Infinite-horizon MDPs Lecture 5* (contd)

MDP components: state space X, action space A, transition probabilities $p_{ij}(a)$

$$J_\pi(x_0) = \lim_{N\to\infty} E\left(\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), x_{k+1}) \right)$$

- x_0: initial state
- π: policy $\pi = \{\mu_0, \mu_1, \mu_2, \dots\}$
- g: single-stage cost "stationary"
- α: discount factor $\alpha \in (0, 1]$

Goal: $J^*(x_0) = \min_{\pi \in \Pi} J_\pi(x_0)$

- J^*: optimal expected cost
- Π: set of admissible policies

Let $\pi^* = \arg\min_{\pi \in \Pi} J_\pi(x_0)$

- π^*: optimal policy

Two popular MDPs:

(I) Stochastic shortest path (SSP)

(i) "Finite state space" $\{1, \dots n\}$ & finite action space

(ii) $\alpha = 1$

(iii) $\exists$ a state "T" s.t. $p_{TT}(a) = 1$ "absorbing"

$g(T, a, T) = 0 \ \forall a$ "cost-free"

(II) Discounted MDP

(i) discount $\alpha < 1$

(ii) $|g(\cdot,\cdot,\cdot)| < M < \infty$

(i) & (ii) $\Rightarrow$ $J_\pi(x_0)$ is finite.

(III) Average-cost MDP : skipped (see Chapter 5 of Vol II, DPOC book)

Main results:

(I) Taking finite horizon to the limit

alert: this is a change of notation from finite horizon chapter

Let $J_N^*(i)$ be the optimal expected cost of a "N-stage" finite horizon MDP, with initial state i & stationary cost $g(i,a,j)$

Then, the infinite horizon optimal expected cost can be obtained by

$$J^*(i) = \lim_{N \to \infty} J_N^*(i)$$

(II) Bellman equation

Assume $X = \{1, \dots, n\}$, transition prob. $p_{ij}(a)$

For a N-stage problem, with J_N^* denoting the optimal cost, the DP algorithm is

$$① \quad J_{k+1}^*(i) = \min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\left(g(i,a,j) + \alpha J_k^*(j)\right)$$

J_{k+1}^*: $k+1$ stage optimal cost

J_k^*: k-stage optimal cost

In the infinite horizon case, the optimal cost J^* satisfies

$$② \quad J^*(i) = \min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\left(g(i,a,j) + \alpha J^*(j)\right)$$

↳ Bellman equation

Eq ① is an algorithm

Eq ② is a system of equations satisfied by J^*

Lecture-6

③ How to get the optimal policy π^*?

Hi, $$J^*(i) = \min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\left(g(i,a,j) + \alpha J^*(j)\right)$$ ← Bellman equation

Make up a policy as follows: For state i, let $a^*(i)$ be the minimizer in the Bellman equation.

Set $\pi(i) = a^*(i)$, $\forall i \in S$.

Then, "$\pi = \pi^*$" i.e., π is a optimal policy.

Stochastic shortest path (SSP) problems

State space = $\{1, \ldots n\}$

In SSP, $\exists$ a special terminal state, say T, that satisfies

$$\underbrace{p_{TT}(a) = 1}_{\text{absorbing}}, \quad \underbrace{g(T, a, T) = 0}_{\text{cost-free}} \quad \forall a$$

Examples:

① A simple example of SSP: Deterministic shortest path

② A finite horizon problem can be easily cast as an SSP.

In the SSP for a finite horizon problem with horizon N:

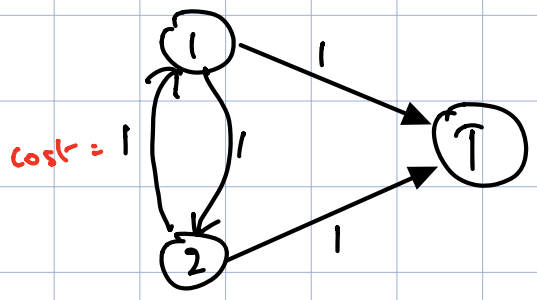
States are tuples: (i, k) (state ↑, stage ↑)

Terminal states: (i, N) $\forall i$

Transitions: $(i,k) \xrightarrow{a} (j, k+1)$ w.p. $p_{ij}(a)$

Costs: H.W.

Proper policies:



Shortest path problem

Improper policy: loop between 1 & 2

Expected cost = ∞.

Proper policy: Goto T from 1 & also 2.

Stationary policy: $\pi: \mathcal{S} \to \mathcal{A}$ which takes the same action, say a, in a state i, irrespective of the stage k in the infinite horizon $\pi = (\mu, \mu, \ldots)$. So, we identify the stationary policy with a mapping from $\mathcal{S}$ to $\mathcal{A}$.

Def: A stationary policy π is **proper**

$$\rho_\pi = \max_{i=1,\ldots n} P\left(i_n \neq T \,\middle|\, i_0 = i, \pi \right) < 1$$

i_n: n^{th} stage; i_0: starting state; actions chosen using π

A policy that ain't proper is "improper" (i.e., $\rho_\pi = 1$)

Assumptions:

(A1) There exists at least one proper policy.

(A2) For every improper policy π, the associated expected cost $J_\pi(i)$ is infinite for at least one state i.

Under a proper policy, Prob(not reaching T) goes down asymptotically.

$$P(i_{2n} \neq T \mid i_0 = i, \pi)$$

$$= P(i_{2n} \neq T \mid i_n \neq T, i_0 = i, \pi)\, P(i_n \neq T \mid i_0 = i, \pi)$$

use Markov property & time homogeneity to conclude this prob $\leq \rho_\pi$

$$\leq \rho_\pi^2$$

Lecture-7*

More generally,

$$P(i_k \neq T \mid i_0 = i, \pi) \leq \rho_\pi^{\lfloor \frac{k}{n} \rfloor}$$

$k < n$: ≤ 1

$k > n$: $P(i_k \neq T \mid i_{\lfloor \frac{k}{n} \rfloor} \neq T, i_0 = i, \pi) \times P(i_{\lfloor \frac{k}{n} \rfloor} \neq T \mid i_0 = i, \pi)$

$$\leq 1 \times \rho_\pi^{\lfloor \frac{k}{n} \rfloor} = \rho_\pi^{\lfloor \frac{k}{n} \rfloor}$$

Let g_k be the cost incurred in kth stage for policy π

$$E(g_k) \leq \rho_\pi^{\lfloor \frac{k}{n} \rfloor} \max_{i=1,\ldots,n} |g(i, \pi(i))|$$

for simplicity, assume single stage cost is a function of current state & action

$$|J_\pi(i)| \le \lim_{N\to\infty} \sum_{k=0}^{N-1} \rho_\pi^{\lfloor \frac{k}{n} \rfloor} \max_{i=1,\dots,n} |g(i,\pi(i))|$$

$$< \infty. \quad (\text{since} \quad \rho_\pi < 1)$$

So, for a proper policy π, the expected cost J_π is finite.

