Lecture-25

Finite horizon Markov decision processes:-

Textbooks:- ① CS6700 in my website

② Tsitsiklis & Bertsekas "Neuro-dynamic Programming", 1996

③ Sutton & Barto, "Intro. to RL"

④ Bertsekas "DP & OC, Vol I & II".

Example:- Machine replacement/repair

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

1 ↓ good, n ↘ worst

Actions = Do nothing, Repair

On repair: machine goes to "1". Cost = R

Operating cost: $g(i)$

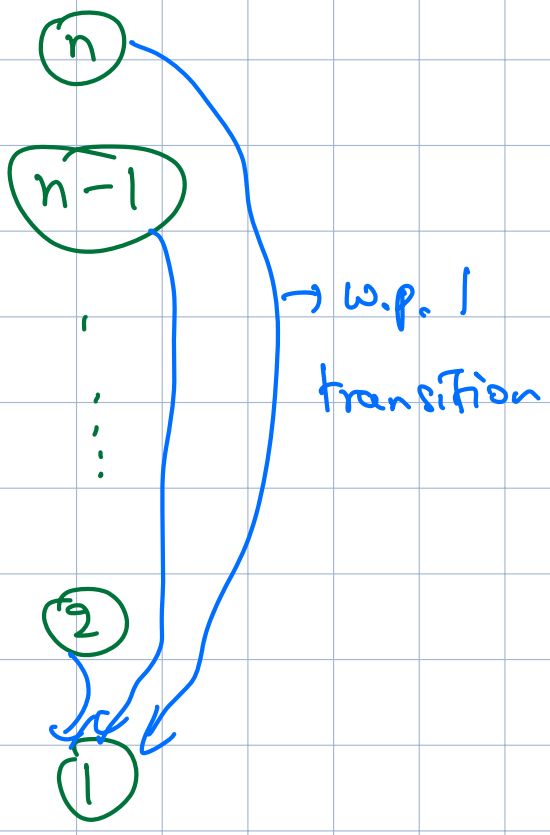
$g(1) \leq g(2) \leq g(3) \leq \ldots \leq g(n)$

State transitions:

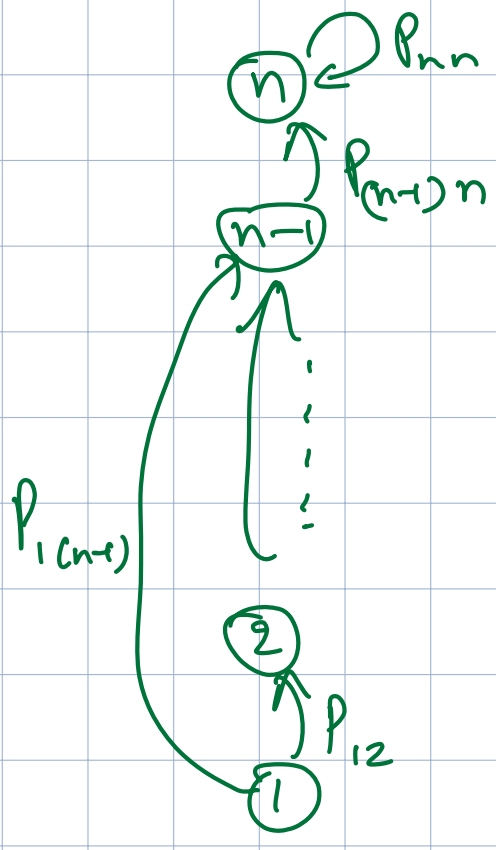
P_{ij}: probability of state transitioning from i to j

