Neuro Dynamic Prog Ch-2

SSPP :-

* There exists a cost free transition state 0.

$$p_{00}(u) = 1, \quad g(0,u,0) = 0 \quad \forall \; u \in A(0).$$

$$\gamma = 1. \quad (\text{No discounting}).$$

Problem :- How to reach terminal state with minimum expected cost ?

Non-terminal states :- $1, 2, \cdots n$.

Let

$$J_\mu(i) = \lim_{N\to\infty} E_\mu \left[\sum_{m=0}^{N-1} g(S_m, \mu(S_m), S_{m+1}) \,\middle|\, S_0 = i \right]$$

Here μ is a stationary policy.
- could be deterministic or randomized.

$$\pi = \{ \mu_0, \mu_1, \mu_2, \cdots \}$$

$$\mu_k(i) \in A(i) \quad \forall k, \forall i.$$

* In a stationary policy,

$$\mu_0 = \mu_1 = \mu_2 = \cdots = \mu.$$

$$\Rightarrow J_\mu(i) = \lim_{N\to\infty} E_\mu \left[\sum_{m=0}^{N-1} g(S_m, \mu(S_m), S_{m+1}) \,\middle|\, S_0 = i \right].$$

Since costs are bounded :-

$$\max_{i,j} |g(i, \mu(i), j)| \leq K < \infty$$

★ Definitions :-

(i) A stationary policy μ is said to be proper if

$$\rho_\mu = \max_{i=1,\dots n} P(S_n \neq 0 \mid S_n = i, \mu) < 1$$

(ii) A stationary policy that is not proper is called improper.

Note :- μ is proper $\Leftrightarrow$ In the Markov chain corresponding to μ, there exists a positive prob. path from each state to terminal state.

$$P(S_n \neq 0 \mid S_0 = i, \mu) = P(S_{2n} \neq 0 \mid S_n \neq 0, S_0 = i, \mu) * P(S_n \neq 0 \mid S_0 = i, \mu)$$

$\leq \rho_\mu$; $P(S_n \neq 0 \mid S_0 = i, \mu) \leq \rho_\mu$

$$\leq \rho_\mu^2$$

More generally,

$$P(S_K \neq 0 \mid S_0 = i, \mu) \leq \rho_\mu^{[K/n]} \qquad i = 1, \dots n.$$

Suppose :- $n < k < 2n$.

$$P(S_K \neq 0 \mid S_0 = i, \mu)$$

$$= P(S_K \neq 0 \mid S_\mu \neq 0, S_0 = i, \mu) * P(S_n \neq 0 \mid S_0 = i, \mu)$$

≤ 1 ; $\leq \mu$.

$$\leq \rho_\mu$$

Page No.
Date / /20

Prob. of reaching terminal state after k states $\longrightarrow 1$ as $k \to \infty$ regardless of the initial state

Recall that :-

$$J_\mu(i) = \sum_{m=0}^{\infty} E_\mu [g(s_m, \mu(s_m), s_{m+1}) \mid s_0 = i]$$

$$|J_\mu(i)| \leq \sum_{m=0}^{\infty} E_\mu [|g(s_m, \mu(s_m), s_{m+1})| \mid s_0 = i]$$

$$= \sum_{m=0}^{N} \sum_{j} \sum_{k} p_{ij}^{m}(\mu(i)) \cdot p_{jk}(\mu(j)) |g(j, \mu(j), k)|$$

If $\sum_k p_{jk}(\mu(j)) |g(i, \mu(j), k)| = |\bar{g}_\mu(j)|$

$$= \sum_{m=0}^{\infty} \sum_{j}$$

31/Jan/2023

Neuro-dynamic
Ch-2

Lec-7

★ SSPP or Episodic problems :-

$S = NT \cup T$

NT = set of non-terminal states.
T = ---- terminal states.

$NT = \{1, 2, \ldots n\}$, $T = \{0\}$

Proper policy :- μ is proper if

$$\rho_\mu = \max_{i \in NT} P\{S_n \neq 0 / S_0 = i\} < 1$$

Let $J = (J(0), J(1), \ldots J(n))$, where $J(0) = 0$

$$(TJ)(i) = \min_{\mu \in A(i)} \sum_{j \in S} p_{ij}(\mu)\,(g(i,u,j) + J(j))$$

$$= \min_{\mu \in A(i)} \left(p_{i0}(u)\, g(i,u,0) + \sum_{j=1}^{n} p_{ij}(u)\,(g(i,u,j) + J(j)) \right)$$

