Lecture 15*

Infinite horizon discounted MDPs

(Ref: DPOC vol. II, Chapter 1&2)

Goal:

$$J^*(i) = \min_{\pi \in \Pi} J_\pi(i), \quad \forall i, \quad \text{where}$$

optimal discounted cost

$$J_\pi(i) = E\left(\sum_{k=0}^{\infty} \alpha^k g(x_k, \pi(x_k), x_{k+1}) \,\Big|\, x_0 = i \right)$$

Single-stage cost

expected cumulative discounted cost a.k.a value function

discount

policy

start state

$0<\alpha<1$ → useful in financial applications as future costs/rewards are discounted.

Let π^* denote the optimal policy i.e., $\arg\min_{\pi \in \Pi} J_\pi(i)$

(A1) The single stage cost $|g(x,a,x')| \le M < \infty \quad \forall x, x' \in \mathcal{X}, a \in \mathcal{A}$

$$(A1) + (0<\alpha<1) \Rightarrow \quad |J_\pi(i)| \le M \sum_{k=0}^{\infty} \alpha^k = \frac{M}{1-\alpha} < \infty$$

"Unlike SSPs, we do not require the existence of a terminal state".

Bellman and another operator

For $J = (J(1), \ldots, J(n))$,

define Bellman operator T as follows:

$$(TJ)(i) = \min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\left(g(i,a,j) + \alpha J(j)\right), \quad \forall i$$

for a stationary policy π,

$$\left(T_\pi J\right)(i) = \sum_{j=1}^{n} p_{ij}(\pi(i))\left(g(i,\pi(i),j) + \alpha J(j)\right), \quad \forall i$$

$$P_\pi = \begin{bmatrix} p_{11}(\pi(1)) & & & \\ & \ddots & & \\ & & \ddots & \\ & & & p_{nn}(\pi(n)) \end{bmatrix}$$

assuming $S = \{1, \dots n\}$

$$g_\pi = \begin{pmatrix} \sum_{j=1}^{n} p_{1j}(\pi(1))\, g(1,\pi(1),j) \\ \vdots \\ \sum_{j=1}^{n} p_{1j}(\pi(n))\, g(n,\pi(n),j) \end{pmatrix}$$

With this notation,

$$T_\pi J = g_\pi + \alpha P_\pi J$$

Remark: If single stage cost is a function of current state & action, ie., $g(i,a)$, then

$$g_\pi = \begin{pmatrix} g(1,\pi(1)) \\ \vdots \\ g(n,\pi(n)) \end{pmatrix}$$

Even in the discounted case, T, T_π are monotone.

Lemma 1: Let $J, J' \in \mathbb{R}^n$ & satisfy
$J(i) \leq J'(i), \forall i$

Then, for any $k = 1, 2, \ldots$

(i) $(T^k J)(i) \leq (T^k J')(i)$, and

(ii) For any stationary policy π,
$$(T_\pi^k J)(i) \leq (T_\pi^k J')(i)$$

Pf: H.W.

The constant shift lemma holds here as well.

Lemma 2:
Stationary π, $\delta \to$ positive scalar, $e \to$ vector of n ones.
Then, $\forall i = 1 \ldots n$, $\forall k = 1, 2, \ldots$, we have

(i) $(T^k(J + \delta e))(i) = (T^k J)(i) + \alpha^k \delta$

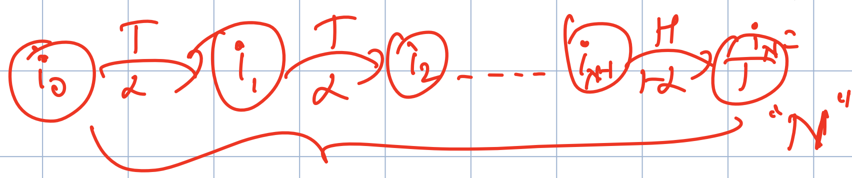
(ii) $(T_\pi^k(J + \delta e))(i) = (T_\pi^k J)(i) + \alpha^k \delta$

Pf: H.W.

"Every discounted problem has an equivalent SSP".

