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

initial state

Expected cost or Value function

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

policy $\pi = \{\mu_0, \mu_1, \mu_2, \ldots\ldots\}$

single-stage cost "stationary"

discount factor $\alpha \in (0,1]$

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

optimal expected cost

set of admissible policies

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

optimal policy

Two popular MDPs:

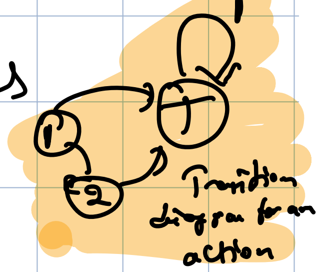
(I) Stochastic shortest path (SSP) aka episodic MDPs

(a) $\alpha = 1$ (b) Finite state-action space
states $= \{1, \ldots n\} \cup \{T\}$

(c) There exist a special state, say "T" that is

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

absorbing $P_{TT}(a) = 1$, $\forall a$



(II) Discounted MDP

(a) $\alpha < 1$

(b) $|g(i,a,j)| \leq M < \infty \quad \forall i,j \in \mathcal{X},\ a \in \mathcal{A}$

(a) & (b) $\Rightarrow$ $J_\pi(x_0)$ 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 J^* is given by

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

infinite horizon optimal cost

Finite horizon optimal cost

(II) Bellman equation

Assume $\mathcal{X} = \{1 \ldots n\}$, transition prob. $P_{ij}(a)$

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

$$J_{k+1}^*(i) = \min_{a \in \mathcal{A}(i)} \sum_{j=1}^{n} P_{ij}(a)\left(g(i,a,j) + \alpha \underbrace{J_k^*(j)}_{\text{optimal cost for a } k\text{-stage problem}}\right)$$

optimal cost in a $(k+1)$-stage problem

optimize action in the current stage

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

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

(or)

Bellman equation

$$\boxed{J^*(i) = \min_{a \in \mathcal{A}(i)} E_{j}\left[g(i,a,j) + \alpha J^*(j)\right],} \quad \forall i$$

DP algorithm is an algorithm, while Bellman equation is a system of eqns that the optimal cost satisfies.

③ How to get the optimal policy π^*?

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

← Bellman equation

For state i, Let $a^*(i)$ be the minimizer in the Bellman equation

Make-up a policy as follows: $\pi(i) = a^*(i), \forall i$

Then, π is ~~the~~ an optimal policy.

Lecture-6 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:

state, stage

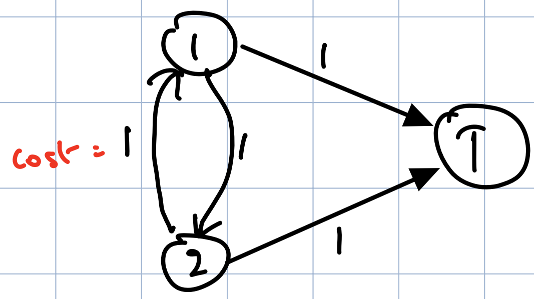
States: (i, k)

Terminal states: $(i, N), \forall i \in \mathcal{X}$

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: Go to T from 1 & also 2.

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

Def: A stationary policy π is proper if $\exists\, m > 0$

$$\rho_\pi = \max_{i=1 \ldots n} P(\underbrace{i_m \neq T}_{\text{m th state visited along a sample path}} \mid \underbrace{i_0 = i}_{\text{for starting in state } i},\ \underbrace{\pi}_{\text{actions governed by } \pi}) < 1$$

Proper policy: $\exists$ a pos. prob. path from any state i to the terminal state.

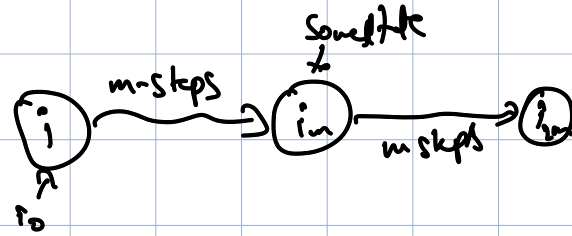
A policy that isn't proper is improper (i.e., $\rho_\pi = 1$)

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.