Notation :

State space $\mathcal{X} = \{1, \dots n, T\}$

For any $J = (J(1), \dots, J(n))$,
define the "Bellman optimality" operator $TJ = (TJ(1), \dots TJ(n))$
as

$$(TJ)(i) = \min_{a \in A(i)} \sum_{j \in \mathcal{X}} p_{ij}(a)\big(g(i,a,j) + J(j)\big), \qquad i = 1, \dots n$$

$\mathcal{X}$ includes terminal state

Alternatively,

$$(TJ)(i) = \min_a E_j\big[g(i,a,j) + J(j)\big]$$

Another operator for a given policy π :

$T_\pi J = (T_\pi J(1), \dots, T_\pi J(n))$, where

$$T_\pi J(i) = \sum_{j \in \mathcal{X}} p_{ij}(\pi(i))\big(g(i,\pi(i),j) + J(j)\big), \qquad i = 1, \dots n$$

For a given π,

$$P^{\pi} = \begin{bmatrix} P_{11}(\pi(1)) & \cdots & P_{1n}(\pi(1)) \\ \vdots & & \\ P_{n1}(\pi(n)) & \cdots & P_{nn}(\pi(n)) \end{bmatrix}$$

P^{π} matrix has positive entries & row sum ≤ 1

$g_{\pi} = (g_{\pi}(1), \ldots, g_{\pi}(n))$, where

$$g_{\pi}(i) = \sum_{j \in S} P_{ij}(\pi(i))\, g(i, \pi(i), j)$$

Using the notation above,

$$T_{\pi} J = g_{\pi} + P^{\pi} J$$

Let $(T^k J)(i) = (T(T^{k-1} J))(i)$, $i = 1 \cdots n$

with $(T^0 J)(i) = J(i)$

Similarly, $(T_{\pi}^k J)(i) = (T_{\pi}(T_{\pi}^{k-1} J))(i)$, $i = 1 \cdots n$

with $(T_{\pi}^0 J)(i) = J(i)$

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

(ii) For any stationary policy π,

$$(T_\pi^k J)(i) \le (T_\pi^k J')(i)$$

Pf:

$$(TJ)(i) = \min_a \sum_{j \in X} p_{ij}(a)\left(g(i,a,j) + J(j)\right)$$

$$\le \min_a \sum_{j \in X} p_{ij}(a)\left(g(i,a,j) + J'(j)\right)$$

$$= (TJ')(i).$$

H.W. Complete the rest of the proof using induction.

H.W. Do the proof for T_π. ■

Another lemma:

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

(ii) $(T_\pi^k(J + \delta e))(i) \le (T_\pi^k J)(i) + \delta$.

Note: Inequalities reversed if $\delta < 0$.

Pf:

$$(T(J+\delta e))(i)$$

$$= \min_a \sum_{j\in X} p_{ij}(a)\,(g(i,a,j) + (J+\delta e)(j))$$

$$= \min_a \sum_{j\in X} p_{ij}(a)\,(g(i,a,j) + J(j) + \delta)$$

$$= \min_a \left[\sum_{j\in X} p_{ij}(a)(g(i,a,j) + J(j)) + \sum_j p_{ij}(a)\,\delta \right]$$

since $\sum_j p_{ij}(a) \le 1$

$$\le \min_a \left[\sum_{j\in X} p_{ij}(a)(g(i,a,j) + J(j)) + \delta \right]$$

$$= \min_a \left[\sum_{j\in X} p_{ij}(a)(g(i,a,j) + J(j)) \right] + \delta$$

$$= (TJ)(i) + \delta$$

H.W. Complete the proof by induction. ■

Properties of T_π:

Proposition 1: Assume (A1) & (A2). Then,

(i) For any proper policy π, the associated cost J_π satisfies

$$\lim_{k\to\infty} (T_\pi^k J)(i) = J_\pi(i), \quad i = 1, \dots, n,$$

for any J.

Also, $J_\pi = T_\pi J_\pi$ ——— (*)

& J_π is the unique solution of (*)

[For computing J_π, one can either apply T_π repeatedly a large number of times (or) solve $J = T_\pi J$]

Policy evaluation: task of computing J_π

(ii) Suppose a stationary policy π satisfies

$$J(i) \geq (T_\pi J)(i), \quad i = 1 \cdots n, \text{ for some } \hat{J}.$$

Then, π is proper.

Properties of Bellman optimality operator T:

Prop 2: Assume (A1) & (A2).

Let $J^*(i) = \min_{\pi \in \Pi} J_\pi(i), \quad \forall i = 1 \cdots n$

$\pi^* \in$ optimal policy

Then,

(i) J^* satisfies

$$J^* = T J^* \quad \text{——— (**)}$$

& J^* is the unique solution to (**)

(ii) For any finite J, we have

$$\lim_{k\to\infty} (T^k J)(i) = J^*(i), \quad i = 1 \dots n$$

(iii) A stationary policy π is optimal if & only if

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

Lecture-8*

<Proof of Prop. 1> Part (i)

First claim: $\lim_{k\to\infty} T_\pi^k J = J_\pi$ for any J

<If within pf> $T_\pi J = g_\pi + P_\pi J$

$$T_\pi^2 J = g_\pi + P_\pi T_\pi J = g_\pi + P_\pi g_\pi + P_\pi^2 J$$

Generalize,

$$T_\pi^k J = P_\pi^k J + \sum_{m=0}^{k-1} P_\pi^m g_\pi \quad \text{---(*)}$$

"π is proper" (Given). So,

$$P(i_k \neq T \mid i_0 = i, \pi) \le e_\pi^{\lfloor \frac{k}{n} \rfloor} \quad \forall i$$

$$\lim_{k\to\infty} P_\pi^k J = 0$$

So, wrtg (*),

$$\lim_{k\to\infty} T_\pi^k J = \lim_{k\to\infty} \sum_{m=0}^{k-1} P_\pi^m g_\pi = J_\pi$$

(See defn of J_π below)

$$J_\pi(i) = \lim_{k\to\infty} E\left(\sum_{m=0}^{k-1} g(x_m, \pi(x_m), x_{m+1}) \,\middle|\, x_0 = i\right)$$

<End of Pf within Pf>

Second claim: $J_\pi = T_\pi J_\pi$ & uniqueness.