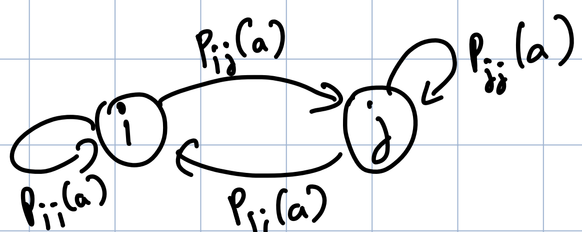
Given discounted MDP on states $\{1, \ldots, n\}$,

form an SSP on states $\{1, \ldots n\} \cup \{T\}$

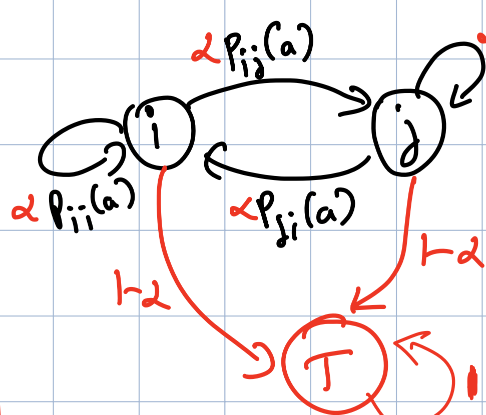


add this extra state & make it cost-free & absorbing.

Idea: In the SSP, w.p. α pick a next state according to transition probabilities of discounted MDP & w.p. $(1-\alpha)$ move to "T" & incurs no cost



Original discounted MDP



"Equivalent SSP"

$$\boxed{\tilde{g}(i,a,j) = \begin{cases} \dfrac{g(i,a,j)}{\alpha} & \text{if } j \neq T \\ 0 & \text{else} \end{cases}}$$

This scaling ain't necessary if cost is of the form $g(i,a)$, i.e., indep. of next state.

Why are these two MDPs equivalent?

1. Note that in the SSP, all policies are proper.

2. In the discounted MDP, the expected kth stage cost is $E\left(\alpha^k g(i,a,j)\right) = \alpha^k \sum_j p_{ij}(a)\, g(i,a,j)$

③ In the SSP, the expected kth stage cost is

$$\left[\sum_j p_{ij}(a)\, g(i,a,j)\right] \times \alpha^k$$

If the terminal state is not hit upto kth stage, then the underlying probabilites will have a α^k multiplier.

"Optimal value is the same for the discounted MDP & the equivalent SSP"

$\tilde{S} \to$ state space in SSP $= \{1 \ldots n\} \cup \{T\}$

Let $\tilde{J}^* \to$ optimal value in SSP, Let $\tilde{P}_{ij}(a)$ denote transition prob. here

$J^* \to$ optimal value in discounted MDP

Bellman equation in SSP:

$$\tilde{J}^*(i) = \min_a \sum_{j \in \tilde{S}} \tilde{P}_{ij}(a)\left(\tilde{g}(i,a,j) + \tilde{J}^*(j)\right)$$

$$= \min_a \sum_{j \in \tilde{S}} \tilde{P}_{ij}(a)\, \tilde{g}(i,a,j) + \sum_{\underbrace{j \in S}_{S=\{1\ldots n\}}} \tilde{P}_{ij}(a)\, \tilde{J}^*(j)$$

$$+ \tilde{P}_{iT}(a)\, \tilde{J}^*(T)$$

$$= \min_a \sum_{j \in \tilde{S}} \tilde{P}_{ij}(a)\, \tilde{g}(i,a,j) + \sum_{j \in S} \alpha P_{ij}(a)\, \tilde{J}^*(j)$$

$$+ (1-\alpha)\, \tilde{J}^* \cancelto{0}{(T)}$$

$$\tilde{J}^*(i) = \min_a \sum_{j \in S} \alpha P_{ij}(a) \frac{g(i,a,j)}{\alpha} + (1-\alpha)\, \cancelto{0}{g(i,a,T)}$$

$$+ \alpha \sum_{j \in S} P_{ij}(a)\, \tilde{J}^*(j) \quad \text{—— } (*)$$

Bellman equation in discounted MDP

$J^* = TJ^*$ for discounted MDP

$$J^*(i) = \min_a \sum_j P_{ij}(a)\left(g(i,a,j) + \alpha J^*(j)\right)$$

So, from (*) $\tilde{J}^*(i) = J^*(i)$ (assuming $J = TJ$ has a unique fixed point → will be shown next)

⇒ optimal values coincide.