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

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

Markov property

$$= P(i_m \neq T \mid i_0 \neq T, \pi)\, P(i_m \neq T \mid i_0 = i, \pi) \leq e_\pi^2$$

More generally,

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

$k < m$: 1

$k > m$:

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

$$\leq 1 \times e_\pi^{\lfloor \frac{k}{m} \rfloor} = e_\pi^{\lfloor \frac{k}{m} \rfloor}$$

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

$$E(g_k) \leq e_\pi^{\lfloor \frac{k}{m} \rfloor} \underbrace{\max_{i=1\ldots n} |g(i, \pi(i))|}_{\text{For simplicity, assume single stage cost doesn't depend on next state}}$$

$$|J_\pi(i)| \le \lim_{N\to\infty} \sum_{k=0}^{N-1} e_\pi^{\lfloor \frac{k}{m} \rfloor} \max |g(i,\pi(i))|$$

$$< \infty$$

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

Lecture-7*

Notation :

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

For any $J = (J(1), \dots, J(n))$,

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

as

For $i=1,\dots n$, $$(TJ)(i) = \min_{a \in \mathcal{A}(i)} \sum_{j=\{1..n,T\}} p_{ij}(a)\, g(i,a,j) + \sum_{j=1}^{n} p_{ij}(a) J(j)$$

operator & it is not tied to any policy

includes terminal state

does not include "T"

Equivalently,

$$(TJ)(i) = \min_{a \in \mathcal{A}(i)} E_j \left(g(i,a,j) + J(j) \right)$$

Another operator for a given policy π :

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

$$T_\pi J(i) = \sum_{j \in \mathcal{S}} p_{ij}(\pi(i)) \left(g(i, \pi(i), j) + J(j) \right),$$

$i=1 \dots n$

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

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

Using g_{π} & P^{π}, we have

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

Applying operator T multiple times:

$$J \xrightarrow{T} TJ \xrightarrow{T} T^2 J \rightarrow \cdots$$

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

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

Similarly for T_{π}.

" T & T_{π} are monotone "

<u>Lemma 1:</u> 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:

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

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

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

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

H.W. Do the proof for T_π. ■

Another lemma: (Constant-shift lemma)

Stationary π, $\delta \to$ any 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) $(T_\pi^k(J + \delta e))(i) \le (T_\pi^k J)(i) + \delta$.

$$\boxed{J + \delta e = \begin{bmatrix} J(1) \\ \vdots \\ J(n) \end{bmatrix} + \begin{bmatrix} \delta \\ \vdots \\ \delta \end{bmatrix}}$$

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

$$= \min_a \sum_{j\in X} p_{ij}(a)\left(g(i,a,j) + \sum_{j=1}^{n} p_{ij}(a)(J+\delta e)(j)\right)$$

$$= \min_a \sum_{j\in X} p_{ij}(a)\, g(i,a,j) + \sum_{j=1}^{n} p_{ij}(a)\left(J(j)+\delta\right)$$

$$= \min_a \left[\sum_j p_{ij}(a)\left(g(i,a,j) + J(j)\right) + \sum_{j=1}^{n} p_{ij}(a)\,\delta \right]$$

≤ 1

Since $\sum_{j=1}^{n} p_{ij}(a) \le 1$

no "J(i)" here

$$\le \left[\min_a \sum_{j\in X} p_{ij}(a)\, g(i,a,j) + \sum_{j=1}^{n} p_{ij}(a) J(j)\right] + \delta$$

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

We showed

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

Rest of the proof $\rightarrow$ a simple induction. $\leftarrow$ H.W.

∎

Properties of T_π:

$\exists$ a proper policy

improper policies have infinite cost

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

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

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

for any J.

Also, $J_\pi = T_\pi J_\pi$ —— (*)

& J_π is the unique solution of (*)

$$\boxed{J(i) = g(i,\pi(i)) + \sum P_{ij}(\pi(i))\, J(j), \ \forall i}$$

Policy evaluation: Task of computing J_π for a given policy π

Using (**), policy evaluation can be done by starting in a "J" & repeatedly applying T_π ($J \xrightarrow{T_\pi} T_\pi J \xrightarrow{T_\pi} T_\pi^2 J \to \cdots$)

Using (*), solve "$J = T_\pi J$" to obtain J_π.

(ii) Suppose a stationary policy π satisfies

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

Then, π is proper.

Properties of Bellman optimality operator T:

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

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

$\pi^* \leftarrow$ optimal policy

$f(x) = x/2$

$\boxed{f(x^*) = x^*}$

$x^* = 0$

$16 \xrightarrow{f} 8 \xrightarrow{f} 4 \xrightarrow{f} 2 \to \cdots 0$

Then,

(i) J^* satisfies

Bellman equation → $$J^* = T J^* \quad \text{—— (**)}$$

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

i.e., does not exist a $\tilde{J} \neq J^*$ satisfying $\tilde{J} = T\tilde{J}$

(ii) For any finite J, we have

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

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

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

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

Lecture-8*

<Proof of Prop. 1> Part (i)

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

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

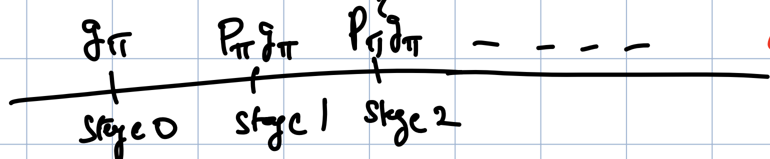
$$T_\pi^2 J = g_\pi + P_\pi T_\pi J = g_\pi + P_\pi g_\pi + P_\pi^2 J$$