<Pf within Pf, again>

$$T_\pi^{k+1} J = g_\pi + P_\pi T_\pi^k J$$

Taking limit as $k\to\infty$ on both sides, and using first claim,

$$J_\pi = g_\pi + P_\pi J_\pi$$

$$J_\pi = T_\pi J_\pi$$

Uniqueness: Suppose J satisfies $J = T_\pi J$

$$J = T_\pi^2 J = \cdots = T_\pi^k J$$

Taking limit as $k\to\infty$

$$J = \lim_{k\to\infty} T_\pi^k J = J_\pi$$

<End of Pf within Pf, again>

<Proof of Prop 1 – part (ii)>

Given: Stationary π satisfying $J \geq T_\pi J$ for some J

$$J \geq T_\pi J \Rightarrow T_\pi J \geq T_\pi^2 J$$

(or) $J \geq T_\pi^2 J$ & so on

$$J \geq T_\pi^k J = P_\pi^k J + \sum_{m=0}^{k-1} P_\pi^m g_\pi$$ (from (*) a couple of pages ago)

If π is not proper, then

$$J_\pi = \lim_{k\to\infty} \sum_{m=0}^{k-1} P_\pi^m g_\pi \text{ diverges.}$$

This leads to a contradiction because $J(i)$ is finite $\forall i$

<End of part (ii)>

<Proof of Prop 2>

(i) To show: J^* satisfies

$$J^* = T J^* \quad \text{——— (**)}$$

& J^* is the unique solution to (**)

<Pf> "T has at most one fixed point."

Suppose J and J' satisfy $J = TJ$, $J' = TJ'$

Select a policy π s.t.

$$TJ(i) = T_\pi J(i), \ \forall i$$

$$\Leftrightarrow \min_{a\in A(i)} \sum_j p_{ij}(a)\left(g(i,a,j)+J(j)\right) = \sum_j p_{ij}(\pi(i))\left(g(i,\pi(i),j)+J(j)\right)$$

"Set $\pi(i)$ to be the minimizer on LHS, $\forall i$"

Similarly, select a policy π' s.t. $TJ' = T_{\pi'} J'$

Notice that $J = TJ = T_\pi J$,

So, $J = J_\pi$. Using part (ii) of Prop 1, we can infer π is proper.

Similarly, $J' = J_{\pi'}$ & π' is proper.

Now, $J = TJ = T^k J \leq T_{\pi'}^k J$, $\forall k \geq 1$

↑ by definition of T, T_π.

Taking limit, $J \leq \lim_{k \to \infty} T_{\pi'}^k J = J_{\pi'} = J'$

So, $J \leq J'$

Swapping J & J' & repeating the arguments,

$J' \leq J$

So, $J = J'$ (or) T has at most one fixed point.

<End of "uniqueness" part>

<Begin of "existence" part>

Let π be a proper policy ((A1) ensures $\exists$ at least one proper policy)

Let π' be another policy s.t.

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

Now, $J_\pi = T_\pi J_\pi \geq T J_\pi = T_{\pi'} J_\pi$

(J_π is unique fixed pt of T_π; $\geq$ by def of T, T_π; $T J_\pi = T_{\pi'} J_\pi$ by construction of π')

So, $J_\pi \geq T_{\pi'} J_\pi \Rightarrow \pi'$ is proper (Prop 1, part (ii))

$$J_\pi \geq T_{\pi'}^2 J_\pi \geq \cdots \geq T_{\pi'}^k J_\pi$$

Taking limits,

$$J_\pi \geq \lim_{k\to\infty} T_{\pi'}^k J_\pi = J_{\pi'}$$

So, we got two proper policies π, π' s.t.

$$J_\pi \geq J_{\pi'}$$

Repeating the arguments above again & again & again, we obtain a sequence of policies $\{\pi_k\}$ s.t.

(i) Each π_k is proper

(ii) $J_{\pi_k} \geq T J_{\pi_k} \geq J_{\pi_{k+1}}, \quad \forall k \geq 1$

State-action spaces finite $\Rightarrow$ # polices is finite

$\Rightarrow$ Some policy, say π, must be repeated in $\{\pi_k\}$

$\Rightarrow$ using (ii), we obtain

$$J_\pi = T J_\pi$$

< End of existence proof)

Claim: $T^k J \to J_\pi$ as $k \to \infty$ (Later we show $J_\pi = J^*$)

<If>: e = n-vector of ones. Fix $\delta > 0$

Pick $\hat{J}$ s.t. $T_\pi \hat{J} = \hat{J} - \delta e$

Such a $\hat{J}$ can be found since

$$\hat{J} = T_\pi \hat{J} + \delta e = (g_\pi + \delta e) + P_\pi \hat{J}$$

i.e., $\hat{J}$ is the expected cost of a policy in an SSP with single stage cost $(g_\pi + \delta e)$

Since π is proper, $\hat{J}$ is unique (Prop 1 part (iii)).

By construction $J_\pi \le \hat{J}$ since $g_\pi \le g_\pi + \delta e$, $\delta > 0$

Lecture 9

Recall J_π satisfies $J_\pi = T J_\pi$

Notice that

$$J_\pi = T J_\pi \le T \hat{J} \le T_\pi \hat{J} = \hat{J} - \delta e \le \hat{J}$$

$$\Rightarrow \quad J_\pi = T^k J_\pi \le T^k \hat{J} \le T^{k-1} \hat{J} \le \hat{J}$$

$T\hat{J} \le \hat{J}$
$T^{k-1}(T\hat{J}) \le T^{k-1}\hat{J}$
$T^k \hat{J} \le T^{k-1}\hat{J}$

$\Rightarrow \quad J_\pi \le T^k \hat{J} \le \hat{J}$

So, $\{ T_k \hat{J} \}$ forms a monotone bounded sequence.

Hence, $\lim_{k \to \infty} T^k \hat{J} = \tilde{J}$, for some $\tilde{J}$

Apply T on both sides of the shaded equation

$$T\left(\lim_{k \to \infty} T^k \hat{J} \right) = T \tilde{J}$$

T is a continuous mapping $\Rightarrow$ T can be taken inside the limit

($TJ = \min_a$ "linear fn(J)" $\Rightarrow$ hence T is continuous)

$$T\tilde{J} = T\left(\lim_{k \to \infty} T^k \hat{J} \right) = \lim_{k \to \infty} T^{k+1} \hat{J} = \tilde{J}$$

So, $T\tilde{J} = \tilde{J}$ (*)

$\Rightarrow \quad \tilde{J} = J_\pi$ since the fixed point of T is unique & $= J_\pi$.