Suppose $\bar{g}(i,u) = \sum_{j \in S} p_{ij}(u) \cdot g(i,u,j)$; $\min_{\mu \in A(i)} \left(\bar{g}(i,u) + \sum_{j=1}^{n} p_{ij}(u) \cdot J(j) \right)$

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

Likewise define an operator T_μ $R^{n+1} \to R^{n+1}$ as follows :-

$$(T_\mu J)(i) = \left(\bar{g}(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j) \right)$$

let $P_\mu = \begin{bmatrix} p_{1,1}(\mu(1)) & p_{1,2}(\mu(1)) & \cdots & p_{1,n}(\mu(1)) \\ p_{2,1}(\mu(2)) & p_{2,2}(\mu(2)) & \cdots & p_{2,n}(\mu(2)) \\ \vdots & & & \\ p_{n,1}(\mu(n)) & p_{n,2}(\mu(n)) & & p_{n,n}(\mu(n)) \end{bmatrix}$

Page No.
Date / /

Note :-

$$\sum_{j=1}^{\Lambda} p_{ij}(\mu(i)) \leq 1 \quad \forall i.$$

Let $\bar{g}_\mu = (\bar{g}(i, \mu(i)),\ i \in NT)$

Then, $T_\mu J = \bar{g}_\mu + P_\mu J$.

Let $(T^k J)(i) = T(T^{k-1} J)(i)$.

$$(T^2 J)(i) = T(TJ)(i)$$

$$= \min_{\mu \in A(i)} \left(\bar{g}(i, u) + \sum_{j=1}^{n} p_{ij}(u) (TJ)(i) \right).$$

$$= \min_{\mu \in A(i)} \left(\bar{g}(i, u) + \sum_{j=1}^{n} p_{ij}(u) \left(\min_{v \in A(j)} \left(\bar{g}(j, v) + \sum_{k=1}^{k} p_{jk}(v) J(k) \right) \right) \right)$$

Note :- $(T^k J)(i)$ = Optimal cost-to-go for a k-stage finite horizon problem with initial state i, one-stage cost $\bar{g}$ and terminal cost J.

★ Monotonicity Lemma :-

For any $J, \bar{J} \in \mathbb{R}^{n+1}$ s.t.

$$J(i) \leq \bar{J}(i) \quad \forall\ i = 1, \ldots, n.$$

$$\& \ J(0) = \bar{J}(0) = 0.$$

and any stationary μ. → what is stationary μ? same μ for both $J, \bar{J}$.

(i) $(T^k J)(i) \leq (T^k \bar{J})(i) \quad \forall\ 1 \leq i \leq n.$

(ii) $(T_\mu^k J)(i) \leq (T_\mu^k \bar{J})(i) \quad \forall\ 1 \leq i \leq n$

Proof :-

Consider $k=1$,

$$(TJ)(i) = \min_{u \in A(i)} \left(\bar{g}(i,u) + \sum_{j=1}^{n} p_{ij}(u) \cdot J(j) \right) \leq$$

$$\leq \min_{u \in A(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) \bar{J}(j) \right\} \leq (T\bar{J})(i)$$

Assume above holds for $k=K$.

$$\Rightarrow (T^K J)(i) \leq (T^K \bar{J})(i) \quad \forall\ i=1,2,\ldots n$$

for $k = K+1$,

$$(T^{K+1}J)(i) = \min_{u \in A(i)} \left(\bar{g}(i,u) + \sum_{j=1}^{n} p_{ij}(u)(T^K J)(j) \right)$$

$$\leq (T^{K+1}\bar{J})(i) \quad \forall\ i = 1, \ldots n.$$

By induction the claim follows.

Lemma :- (1)

$\forall\ k \geq 0$ Vectors J, stationary μ. and $r \geq 0$,

(i) $T^k(J + re)(i) \leq (T^k J)(i) + r$, $i = 1, \ldots, n$

(ii) $T_\mu^k(J + re)(i) \leq (T_\mu^k J)(i) + r$, $i = 1, \ldots n$.

$J = \bar{J}(i)$, $i = 0, \ldots$

$J(0) = 0$.

The inequalities are reversed for $r < 0$.

Here, $e = (1, 1, \ldots 1)^T$ is the $(n+1)$ vector of all i's.

Page No.
Date / /

Note that :-

$$T(J+re)(i) = \min_{\mu \in A(i)} \left(\bar{g}(i,u) + \sum_{j=1}^{n} p_{ij}(u)\left(J + re\right)(j) \right)$$

$$= \min_{\mu \in A(i)} \bar{g}(i,u) + \sum_{j=1}^{n} p_{ij}(u) J(j) + \underbrace{\sum_{j=1}^{n} p_{ij}(u)\, r e(j)}_{\leq r}$$

$$\leq (TJ)(i) + r.$$

Rest can be shown by induction.