Generalizing, we obtain

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

Note: $\lim_{k\to\infty} \sum_{m=0}^{k-1} P_\pi^m g_\pi = J_\pi$ --- (I)

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



$\leftarrow$ infinite horizon sum of these $P_\pi^m g_\pi$ is J_π

π is proper $\quad P(i_k \neq T \mid i_0 = i, \pi) \leq \rho_\pi^{\lfloor k/m \rfloor}, \; \forall i$

$$\lim_{k \to \infty} P_\pi^k J = 0 \quad \text{— (II)}$$

Using (I) & (II) in (+), we obtain $\lim_{k \to \infty} T_\pi^k J = J_\pi$

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

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

Taking limit as $k \to \infty$ on both sides, we get

$$\lim_{k \to \infty} T_\pi^{k+1} J = \lim_{k \to \infty} \left(g_\pi + P_\pi T_\pi^k J \right)$$

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

So, $\quad J_\pi = T_\pi J_\pi$

Uniqueness: Suppose $\tilde{J}$ satisfies $\tilde{J} = T_\pi \tilde{J}$

$$\tilde{J} = T_\pi \tilde{J} = T_\pi^2 \tilde{J} = \cdots\cdots = T_\pi^k \tilde{J}$$

Taking limits, $\quad \tilde{J} = \lim_{k \to \infty} T_\pi^k \tilde{J} = J_\pi$

So, J_π is the unique fixed point.

<End of pf within pf, again>

<Proof of Prop 1 – part (ii)>

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

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

$$\implies J \geq T_\pi^2 J \quad \text{\& so on}$$

$$J \geq T_\pi^k J = P_\pi^k J + \sum_{m=0}^{k-1} P_\pi^m g_\pi$$ (From eq (+) in part (i) proof)

If π is not proper, then

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

This lead to a contradiction since $\underset{\text{finite}}{J} \geq \lim_{k\to\infty} T_\pi^k J = \underset{\text{infinite for an improper policy}}{J_\pi}$

< End of part (ii) >

<Proof of Prop 2>

(i) To show: J^* satisfies

$$J^* = T J^* \quad \text{——} \quad (**)$$

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

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

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

We will show $J \geq J'$ by an interchangeable argument which would imply $J = J'$

Select a policy π s.t. $T_\pi J = TJ$

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

" Set $\pi(i)$ to be the minimizing action on the RHS"

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

We have $J = TJ = T_\pi J$

If $J = T_\pi J$, then $J = J_\pi$. Also, π is proper (using part (ii) after observing $J = T_\pi J$)

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

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

(compare $TJ \leq T_{\pi'} J$ for any π')

$\min_a \sum p_{ij}(a)(g(i,a,j)+J(j)) \rightarrow \leq \sum p_{ij}(\pi'(i))(g(\cdot\cdot)+J(j))$

$J \leq T^k_{\pi'} J \xrightarrow[\text{limits}]{\text{Taking}} J \leq \lim_{k\to\infty} T^k_{\pi'} J = J_{\pi'} = J'$

We got $J \leq J'$

Repeating the arguments with J & J' swapped leads to $J' \leq J$

So, $\boxed{J = J'}$ (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.

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

$J^* + \delta e = T(J^* + \delta e) = \lim_{k\to\infty} T^k(J^* + \delta e) = J^* \Rightarrow \delta = 0$

Now, $J_\pi \underset{\text{Prop 1}}{=} T_\pi J_\pi \underset{\text{def of } T, T_\pi}{\geq} T J_\pi \underset{\text{by construction}}{=} T_{\pi'} J_\pi$

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

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

Also, π' is proper from Prop 1, part(ii) since $J_\pi \geq T_{\pi'} J_\pi$ & π is proper.

So, we have two proper policies satisfying

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

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

(i) Each π_k is proper

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

No. of policies is finite, because we assumed that the state & action spaces are finite.

$J_{\pi_1} \geq T J_{\pi_1} \geq J_{\pi_2} \cdots\cdots\cdots$ ← in this sequence, a policy has to repeat

For that repeat policy, say $\tilde{\pi}$, we get $J_{\tilde{\pi}} = T J_{\tilde{\pi}}$

$\Rightarrow$ fixed point exists.