"The sandwich principle"

We will show $\lim_{k \to \infty} T^k (J_\pi - \delta e) = J_\pi$

To see this,

See a lemma on p. 9 above

$$J_\pi - \delta e = TJ_\pi - \delta e \overset{\uparrow}{\leq} T(J_\pi - \delta e) \leq TJ_\pi = J_\pi$$

$$J_\pi - \delta e \leq T(J_\pi - \delta e) \leq J_\pi$$

Using "$T(J_\pi - \delta e) \leq T^2(J_\pi - \delta e) \leq \cdots \cdots \leq T^k(J_\pi - \delta e)$"

$$J_\pi - \delta e \leq T^k(J_\pi - \delta e) \leq J_\pi$$

So, $\{T^k(J_\pi - \delta e)\}$ is a monotone bounded sequence, implying

$$\lim_{k\to\infty} T^k(J_\pi - \delta e) = J_\pi$$

→ use an argument similar to (*)

For any J, we can find a $\delta > 0$ such that

$$J_\pi - \delta e \leq J \leq \hat{J}$$

→ cost of policy π with single stage cost $g_\pi + \delta e$

$$T^k(J_\pi - \delta e) \leq T^k J \leq T^k \hat{J}$$

Taking limits as $k \to \infty$,

$$J_\pi \leq \lim_{k\to\infty} T^k J \leq J_\pi$$

So, $\lim_{k\to\infty} T^k J = J_\pi$ for any J.

What remains: To show $J_\pi = J^*$

Take any policy π', we have

$$T_{\pi'}^k J_0 \geq T^k J_0, \quad \text{— (**)}$$

where J_0 is a arbitrary n-vector.

Take limits as $k \to \infty$ on both sides of (**) to obtain

$$J_{\pi'} \geq J_\pi \quad \text{—— True for any } \pi'$$

So, $J_\pi = J^*$.

< End of part (b), we showed (i) $J^* = TJ^*$
(ii) $\lim_{k\to\infty} T^k J = J^*$ for any J >

< Start of part (c):
π is optimal $\iff$ $T_\pi J^* = TJ^*$ >

($\Rightarrow$) Suppose π is optimal

$$J_\pi = J^*$$

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

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

($\Leftarrow$) Suppose $J^* = TJ^* = T_\pi J^*$

Using $J^* = T_\pi J^*$, we can infer π is proper (Prop 1 part (ii))

For a proper π, if we have

$$J^* = T_\pi J^*, \text{ then}$$

$$J^* = J_\pi \quad \text{(Prop 1 part (i))}$$

$\Rightarrow \pi$ is optimal.

<End of part (c)>

Lecture-10*

Example: Time to termination

Consider an SSP where

$$g(i,a) = 1 \quad i = 1, \dots, n$$

Goal: get to terminal state asap.

Let $J^* \to$ optimal expected cost satisfies

Bellman equation: $J^*(i) = TJ^*(i)$, which is

$$J^*(i) = \min_{a \in A(i)} \left[1 + \sum_{j=1}^{n} p_{ij}(a)\, J^*(j) \right], \forall i$$

Special Case: Only one action in each state ← Markov chain

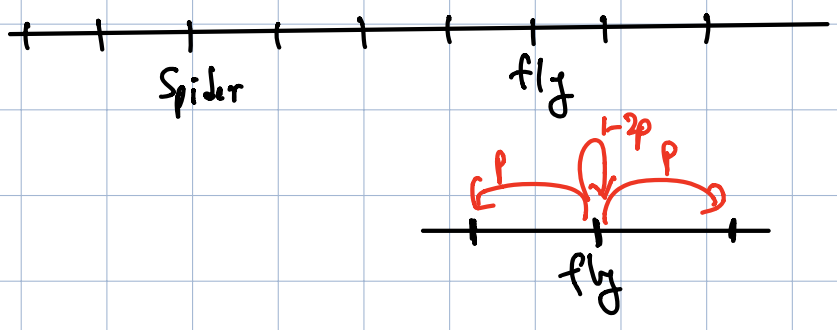
$J^*(i)$ ← expected first passage time to terminal state "T".

Let m_i denote this time (instead of $J^*(i)$)

$$m_i = 1 + \sum_{j=1}^{n} p_{ij} m_j, \quad i = 1 \cdots n$$

Another example: Fly & a spider

Spider & fly move on a line.



Spider: (i) if at a distance > 1 from fly, then jump 1 unit towards fly.

(ii) if distance $= 1$

jump towards fly

don't jump.

If fly & spider in same position, then you know what happens.

Goal: help spider catch the fly within "min" expected time.

SSP formulation:

State = distance between spider fly

States = $\{0, 1, \ldots n\}$ (n ↑ initial distance)

Terminal state = 0

let $P_{ij}(M)$ & $P_{ij}(\bar{M})$

($P_{ij}(M)$ ↑ spider moves; $P_{ij}(\bar{M})$ ↓ spider doesn't move)

Cost: = 1 in all states before hitting terminal "0" state.

remaining transition probabilities: P_{ij} $i \geq 2$

For $i \geq 2$,

$P_{i,i} = p$, $P_{i,i-1} = 1-2p$, $P_{i,i-2} = p$

In state 1,

Spider jumps: $P_{11}(M) = 2p$, $P_{10}(M) = 1-2p$

Spider doesn't jump: $P_{12}(\bar{M}) = p$, $P_{11}(\bar{M}) = 1-2p$, $P_{10}(\bar{M}) = p$

Bellman equation for $i \geq 2$

$$J^*(i) = 1 + p\,J^*(i) + (1-2p)\,J^*(i-1) + p\,J^*(i-2) \quad (*)$$

$$J^*(0) = 0$$

For state = 1 :

$$J^*(1) = 1 + \min\left[2p\,J^*(1) + (1-2p)\,J^*(0),\ p\,J^*(2) + (1-2p)\,J^*(1) + p\,J^*(0) \right]$$

$$J^*(1) = 1 + \min\left[\underbrace{2p\,J^*(1)}_{\text{jump}},\ \underbrace{p\,J^*(2) + (1-2p)J^*(1)}_{\text{don't jump}} \right] \quad \text{— (**)}$$