Lecture-16*

Prop 1: (VI converges)

Assume (A1).

For any finite J, the optimal cost satisfies

$$J^*(i) = \lim_{N\to\infty} (T^N J)(i), \quad \forall i$$

Corollary: For a stationary policy π, we have

$$J_\pi(i) = \lim_{N\to\infty} (T_\pi^N J)(i), \quad \forall i \text{ for any finite } J.$$

Pf) Given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ and a state $i \in X$,

$$J_\pi(i) = \lim_{N\to\infty} E\left(\sum_{l=0}^{N-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

For convenience, drop the dependence on i on RHS.

$$= E\left(\sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

$$+ \underbrace{\lim_{N\to\infty} E\left(\sum_{l=L}^{N-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)\Big)}_{\text{term } B} \quad \text{— (1)}$$

Since $|g(\cdot,\cdot,\cdot)| \leq M$ by assumption,

$$|\text{term } B| \leq M \sum_{l=L}^{\infty} \alpha^l = \frac{M\alpha^L}{1-\alpha} \quad \text{— (2)}$$

$$\Downarrow$$

$$-\frac{M\alpha^L}{1-\alpha} \leq \text{term}B \leq \frac{M\alpha^L}{1-\alpha}$$

$$E\left(\sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

$$= J_\pi(i) - \text{term } B$$

$$A = J_\pi + C$$
$$|C| \leq M$$
$$J_\pi - M \leq A \leq J_\pi + M$$

$$E\left(\alpha^L J(x_L) + \sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

Adding this to both sides

$$= J_\pi(i) - \text{term } B + E\left(\alpha^L J(x_L)\right)$$

Using (2),

$$J_\pi(i) - \frac{M\alpha^L}{1-\alpha} - \alpha^L \max_{j\in X} |J(j)|$$

applying T_π L times on J

i.e., $\to = T_\pi^L J$

$$\leq E\left(\alpha^L J(x_L) + \sum_{l=0}^{L-1} \alpha^l g(x_l, \mu_l(x_l), x_{l+1})\right)$$

$$\leq J_\pi(i) + \frac{M\alpha^L}{1-\alpha} + \alpha^L \max_{j\in X} |J(j)| \quad \text{— (3)}$$

Taking minimum over π on all sides of ③, we obtain $\forall i \in X$ and any $L > 0$ that

$$J^*(i) - \frac{M\alpha^L}{1-\alpha} - \alpha^L \max_{j \in X} |J(j)|$$

$$\leq (T^L J)(i) \quad \text{because } \min_{\pi} T_\pi^L = T^L$$

$$\leq J^*(i) + \frac{M\alpha^L}{1-\alpha} + \alpha^L \max_{j \in X} |J(j)| \quad \text{— ④}$$

Taking $L \to \infty$ on all sides of ④, we obtain (Note: $\alpha \in (0,1)$)

$$J^*(i) \leq \lim_{L \to \infty} (T^L J)(i) \leq J^*(i)$$

& the claim follows. ■

Corollary: The claim follows by considering an MDP where the only feasible action in a state i is $\pi(i)$, $\forall i$, and invoking Prop 1. (Note $T = T_\pi$ for this MDP).

Prop 2: (Bellman equation)

The optimal discounted cost J^* satisfies

$$J^* = T J^* \quad , \text{ie.,}$$

Bellman equation

$$\boxed{J^*(i) = \min_a \sum_j P_{ij}(a) \left(g(i,a,j) + \alpha J^*(j)\right)}$$

Also, J^* is the unique fixed point of T.