< End of existence proof >

Lecture 9

Claim: $T^k J \to J_{\tilde{\pi}}$ as $k \to \infty$ (later we show $J_{\tilde{\pi}} = J^*$)

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

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

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

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

$$J_{\tilde{\pi}} = g_{\tilde{\pi}} + P_{\tilde{\pi}} J_{\tilde{\pi}} = T_{\tilde{\pi}} J_{\tilde{\pi}}$$

$$\boxed{\hat{J} = (g_{\tilde{\pi}} + \delta e) + P_{\tilde{\pi}} \hat{J}} \quad \text{—} (*)$$

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

$J_{\tilde{\pi}}$ is expected cost in an SSP with single stage cost $g_{\tilde{\pi}}$

$\hat{J}$ is same for SSP with $g_{\tilde{\pi}} + \delta e$

$\hat{J}$ is the unique soln to $(*)$ since $\tilde{\pi}$ is proper.

By construction, $J_{\tilde{\pi}} \leq \hat{J}$ since $g_{\tilde{\pi}} \leq g_{\tilde{\pi}} + \delta e$

Recall J_{π} satisfies $J_{\tilde{\pi}} = T J_{\tilde{\pi}}$ ← from previous proof

Notice that

T is monotone

T is Bellman optimality operator $\Rightarrow T \leq T_{\pi}\ \forall \pi$.

trivial

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

By construction of $\hat{J}$

$$\Rightarrow \underbrace{J_{\tilde{\pi}} = T^k J_{\tilde{\pi}} \leq T^k \hat{J}}_{\text{repeated application of } T \text{ in earlier inequality}} \leq T^{k-1} \hat{J} \leq \hat{J}$$

$$\boxed{\begin{gathered} T\hat{J} \leq \hat{J} \\ T^{k-1}(T\hat{J}) \leq T^{k-1}\hat{J} \\ T^k \hat{J} \leq T^{k-1} \hat{J} \end{gathered}}$$

$$\boxed{\begin{gathered} J_1 \leq J_2 \\ \Rightarrow T J_1 \leq T J_2 \\ \text{contr. lemma} \end{gathered}}$$

$\Rightarrow \quad J_{\pi} \leq T^k \hat{J} \leq \hat{J}$ are both finite

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

Hence, $\boxed{\lim_{k \to \infty} T^k \hat{J} = \tilde{J}}$, 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 of J" $\Rightarrow$ 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,

$J_{\tilde{\pi}} = T J_{\tilde{\pi}}$

See a lemma on p. 9 above

$$J_{\tilde{\pi}} - \delta e = T J_{\tilde{\pi}} - \delta e \leq T(J_{\tilde{\pi}} - \delta e) \leq T J_{\tilde{\pi}} = J_{\tilde{\pi}}$$

T is monotone, $J_{\tilde{\pi}} - \delta e \leq J_{\tilde{\pi}}$

$$J_{\tilde{\pi}} - \delta e \leq T(J_{\tilde{\pi}} - \delta e) \leq J_{\tilde{\pi}}$$

Using "$T(J_{\tilde{\pi}} - \delta e) \leq T^2(J_{\tilde{\pi}} - \delta e) \leq \dots \dots \leq T^k(J_{\tilde{\pi}} - \delta e)$",

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

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

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

use an argument similar to (*)

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

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

$\hat{J}$: cost of policy π with single stage cost $g_{\tilde{\pi}} + \delta e$

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

Taking limits,

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

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

for any finite J

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

Take any policy π'. We have

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

where J_0 is a arbitrary n-vector.

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

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

So, $J_{\tilde{\pi}} = J^*$.

< 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 $\Longleftrightarrow$ $T_\pi J^* = TJ^*$ >

($\Rightarrow$) Suppose π is optimal

$$J_\pi = J^*$$

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

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

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

$J^* = T_\pi J^* \Rightarrow$ From Prop 1 part(ii), π is proper

For a proper π, we have

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

$$\Rightarrow \quad J^* = J_\pi$$

$\Rightarrow$ π is optimal.

< End of part (c) >

Lecture - 10*

Example: Time to termination

Consider an SSP where

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

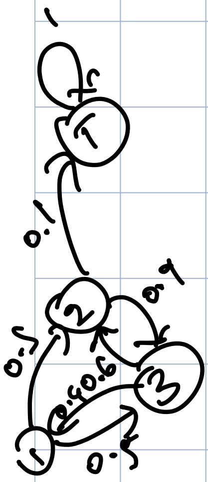
Goal: get to terminal state asap.

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

Bellman equation: $J^*(i) = TJ^*(i), \ \forall i$

$$J^*(i) = \min_a \left[1 + \sum_{j=1}^{n} P_{ij}(a) J^*(j) \right], \ i = 1 \dots n$$

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



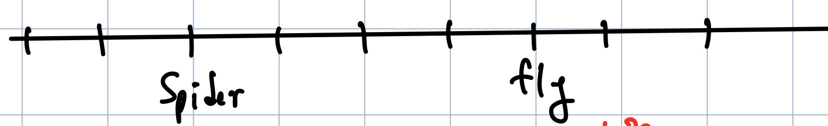
$J^*(i)$ ← expected first passage time to terminal state "T".

