Theoretical foundations of RL

Not: Deep RL

Today: Theory of Markov decision processes (MDPs)

Tommorow: Reinforcement learning (RL)

Finite horizon MDPs:-

Example 1: Machine replacement/repair

States = $\{1, 2, \ldots n\}$

1 ↓ good, n ↓ worst

Actions = Do nothing, Repair

On repair, m/c goes to "1". Cost = R.

Operating cost: $g(i)$, $i = 1 \ldots n$

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

References:-

(1) My webpage ↳ CS6700

(2)

(i) "Neuro-dynamic Programming", Tsitsiklis & Bertsekas

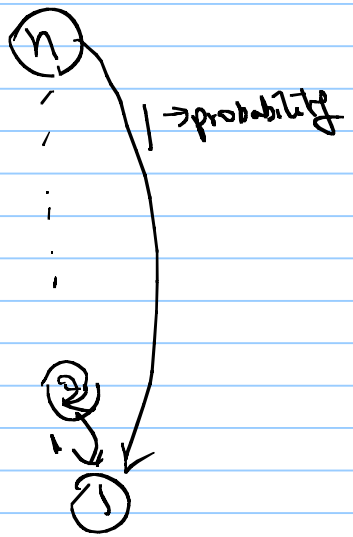
(ii) "Introduction of RL" Barto & Sutton

(iii) Bertsekas "Dynamic programming & optimal control" Vol I & II.

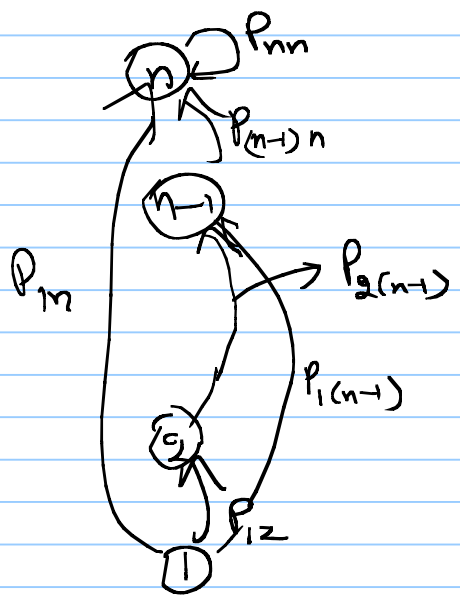
State transitions:- P_{ij} → Probability of state transitioning from i to j.

$P_{ij} = 0$ if $j < i$

Action: repair



Action: Do nothing



Example 2: Chess match

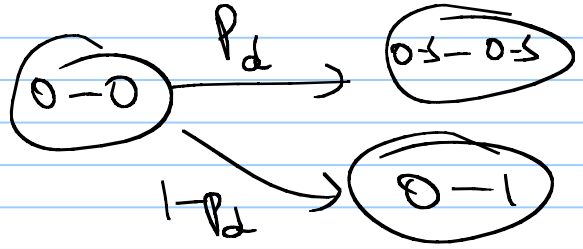
Playing styles (actions): Timid, Bold

Timid → draw: P_d, lose: $1-P_d$

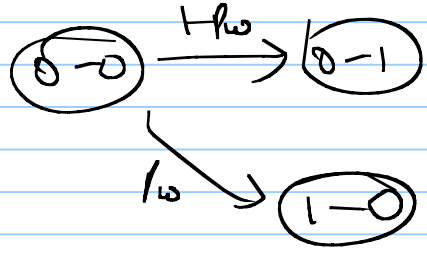
Bold → win: P_w, lose: $1-P_w$

$P_d > P_w$

Game 1

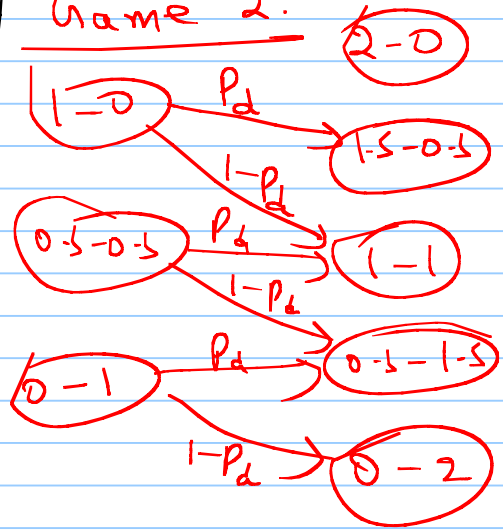


Timid



Bold

Game 2:



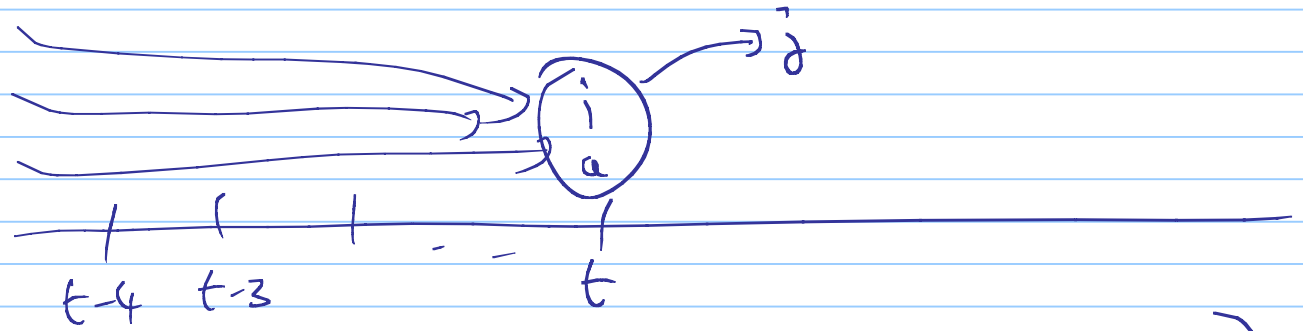
Timid

H.W: Transition diagram for bold play

Framework

S: State space, A: Action space

$$p_{ij}(a) = P(S_{k+1} = j \mid S_k = i, a_k = a)$$



Controlled Markov property

$$P(S_{k+1} = j \mid S_k = i, a_k = a, S_{k-1}, a_{k-1}, \dots, S_0) = p_{ij}(a)$$