Assumptions :- (A) $\exists$ at least 1 proper policy.

(B) $\forall$ improper μ, $J^{\mu}(i) = \infty$ for at least one state $i \in NT$.

× × ×

Aside :-

$$J_{\mu}(i) = E_{\mu}\left[\sum_{k=0}^{\infty} g(x_k, \mu(x_k), x_{k+1}) \,\middle|\, x_0 = i \right]$$

$$J^*(i) = \min_{\mu} J_{\mu}(i) \quad \forall\, i = 1, \ldots n.$$

We shall show that $T^k J \longrightarrow J^*$

$T_{\mu}^k J \longrightarrow J_{\mu}$ as $k \to \infty$.

$$(TJ)(i) = \min_{\text{of } \mu \in A(i)} \left(\bar{g}(i,\mu) + \sum_{j=1}^{n} p_{ij}(u) J(j) \right)$$

$$\underbrace{J^*(i) = (TJ^*)(i)}_{\text{We shall show this}} = \min_{u \in A(i)} \left(\bar{g}(i,u) + \sum_{j=1}^{n} p_{ij}(u) J^*(j) \right)$$

Bellman Equation for Optimality

Similarly, $J_{\mu}(i) = (T_{\mu} J_{\mu})(i)$

Bellman eqn for μ.

★ Proposition 1:

(a) For a proper policy μ, the associated cost vector J_μ satisfies $(T_\mu^k J)(i) \to J_\mu(i)$ as $k \to \infty$ $\forall i = 1, \ldots n$

for every vector J also,

$J_\mu = T_\mu J_\mu$ and J_μ is the unique soln to this equation.

(b) A stationary policy μ satisfying for some vector J,

$J(i) \geq (T_\mu J)(i)$ $\forall i = 1, \ldots n$. is proper.

Proof:-

(a) Recall that:-

$$T_\mu J = g_\mu + P_\mu J$$

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

$$= g_\mu + P_\mu (g_\mu + P_\mu J)$$

$$\Rightarrow T_\mu^2 J = g_\mu + P_\mu g_\mu + P_\mu^2 J$$

Proceeding similarly:-

$$T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu.$$

Note:- $(P_\mu^k J)(i) = \sum_{j=1}^{n} P(x_k = j \mid x_0 = i, \mu)\, J(j)$

$$\leq \sum_{j=1}^{n} P(x_k = j \mid x_0 = i, \mu) \left(\max_{j=1,\ldots n} J(j) \right)$$

$$= P(x_k \neq 0 \mid x_0 = i, \mu) \max_{j=1,\ldots n} J(j).$$

$$\leq P_\mu^{(k/n)} \cdot \max_{j=1,\ldots n} J(j)$$

Page No.
Date / /20

Thus as $k \to \infty$ the term $\to 0$ $\because \rho_\mu < 1$

$$\therefore \lim_{k\to\infty} [T_\mu{}^k J] = \lim_{k\to\infty} \left\{ \sum_{m=1}^{k-1} P_\mu^m g_\mu \right\} = J_\mu.$$

By definition:

$$T_\mu{}^{k+1} J = g_\mu + P_\mu (T_\mu{}^k J)$$

Letting $k \to \infty$ on either side

$$J_\mu = g_\mu + P_\mu J_\mu \quad \text{or} \quad J_\mu = T_\mu J_\mu.$$

Proof of Uniqueness — By contradiction

Suppose $\bar{J}$ is one more fixed pt. of T_μ.

$\Rightarrow T_\mu \bar{J} = \bar{J}$

$\Rightarrow T_\mu^2 \bar{J} = T_\mu \bar{J} = \bar{J}$

$\vdots$

$T_\mu^n \bar{J} = \bar{J}$

$\vdots$

$\lim_{n\to\infty} (T_\mu^n \bar{J}) = \bar{J}$

$\|$

J_μ

Thus $J_\mu = \bar{J}$

(b) Given that for a stationary μ & same J.

$$J(i) \geq (T_\mu J)(i) \quad \forall\, i = 1, \ldots n.$$

or $J \geq T_\mu J$

Applying T_μ repeatedly on either side

$$T_\mu J \geq T_\mu^2 J \geq T_\mu^3 J \geq \cdots\cdots T_\mu^k J.$$

$$= P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu \to J_\mu$$

$\Rightarrow J(i) \geq J_\mu(i) \quad \forall\, i.$

Now if μ is not proper $J_\mu(i) = \infty$ for at least one i $\Rightarrow J(i) = \infty$ for that i.