From (*),

$$J^*(2) = 1 + p\,J^*(2) + (1-2p)\,J^*(1)$$

$$\Rightarrow J^*(2) = \frac{1}{1-p} + \frac{(1-2p)\,J^*(1)}{1-p} \quad \text{— (***)}$$

Substitute (***) in (**),

$$J^*(1) = 1 + \min\left[2p\,J^*(1),\ \frac{p}{1-p} + \frac{p(1-2p)\,J^*(1)}{1-p} + (1-2p)\,J^*(1) \right]$$

which is the same as

$$J^*(1) = 1 + \min\left[\underbrace{2p\,J^*(1)}_{(A)},\ \underbrace{\frac{p}{1-p} + \frac{(1-2p)\,J^*(1)}{1-p}}_{(B)} \right]$$

Case (A) ≤ (B) :

$$J^*(1) = 1 + 2p\,J^*(1) \quad \&$$

$$2p\,J^*(1) \le \frac{p}{1-p} + \frac{(1-2p)\,J^*(1)}{1-p} \qquad (***)$$

In this case, $J^*(1) = \frac{1}{1-2p}$

Substitute this in $(***)$ to obtain

$$\frac{2p}{1-2p} \le \frac{p}{1-p} + \frac{1}{1-p} \iff p \le \frac{1}{3}$$

optimal to → jump.

Case (A) > (B)

$$J^*(1) = 1 + \frac{p}{1-p} + \frac{(1-2p)\,J^*(1)}{1-p}, \text{ and } (****)$$

$$2p\,J^*(1) > \frac{p}{1-p} + \frac{(1-2p)\,J^*(1)}{1-p}$$

From $(***)$, $J^*(1) = \frac{1}{p}$

Substitute this in the constraint to obtain

$$2 > \frac{p}{1-p} + \frac{1-2p}{p(1-p)} \iff p > \frac{1}{3}$$

↑ optimal to not jump.

$$J^*(1) = \begin{cases} \frac{1}{1-2p} & \text{when } p \le \frac{1}{3} \\ \frac{1}{p} & \text{else} \end{cases}$$

Using $J^*(1)$, we obtain $J^*(2)$ & so on for $J^*(i)$, $i \geq 3$.

Coming next!

Value iteration!

Start with J_0

$$J_0 \xrightarrow{T} J_1 \xrightarrow{T} J_2 \cdots\cdots$$

$$J_{k+1} = T J_k$$

We know under (A1) & (A2) (∃ at least one proper policy), (improper policies have ∞ cost)

$$\lim_{k \to \infty} T^k J_0 = J^*$$

Special case: "All policies are proper"

Then, VI converges at a geometric rate.

$$\| J_k - J^* \|_\infty \leq e^k \| J_0 - J^* \|_\infty$$

$0 < e < 1$

← need theory of contraction mappings.

Lecture-11* "Assume all policies are proper".

We will show that the Bellman optimality operator T is a contraction w.r.t. a weighted max-norm.

In particular, $\exists$ a vector $\xi = (\xi(1), \ldots, \xi(n))$ s.t. $\xi(i) > 0 \ \forall i = 1, \ldots, n$, and a scalar $\rho \in (0,1)$ such that

$$\|TJ - T\bar{J}\|_\xi \leq \rho \|J - \bar{J}\|_\xi, \quad \forall J, \bar{J} \in \mathbb{R}^n$$

(ρ: modulus of contraction)

Here $\|J\|_\xi = \max_{i=1,\ldots,n} \frac{|J(i)|}{\xi(i)}$

(weighted max-norm)

Also, for any stationary proper policy π,

$$\|T_\pi J - T_\pi \bar{J}\|_\xi \leq \rho \|J - \bar{J}\|_\xi \quad \forall J, \bar{J} \in \mathbb{R}^n$$

Pf: Need: a vector ξ such that T is a contraction wrt $\|\cdot\|_\xi$

Idea: Make up a MDP, solve it & get ξ.

& then show such a ξ helps in contraction business.

Consider a new SSP where transition probabilities are the same as in the original SSP, but transition costs $= -1$ except at terminal state.

$g(T, a, T) = 0 \quad \forall a \quad \& \quad g(i, a, j) = -1$ else.

Let $\hat{J} \to$ optimal cost in this new SSP.

Then, using Bellman equation

$$\hat{J}(i) = -1 + \min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\, \hat{J}(j)$$

$$= T\hat{J}(i)$$

$$\leq T_\pi \hat{J}(i) \quad \text{for any } \pi \text{ (proper stationary)}$$

$$= -1 + \sum_{j=1}^{n} p_{ij}(\pi(i))\, \hat{J}(j)$$

Let $\xi(i) = -\hat{J}(i)$, $i = 1 \cdots n$

Note: ① $\xi(i) \geq 1 \quad \forall i$

② $$-\hat{J}(i) \geq 1 + \sum_{j=1}^{n} p_{ij}(\pi(i)) \left(-\hat{J}(j)\right)$$

$$\Leftrightarrow \quad \xi(i) \geq 1 + \sum_{j=1}^{n} p_{ij}(\pi(i))\, \xi(j)$$

$$\sum_{j=1}^{n} p_{ij}(\pi(i))\, \xi(j) \leq \xi(i) - 1 \leq \rho\, \xi(i), \quad \text{—(*)}$$

where $\rho = \max_{i=1,\ldots,n} \left(\frac{\xi(i)-1}{\xi(i)} \right) < 1.$

How, for any π, J, $\bar{J}$,

$$T_\pi J(i) = \sum_j p_{ij}(\pi(i))\big(g(i,\pi(i),j) + \gamma J(j)\big)$$

$$\left|(T_\pi J)(i) - (T_\pi \bar{J})(i)\right|$$

$$= \left|\sum_{j=1}^{n} p_{ij}(\pi(i))\big(J(j) - \bar{J}(j)\big)\right|$$

$$\le \sum_{j=1}^{n} p_{ij}(\pi(i))\left|J(j) - \bar{J}(j)\right| = \sum_{j=1}^{n} p_{ij}(\pi(i))\xi(j)\frac{\left|J(j) - \bar{J}(j)\right|}{\xi(j)}$$

$$\le \left(\sum_{j=1}^{n} p_{ij}(\pi(i))\xi(j)\right)\left(\max_{j=1,\dots,n} \frac{\left|J(j) - \bar{J}(j)\right|}{\xi(j)}\right)$$

using (*)

$$\le \rho\,\xi(i)\left(\max_{j=1,\dots,n} \frac{\left|J(j) - \bar{J}(j)\right|}{\xi(j)}\right)$$

holds for any i

$$\frac{\left|(T_\pi J)(i) - (T_\pi \bar{J})(i)\right|}{\xi(i)} \le \rho\left(\max_{j=1,\dots,n} \frac{\left|J(j) - \bar{J}(j)\right|}{\xi(j)}\right)$$

$$\max_{i=1,\dots,n} \frac{\left|(T_\pi J)(i) - (T_\pi \bar{J})(i)\right|}{\xi(i)} \le \rho\left(\max_{j=1,\dots,n} \frac{\left|J(j) - \bar{J}(j)\right|}{\xi(j)}\right)$$

$$\|T_\pi J - T_\pi \bar{J}\|_\xi \le \rho\,\|J - \bar{J}\|_\xi$$

So, T_π is a contraction w.r.t. $\|\cdot\|_\xi$ with modulus ρ.