Single-stage cost: $g_k(i, a, j)$

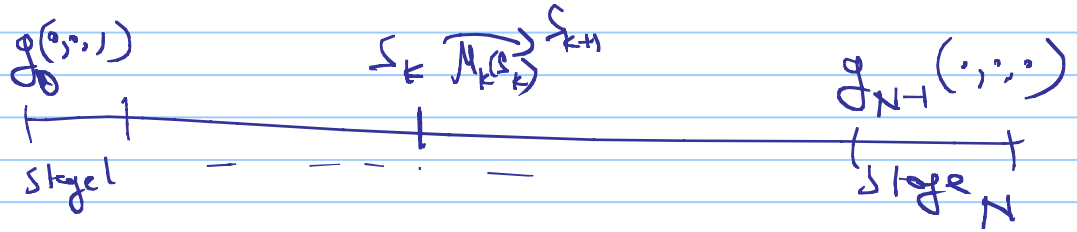
Policy π: $\{\mu_0, \cdots \mu_{N-1}\}$ $\quad \mu_i : S \to A$

S: state space, A: Action space

Performance metric

$$J_\pi(s_0) = E\left[g_N(s_N) + \sum_{k=0}^{N-1} g_k(s_k, \mu_k(s_k), s_{k+1}) \right]$$

g_N: Terminal cost



Goal: $J_{\pi^*}(s_0) = \min_\pi J_\pi(s_0)$

Optimality principle & dynamic programming (DP) algorithm

$$\pi^* = (\mu_0^*, \dots, \mu_{N-1}^*)$$

$$\min_{\pi^i = (\mu_i, \dots, \mu_{N-1})} E\left(g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1}) \right)$$

$(N-i)$ stage tail subproblem

For this tail problem, $\{\mu_i^*, \dots \mu_{N-1}^*\}$ is optimal.

DP algorithm:- Solve tail sub-problems, starting at the end, & progressing backwards.

DP algorithm

$$J_N(s_N) = g_N(s_N), \quad \forall s_N \in S.$$

For $k = 0, 1, \ldots, N-1$

$$J_k(s_k) = \min_{a_k \in A(s_k)} E_{s_{k+1}} \left[g_k(s_k, a_k, s_{k+1}) + J_{k+1}(s_{k+1}) \right]$$

$\forall s_k \in S$

Example:- Applying DP algo to "Machine replacement" example

Suppose terminal cost is zero. $S = \{1, \ldots, n\}$

$$J_N(i) = 0$$

$$J_k(i) = \min \left[\underbrace{R + g(i) + J_{k+1}(1)}_{\text{Repair}}, \underbrace{g(i) + \sum_{j=1}^{n} P_{ij} J_{k+1}(j)}_{\text{Do nothing}} \right]$$

Chess - example (revisited)

State = net-score = (#wins − #losses) N stages.

$$J_k(S_k) = \max \left(\underbrace{P_d J_{k+1}(S_k) + (1 - P_d) J_{k+1}(S_k - 1)}_{\text{Timid}}, \underbrace{P_w J_{k+1}(S_k + 1) + (1 - P_w) J_{k+1}(S_k - 1)}_{\text{Bold}} \right)$$

Play bold if

$$\frac{P_w}{P_d} \geq \frac{J_{k+1}(S_k) - J_{k+1}(S_k - 1)}{J_{k+1}(S_k + 1) - J_{k+1}(S_k - 1)}$$

Initial condition:

$$J_N(S_N) = \begin{cases} 1 & \text{if } S_N > 0 \\ P_w & \text{if } S_N = 0 \\ 0 & \text{if } S_N < 0 \end{cases}$$

S_{N-1}	$J_{N-1}(S_{N-1})$	
≥ 1	1	Best play does not matter
1	$\max(P_d + (1-P_d)P_w,\ P_w + (1-P_w)P_w) = P_d + (1-P_d)P_w$	So, timid play is best
0	P_w	bold
-1	P_w^2	bold

S_{N-1}	$J_{N-1}(S_{N-1})$	
≤ -1	0	Best play: does not matter

$$J_{N-2}(0) = \max\Big(\underbrace{P_d P_w + (1-P_d) P_w^2}_{\text{timid}},\ \underbrace{P_w\big(P_d + (1-P_d)P_w\big) + (1-P_w)P_w^2}_{\text{bold}} \Big)$$

$$= P_w\big(P_d + (1-P_d)P_w\big) + (1-P_w)P_w^2$$

Example: Job-scheduling

N jobs. Job i takes T_i time to complete

If job i is completed at time t, then the reward is $\boxed{\alpha^t R_i}$

$0 < \alpha < 1$

Schedule $L = \{i_0, i_1, \ldots, i_{k-1}, i, j, i_{k+2}, \ldots, i_{N-1}\}$

$L' = \{\ldots\ j, i, \ldots\}$

$$J_L = E\left[\alpha^{t_0} R_{i_0} + \cdots + \alpha^{t_{k-1}} R_{i_{k-1}} + \alpha^{t_{k-1}+T_i} R_i + \alpha^{t_{k-1}+T_i+T_j} R_j + \cdots + \alpha^{t_{N-1}} R_{i_{N-1}}\right]$$

$$J_{L'} = E\left[\alpha^{t_0} R_{i_0} + \cdots + \alpha^{t_{k-1}} R_{i_{k-1}} + \alpha^{t_{k-1}+T_j} R_j + \alpha^{t_{k-1}+T_j+T_i} R_i + \cdots + \alpha^{t_{N-1}} R_{i_{N-1}}\right]$$

So, L better than L' if

$$E\left[\alpha^{t_{k-1}+T_i} R_i + \alpha^{t_{k-1}+T_i+T_j} R_j\right] \geq E\left[\alpha^{t_{k-1}+T_j} R_j + \alpha^{t_{k-1}+T_j+T_i} R_i\right]$$

(or)

$$\frac{E(\alpha^{T_i})\, R_i}{1 - E(\alpha^{T_i})} \geq \frac{E(\alpha^{T_j})\, R_j}{1 - E(\alpha^{T_j})}$$

Optimal policy is an index-based policy

Infinite horizon discounted cost MDPs

States $= \{1, \dots n\}$

$\pi = \{\mu_0, \mu_1, \dots\}$ $\quad \mu_i : S \to A$

Stationary policy $\pi = \{\mu, \dots, \mu\}$ $\quad \mu : S \to A$.

$$J_\pi(i) = E\left(\sum_{k=0}^{\infty} \alpha^k \, g(s_k, \pi(s_k), s_{k+1}) \,\middle|\, s_0 = i \right]$$

- i: initial state.
- α^k: discount factor, $0 < \alpha < 1$
- g: single-stage cost (same for all stages)
- π: policy