Converse happens since each $J(i) \in \mathbb{R} \Rightarrow \mu$ is proper.

Proposition 3 :-

(a) The optimal cost vector J^* satisfies the Bellman Equation $J^* = TJ^*$ & moreover J^* is the unique soln to the eqn.

(b) We have $\lim_{k \to \infty} (T^k J)(i) = J^*(i) \quad \forall \; i = 1, \ldots n$ for every vector J.

(c) A stationary play μ is optimal if & only if $T_\mu J^* = TJ^*$.

6/2/2023

Page No.
Date / /20

Lec-9*

Proposition : 2

(a) The optimal cost vector J^* satisfies the Bellman eqn $J^* = TJ^*$. Moreover, J^* is the unique soln to this eqn.

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

for every vector J.

(c) A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$

Proof (a), (b) :-

We first show that T has atmost one fixed point. Suppose these are two fixed points J and J' and suppose μ and μ' are such that :-

$J = TJ = T_\mu J$ and.

$J' = TJ' = T_{\mu'} J'$

$$TJ(i) = \min_{\mu \in A(i)} \sum_{j \in S} p_{ij}(\mu) \left(g(i, \mu, j) + J(j) \right).$$

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

By Proposition-1, μ and μ' are proper policies and $J = J_\mu$ and $J' = J_{\mu'}$

Now, $J = TJ = T^2 J = \cdots = T^k J$.

Also, $J = TJ \le T_{\mu'} J \le T_{\mu'}^2 J \le \cdots \le T_{\mu'}^k J$.

$(J \leq T_{\mu'} J \leq T_{\mu'}^2 J \leq \cdots)$

Thus, $J = T^k J \leq T_{\mu'}^k J \quad \forall\ k \geq 1.$

Thus as $k \to \infty$, we obtain $J \leq \lim_{k\to\infty} T_{\mu'}^k J = J_{\mu'} = J'$

A symmetric argument then shows that $J' \leq J$. In other words $J = J'$ or $J_\mu = J_{\mu'}$.

Thus "T" has atmost one fixed point.

* We now show that T has at least one fixed point:-

Let μ be a proper policy. By Assumption (A), $\exists$ a proper policy.
Suppose μ' be another policy s.t.

$$\boxed{T_{\mu'} J_\mu = T J_\mu}$$

$$(T_{\mu'} J_\mu)(i) = \sum_{j \in S} p_{ij}(\mu'(i)) \left(g(i, \mu'(i), j) + J_\mu(j)\right)$$

$$= \min_{\mu \in A(i)} \sum_{j \in S} p_{ij}(\mu) \left(g(i, \mu, j) + J_\mu(j)\right)$$

$$= (T J_\mu)(i).$$

Then, $T_\mu J_\mu \geq T J_\mu = T_{\mu'} J_\mu$

Since $J_\mu \geq T_{\mu'} J_\mu$, by Proposition-1, μ' is a proper policy and $J_\mu \geq T_{\mu'} J_\mu \geq T_{\mu'}^2 J_\mu \geq \lim_{k\to\infty} T_{\mu'}^k J_\mu = J_{\mu'}$.

Repeating this argument, we obtain a seq $\{\mu^k\}$ s.t. each μ^k is proper and $J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq T J_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k} \geq \cdots \geq \lim_{n\to\infty} T_{\mu^{k+1}}^n$

$= J_{\mu^{k+1}}$.

Since the set of proper policies is finite, some policy μ must get repeated in the sequence.

Thus, $J_\mu = TJ_\mu$.

$\Rightarrow$ J_μ is a fixed point of T and by the uniqueness property, it is the unique fixed pt. of T.

* Next we show, that $J_\mu = J^*$ and $T^k J \to J^*$ as $k \to \infty$ $\forall\, J \in \mathbb{R}^n$.

Let $e = (1, 1, \ldots, 1)$ & $\delta > 0$ be a scalar.

Also let $\hat{J} \in \mathbb{R}^n$ be s.t $T_\mu \hat{J} = \hat{J} - \delta e$

rewrite

(Note :- $\hat{J} = T_\mu \hat{J} + \delta e$

$$\hat{J} = g_\mu + \delta e + P_\mu \hat{J}$$

Single state cost.

$\hat{J}$ is the cost vector corresponding to μ with g_μ replaced with $g_\mu + \delta e$.

Since μ is proper, $\hat{J}$ is unique. Further $J_\mu \le \hat{J}$.