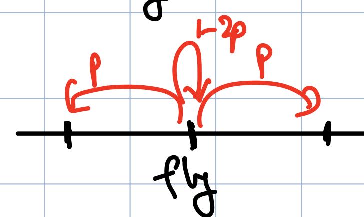
Let m_i denote this time (instead of $J^*(i)$)

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

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 = $\{0, 1, \ldots, n\}$ initial distance b/w spider & fly

Terminal state = 0

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

spider moves — spider doesn't move

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

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

For $i \geq 2$, $P_{i,i} = p$, $P_{i,i-1} = 1-2p$, $P_{i,i-2} = p$

For $i=1$, "Spider jumps" action M

$P_{11}(M) = 2p$ $\qquad$ $P_{10}(M) = 1-2p$

"Spider doesn't jump" action $\overline{M}$

$P_{12}(\bar{M}) = p$, $\qquad$ $P_{11}(\bar{M}) = 1-2p$, $P_{10}(\bar{M}) = p$

Bellman equation for $i \geq 2$

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

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

For state = 1 :

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

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

From $(*)$,

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

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

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

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

Case $(A) \leq (B)$:

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

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

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

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 $(A) > (B)$

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

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

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

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.

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

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

Lecture-11

Coming next!

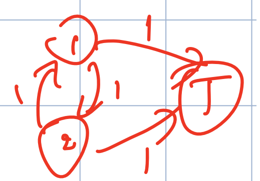
Value iteration (VI)

Start with J_0

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

$J_1 = TJ_0$
$J_2 = TJ_1 = T^2 J_0$
& so on ...

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



We already know very clearly that

$$\lim_{k\to\infty} T^k J_0 = J^* \quad \text{under (A1) \& (A2)}$$

(A1): $\exists$ 1 proper policy

(A2): improper policies have ∞ cost

"Can do VI for policy evaluation"

$$J_0 \xrightarrow{T_\pi} J_1 \xrightarrow{T_\pi} J_2 \xrightarrow{T_\pi} \cdots\cdots \xrightarrow[\text{as } k\to\infty]{} J_\pi$$

Special Case: "All policies are proper"

Then, VI converges at a geometric rate.

# of iterations of VI

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

$0 < \rho < 1$ constant

← need theory of contraction mappings.

$\|\cdot\|_\xi \to$ weighted-max norm

Note: Can find a k s.t. $\|J_k - J^*\| \leq \epsilon$, for any $\epsilon > 0$

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 $0 < \rho < 1$ such that

modulus of contraction

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

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

weighted max-norm

Also, for any stationary proper policy π,

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

Same ρ for $\forall \pi$ T & T_π'

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 with same "$p_{ij}(a)$", $\forall i, j, a$

but, different transition costs.

Transition cost $= -1$ everywhere except terminal state

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

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

Then, using Bellman equation

$$\hat{J}(i) = -1 + \min_{a} \sum_{j \in X} P_{ij}(a)\, \hat{J}(j)$$

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

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

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

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

Note: ① $\xi(i) \geq 1$

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

$$\xi(i) \geq 1 + \sum_{j} P_{ij}(\pi(i))\, \xi(j) \quad \text{—}(*)$$

$$\underbrace{\sum_{j} P_{ij}(\pi(i))\, \xi(j) \leq \xi(i) - 1}_{} \leq \rho\, \xi(i), \quad \text{—} (**)$$

follows from $(*)$

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

Now, for any π, J, $\bar{J}$,

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

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

triangle inequality

$$\leq \sum_{j=1}^{n} P_{ij}(\pi(i))\,|J(j) - \bar{J}(j)|$$

$$= \sum_j P_{ij}(\pi(i))\,\xi(j)\left(\frac{|J(j) - \bar{J}(j)|}{\xi(j)}\right)$$

$|a-b| \leq 10$
$a \leq b + 10$
$b \leq a + 10$

$$\leq \left(\sum_j P_{ij}(\pi(i))\,\xi(j)\right)\left(\max_{j=1\ldots n} \frac{|J(j) - \bar{J}(j)|}{\xi(j)}\right)$$

$$|(T_\pi J)(i) - (T_\pi \bar{J})(i)| \leq \rho\, \xi(i) \underbrace{\max_{j=1\ldots n} \frac{|J(j) - \bar{J}(j)|}{\xi(j)}}_{(I)}$$