Next: to show that T is a contraction wrt. $\|\cdot\|_\xi$

We have shown already that

$$(T_\pi J)(i) \le (T_\pi \bar{J})(i) + \rho\, \xi(i) \max_{j=1,\dots,n} \frac{|J(j)-\bar{J}(j)|}{\xi(j)}$$

Take minimum over π on both sides to obtain

$$(TJ)(i) \le (T\bar{J})(i) + \rho\, \xi(i) \max_{j=1,\dots,n} \frac{|J(j)-\bar{J}(j)|}{\xi(j)}$$

Interchange J & $\bar{J}$ to obtain

$$(T\bar{J})(i) \le (TJ)(i) + \rho\, \xi(i) \max_{j=1,\dots,n} \frac{|J(j)-\bar{J}(j)|}{\xi(j)}$$

$$|(TJ)(i) - (T\bar{J})(i)| \le \rho\, \xi(i) \max_{j=1,\dots,n} \frac{|J(j)-\bar{J}(j)|}{\xi(j)}$$

$$\max_{i=1,\dots,n} \frac{|(TJ)(i) - (T\bar{J})(i)|}{\xi(i)} \le \rho \max_{j=1,\dots,n} \frac{|J(j)-\bar{J}(j)|}{\xi(j)}$$

Or, $$\|TJ - T\bar{J}\|_\xi \le \rho\, \|J - \bar{J}\|_\xi$$

Thus, T is a contraction wrt $\|\cdot\|_\xi$ with modulus ρ.

■

Value iteration (VI):

1) Choose some J_0

2) Repeatedly apply T

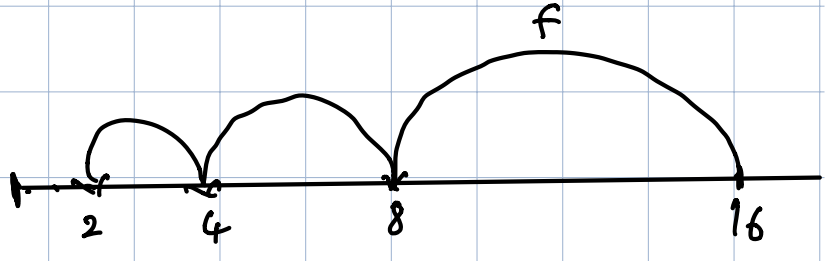
$$J_0 \xrightarrow{T} J_1 \xrightarrow{T} J_2 \xrightarrow{T} \cdots$$

$$J_{k+1} = T J_k$$

We know $\lim_{k \to \infty} T^k J_0 = J^*$ for any J_0.

Contraction mappings: A quick detour

$$f(x) = \frac{x}{2}$$



$$\lim_{n \to \infty} f^n(x) = 0$$

$f(0) = 0$

zero is a fixed point

& I get to the fixed point by repeatedly applying f.

Vector space $\mathcal{X}$.

Norm $\|\cdot\|$ satisfies 3 properties:

(i) $\|x\| = 0$ iff $x = 0$

(ii) $\|x + y\| \le \|x\| + \|y\|$, $\forall x, y \in \mathcal{X}$.

(iii) $\|cx\| = |c| \, \|x\|$, $\forall x \in \mathcal{X}$ & scalar c.

Contraction mapping!

$F : \mathcal{X} \to \mathcal{X}$ is a contraction mapping if $\exists \rho \in (0,1)$ s.t.

$$\|F(x) - F(y)\| \le \rho \|x - y\|, \quad \forall x, y \in \mathcal{X}.$$

The space $\mathcal{X}$ is complete under the norm $\|\cdot\|$ if every Cauchy sequence $\{x_k\} \subset \mathcal{X}$ converges.

[A sequence $\{x_k\}$ is Cauchy if $\exists N$ s.t. $\|x_m - x_n\| \le \epsilon$ $\forall m, n \ge N$]

Facts (i) If $\mathcal{X}$ is complete & F is a contraction wrt $\|\cdot\|$ with modulus ρ, then "F has a unique fixed point i.e., an x^* s.t. $F(x^*) = x^*$"

(ii) $x_{k+1} = F(x_k)$. Then $x_k \to x^*$ as $k \to \infty$.

Lecture-12

Contraction mappings in MDPs: (Ref: Sec 1.5 of DPOC-Vol II)

$$\|J\|_\xi = \max_{x \in X} \frac{|J(x)|}{\xi(x)}$$

where $\xi(x) > 0 \quad \forall x \in X$.

Let $B(X)$ denote all functions J s.t. $\|J\|_\xi < \infty$.

Simple case: finite state space.

Prop 3: Let $F : B(X) \to B(X)$ be a contraction mapping with modulus $\rho \in (0,1)$.

Assume $B(X)$ is complete. Then,

(i) There exists a unique $J^* \in B(X)$ s.t.

$$J^* = F J^*$$

(ii) $\lim_{k \to \infty} F^k J_0 = J^*$ ← VI converges

Also,

$$\|F^k J_0 - J^*\|_\xi \le \rho^k \|J_0 - J^*\|_\xi .$$

error bound for VI

Pf: See next page

Fix some $J_0 \in B(X)$.