Thus, $J_\mu = TJ_\mu \le T\hat{J} \le T_\mu \hat{J} = \hat{J} - \delta e \le \hat{J}$

$\Rightarrow J_\mu = T^k J_\mu \le T^k \hat{J} \le T^{k-1}\hat{J} \le \ldots \le \hat{J} \quad \forall\, k \ge 1.$

Thus, $\{T^k \hat{J}\}$ forms a bounded monotone sequence and it converges.

$T^k \hat{J} \to \breve{J}$ as $k \to \infty$ for some $\breve{J} \in \mathbb{R}^n$

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

We know that J_μ is a unique fixed point of T, i.e., $J_\mu = TJ_\mu$.
Since $\tilde{J}$ is also a fixed pt., it must happen that $\tilde{J} = J_\mu$.

Also, $J_\mu - \delta e = TJ_\mu - \delta e \le T(J_\mu - \delta e) \le TJ_\mu = J_\mu$.

seen in a prev. class.

$\Rightarrow T(J_\mu - \delta e) \le T^2(J_\mu - \delta e)$.

Thus $T^k(J_\mu - \delta e)$ is monotonically increasing and bounded above.

Also, $\lim_{k\to\infty} T^k(J_\mu - \delta e) = J_\mu$.

For any J, we can find $\delta > 0$ s.t.

$$J_\mu - \delta e \le J \le \hat{J}$$

$\hat{J}$: cost-to-go vector for policy μ but with single stage costs $g_\mu + \delta e$.

By monotonicity of T,

$$T^k(J_\mu - \delta e) \le T^k J \le T^k \hat{J}, \quad k \ge 1$$

In the limit,

$$J_\mu = \lim_{k\to\infty} T^k(J_\mu - \delta e) \le \lim_{k\to\infty} T^k J \le \lim_{k\to\infty} T^k \hat{J} \le J_\mu.$$

To show that $J_\mu = J^*$, take any policy $\pi = \{\mu_0, \mu_1 \ldots\}$

Then, $T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{k-1}} J_0 \geq T^k J_0$.

Where J_0 is an arbitrary vector. Taking lim sup as $k \to \infty$ on either side, we obtain

$$J_\pi \geq J_\mu \Rightarrow J_\mu \text{ is optimal.}$$

$\Rightarrow \mu$ is optimal stationary policy and $J_\mu = J^*$

(c) Note :- If μ is optimal, then $J_\mu = J^*$.

By Assumption (A)..(B), μ is proper.

Thus by prop. 1, $T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*$.

Conversely let $J^* = TJ^* =$ Then μ is proper and $J^* = J_\mu$

$\Rightarrow \mu$ is optimal.

9/2/2023

Lec-10

★ Recall:-

$$(TJ)(i) = \min_{\mu \in A(i)} \sum_{j \in S} p_{ij}(u)\left(g(i,u,j) + J(j)\right)$$

$i = 1, \ldots, n.$

we show that T is a contraction w.r.t. some norm.

$\exists$ a vector $\xi = (\xi(1), \ldots, \xi(n))$ s.t. $\xi(i) > 0$.

$\forall\ i = 1 \ldots n$ and a scalar $\beta \in (0,1)$ s.t.

$$\|J\|_{\xi} \triangleq \max_{i=1\ldots n} \left(\frac{|J(i)|}{\xi(i)}\right)$$

and $\|TJ - T\bar{J}\|_{\xi} \leq \beta \|J - \bar{J}\|_{\xi}$

$\forall\ J, \bar{J} \in \mathbb{R}^n$.

Aside:-

$(TJ) = ((TJ)(1), \ldots (TJ)(n))$

$(T\bar{J}) = ((T\bar{J})(1), \ldots (T\bar{J})(n))$

we say that T is a contraction w.r.t. some norm $\|\cdot\|$ if $\|TJ - T\bar{J}\| \leq \alpha \|J - \bar{J}\|$ for some $\alpha \in (0,1)$

★ Proposition-3:- Assume all stationary policies are proper. Then $\exists$ a vector ξ with $\xi(i) > 0\ \forall\ i$ s.t. the mappings T and T_μ are contractions w.r.t. $\|\cdot\|_{\xi}$.

Proof:-

Consider a new SSPP where transition probabilities are same as original but transition costs are all equal to -1 except that upon termination costs are 0.

Page No.
Date / /20

Let $\hat{J}(i)$ = optimal cost to go from state i in the new problem.

Since, $J^* = TJ^*$, thus.

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

$$\leq -1 + \sum_{j=1}^{n} p_{ij}(\mu(i)) \cdot \hat{J}(j).$$

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

Then $\forall i$, $\xi(i) \geq 1$ and for all stationary μ.