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

This inequality holds for any state i

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

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

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

$TJ(i) = \min_a \sum P_{ij}(a)(g(i,a,j) + J(j))$

$(\hat{T}_\pi J)(i) = \sum P_{ij}(\pi(i))(g(i,\pi(i),j) + J(j))$

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

From (I), we have shown already that

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

Take minimum over π on both sides to obtain

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

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

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

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

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

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

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

Value iteration (VI):

1) Choose some J_0

2) Repeatedly apply T

$$J_0 \xrightarrow{T} J_1 \xrightarrow{T} J_2 \xrightarrow{T} \cdots \text{ — }.$$

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

$$\|TJ - T\bar{J}\|_\xi \le e\|J - \bar{J}\| \Rightarrow \|T^k J - T^k \bar{J}\|_\xi \le e^k \|J - \bar{J}\|_\xi$$

For the "all policies are proper" case,

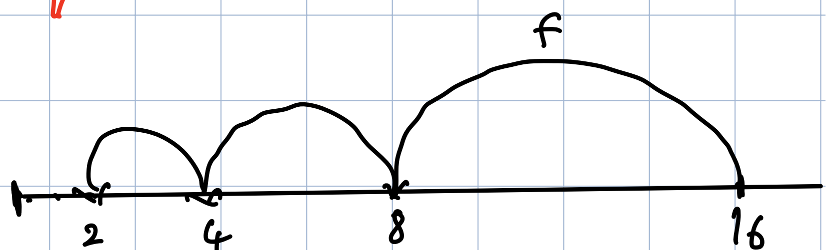
$$\|J_k - J^*\|_\xi = \|T^k J_0 - T^k J^*\|_\xi \le e^k \|J_0 - J^*\|_\xi$$

→ VI error bound if all policies are proper

Lecture-12*

Contraction mappings: A quick detour

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



For any x, $\lim_{n \to \infty} f^n(x) = 0$, $\quad f(0) = 0$

fixed point

Repeated application of f leads to the fixed point.

Vector space $\mathcal{X}$.

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

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

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

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

Contraction mapping!

input space

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

modulus of contraction

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

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 $\forall \epsilon > 0$ $\exists N_\epsilon$ s.t.
$\|x_m - x_n\| \le \epsilon \quad \forall m, n \ge N_\epsilon$

$x_1, x_2, x_3, \text{------} x_N$

$1, 1/2, 1/4, \text{----}$

$\downarrow x_m \quad \downarrow x_n$

Convergence $\to$ $\forall \epsilon > 0, \exists N$ s.t. $\|x_n - x^*\| < \epsilon \ \forall n > N$ | $\lim_{n\to\infty} x_n = x^*$

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 $k \to \infty$.

$\{\frac{1}{n}\}$ $\lim_{n\to\infty} \frac{1}{n} = 0$

If the underlying space is restricted to $(0,1]$, then $\{\frac{1}{n}\}$ is Cauchy, but doesn't converge

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

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

Weighted max-norm

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

If $\xi(x) = 1 \ \forall x$

Then, $\|J\|_\xi = \|J\|_\infty$

$\|\cdot\|_\infty \rightarrow$ sup-norm or max-norm

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

Simple case: finite state space.

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

Assume $B(\mathcal{X})$ is complete. Then,

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

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

VI converges (asymptotically)

(ii) $\lim_{k \to \infty} F^k J_0 = J^*$, for any $J_0 \in B(\mathcal{X})$

Also,
$$\underbrace{\|F^k J_0 - J^*\|_\xi \leq \rho^k \|J_0 - J^*\|_\xi}_{\text{error bound for VI that holds } \forall k \geq 1}.$$

Pf: See next page

Fix some $J_0 \in B(\mathcal{X})$.

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

$$\| J_{k+1} - J_k \|_\xi = \| F J_k - F J_{k-1} \|_\xi \leq \rho \, \| J_k - J_{k-1} \|_\xi$$

F is a ρ-contraction

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

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

Telescoping sum

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

Δ^{le} ineq. $\rightarrow$
$$\leq \sum_{i=1}^{m} \| J_{k+i} - J_{k+i-1} \|_\xi$$

using $(*)$ $\rightarrow$
$$\leq \rho^k \underbrace{(1 + \rho + \rho^2 + \cdots\cdots + \rho^{m-1})}_{\leq \sum \rho^i = \frac{1}{1-\rho}} \| J_1 - J_0 \|_\xi$$

$$\boxed{\| J_{k+m} - J_k \|_\xi \leq \frac{\rho^k}{1-\rho} \| J_1 - J_0 \|_\xi} \quad \text{---} \; (**)$$

Is $\{J_k\}$ a Cauchy sequence? Yes. (make RHS $\leq \epsilon$ in $(**)$)

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

J^* is only a limit of VI & we haven't shown $FJ^* = J^*$ yet.

To show: J^* is a fixed point.