Do $J_{k+1} = F J_k$ starting with J_0

$$\| J_{k+1} - J_k \|_\infty \leq \rho \| J_k - J_{k-1} \|_\infty$$

($\uparrow$ F is a ρ-contraction)

$$\Rightarrow \quad \| J_{k+1} - J_k \|_\infty \leq \rho^k \| J_1 - J_0 \|_\infty \quad \text{— (*)}$$

So, $\forall k \geq 0$, $m \geq 1$, we have

$$\| J_{k+m} - J_k \|_\infty = \| (J_{k+m} - J_{k+m-1}) + (J_{k+m-1} - J_{k+m-2}) + \cdots + (J_{k+1} - J_k) \|_\infty$$

triangle ineq $\rightarrow$

$$\leq \sum_{i=1}^{m} \| J_{k+i} - J_{k+i-1} \|_\infty$$

using (*) $\rightarrow$

$$\leq \rho^k (1 + \rho + \cdots + \rho^{m-1}) \| J_1 - J_0 \|_\infty$$

$$\| J_{k+m} - J_k \|_\infty \leq \frac{\rho^k}{1-\rho} \| J_1 - J_0 \|_\infty$$

Is $\{J_k\}$ a Cauchy sequence? Yes.

Since $B(X)$ is complete, $J_k \to J^*$ and $J^* \in B(X)$.

To show: J^* is a fixed point.

$$\|FJ^* - J^*\|_\xi \le \|FJ^* - J_k\|_\xi + \|J_k - J^*\|_\xi$$

why $J_k = FJ_{k-1}$ & F is contraction $\rightarrow$ $\le e\|J^* - J_{k-1}\|_\xi + \|J_k - J^*\|_\xi$

$\longrightarrow 0$ as $k \to \infty$ since $J_k \to J^*$.

So, $FJ^* = J^*$.

To show: uniqueness of J^*.

Assume $\tilde{J}$ is another fixed point ($F\tilde{J} = \tilde{J}$).

$$\|J^* - \tilde{J}\|_\xi = \|FJ^* - F\tilde{J}\|_\xi$$

$$\le e\|J^* - \tilde{J}\|_\xi$$

$\Rightarrow J^* = \tilde{J}$ (Why? because $e < 1$)

To show: error bound

$$\|F^k J_0 - J^*\|_\xi = \|F^k J_0 - FJ^*\|_\xi$$

$$\le e\|F^{k-1}J_0 - J^*\|_\xi$$

& repeat to infer

$$\|F^k J_0 - J^*\|_\xi \le e^k\|J_0 - J^*\|_\xi$$

■

Back to SSPs:

Assuming all policies are proper, $\exists \xi$ s.t. $\xi(i) > 0 \ \forall i$ and

$$\| TJ - T\bar{J} \|_\xi \leq \rho \, \| J - \bar{J} \|_\xi , \quad \forall J, \bar{J} \in \mathbb{R}^n$$

Using the error bound from the second claim in Prop 3, we obtain

$$\| J_k - J^* \|_\xi \leq \rho^k \, \| J_0 - J^* \|_\xi ,$$

where $J_k = T J_{k-1}$ & J_0 is the initial vector for value iteration.

Remark: A similar error bound holds for a evaluating a proper policy π using VI.

Gauss-Seidel variant of VI:

Note: In VI, we do $J_{k+1} = T J_k$, i.e., we apply the operator T for all states i

Alternative: update one state at a time & use recent updates of states that are already updated.

Gauss-Seidel VI update:

$$(FJ)(1) = \min_{a \in A(1)} \sum_j p_{1j}(a)\big(g(1,a,j) + J(j)\big)$$

same as $(TJ)(1)$

For $i = 2, \dots, n$,

$$(FJ)(i) = \min_{a \in A(i)} \Bigg[\sum_j p_{ij}(a)\, g(i,a,j) + \sum_{j=1}^{i-1} p_{ij}(a)(FJ)(j) + \sum_{j=i}^{n} p_{ij}(a)\, J(j) \Bigg]$$

← For $j = 1, \dots, i-1$ use previous updates

Remark: Gauss-Seidel VI converges to J^* & is faster than regular VI. See 2.2.2 of DPOC-Vol II.

$\|J_k^{VI} - J^*\|_\infty \geq \|J_k^{Gauss} - J^*\|_\infty$ ← See Prop 2.2.3 there.

Asynchronous VI

(i) Start with an arbitrary J_0

(ii) In kth iteration, pick an index $i_k \in \{1, \dots, n\}$ and do

$$J_{k+1}(i) = \begin{cases} (TJ_k)(i) & \text{if } i = i_k \\ J_k(i) & \text{else} \end{cases}$$

Assuming all states in $\{1 \dots n\}$ are picked infinitely often, It can be shown that

$$J_k \to J^* \text{ as } k \to \infty.$$

(See the same section of VI)

Lecture 13*

Q-learning as a form of VI

(Usual Q-learning doesn't require the model. But we are following DProc Vol II nomenclature)

When the model is known, i.e., we can compute T, Q-learning is equivalent to VI.

However, in the case when T is not computable directly, such as in a typical RL setting, there is a natural extension of "Q-learning" available.

Consider an SSP

$$\text{VI}: \quad J_{k+1} = T J_k$$

$$J_{k+1}(i) = \min_{a \in A(i)} \sum_j p_{ij}(a) \left(g(i,a,j) + J_k(j) \right)$$

$$J_{k+1}(i) = \min_{a \in A(i)} Q_{k+1}(i,a), \quad \text{— ①}$$

where $$Q_{k+1}(i,a) = \sum_j p_{ij}(a) \left(g(i,a,j) + \min_{b \in A(j)} Q_k(j,b) \right)$$ — ②

with $J_0(i) = \min_{a \in A(i)} Q_0(i,a)$ — ③

①-③ : Q-learning update

Why Q-learning works?

Define $Q^*(i,a) = \sum_j p_{ij}(a)\left(g(i,a,j) + J^*(j)\right)$

(Q-factors / Q-values)

Take action a in state i & from the next state onwards follow the optimal policy.

Using Bellman equation $J = TJ$:

$$J^*(i) = \min_{a \in A(i)} Q^*(i,a)$$

Thus, Q^* can be alternatively defined by

$$Q^*(i,a) = \sum_j p_{ij}(a)\left(g(i,a,j) + \min_{b \in A(j)} Q^*(j,b)\right)$$

Q-Bellman equation

The operator underlying Q-Bellman equation is

$$(FQ)(i,a) = \sum_j p_{ij}(a)\left[g(i,a,j) + \min_{b \in A(j)} Q(j,b)\right]$$

It can be shown that FQ is a contraction mapping when all policies are proper

So, VI for Q-Bellman equation is start with $Q_0(\cdot,\cdot)$ & keep applying $(FQ)(\cdot,\cdot)$ operator

$$Q_0 \xrightarrow{F} Q_1 \xrightarrow{F} \cdots \longrightarrow Q^*$$

$\uparrow$ eventually

Once you have Q^*, we can obtain J^* using

$$J^*(i) = \min_a Q^*(i,a)$$

Eq (1)-(3) is just VI on Q-values.