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

$$\text{or } \xi(i) \geq 1 + \sum_{j=1}^{n} p_{ij}(\mu(i))\,\xi(j).$$

$$\text{or } \sum_{j=1}^{n} p_{ij}(\mu(i))\,\xi(j) \leq \xi(i) - 1 \leq \beta \cdot \xi(i), \ \forall i.$$

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

Now for any stationary policy μ, state i, vectors $J, \bar{J} \in \mathbb{R}^n$.

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

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

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

$$= \|J - \bar{J}\|_\xi$$

$$\leq \beta\, \xi(i)\, \|J - \bar{J}\|_\xi$$

$$\Rightarrow \frac{|(T_\mu J)(i) - (T_\mu \bar{J})(i)|}{\xi_q(i)} \leq \beta \|J - \bar{J}\|_{\xi_q}$$

$$\forall\ i = 1, \ldots n.$$

$$\therefore \max_{i=1,\ldots n} \left(\frac{|(T_\mu J)(i) - (T_\mu \bar{J})(i)|}{\xi_q(i)} \right)$$

$$\leq \beta \|J - \bar{J}\|_{\xi_q}$$

or

$$\|T_\mu J - T_\mu \bar{J}\|_{\xi_q} \leq \beta \|J - \bar{J}\|_{\xi_q}$$

Note also that

$$(T_\mu J)(i) \leq (T_\mu \bar{J})(i) + \beta \xi_q(i) \max_{j=1,\ldots n} \left(\frac{|J(j) - \bar{J}(j)|}{\xi_q(j)} \right)$$

Taking min over μ on both sides

$$(TJ)(i) \leq (T\bar{J})(i) + \beta \xi_q(i) \max_{j=1,\ldots n} \left(\frac{|J(j) - \bar{J}(j)|}{\xi_q(j)} \right)$$

Interchanging J and $\bar{J}$, we also obtain :-

$$(T_\mu \bar{J})(i) \leq (T_\mu J)(i) + \beta \xi_q(i) \max_{i=1,\ldots n} \left(\frac{|J(j) - \bar{J}(j)|}{\xi_q(j)} \right)$$

Again taking min over μ on either side :-

$$(T\bar{J})(i) \leq (TJ)(i) + \beta \xi_q(i) \max_{j=1,\ldots n} \left(\frac{|J(j) - \bar{J}(j)|}{\xi_q(j)} \right)$$

It is also easy to see that :-

$$\|TJ - T\bar{J}\|_{\infty} \leq \beta \|J - \bar{J}\|_{\infty}$$

★ Numerical approaches for solving Bellman equation

1. Value Iteration :-

(a.) Start with an arbitrary $J_0 = J \in \mathbb{R}^n$.
(b.) Iterate as $J_{k+1} = TJ_k$, $k = 0, 1, 2, \ldots$

In other words,

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

Now, $J_k \longrightarrow J^*$ as $k \to \infty$

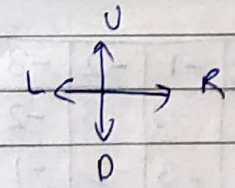
where

$$J^*(i) = \min_{\mu \in A(i)} \sum_{j \in S} p_{ij}(\mu) \left(g(i, \mu, j) + J^*(j) \right)$$

$i = 1, \ldots n$

Ex:- Grid World (Sutton's Ch. 4).

▨	1	2	3
4	5	6	7
8	9	10	11
12	13	14	▨



* Non-terminal state = $\{1, 2, \ldots 14\}$.

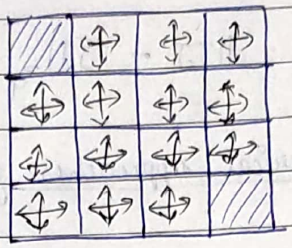
* Terminal states = ▨

* Rewards = -1 in non-terminal states.

