. In general we can have policies that are time-dependent

$$V_\pi(i) \geq \sum_a P_a\left[C(i,a) + \alpha \sum P_{ij}(a) V^*(j)\right]$$

$$\geq \min_a \left[C(i,a) + \alpha \sum P_{ij}(a) V^*(j)\right]$$

$$\Rightarrow \inf_\pi V_\pi(i) = V^*(i) \geq \text{———}$$

(UB)

Let a_0 be s.t.

$$a_0 = \operatorname{argmin}\left[C(i,a) + \alpha \sum P_{ij}(a) V^*(j)\right]$$

Fix $\epsilon > 0$. π: Play a_0 at time 0, then from state j at time 1, play π_j

s.t. $V_{\pi_j}(j) \leq V^*(j) + \epsilon$, for $\epsilon > 0$

because $V^*(j)$ is the infimum

$V_\pi(i) = C(i,a_0) + \alpha \sum P_{ij}(a_0) V_{\pi_1}(j)$

$\leq C(i,a) + \alpha \sum P_{ij}(a_0) V^*(j) + \alpha\epsilon$

~~$V_\pi(i) \leq$~~

$\therefore V^*(i) \leq V^\pi(i) \leq \min_a [C(i,a) + \alpha \sum P_{ij}(a) V^*(j)] + \alpha\epsilon$

$\therefore V^*(i) \leq \min_a [C(i,a) + \alpha \sum P_{ij}(a) V^*(j)] + \alpha\epsilon \; \forall \epsilon$

$\Rightarrow$ upper bound follows letting $\epsilon \downarrow 0$

$\therefore$ LB + UB $\Rightarrow V^*$ satisfies Bellman equation

$B(S)$: set of bounded value functions on S.

Def: For stationary policy f, $T_f : B(S) \longrightarrow B(S)$

$(T_f U)(i) = C(i, f(i)) + \alpha \sum P_{ij}(a) U(j)$, with $a = f(i)$

U is a mapping from $S \longrightarrow \mathbb{R}$

$(T_f U)$ is also a mapping from $S \longrightarrow \mathbb{R}$

Lemma For $u, v \in B(S)$, stationary policy f.

(i) $u \leq v \Longrightarrow T_f u \leq T_f v$

(ii) $T_f v_f = v_f$

(iii) $T_f^n u \longrightarrow v_f \;\; \forall \; u$

EXERCISE

Note (I) (i) is an elementwise inequality. The intuition is that u, v are like terminal costs

(II) v_f is the value function associated with the policy f.
Hence, v_f is a fixed point of T_f.

(III) Every element of $B(S)$ can be thought out of as a vector
If we take any vector $u \in B(S)$ and keep on applying T_f to it, there will be convergence towards v_f, which is the fixed point of T_f.

$\star \; U_n \longrightarrow U \iff U_n(i) \longrightarrow U(i)$ uniformly in i

$(T_f^2 U)(i) = C(i, f(i)) + \alpha \sum P_{ij}(f(i)) (T_f(U))(j)$

$= C(i, f(i)) + \alpha \sum P_{ij}(f(i)) \left[C(i, f(i)) + \alpha \sum P_{jk} f(j) U(k) \right]$

$= C(i, f(i)) + \alpha \sum P_{ij} f(i) \, C(j, f(j)) + \alpha^2 \text{----}$

expected cost at the second time-step

Thm: Let π^* be stationary policy that attains the $\min_a [C(i,a) + \alpha \sum P_{ij}(a) V^*(j)]$

Then $V_{\pi^*} = V^*$ (i.e. π^* is optimal)

The optimality is over the space of all policies

Proof sketch Apply T_{π^*} to V^*

Claim: $T_{\pi^*} V^* = V^*$

$$(T_{\pi^*})^n V^* = V^*$$

$$\downarrow n \uparrow \infty$$

$$V_{\pi^*}$$