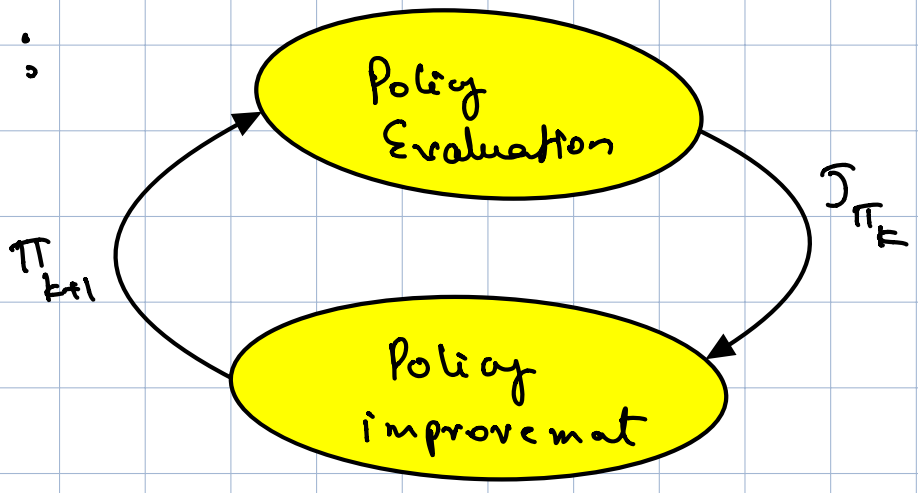
PTO

Policy iteration (PI)

VI can possibly take infinite # of iterations to converge

PI is a method "guaranteed" to converge within a finite # of iterations.

PI:



$J_{\pi_0} \geq J_{\pi_1} \cdots\cdots$

↑ ineq strict for at least one state

, if not, i.e., $J_{\pi_k} = J_{\pi_{k+1}}$, then, we have found the optimal policy.

PI algorithm:

Step 1: Start with a proper policy π_0

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

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

(here $J(1) \ldots J(n)$ are the unknowns & solving $(*)$ given J_{π_k})

Step 3: Policy improvement

Find a new policy π_{k+1} by

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

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

If $J_{\pi_{k+1}}(i) < J_{\pi_k}(i)$ for at least one state i,

then go to step 2 & repeat.

Else, Stop. In this case, $J_{\pi_{k+1}} = J_{\pi_k} = J^*$ (we will show this)

Step 1 can be modified by starting with a finite J & going to Step 3 directly. That would give π, & from thereon do step 2 & 3 in tandem until convergence.

Claim: Policy improvement does just that.

Let π, π' be two proper policies s.t.

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

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

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

Pf: We know that $J_\pi = T_\pi J_\pi$

Also, $T_{\pi'} J_\pi = T J_\pi$ ← given

So,

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

because $T_\pi J_\pi \geq T J_\pi$

$$\geq \sum_j p_{ij}(\pi'(i)) \left(g(i, \pi'(i), j) + J_\pi(j) \right)$$

$$= (T_{\pi'} J_\pi)(i)$$

So, $J_\pi \geq T_{\pi'} J_\pi \geq \cdots \geq T_{\pi'}^k J_\pi \geq \cdots \geq \lim_{k \to \infty} T_{\pi'}^k J_\pi = J_{\pi'}$

(eqn (4))

$$\Rightarrow J_\pi \geq J_{\pi'}$$

Now, if $J_\pi = J_{\pi'}$, then from eqn (A),

$$J_\pi = T_{\pi'} J_\pi \quad \text{— (1)}$$

We also have $T_{\pi'} J_\pi = T J_\pi$ — (2)

Combining (1) & (2),

$$J_\pi = T J_\pi$$

So, π is optimal.

Thus, if π isn't optimal, then $J_{\pi'}(i) < J_\pi(i)$ for at least one state i

■

Claim: If the number of proper policies is finite, the PI converges in a finite number of steps.

Why? Policy improvement guarantees a better policy if the latter isn't optimal.

Modified PI:

Let $\{m_0, m_1, \ldots\}$ be positive integers.

Let $J_1, J_2, \ldots$ and $\pi_0, \pi_1, \ldots$ be computed as follows:

Policy improvement $\rightarrow$ $T_{\pi_k} J_k = T J_k$

Policy evaluation $\rightarrow$ $J_{k+1} = T_{\pi_k}^{m_k} J_k$

by m_k steps of VI

Two special cases:

If $m_k = \infty$, Modified PI = PI since $T_{\pi_k}^{\infty} J_k = J_{\pi_k}$

If $m_k = 1$, Modified PI = VI

since $J_{k+1} = T_{\pi_k}^{m_k} J_k = T_{\pi_k} J_k = T J_k$

Recommended: Choose $m_k > 1$

Convergence: Modified PI converges

"proof skipped" See Ch. 2 of Bertsekas DPOC-Vol II

Lecture - 14*

Asynchronous PI:

Let $J_1, J_2, J_3, \ldots$: sequence of optimal cost estimates

$\pi_1, \pi_2, \pi_3, \ldots$: corresponding sequence of policies

Given (J_k, π_k), we select a subset S_k of states and generate (J_{k+1}, π_{k+1}) in one of the following two ways:

(Way 1)

$$J_{k+1}(i) = \begin{cases} (T_{\pi_k} J_k)(i) & \text{if } i \in S_k \\ J_k(i) & \text{else} \end{cases} \quad (**)$$

Set $\pi_{k+1} = \pi_k$

(Way 2)

$$\pi_{k+1}(i) = \begin{cases} \underset{a \in A(i)}{\arg\min} \sum_j p_{ij}(a)\big(g(i,a,j) + J_k(j)\big), & \text{if } i \in S_k \\ \pi_k(i) & \text{else} \end{cases} \quad (+)$$

Set $J_{k+1} = J_k$

Special Cases of Async-PI:

① If S_k = entire state space, (*) is performed infinite # of times before one (+) operation, then we get **regular PI**

② If S_k = entire state space, (*) performed m_k times before (+), then we get **modified PI**

③ If S_k = entire state space, one (*) operation followed immediately by one (+) operation, then we get **Value iteration**

④ If $|S_k| = 1$, and one (*) operation followed immediately by one (+) operation, then we get **Asynchronous VI**

Remark: Async-PI can be shown to converge. See DPOC-Vol II for details.

Aside: Value iteration with Q-factors leads finally to Q-learning (in an RL setting)

PI leads to "Actor-critic methods".