Apply value iteration :-

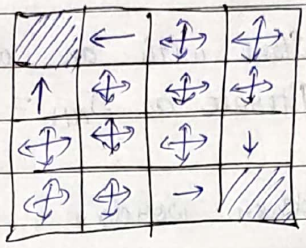
$k=0$

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0



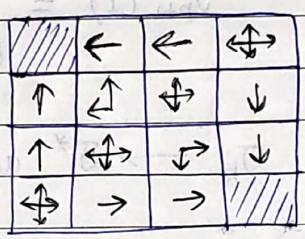
$k=1$

0	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	0



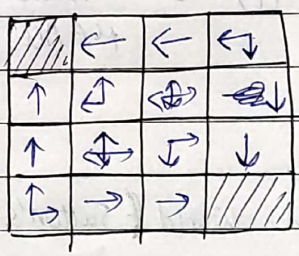
$k=2$

0	-1	-2	-2
-1	-2	-2	-2
-2	-2	-2	-1
-2	-2	-1	0



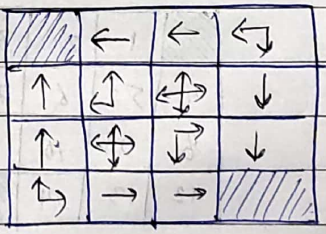
$k=3$

0	-1	-2	-3
-1	-2	-3	-2
-2	-3	-2	-1
-3	-2	-1	0



$k=4$

0	-1	-2	-3
-1	-2	-3	-2
-2	-3	-2	-1
-3	-2	-1	0



Converged.

Page No.
Date / /20

★ Policy iteration :- Iterate over policies.

updates policies via a 2-step procedure

outer loop.
{ policy update (policy improvement).

Inner loop.
{ find value of new policy (policy evaluation)
}

Go to outer loop.

}.

★ Procedure :-

1. Start with a proper policy μ_0.

2. Policy evaluation :- Given policy μ_k, Compute $J^{\mu_k}(i)$. $i \in S$.

as the sol^n to :

$$J(i) = \sum_{j=1}^{n} p_{ij}(\mu_k(i)) \left(g(i, \mu_k(i), j) + J(j) \right)$$

$i = 1, \dots n.$

3. Policy improvement : Find a new policy μ_{k+1} s.t.

$$\mu_{k+1}(i) = \underset{\mu \in A(i)}{\operatorname{argmin}} \left\{ \sum_{j=1}^{n} P_{ij}(u) \left[g(i, u, j) + J^{\mu_k}(j) \right] \right\}$$

$i = 1, \dots n.$

[OR] $T_{\mu_{k+1}} J^{\mu_k} = T J^{\mu_k}$

$$\left(T_{\mu_{k+1}} J^{\mu_k}\right)(i) = \sum_{j=1}^{n} P_{ij}(\mu_{k+1}(i)) \left[g(i, \mu_{k+1}(i), j) + J^{\mu_k}(i) \right]$$

$$= \min_{u \in A(i)} \sum P_{ij}(u) \left[g(i, \mu_{k+1}(i), j) + J^{\mu_k}(j) \right]$$

Lec:

Result:

Suppose $T: \mathbb{R}^d \to \mathbb{R}^d$ is a contraction.

Thus, $\|Tx - Ty\| \leq \alpha \|x - y\|$, for any $x, y \in \mathbb{R}^d$ and $0 < \alpha < 1$

Note that,

$$\|T^2x - T^2y\| \leq \alpha \|Tx - Ty\| \leq \alpha^2 \|x - y\|$$

where $T^2 = T \circ T$

Proceeding similarly, $\|T^n x - T^n y\| \leq \alpha^n \|x - y\|$

Thus the sequence $\{T^n x\}$ is cauchy. Thus it has a limit $\bar{y}$

$\Rightarrow T^n x \longrightarrow \bar{y}$ as $n \to \infty$.

We say that $\{x_n\}$ is cauchy if given any $\epsilon > 0$

$$\exists\ N > 0,\ \forall\ m, n > N$$

$$\|x_m - x_n\| < \epsilon$$

$$\Rightarrow \lim_{n \to \infty} T^n x = \bar{y}$$

applying T on both sides we obtain :-

$$T\left(\lim_{n\to\infty} T^n x\right) = T\bar{y}$$

$$\| \quad \text{(by continuity of } T\text{)}$$

$$\lim_{n\to\infty} T^{n+1}x = T\bar{y}$$

$$\underbrace{\lim_{n\to\infty} T^{n+1}x}_{\bar{y}} = T\bar{y}$$

$\Rightarrow T\bar{y} = \bar{y}$ or that $\bar{y}$ is a fixed pt. of T.

Suppose $\bar{z}$ is one more fixed pt of T, i.e.

$T\bar{z} = \bar{z}$. By contraction property

$$\|T\bar{z} - T\bar{y}\| \leq \alpha \|\bar{z} - \bar{y}\|$$

Since, $T\bar{z} = \bar{z}$ & $T\bar{y} = \bar{y}$, we have

$$\|\bar{z} - \bar{y}\| \leq \alpha \|\bar{z} - \bar{y}\|$$

Cannot happen unless $\bar{z} = \bar{y}$

$\Rightarrow T$ has a unique fixed pt.