Goal: $J^*(i) = \min_{\pi} J_\pi(i)$, $\forall i \in S$

$$J^* = [J^*(1), \ldots, J^*(n)]$$

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

Bellman-operator $TJ = ((TJ)(1), \ldots (TJ)(n))$

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

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

Let $P_\pi = \left[\left[p_{ij}(\pi(i)) \right] \right]_{i,j=1}^{n} = \begin{bmatrix} p_{11}(\pi(1)) & & \\ & \ddots & \\ & & p_{nn}(\pi(n)) \end{bmatrix}$

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

$$\boxed{T_\pi J = g_\pi + \alpha P_\pi J}$$

Properties of T, T_π : (I) For any n-vectors J, J' s.t. $J(i) \le J'(i), \forall i = 1-n$ and stationary policy π, we have

(1) $(T^k J)(i) \le (T^k J')(i), \forall i$

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

(II) Let e = n-vector of one's. For any J and scalar r,

(1) $(T^k(J + re))(i) = (T^k J)(i) + \alpha^k r, \forall i: \forall k \ge 1$

(2) $(T_\pi^k(J + re))(i) = (T_\pi^k J)(i) + \alpha^k r$, — " —

Proposition (Basis for Value-iteration)

For any bounded $J : S \to \mathbb{R}$,

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

Proof: for any policy $\pi = \{\mu_0, \mu_1, \cdots\}$ and state $i \in S$,

$$J_\pi(i) = \lim_{N \to \infty} E\left(\sum_{k=0}^{N-1} \alpha^k g(s_k, \mu_k(s_k), s_{k+1}) \,\middle|\, s_0 = i \right)$$

$$= E\left(\sum_{k=0}^{K-1} \alpha^k g(s_k, \mu_k(s_k), s_{k+1}) \right) + \underbrace{\lim_{N \to \infty} E\left(\sum_{k=K}^{N-1} \alpha^k g(s_k, \mu_k(s_k), s_{k+1}) \,\middle|\, s_K = j \right)}_{\text{term-(B)}}$$

Since $|g(s_k, \mu_k(s_k), s_{k+1})| \le M \quad \forall s_k, s_{k+1} \in S, \ \mu_k(s_k) \in A(s_k)$

$$|\text{term-LB}| \le M \sum_{k=K}^{\infty} \alpha^k \le \frac{\alpha^K M}{1-\alpha}$$

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

$$E\left[\sum_{k=0}^{K-1} \alpha^k g(s_k, \mu_k(s_k), s_{k+1}) \,\middle|\, s_0 = i\right]$$

$$E\left(\alpha^K J(s_k) + \sum_{k=0}^{K-1} \alpha^k g(s_k, \mu_k(s_k), s_{k+1}) \,\middle|\, s_0 = i\right)$$

$$= J_\pi(i) - \text{term}(B) + E(\alpha^K J(s_k))$$

$$J_\pi(i) - \frac{\alpha^k M}{1-\alpha} - \alpha^k \max_j |J(j)|$$

$$\leq E\left[\alpha^k J(s_k) + \sum_{k=0}^{k+1} \alpha^k g(s_k, \mu_k(s_k), s_{k+1}) \,\middle|\, s_0 = i\right]$$

$$\leq J_\pi(i) + \frac{\alpha^k M}{1-\alpha} + \alpha^k \max_j |J(j)|$$

Take min over π on both sides,

$$J^*(i) - \frac{\alpha^k M}{1-\alpha} - \alpha^k \max_j |J(j)| \leq (T^k J)(i)$$

$$\leq J^*(i) + \frac{\alpha^k M}{1-\alpha} + \alpha^k \max_j |J(j)|$$

Claim follows by taking $k \to \infty$ on both sides, i.e., $\lim_{k\to\infty} (T^k J)(i) = J^*(i)$ ∎

Value iteration

Fix J_0.

Repeatedly apply T i.e.,

$$J_0 \xrightarrow{T} J_1 = TJ_0 \xrightarrow{T} J_2 = TJ_1 + \cdots\cdots \rightarrow J^*$$

Corollary: For any stationary policy π,

$$J_\pi(i) = \lim_{N \to \infty} \left(T_\pi^N J\right)(i),$$ where J is any bounded n-vector.

Bellman optimality equation

from the proof above,

$$J^*(i) - \frac{\alpha^k M}{1-\alpha} - \alpha^k \max_j |J(j)| \le (T^k J)(i) \le J^*(i) + \frac{\alpha^k M}{1-\alpha} + \alpha^k \max_j |J(j)|$$

Apply T to both sides

$$(TJ^*)(i) - \frac{\alpha^{k+1} M}{1-\alpha} - \alpha^{k+1} \max_j |J(j)| \le (T^{k+1} J)(i)$$

$$\le (TJ^*)(i) + \frac{\alpha^{k+1} M}{1-\alpha} + \alpha^{k+1} \max_j |J(j)|$$

Taking $k \to \infty$ on both sides, we obtain

$$\boxed{J^*(i) = TJ^*(i), \quad \forall i}$$

Uniqueness: Suppose J' satisfies $J' = TJ'$

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

Corollary: For any Stationary π,

$$J_\pi = T_\pi J_\pi \quad \& \quad J_\pi \text{ is the unique solution}$$

Illustration of Value Iteration

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

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

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

$J_0 = (0,0)$, $\alpha = 0.9$

Do Value iteration for 2 steps.

$J_1 = (0.5, 1)$

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

$J_2 = (1.287, 1.562)$

Prelude to policy Iteration

$\pi_0 \xrightarrow{P.E.} J_{\pi_0} \xrightarrow{P.I.} \pi' \xrightarrow{P.E.} J_{\pi'} \xrightarrow{} \cdots \xrightarrow{} \pi^*$ in finite steps

Claim: A policy π is optimal if and only if $\pi(i)$ attains the minimum in the Bellman equation, $\forall i \in S$, i.e.,

$$\boxed{T_\pi J^* = T J^*} \quad \text{i.e.,} \quad g(i,\pi(i)) + \alpha \sum_{j=1}^{n} P_{ij}(\pi(i)) J^*(j) = \min_a g(i,a) + \alpha \sum_{j=1}^{n} P_{ij}(a) J^*(j), \quad \forall i \in S$$

Pf: Assume $TJ^* = T_\pi J^*$

we know $J^* = TJ^*$

So, $J^* = TJ^* = T_\pi J^* \Rightarrow J^* = J_\pi \Rightarrow \pi$ is optimal.