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

Action: repair



Action: Do nothing



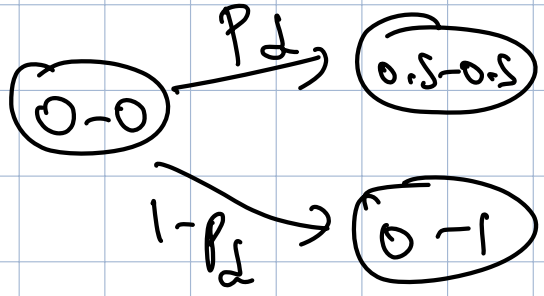
Example 2: Chess match

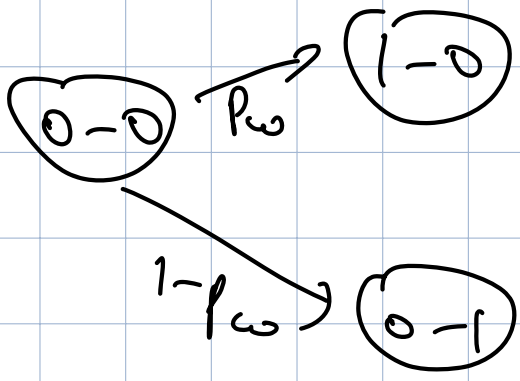
Playing styles (actions): Timid, Bold

Timid	draw p_d	lose $1-p_d$
Bold	win p_w	lose $1-p_w$

$p_d > p_w$

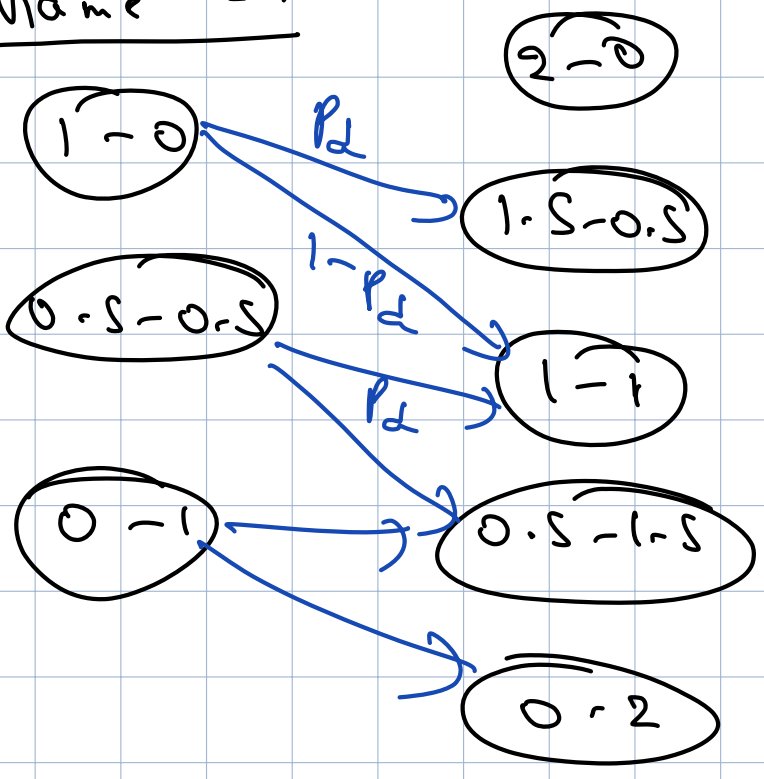
Game 1

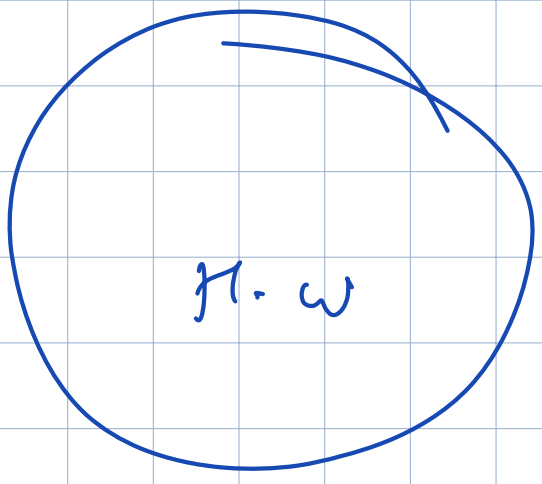




Timid

Game 2:





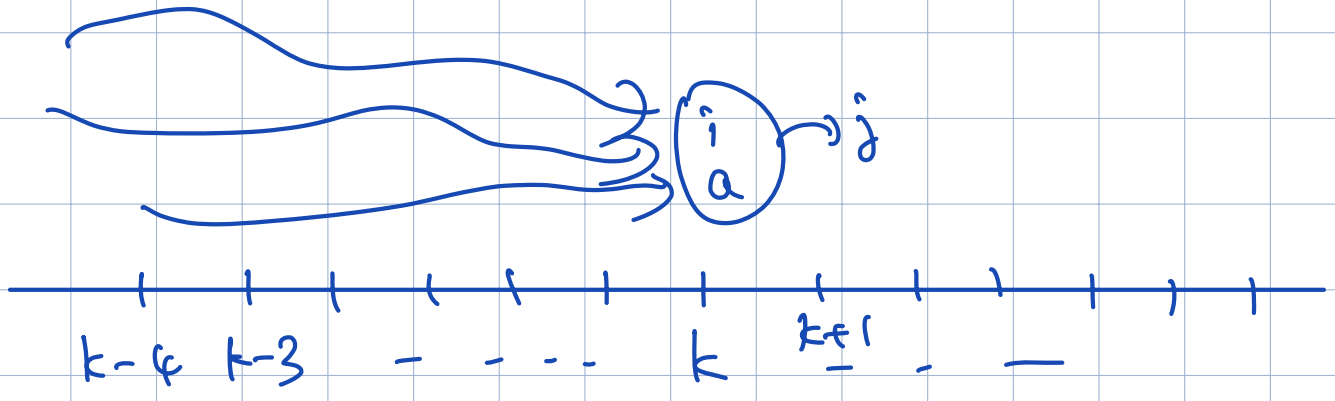
Timid play

Bold play

Framework

S: State space A: action space

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



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

Policy π: $\{\mu_0, \cdots \mu_{N-1}\}$

N ← # of slots, $\mu_i : S \to A$ (S: to 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: Find π^* s.t.

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

Optimality principle:

$$\pi^* = (\mu_0^*, \mu_1^*, \cdots \mu_{N-1}^*)$$

A tail sub-problem

$$\min_{\pi^i = (\mu_i, \cdots, \mu_{N-1})} E\left(g_N(s_N) + \sum_{k=i}^{N-1} g_k(s_k, \mu_k(s_k), s_{k+1})\right)$$

Claim: For this tail sub-problem

$\{\mu_i^*, \cdots \mu_{N-1}^*\}$ is optimal

[redacted]

DP algorithm:

Idea: Solve tail sub-problems & progress backwards

Algorithm:

Set $J_N(S_N) = g_N(S_N) \quad \forall S_N \in S.$

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

{

$$J_k(S_k) = \min_{a_k \in A} E_{S_{k+1}}\left[g_k(S_k, a_k, S_{k+1}) + J_{k+1}(S_{k+1}) \right], \quad \forall S_k \in S.$$

}

Apply DP algo to "Machine Example".

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

P.T.O.

Suppose terminal cost is zero

$$J_N(i) = 0$$

$$J_k(i) = \min\{ \underbrace{R + g(1) + J_{k+1}(1)}_{\text{Repair}},$$

$$\underbrace{g(i) + \sum_{j=i}^{n} p_{ij} J_{k+1}(j)}_{\text{Do nothing}} \}$$

Chess-example (re-visited)

State = net-score (#wins − #losses)

$$J_k(S_k) = \max\Big(\underbrace{p_d J_{k+1}(S_k) + (1-p_d) J_{k+1}(S_k - 1)}_{\text{Timid}},$$

$$\underbrace{p_w J_{k+1}(S_k + 1) + (1-p_w) J_{k+1}(S_k - 1)}_{\text{Bold}} \Big)$$

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}	$J_{N-1}(s_{N-1})$	Best play
≥ 1	1	does not matter
1	$\max(\underbrace{p_d + (1-p_d)p_w}_{\text{timid}}, \underbrace{p_w + (1-p_w)p_w}_{\text{Bold}})$	Timid play is best
0	p_w	Bold play is best
-1	p_w^2	"