Pf) Eq ④ in the proof above is

$$J^*(i) - \frac{M\alpha^L}{1-\alpha} - \alpha^L \max_{j \in X} |J(j)|$$

$$\leq (T^L J)(i)$$

$$\leq J^*(i) + \frac{M\alpha^L}{1-\alpha} + \alpha^L \max_{j \in X} |J(j)| \quad \text{--- ⑤}$$

Applying operator T on all sides,

$$TJ^*(i) - \frac{M\alpha^{L+1}}{1-\alpha} - \alpha^{L+1} \max_{j \in X} |J(j)| \leq (T^{L+1} J)(i)$$

$$\leq TJ^*(i) + \frac{M\alpha^{L+1}}{1-\alpha} + \alpha^{L+1} \max_{j \in X} |J(j)|$$

we used constant shift lemma.

Taking $L \to \infty$ on all sides of the equation above,

$$T J^*(i) = J^*(i) \quad , \forall i. \Rightarrow$$ J^* is a fixed point of T

Uniqueness: Let J' be another fixed point of T.

$$J' = T J' = T^2 J' = \cdots\cdots = \lim_{N \to \infty} T^N J' = J^*$$

$\therefore$ T has a unique fixed point ∎

Corollary: For a stationary policy π, the associated cost J_π satisfies

$$J_\pi = T_\pi J_\pi \quad \text{(or)}$$

$$J_\pi(i) = \sum_j p_{ij}(\pi(i)) \left(g(i, \pi(i), j) + \alpha J_\pi(j) \right), \quad \forall i$$

Also, J_π is the unique fixed point of T_π.

Pf) Follows from Prop 2. ∎

Necessary & sufficient condition for optimal policy:

<u>Prop 3:</u> A stationary policy π is optimal if and only if $\pi(i)$ attains the minimum in the Bellman equation, $\forall i \in \mathcal{X}$. Or, equivalently,

$$TJ^* = T_\pi J^*$$

<u>Pf></u> Assume $TJ^* = T_\pi J^*$ — ①

We know $J^* = TJ^*$ — ②

$$J^* = T_\pi J^*$$

$$\Rightarrow \quad J^* = J_\pi \Rightarrow \pi \text{ is optimal}$$

Converse: π is optimal

$$\Rightarrow \quad J^* = J_\pi \quad \Rightarrow \quad J^* = T_\pi J^* \quad \text{— ①}'$$

From BE, $J^* = TJ^*$ — ②'

$$①' + ②' \Rightarrow \quad T_\pi J^* = TJ^*$$

$\blacksquare$

Contraction property of T and T_π:

Max-norm: $\|J\|_\infty = \max_{i=1\sim n} |J(i)|$

We will show that T, T_π are α-contractions in $\|\cdot\|_\infty$.

$\Rightarrow$ T is a contraction in $\|\cdot\|_\infty$-norm with modulus α

Prop 4: For any two bounded J, J', and $\forall k \geq 1$

$$\|T^k J - T^k J'\|_\infty \leq \alpha^k \|J - J'\|_\infty \quad — (*)$$

$\hookrightarrow$ modulus of contraction

Pf: Let $c = \max_{i=1 \sim n} |J(i) - J'(i)|$

$$J(i) - c \leq J'(i) \leq J(i) + c \quad — ①$$

Apply T "k" times on all sides of ① to get

holds $\forall i$ $\rightarrow$ $(T^k J)(i) - \alpha^k c \leq (T^k J')(i) \leq (T^k J)(i) + \alpha^k c$

$$\Rightarrow \left|(T^k J)(i) - (T^k J')(i)\right| \leq \alpha^k c, \ \forall i$$

$$\max_{i=1 \sim n} \left|(T^k J)(i) - (T^k J')(i)\right| \leq \alpha^k c$$

$$\Rightarrow \|T^k J - T^k J'\|_\infty \leq \alpha^k \|J - J'\|_\infty$$

∎

Corollary: For any stationary π & bounded J, J', and $\forall k \geq 1$

$$\|T_\pi^k J - T_\pi^k J'\|_\infty \leq \alpha^k \|J - J'\|_\infty$$

Value Iteration: Start with J_0 & repeatedly apply T.