Converse: π is optimal $\Rightarrow J^* = J_\pi \Rightarrow J^* = T_\pi J^*$

from BE, $J^* = TJ^*$, So, $\boxed{T_\pi J^* = TJ^*}$

Example:- Machine replacement

$S = \{1, \ldots n\}$ $\quad g(1) \leq \cdots \leq g(n)$ $\quad P_{ij}$: transition probabilities

Actions: Do nothing, repair

Bellman equation $J^*(i) = \min\left(\underbrace{R + g(1) + \alpha J^*(1)}_{\text{Repair}}, \underbrace{g(i) + \alpha \sum_{j=i}^{n} P_{ij} J^*(j)}_{\text{Do nothing}} \right)$

Assume $P_{ij} \le P_{(i+1)j}$ if $i < j$, Also, $P_{ij} = 0$ if $j < i$

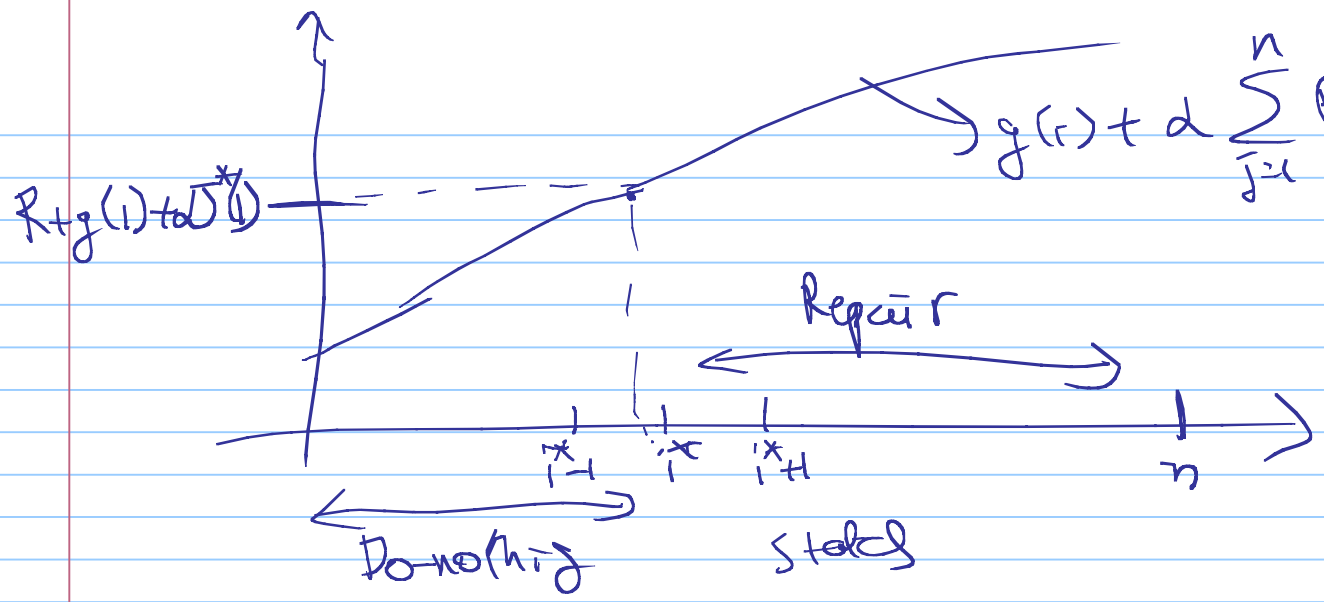
So, $\sum_{j=1}^{n} P_{ij} J(j) \le \sum_{j=1}^{n} P_{(i+1)j} J(j)$, $i = 1, \ldots n-1$

for J that is monotonically non-decreasing

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

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

So, $g(i) + \alpha \sum_{j=1}^{n} P_{ij} J^*(j)$ is non-decreasing in "i".



Optimal policy is "threshold" - based

Policy iteration algorithm

Step 1: Fix an initial policy π_0

Step 2: Find the cost-to-go J_{π_k} corresponding to π_k] Policy Evaluation

Step 3: Obtain a new policy π_{k+1} by solving

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

} Policy Improvement

If $\pi_{k+1} = \pi_k$, stop, else goto step 2.

Proposition: Let π and π' be two stationary policies that satisfy

$$T_{\pi'} J_\pi = T J_\pi \quad \text{or,}$$

$$g(i,\pi'(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\pi'(i)) J_\pi(j) = \min_a \left(g(i,a) + \alpha \sum_{j=1}^{n} p_{ij}(a) J_\pi(j) \right), \quad i = 1 \dots n$$

Then, $J_{\pi'}(i) \leq J_\pi(i), \quad i = 1 \dots n$

Further, if π is not optimal, then $J_{\pi'}(i) < J_\pi(i)$ for at least one state i.

Proof: $J_\pi = T_\pi J_\pi$. Also, $T_{\pi'} J_\pi = T J_\pi$

$$J_\pi(i) = g(i,\pi(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\pi(i)) J_\pi(j)$$

$$\geq g(i,\pi'(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\pi'(i)) J_\pi(j) = (T_{\pi'} J_\pi)(i)$$

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

Uniqueness: Skipped.

Policy iteration: 1 Verification

$S = \{1, 2\}$, $A = \{a, b\}$ $P(a) = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$ $P(b) = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$

$g(1,a) = 2$, $g(1,b) = 0.5$, $g(2,a) = 1$, $g(2,b) = 3$, $\alpha = 0.9$

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

$$J_{\pi_0}(1) = g(1,a) + \alpha P_{11}(a) J_{\pi_0}(1) + \alpha P_{12}(a) J_{\pi_0}(2)$$

$$J_{\pi_0}(2) = g(2,b) + \alpha P_{21}(b) J_{\pi_0}(1) + \alpha P_{22}(b) J_{\pi_0}(2)$$

Using MDP data, we have

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

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

Solving, we obtain $\boxed{J_{\pi_0}(1) = 24.12, \; J_{\pi_0}(2) = 25.96}$

Policy improvement

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

$$(T J_{\pi_0})(1) = \min\{\underbrace{2 + 0.9\left(\frac{3}{4} \times 24.12 + \frac{1}{4} \times 25.96\right)}_{\text{action a}}, \underbrace{0.5 + 0.9\left(\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96\right)}_{\text{action b}}\}$$

$$= \min(24.12, 23.45) = 23.45$$

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

$$= \min(23.12, 25.95) = 23.12$$

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

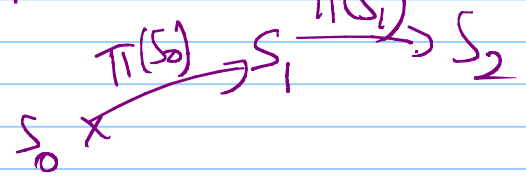
Policy evaluation:- $J_{\pi_1}(1) = 7.33$, $J_{\pi_1}(2) = 7.67$