Another way,

$$\|T^n x - T^n y\| \leq \alpha^n \|x - y\|$$

Suppose we have $\bar{y}$ in place of $y \Rightarrow T^n\bar{y} = \bar{y}$

$$\Rightarrow \|T^n x - \bar{y}\| \leq \alpha^n \|x - \bar{y}\| \Rightarrow T^n x \to \bar{y} \text{ as } n \to \infty$$

★ Policy iteration

Procedure:-

1. Start with a proper policy μ_0.

2. Policy evaluation :- Given policy μ_k, ~~complete~~ compute $J^{\mu_k}(i)$, $i \in S$, as the solution to

$$J(i) = \sum_{j=1}^{n} p_{ij}(\mu_k(i)) \left(g(i, \mu_k(i), j) + J(j)\right)$$

$i = 1, \ldots, n$.

3. Policy improvement :- Find a new policy μ_{k+1} s.t.

$$\mu_{k+1}(i) = \underset{\mu \in A(i)}{\operatorname{argmin}} \sum_{j \in S} p_{ij}(\mu) \left(g(i, u, j) + J^{\mu_k}(j)\right),$$

$i \in S$.

In other words, we find μ_{k+1} s.t.

$$T_{\mu_{k+1}} J^{\mu_k} = T J^{\mu_k}.$$

Go back to (2), find $J^{\mu_{k+1}}$ and repeat process if $J^{\mu_{k+1}}(i) < J^{\mu_k}(i)$ for at least one $i \in S$. If $J^{\mu_{k+1}}(i) = J^{\mu_k}(i)$. $\forall i \in S$, then stop and output the optimal policy-value pair as (μ_k, J^{μ_k}).

Date / /20

Proposition :- The policy iteration algorithm generates an improving sequence of proper policies, i.e., $J^{\mu_{k+1}}(i) \leq J^{\mu_k}(i)$ $\forall\, i \in S$, and $\forall\, k$. Further, it terminates with an optimal policy in a finite no. of steps.

Proof :- Given a proper policy μ, new policy $\bar{\mu}$ is optimal as $T_{\bar{\mu}} J^{\mu} = TJ^{\mu}$. Then,

$$J^{\mu} = T_{\mu} J^{\mu} \geq TJ^{\mu} = T_{\bar{\mu}} J^{\mu}$$

$$\Rightarrow J^{\mu} \geq T_{\bar{\mu}} J^{\mu}.$$

Repeatedly applying $T_{\bar{\mu}}$ on either side we get,

$$J^{\mu} \geq T_{\bar{\mu}} J^{\mu} \geq T_{\bar{\mu}}^2 J^{\mu} \geq \ldots \geq T_{\bar{\mu}}^k J^{\mu} \geq \ldots \geq \lim_{k \to \infty} \left(T_{\bar{\mu}}^k J^{\mu}\right)$$

$$\Rightarrow J^{\bar{\mu}}(i) \leq J^{\mu}(i) \qquad \forall\, i \in S.$$

Since μ is proper, it follows that $\bar{\mu}$ is also proper. Since $\not\exists\, i$ s.t. $J^{\bar{\mu}}(i) = \infty$.

$\Rightarrow J^{\mu}(i) = \infty$ which cannot happen since μ is proper.

If μ is not optimal, then, $J^{\bar{\mu}}(i) < J^{\mu(i)}$ for at least one $i \in S$.

Else, $J^{\mu} = J^{\bar{\mu}} = T_{\bar{\mu}} J^{\bar{\mu}} = T_{\bar{\mu}} J^{\mu} = TJ^{\mu} = TJ^{\bar{\mu}}$

$$\Rightarrow J^{\mu} = TJ^{\mu}$$
$$\& \quad J^{\bar{\mu}} = TJ^{\bar{\mu}}$$

$$\Rightarrow J^{\mu} = J^{\bar{\mu}} = J^*$$

In this case, μ is optimal and so is $\bar{\mu}$ & the optimal cost is $J^{\mu} = J^{\bar{\mu}}$.

Since, no. of proper policies is finite, convergence will happen in a finite no of steps.

★ Modified PI :-

- Start with μ_0.
- PE :- Run value iteration for a-priori chosen, m_k steps at iteration k.
  i.e. Solve $J_k = T_{\mu_k}^{m_k} J_k$ for given m_k.
- PI :- Find μ_{k+1} s.t.

$$T_{\mu_{k+1}} J_k = T J_k.$$

★ Syllabus for Mid term I :-

(15 Marks) (4 ques) (3+1).

- Intro to RL, Multi-armed Bandits (Sutton + Barto).
- Finite horizon problems (Bertsekas, Vol. I Optimal Control & DP.) (Ch-1)
- SSPP (Bertsekas, Vol. II → Ch-2) ....