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

$$\le e \| J^* - J_{k-1} \|_\xi + \| J_k - J^* \|_\xi$$

$$\xrightarrow{k \to \infty} 0 \quad \text{since } J_k \to J^*$$

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

To show: uniqueness of J^*.

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

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

$$\| J^* - \tilde{J} \|_\xi \le \underbrace{e}_{\in (0,1)} \| J^* - \tilde{J} \|_\xi$$

$$\Rightarrow \quad J^* = \tilde{J}$$

To show: error bound

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

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

& repeat to infer

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

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.

Lecture-13*

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 $i = 2, \ldots, n$,

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

For $j = 1 \ldots i-1$ use previous updates

Remark: Gauss-Seidel VI converges faster than regular VI

$$\| J_k^{VI} - J^* \|_\xi \geq \| J_k^{GS\text{-}VI} - J^* \|_\xi$$

Proof in Prop 2.2.5 of Vol. II.

Asynchronous VI

Synchronous VI: Do $J_{k+1}(i) = TJ_k(i)\ \forall i$

(i) Start with an arbitrary J_0

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

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

If all states are picked infinitely often, then

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

(Proof: Check Vol. II same section)

Lecture 13*

Q-learning as a form of VI

(usual Q-learning doesn't require the model. But we are following DPOC 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

VI : $J_{k+1} = T J_k$, with fixed J_0

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

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

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

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

①–③ : Q-learning update

Why Q-learning works?

Define $Q^*(i,a) = \sum_j P_{ij}(a)\,\underbrace{\left(g(i,a,j) + J^*(j)\right)}_{\text{Take action } a \text{ in state } i \text{ \& from the next state onwards follow the optimal policy.}}$ (Q-factors/Q-values)

Bellman Equation ($J^*(i) = TJ^*(i), \forall i \Leftrightarrow J^*(i) = \min_a E\left(g(i,a,j) + J^*(j)\right)$

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

Thus, Q^* can be alternatively defined by

$$\underbrace{Q^*(i,a) = \sum_j P_{ij}(a)\left(g(i,a,j) + \min_{b \in A(j)} Q^*(j,b)\right)}_{\text{Q-Bellman Equation}}$$

Q-Bellman Equation $\Leftrightarrow Q^*(i,a) = E\left(g(i,a,j) + \min_b Q^*(j,b)\right)$

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

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

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.

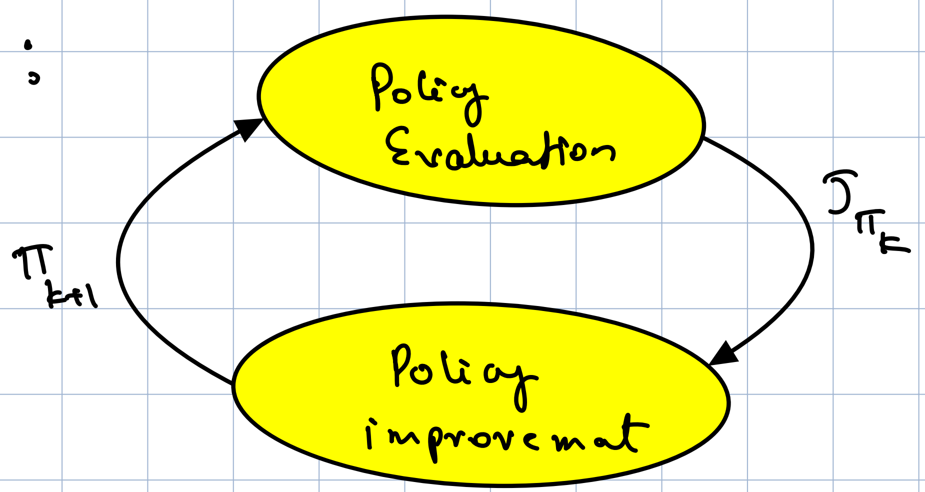
P T O

Policy iteration (PI)

VI can possibly take infinite # of iterations to converge

PI is a method "guaranteed" to converge within a finite # of iterations, for finite state-action spaced MDPs.

PI :



$$J_{\pi_0} \geq J_{\pi_1} \text{ - - - - - -}$$

ineq strict for at least one state, "$\exists i$ s.t. $J_{\pi_0}(i) > J_{\pi_1}(i)$"

, 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 $J = T_{\pi_k} J$

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

Current Policy π_k

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

Step 3: Policy improvement

Find a new policy π_{k+1} by

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

known (from step 2)

Bellman operator

next/improved policy

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

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)

~~Remark: (Finding a proper policy to start with)~~

~~Step 1 can be modified by starting with a finite J & going to step 3 directly. That would give a proper policy π_0, ($T_{\pi_0} J = TJ \Rightarrow \pi_0$ is proper) & from thereon do step 2 & 3 in tandem until convergence.~~

This construction won't work for finding a proper policy

Lecture-14

Claim: Policy improvement does just that.