Policy improvement:- $\pi_2 = \pi_1$

Tomorrow:- Reinforcement learning (RL)

Task: Policy evaluation Fix π

Sample path:



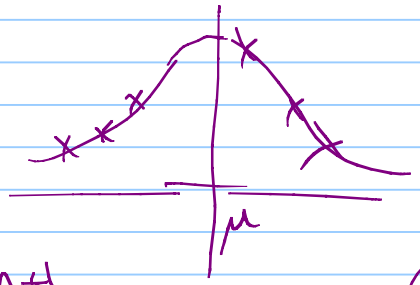
s_1 sampled from $P_{s_0 s_1}(\pi(s_0))$

TD-learning → Policy evaluation & Q-learning → Control (to find the optimal policy)

Today: Reinforcement Learning

Mean estimation:- Random variable v with mean μ and finite variance.

$\{v_1, v_2, \ldots, v_n\}$ i.i.d. samples from the distribution of v.



$$r_n = \frac{1}{n}\sum_{i=1}^{n} v_i$$

$$r_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} v_i = \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right) + \frac{1}{n+1}v_{n+1} = \frac{n}{n+1}r_n + \frac{v_{n+1}}{n+1}$$

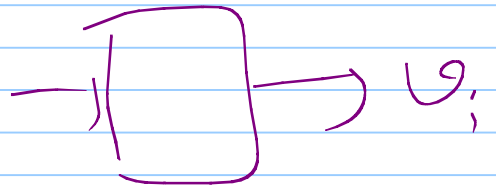
$$\boxed{r_{n+1} = r_n + \frac{1}{n+1}(v_{n+1} - r_n)}$$

More generally, $r_{n+1} = r_n + \eta_n(v_{n+1} - r_n)$,

where $\sum_n \eta_n = \infty$, $\sum_n \eta_n^2 < \infty$ η_n : step-size

Then, $r_n \to \mu$ w.p. 1 as $n \to \infty$

Want to estimate $E[v]$



$$r_{n+1} = r_n + \eta_n (v_{n+1} - r_n)$$

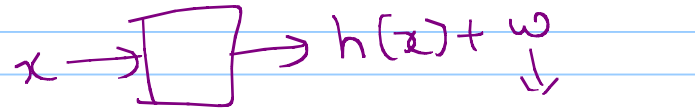
$h(x) = f(x) - x$

$x_{n+1} = x_n + \eta_n (f(x_n) + w_n - x_n)$

$= (1-\eta_n) x_n + \eta_n (f(x) + w_n)$

Robbins Monro Stochastic approximation scheme (1951)

Generally, if you want to solve $h(x) = 0$



zero-mean

$$x_{n+1} = x_n + \eta_n (h(x_n) + w_n)$$

$x_n \xrightarrow[n\to\infty]{} x^*$ w.p.1, where $h(x^*) = 0$

assume $E w_n = 0$, $E w_n^2 < \infty$

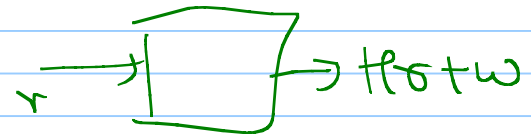
Special cases:- ① $\min_x f(x)$ Then $h = f'$ or in higher dimensions $h = \nabla f$

② Fixed-point iteration. Want: $f(x^*) = x^*$
Then, set $h(x) = f(x) - x$

Aim: To solve $\boxed{Hr = r}$ $H: \mathbb{R}^n \to \mathbb{R}^n$, $r \in \mathbb{R}^n$. e.g. $H = T_\pi$

A direct method: $r_{n+1} = Hr_n$, with r_0 set arbitrarily

Alternately, $r_{n+1} = (1-\eta) r_n + \eta (Hr_n)$
(η: stepsize)



$r_{n+1} = (1-\eta) r_n + \eta (Hr_n + w_n)$

Assuming H is "well-behaved", & noise w_n is zero-mean & bounded variance,

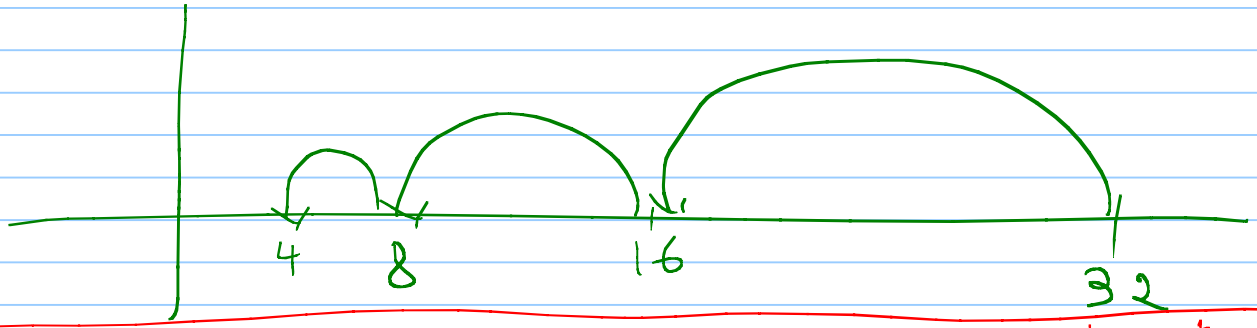
One can show $r_n \xrightarrow[n\to\infty]{} r^*$ w.p.1, where $\boxed{Hr^* = r^*}$

Brief introduction to contraction mappings:-

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

$\boxed{f(0) = 0}$

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



Def: $f: \mathbb{R} \to \mathbb{R}$ is a contraction mapping with modulus β if

$$|f(x) - f(y)| \le \beta |x - y|, \text{ for any } x, y$$

with $0 < \beta < 1$

Contraction properties of T, T_π in a discounted MDP setting:

Max-norm:- $\|\cdot\|_\infty$ on $\mathbb{R}^n$ is defined by

$$\|J\|_\infty = \max_{i=1,\dots,n} |J(i)|$$

Proposition: For any two bounded J, J', we have

$$\|TJ - TJ'\|_\infty \leq \alpha \|J - J'\|_\infty$$

(α: discount factor)

$$\|T_\pi J - T_\pi J'\|_\infty \leq \alpha \|J - J'\|_\infty, \text{ for any stationary policy } \pi.$$

Further, $\forall k \geq 1$, $\|T^k J - T^k J'\|_\infty \leq \alpha^k \|J - J'\|_\infty$ &

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

Corollary:

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

Stochastic iterative algorithm for finding fixed-point of a contraction mapping:-

Goal:- Solve $Hr^* = r^*$, $H: \mathbb{R}^n \to \mathbb{R}^n$, $r \in \mathbb{R}^n$

Fixed-point iteration:- $r_{t+1}(i) = (1-\eta_t) r_t(i) + \eta \left((Hr_t)(i) + w_t(i) \right)$ (noise),

Assumptions: $\mathcal{F}_t = \{ r_0(i), \ldots, r_t(i), w_0(i), \ldots, w_{t-1}(i), i = 1, \ldots n \}$ $i = 1, \ldots n$

history

(A1) $E(w_t(i) | \mathcal{F}_t) = 0$, $E(w_t^2(i) | \mathcal{F}_t) \leq A + B \| r_t \|^2$

(A2) $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$. e.g. $\eta_t = \frac{c}{t}$

(A3) H is a max-norm pseudo-contraction, i.e.,

$$\| Hr - r^* \|_\infty \le \beta \| r - r^* \|_\infty, \quad \forall r$$

where $0 < \beta < 1$ & $Hr^* = r^*$.

Under (A1) – (A3)