Error-bound for VI:

$$\|T^k J_0 - J^*\|_\infty \leq \alpha^k \|J_0 - J^*\|_\infty$$

Why? Set $J' = J^*$ in $(*)$ & note $T^k J^* = J^*$.

Example: Machine replacement

Recall n-states $1 \ldots n$

Operating cost $g(i)$

$g(1) \le g(2) \ldots\ldots\ldots \le g(n)$

$P_{ij} \to$ transition probabilites (do nothing action)

actions: do nothing & repair (Repair cost R)

Goal: minimize infinite horizon discounted cost
($\alpha \leftarrow$ discount factor)

Bellman equation: $J^*(i) = T J^*(i) = \min_a E\big(\overset{\text{single stage cost}}{g(i,a,j)} + \alpha J^*(j)\big)$

$$J^*(i) = \min \Big\{ \underbrace{R + g(1) + \alpha J^*(1)}_{\text{repair}},$$

$$\underbrace{g(i) + \alpha \sum_{j=1}^{n} P_{ij} J^*(j)}_{\text{do nothing}} \Big\}$$

Optimal action: repair if

$$R + g(1) + \alpha J^*(1) < g(i) + \alpha \sum_{j=1}^{n} P_{ij} J^*(j)$$

& "do nothing" otherwise

Assume . (B1) $P_{ij} = 0$ if $j < i$ $\leftarrow$ machine won't get better if we don't repair

(B2) $P_{ij} \le P_{(i+1)j}$ if $i < j$ $\leftarrow$ e.g. $j = 10$ $P_{5,10} \le P_{8,10}$

Suppose J is monotone non-decreasing, i.e.,

$J(1) \le J(2) \dots \le J(n)$

Then,

Because of (B2) & monotone J $\rightarrow$ $\boxed{\sum_{j=1}^{n} P_{ij} J(j) \le \sum_{j=1}^{n} P_{(i+1)j} J(j)}$, $i = 1 \dots n-1$

Since $g(i)$ is non-decreasing, we have

$(TJ)(i)$ is non-decreasing in i, if J is non-decreasing. since $TJ(i) = \min\left(R + g(1) + \alpha J^*(1),\ g(i) + \alpha \sum P_{ij} J^*(j)\right)$

$\Rightarrow$ $(T^k J)(i)$ is non-decreasing in i, $\forall k$

$\Rightarrow$ $\lim_{k \to \infty} T^k J(i) = J^*(i)$ is non-decreasing in i

So, the function " $g(i) + \alpha \sum_{j \ge i} P_{ij} J^*(j)$ "

is non-decreasing in i

Set of states $S_R = \left\{ i \;\middle|\; R + g(1) + \alpha J^*(1) \le g(i) + \alpha \sum_j P_{ij} J^*(j) \right\}$

$$\boxed{\sum_{j=1}^{n} P_{ij} J(j) \le \sum_{j=1}^{n} P_{(i+1)j} J(j)},$$

$n=3, \quad i=1, \quad i+1=2 \qquad J(j) \ge 1.$

$$P_{11} + P_{12} + P_{13} = 1 \qquad P_{ij} = 0 \text{ if } j < i$$

$$\le \cancel{P_{21}}^{\,0} + P_{22} + P_{23} = 1$$

$J_1 \le J_2 \le J_3$

$P_{11} + P_{12} + P_{13} = 1$

$P_{22} + P_{23} = 1$

$$P_{11} J_1 + P_{12} J_2 + P_{13} J_3$$

$$\le P_{22} J_2 + P_{23} J_3$$

$$P_{11} J_1 + P_{12} J_2 + P_{13} J_3$$

$$\le P_{11} J_2 + P_{12} J_2 + P_{13} J_3$$

$$P_{11} J_1 \le (P_{22} - P_{12}) J_2 + (P_{23} - P_{13}) J_3$$

$$(P_{22} - P_{12}) J_1 + (P_{23} - P_{13}) J_1 \le \underbrace{(P_{22} - P_{12}) J_2 + (P_{23} - P_{13}) J_3}$$