Let π, π' be two proper policies s.t.

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

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

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

Pf: We know that $J_\pi = T_\pi J_\pi$ → J_π is a fixed point of T_π

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

So, This is $J_\pi = T_\pi J_\pi$

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

$T_\pi J_\pi \ge T J_\pi$ →

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

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

eqn (A) → $\boxed{J_\pi \ge T_{\pi'} J_\pi} \ge T_{\pi'}^2 J_\pi \ge \cdots \ge T_{\pi'}^k J_\pi$

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

$\Rightarrow J_\pi \ge J_{\pi'}$ — $(**)$ So, $(*)$ is proved.

To prove: π is optimal if the inequality in (**) isn't strict for at least one state

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

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

$$T_{\pi'} J_\pi = T J_\pi \quad \text{by construction} \; — (2)$$

Using (1) & (2), we get

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

$\Rightarrow$ π is optimal since T is the Bellman operator, which has a unique fixed point

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

$\blacksquare$

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 → $T_{\pi_k} J_k = T J_k$ → as in PI

Policy evaluation by m_k steps of VI → $J_{k+1} = T_{\pi_k}^{m_k} J_k$ → Approx. policy evaluation by applying T_{π_k} "m_k" times

Two special cases:

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

If $m_k = 1$, Modified PI = VI

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

Recommended: Choose $m_k > 1$

Convergence: Modified PI converges

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

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:

Policy evaluation

(Way 1)

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

(Way 2)

(+) is like policy improvement on subset S_k

$$\pi_{k+1}(i) = \begin{cases} \underset{a \in A(i)}{\arg\min} \overbrace{\sum_j P_{ij}^{(a)}\left(g(i, a, j) + J_k(j)\right)}^{TJ_k}, & \text{if } i \in S_k \\ \pi_k(i) & \text{else} \end{cases} \qquad (+)$$

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.

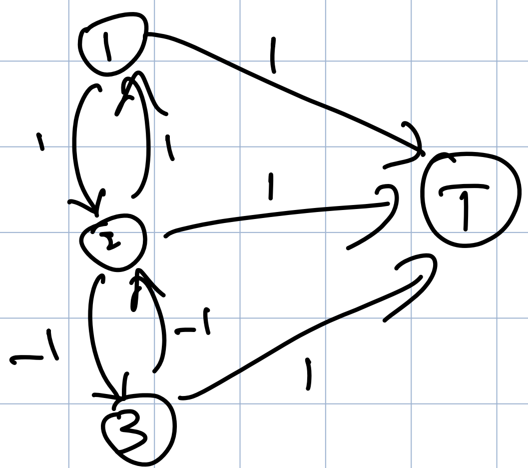
Aside! Value iteration with Q-factors leads finally to Q-learning (in an RL setting)

PI leads to "Actor-critic methods".

Discussion of some problems

① SSP $\exists$ one proper policy
Each improper π has $J_\pi(i) = \infty$ for at least one i

Claim: $\nexists$ improper π' s.t. $J_{\pi'}(i) = -\infty$ for some j.



→ This is not a valid counterexample

Claim is indeed true.

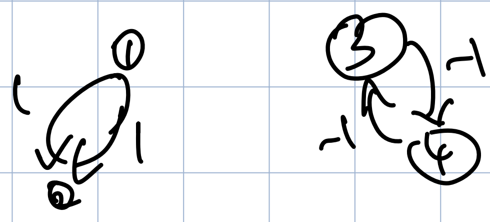
Proof by contradiction.

$\exists$ Proper π_1 & $\exists$ improper π_2 s.t. $J_{\pi_2}(j) = -\infty$.

Manufacture π_3 that follows π_1 everywhere except j.

$\nexists$ any state i s.t. $J_{\pi_2}(i) = \infty$. → violates assumption.

"$\exists$ one proper policy" ← why is this needed?



T

---

(2) f is a contraction mapping with modulus α.

$$f(x^*) = x^*$$

$$\boxed{f(x) \le x \implies x^* \le x} \quad \text{— (1)}$$

<u>One proof for (1):</u>

$$f(f(x)) = f^2(x) \le f(x) \le x$$ ← requires monotonicity

$$\implies \lim_{k \to \infty} f^k(x) \le x$$

$$\implies x^* \le x$$

Monotonicity of f: $x \le y \implies f(x) \le f(y)$

H.W. Think of a counterexample.

$$J_{N-1}(x) = \min_{a} E\Big(g(\quad) + \underbrace{J_N(x')}\Big)$$

$$\leq \min_{a} E\Big(g(\quad) + J_N'(x')\Big)$$

$$= J_{N-1}'(x)$$