$r_t \to r^*$ w.p.1 as $n \to \infty$

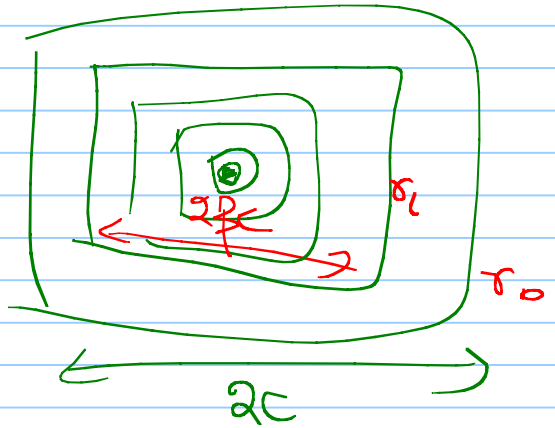
Full-contraction

$\| Hr - Hr' \|_\infty \le \beta \| r - r' \|_\infty \quad \forall r, r'$

$$r_{t+1}(i) = (1-\eta_n) r_t(i) + \eta_n \left((Hr_t)(i) + w_t(i) \right)$$

$H0 = 0$ $\quad r^* = 0$

$|r_0(i)| \le C$

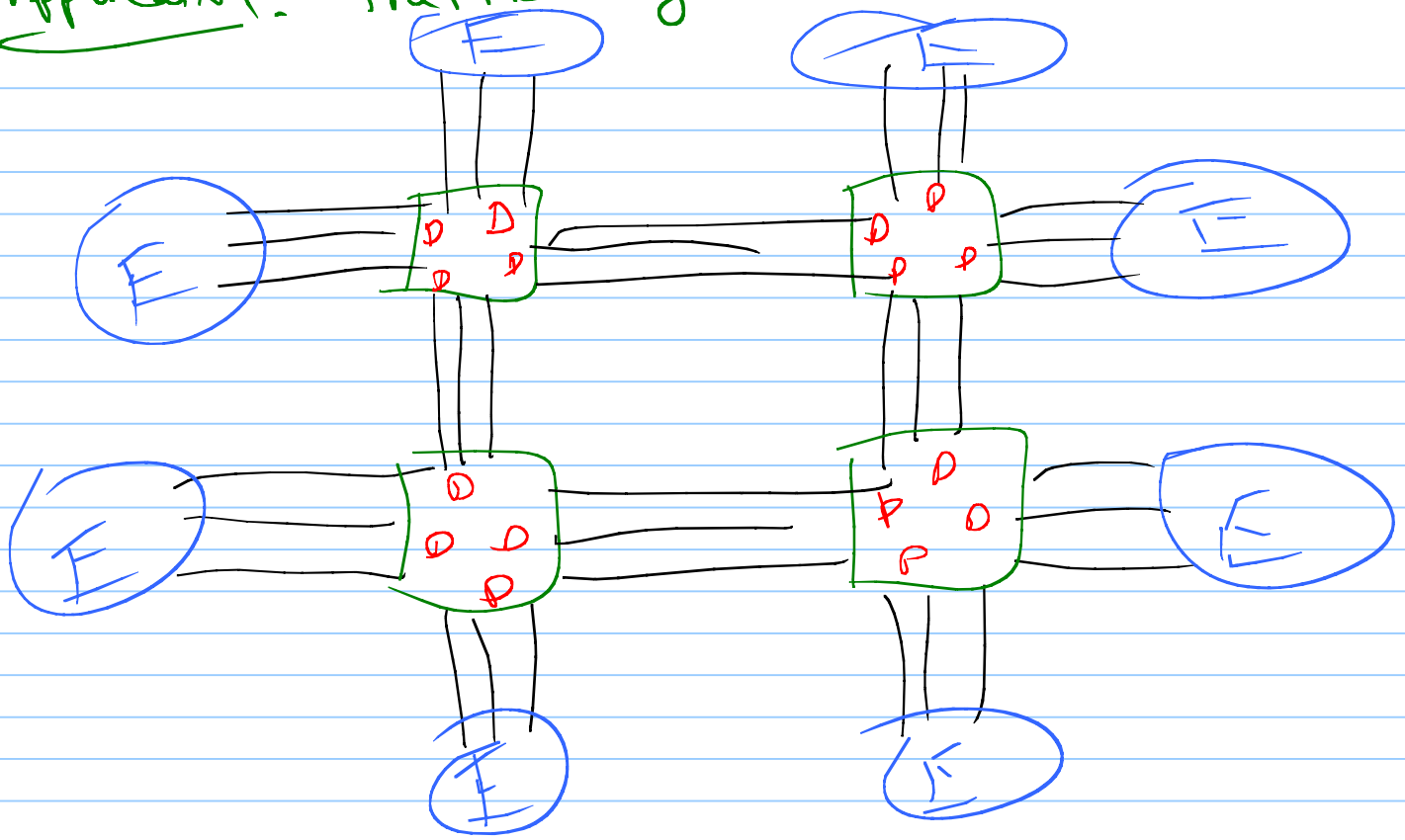


$$r_{t+1}(i) = (Hr_t)(i)$$

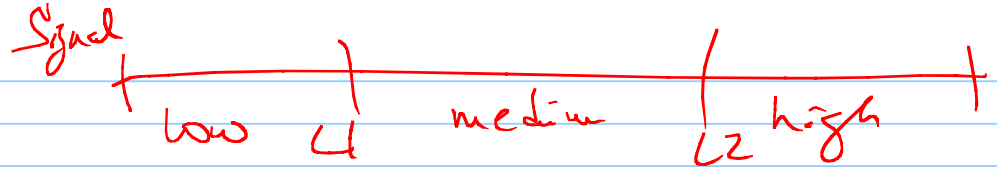
Now, with $r_{t+1}(i) = (1-\eta_t) r_t(i) + \eta_t \left(Hr_t(i) + w_t(i) \right)$

Similar behaviour, but shrinkage is not immediate

Application: Traffic signal control



Discounted MDP



(N lanes with signals (across junction)

State = $S_t = (q_1(t), \ldots, q_N(t), t_1(t), \ldots t_N(t))$

elapsed time

Actions = feasible sign configurations

Single stage cost = $g(S_t, a_t) = \alpha_1 \left(\sum_{i=1}^{N} q_i(t) \right) + \alpha_2 \left(\sum_{i=1}^{N} t_i(t) \right)$

normalized to be in $(0,1]$

Temporal-difference (TD) learning for policy evaluation :-

Single stage cost: $g(i,j)$

Fix a policy π.

Want to estimate: $J_\pi(i) = E\left[\sum_{m=0}^{\infty} \alpha^m g(i_m, i_{m+1}) \,\middle|\, i_0 = i\right]$

Bellman equation :- $J_\pi = T_\pi J_\pi$

$$J_\pi(i) = E\left(g(i,\bar{i}) + \alpha J_\pi(\bar{i})\right)$$

TD(0) algorithm :-

$$J_{t+1}(i) = (1-\eta_t) J_t(i) + \eta_t\left(g(i,\bar{i}) + \alpha J_t(\bar{i})\right)$$

$\bar{i} \sim P_{ij}(\pi(i))$

$$J_{t+1}(i) = J_t(i) + \eta_t \Big(\underbrace{g(i,\tilde{i}) + \alpha J_t(\tilde{i})}_{\text{Proxy for}} - J_t(i) \Big)$$

$$E\big(g(i,\tilde{i}) + \alpha J_t(\tilde{i})\big) = (T_\pi J_t)(i)$$

$$J_{t+1}(i) = J_t(i) + \eta_t \Big((T_\pi J_t)(i) - J_t(i) + \underbrace{g(i,\tilde{i}) + \alpha J_t(\tilde{i}) - (T_\pi J_t)(i)}_{w_t(i)} \Big)$$

$$J_\pi(i) = E\left(g(i,\bar{i}) + \alpha J_\pi(\bar{i}) \right)$$

$$= E\left(g(i,\bar{i}) + \alpha g(\bar{i},\bar{\bar{i}}) + \alpha^2 J_\pi(\bar{\bar{i}}) \right)$$

$\bar{i}$ = next state

$\bar{\bar{i}}$ = next-to-next state

$TD(\lambda)$, $\lambda \in (0,1)$:

$$J_\pi(i) = E\left(\sum_{m=0}^{\ell} \alpha^m g(i_m, i_{m+1}) + \alpha^{\ell+1} J_\pi(i_{\ell+1}) \right)$$

Fix $\lambda < 1$, and multiply $(1-\lambda)\lambda^\ell$ & sum over ℓ:

$$J_\pi(i) = (1-\lambda)\, E\left(\sum_{l=0}^{\infty} \lambda^l \left(\sum_{m=0}^{l} \alpha^m g(i_m, i_{m+1}) + \alpha^{l+1} J_\pi(i_{l+1}) \right) \right)$$

Interchange the two summations & we $(1-\lambda) \sum_{l=m}^{\infty} \lambda^l = \lambda^m$, to obtain

$$J_\pi(i) = E\left[(1-\lambda) \sum_{m=0}^{\infty} \alpha^m g(i_m, i_{m+1}) \sum_{l=m}^{\infty} \lambda^l + \sum_{l=0}^{\infty} \alpha^{l+1} J_\pi(i_{l+1}) (\lambda^l - \lambda^{l+1}) \right]$$

$$= E\left(\sum_{m=0}^{\infty} \lambda^m \left(\alpha^m g(i_m, i_{m+1}) + \alpha^{m+1} J_\pi(i_{m+1}) - \alpha^m J_\pi(i_m) \right) \right) + J_\pi(i)$$

Letting $d_m = g(i_m, i_{m+1}) + \alpha J_\pi(i_{m+1}) - J_\pi(i_m)$

$$J_\pi(i) = E\left(\sum_{m=0}^{\infty} (\alpha\lambda)^m d_m \right) + J_\pi(i)$$

TD(λ) update rule :-

$$J_{t+1}(i) = J_t(i) + \eta_t \left(\sum_{m=0}^{\infty} (\alpha\lambda)^m \; d_m \right)$$

For $\lambda = 0$, $d_0 = g(i, \tilde{i}) + \alpha J_\pi(\tilde{i}) - J_\pi(i)$

& we recover TD(0)

Stochastic Shortest path

$S = \{1, \dots, n, T\}$ $\quad p_{TT}(a) = 1 \quad g(T, a, T) = 0$

$$J_\pi(i) = E\left(\sum_{t=0}^{\infty} g(i_t, \pi(i_t), i_{t+1}) \,\middle|\, i_0 = i \right)$$

$$J_\pi(i) = E\left(g(i, \pi(i), \bar{i}) + J_\pi(\bar{i}) \right)$$

π: proper if it ensures that terminal state T is reached with positive probability.