$$P_{11} J_1 \le \text{RHS}$$

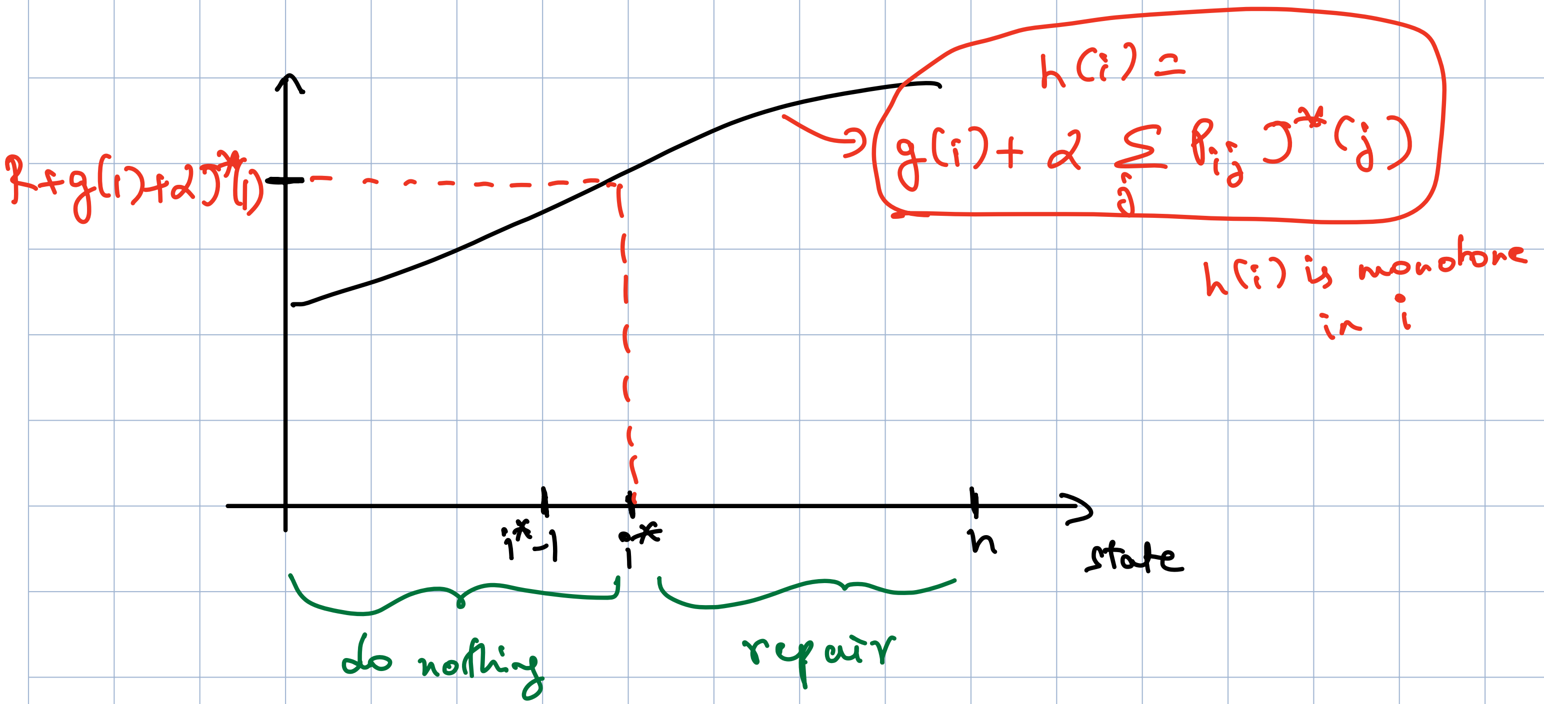
S_R = set of states where it is optimal to repair

$$i^* = \begin{cases} \text{smallest state in } S_R & \text{if } S_R \neq \phi \\ n+1 & \text{else} \end{cases}$$

Optimal policy

$$= \begin{cases} \text{repair} & \text{if } i \geq i^* \\ \text{do nothing} & \text{else} \end{cases}$$

A threshold-based optimal policy



H.W. Think about policy iteration for this problem. In particular, if we start with a threshold-based policy & do policy improvement, then does it lead to another threshold policy?

If yes, then PI converges to optimal policy in at most n iterations.

Illustrative example for VI :

MDP $\mathcal{X} = \{1, 2\}$ $\mathcal{A} = \{a, b\}$

$$P(a) = \begin{bmatrix} P_{11}(a) & P_{12}(a) \\ P_{21}(a) & P_{22}(a) \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix}$$

$$P(b) = \begin{bmatrix} P_{11}(b) & P_{12}(b) \\ P_{21}(b) & P_{22}(b) \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Costs: $g(1,a) = 2$, $g(1,b) = 0.5$
$g(2,a) = 1$, $g(2,b) = 3$

discount $\alpha = 0.9$

$J_0 = (0, 0)$

$$(TJ)(i) = \min\left\{ g(i,a) + \alpha \sum_{j=1}^{2} P_{ij}(a) J(j),\ g(i,b) + \alpha \sum_{j=1}^{2} P_{ij}(b) J(j) \right\}$$

$J_1 = T J_0 = (0.5, 1)$

$J_2 = (1.28, 1.56)$

& so on.

PI algorithm:

Step 1: Start with a policy π_0

Step 2: (Policy Evaluation) Evaluate π_k, i.e., compute J_π by solving $J = T_{\pi_k} J$

$$\Leftrightarrow \quad J(i) = \sum_j p_{ij}(\pi_k(i))\left(g(i,\pi_k(i),j) + \alpha J(j)\right), \quad \forall i \quad (*)$$

(here $J(1), \dots J(n)$ are the unknowns & solving $(*)$ gives J_{π_k})

Step 3: Policy improvement

Find a new policy π_{k+1} by

$$T_{\pi_{k+1}} J_{\pi_k} = T J_{\pi_k}$$

$$\Leftrightarrow \quad \pi_{k+1}(i) = \arg\min_{a \in A(i)} \sum_j p_{ij}(a)\left(g(i,a,j) + \alpha J_{\pi_k}(j)\right)$$

If $J_{\pi_{k+1}}(i) < J_{\pi_k}(i)$ for at least one state i, then go to step 2 & repeat.

Remark: Policy improvement claim holds even in the discounted setting.

Policy improvement claim:

Let π, π' be two policies s.t.

$$T_{\pi'} J_\pi = T J_\pi$$

Then, $J_{\pi'}(i) \le J_\pi(i) \quad \forall i$

with strict inequality for at least one of the states if π is not optimal.

Pf: Follows by a parallel argument to the proof in SSP case.

Lecture - 17

PI example:

$S = \{1, 2\}$, $A = \{a, b\}$,

$$P(a) = \begin{bmatrix} P_{11}(a) & P_{12}(a) \\ P_{21}(a) & P_{22}(a) \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix}$$

$$P(b) = \begin{bmatrix} P_{11}(b) & P_{12}(b) \\ P_{21}(b) & P_{22}(b) \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}$$

Costs: $g(1,a) = 2$, $g(1,b) = 0.5$
$g(2,a) = 1$, $g(2,b) = 3$

discount $\alpha = 0.9$

Initialization! $\pi_0(1) = a$, $\pi_0(2) = b$

Policy evaluation:

Finding J_{π_0} :

Two equations in two unknowns $J_{\pi_0}(1)$ & $J_{\pi_0}(2)$

$$\begin{cases} J_{\pi_0}(1) = g(1,a) + \alpha P_{11}(a) J_{\pi_0}(1) + \alpha P_{12}(a) J_{\pi_0}(2) \\ J_{\pi_0}(2) = \underbrace{g(2,b) + \alpha P_{21}(b) J_{\pi_0}(1) + \alpha P_{22}(b) J_{\pi_0}(2)} \end{cases}$$