TD(0): $J_{t+1}(\bar{i}) = J_t(i) + \eta_t \underbrace{\left(g(i, \pi(i), \bar{i}) + J_t(\bar{i}) - J_t(i) \right)}_{\text{temporal-difference}}$

For $\bar{i} = 1, \dots n$

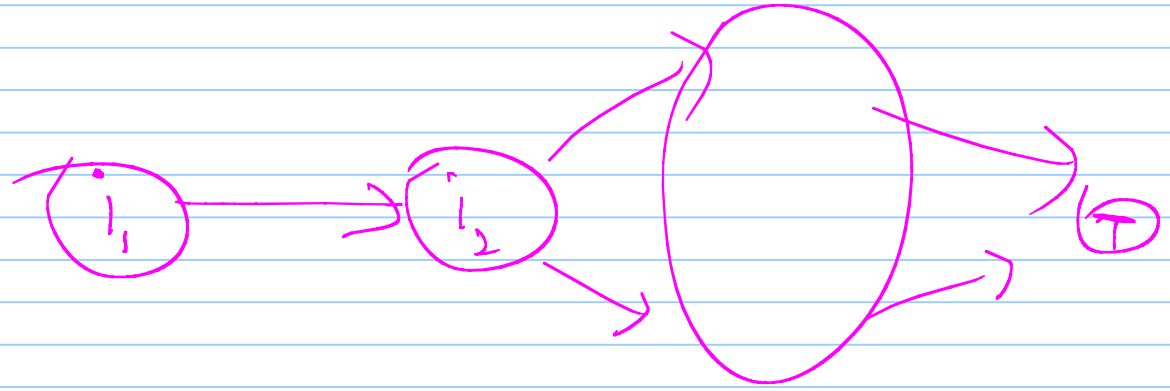
TD(1) :- Simulate episodes of the underlying SSP using π.

$$(i_0^m, i_1^m, \ldots\ldots, i_N^m), \quad i_N = T, \quad i_0 = i$$

$$\hat{J}(m) = g(i_0^m, \pi_1(i_0^m), i_1^m) + \ldots\ldots + g(i_{N-1}^m, \pi(i_{N-1}^m), i_N^m)$$

$$\tilde{J}(i) = \sum_{m=1}^{M} \hat{J}(m), \qquad M = \#\text{ of episodes simulated}$$

TD(0) vs TD(1) :- A toy example



TD(1) : Start in i_1, simulate episode, use total cost, say $\hat{J}(m)$, to estimate $J_\pi(i_1)$

TD(0) : Start in i_1, simulate one transition and use "$g(i_1, i_2) + J(i_2) - J(i_1)$" to estimate $J_\pi(i_1)$

Q-learning: SSP setting

Let J^* be the optimal cost-to-go vector

$$Q^*(i,a) = \sum_{j=1,\dots,n,T} P_{ij}(a)\left(g(i,a,j) + J^*(j)\right), \quad i = 1,\dots,n$$

Bellman's equation:

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

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

Q-Bellman equation:

$$Q^*(i,a) = \sum_{j=1,\dots,n,T} P_{ij}(a)\left(g(i,a,j) + \min_b Q^*(j,b)\right)$$

$$\boxed{Q^*(i,a) = E\left(g(i,a,j) + \min_b Q^*(j,b)\right)}$$

$$Q_{t+1}(i,a) = Q_t(i,a) + \eta_t \left(g(i,a,\bar{i}) + \min_b Q_t(\bar{i},b) - Q_t(i,a) \right)$$

Define the operator

$$(HQ)(i,a) = \sum_{\bar{j}=1,\dots,n_i T} P_{ij}(a) \left(g(i,a,j) + \min_b Q(j,b) \right)$$

Under some conditions, HQ is a contraction mapping.

$$Q_{t+1}(i,a) = Q_t(i,a) + \eta_t \left((HQ_t)(i,a) + w_t(i,a) \right)$$

$$w_t(i,a) = g(i,a,\bar{i}) + \min_b Q_t(\bar{i},b) - \sum_{\bar{j}=1,\dots,n_i T} P_{ij}(a) \left(g(i,a,j) + \min_b Q_t(j,b) \right)$$

$Q_t \to Q^*$ w.p.1 as $n \to \infty$ under conditions like (A1) – (A3)

Issue of exploration:-

For Q-learning to converge, all state-action pairs (i,a) have to be visited infinitely often.

Greedy policy: $\pi_{t+1}(i) = \min_a Q_t(i,a)$

ϵ-Greedy policy:- $\pi_{t+1}(i) = \begin{cases} \text{greedy action w.p. } (1-\epsilon) \\ \text{random action w.p. } \epsilon \end{cases}$

$\epsilon \rightarrow$ small number e.g. 0.05

Linear function approximation

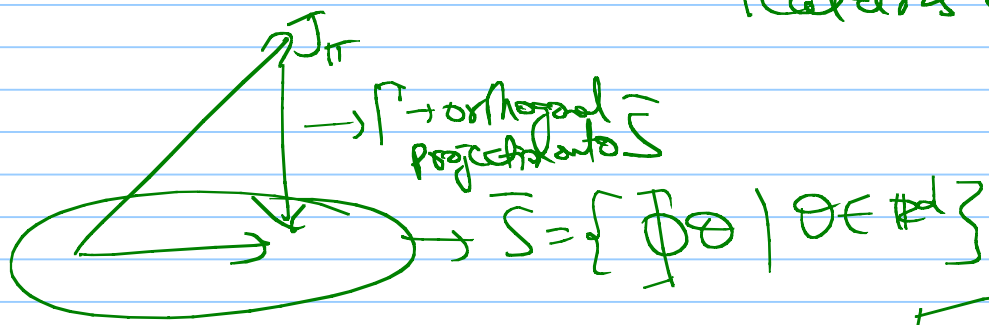
Fix a policy π.

$$J_\pi(i) \approx \phi(i)^T \theta$$

$\phi(i)$: features $\in \mathbb{R}^d$

$\theta \in \mathbb{R}^d$

$d << |S|$



$$J_\pi \approx \Phi\theta$$

Cannot solve $\boxed{J_\pi = T_\pi J_\pi}$

Approximate solution $\Phi\theta = \Gamma(T_\pi \Phi\theta)$ → has a unique solution since ΓT_π is a contraction.

$$\theta_{t+1} = \theta_t + \eta_t \left(g(i, \pi(i), \bar{i}) + \phi(\bar{i})^T \theta_t - \phi(i)^T \theta_t \right) \phi(i)$$

TD(0) with linear function approximation

Q-learning with linear function approximation

$$Q(i,a) \approx \phi(i,a)^T \theta$$

$$\theta_{t+1} = \theta_t + \eta_t \left(g(i,a,\bar{i}) + \min_b \phi(\bar{i},b)^T \theta_t - \phi(i,a)^T \theta_t \right) \times \phi(i,a)$$

Greedy action $\min_a \theta_t^T \phi(i,a)$