Using MDP data, we have $\qquad J_{\pi_0} = T_{\pi_0} J_{\pi_0}$

$$J_{\pi_0}(1) = 2 + 0.9 \times \frac{3}{4} \times J_{\pi_0}(1) + 0.9 \times \frac{1}{4} \times J_{\pi_0}(2)$$

$$J_{\pi_0}(2) = 3 + 0.9 \times \frac{1}{4} \times J_{\pi_0}(1) + 0.9 \times \frac{3}{4} \times J_{\pi_0}(2)$$

Solving, $\quad J_{\pi_0}(1) = 24.12$, $\quad J_{\pi_0}(2) = 25.96$

Policy improvement:

$$T_{\pi_1} J_{\pi_0} = T J_{\pi_0}$$

$$(T J_{\pi_0})(1) = \min \Big\{ \underbrace{2 + 0.9\left(\frac{3}{4} \times 24.12 + \frac{1}{4} \times 25.96\right)}_{\text{action } a},$$

$$\underbrace{0.5 + 0.9\left(\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96\right)}_{\text{action } b} \Big\}$$

$\boxed{\pi_1(1) = b}$

$$= \min\{24.12, 23.45\} = 23.45 \leftarrow \text{for action } b$$

$$(TJ_{\pi_0})(2) = \min\Big\{ \underbrace{1 + 0.9\left(\tfrac{3}{4}\times 24.12 + \tfrac{1}{4}\times 25.96\right)}_{\text{action } a},$$

$$\underbrace{3 + 0.9\left(\tfrac{1}{4}\times 24.12 + \tfrac{3}{4}\times 25.96\right)}_{\text{action } b}\Big\}$$

$\pi_1(2) = a$

$$= \min\{23.12,\ 25.95\} = 23.12 \quad \leftarrow \text{for action } a$$

$\pi_1(1) = b, \quad \pi_1(2) = a$

Policy evaluation: J_{π_1} ?

$$\boxed{J_{\pi_1}(1) = 7.33, \quad J_{\pi_1}(2) = 7.67}$$

Optimal cost J^*

Policy improvement: $T_{\pi_2} J_{\pi_1} = T J_{\pi_1}$

$$\boxed{\pi_2(1) = b, \quad \pi_2(2) = a}$$

So, stop & output $\pi_2 \rightarrow$ optimal policy π^*

Linear programming

Want to solve $J^* = T J^*$

Idea: is to form a linear optimization problem whose solution is J^*

How? We know $\lim_{N \to \infty} T^N J = J^*$ for any J.

Suppose $J \leq TJ$

Then, $J \leq T^2 J$

$\vdots$

$J \leq T^k J$

$\Rightarrow$ $J \leq J^* = TJ^*$ i.e., $J(1) \leq J^*(1)$ $\vdots$ $J(n) \leq J^*(n)$

J^* is the largest solution that satisfies $J \leq TJ$

$x \leq \min_{y \in y_1 \ldots y_k} f(y)$

$x \leq f(y_1)$

$x \leq f(y_2)$

$\vdots$

$x \leq f(y_k)$

Optimization problem:

constraints: $J \leq TJ$

objective: $\max J$

More precisely, $J \leq TJ$

$\Updownarrow$

$$\forall i, \quad J(i) \leq g(i,a) + \alpha \sum_{j=1}^{n} P_{ij}(a) J(j)$$

$\Rightarrow$ for simplicity, assume single stage cost doesn't depend on next state.

So, the LP formulation is:

Variables: $\lambda_1, \ldots, \lambda_n$

Objective: $\max \sum_i \lambda_i$

linear fn of $\{\lambda_1, \ldots, \lambda_n\}$

linear functions

subject to

$J \le TJ$ constraint

$$\lambda_i \le g(i,a) + \alpha \sum_j P_{ij}(a)\lambda_j, \quad \text{for } i = 1, \ldots, n, \text{ and } a \in A(i)$$ (*)

Remark! Assuming $A(i) = A \ \forall i$ & $|A| = q$, we have $n \times q$ constraints in the LP (*)

variables $= n$ ← cardinality of the state space

On problems with a large state space, LP is not practical.

Remark! Can use LP approach for solving SSPs as well.

H.W.: Write down the LP for the 2-state 2-action example used for VI/PI above.