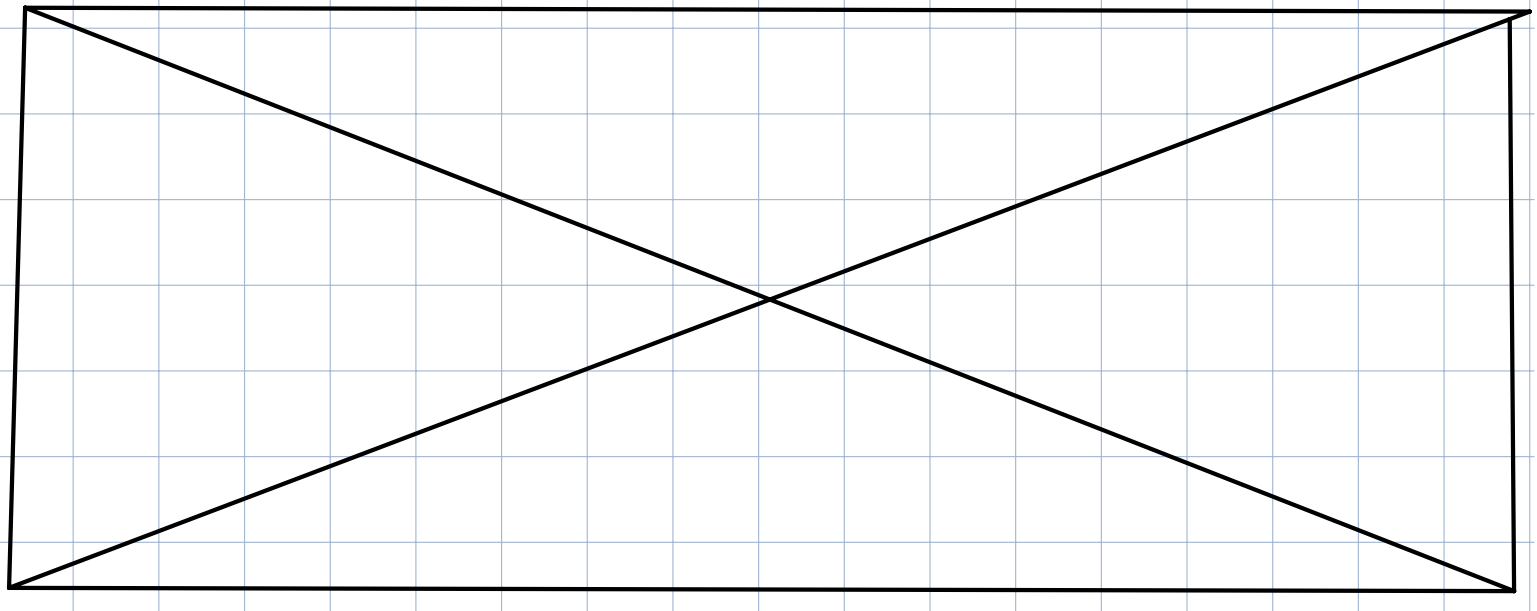
<-1	0	Leave it

For the 2-game match

$$J_{N-2}(0) = \max\Big(\underbrace{p_d p_\omega + (1-p_d) p_\omega^2}_{\text{Timid}},$$

$$\underbrace{p_\omega \left(p_d + (1-p_d) p_\omega \right) + (1-p_\omega) p_\omega^2}_{\text{Bold}} \Big)$$

Check $\Rightarrow$ $=$ $p_\omega \left(p_d + (1-p_d) p_\omega \right) + (1-p_\omega) p_\omega^2$



Lecture-21

Infinite horizon discounted cost MDPs:

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

Stationary deterministic policy (SDP)

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

Stationary policy $\pi = \{\mu, \mu, \dots\}$

Start state

$$J_\pi(i) = E\left(\sum_{k=0}^{\infty} \alpha^k g(S_k, \pi(S_k), S_{k+1}) \,\middle|\, S_0 = i \right]$$

α^k: discount factor $0 < \alpha < 1$

$g(S_k, \pi(S_k), S_{k+1})$: single-slot cost; π: policy

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

Let π^* be an optimal policy corresponding to J^*

states

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

Bellman operator

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

$$TJ = (TJ(1), \ldots, TJ(n)), \text{ where}$$

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

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

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

$$= \left[\left[P_{ij}(\pi(i)) \right]\right]_{i,j=1\cdots n}$$

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

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

Properties of T, T_π:

For any J, J' s.t. $J(i) \le J'(i)$ $\forall i$

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

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

$\forall k \ge 1$

Proposition:- (Basis for Value Iteration)

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

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

Remark:-

$$J_0 \xrightarrow{T} J_1 = TJ_0 \xrightarrow{T} J_2 = T^2 J_0$$

$$\downarrow r$$

$$\vdots \quad \text{as } N \to \infty$$

$$T^N J_0 \to J^*$$

Corollary: $J_\pi(i) = \lim_{N \to \infty} (T_\pi^N J)(i), \quad \forall i \in S$

for any bounded J.

Value iteration for solving a discounted MDP:

Fix J_0

Repeatedly apply T

$$J_0 \xrightarrow{T} J_1 \rightarrow \cdots \rightarrow J^*$$

Bellman optimality equation

The optimal cost $J^*(i) = \min_{\pi} J_{\pi}(i), \forall i$

Satisfies

Fixed-point equation

$$J^*(i) = TJ^*(i)$$

(or)

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

Claim:- J^* is the unique fixed point

Pf: Suppose not.

$\exists J'$, $J' = TJ'$

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

$$= J^*$$

Corollary: For any stationary π,

$J_\pi = T_\pi J_\pi$ & J_π is the unique solution.

Illustration of Value Iteration:-

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

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

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

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

$g(2,a) = 1$, $g(2,b) = 3$

$\alpha = 0.9$

$J_0 = (0, 0)$

Do VI for 2 steps

$J_1 = (0.5, 1)$

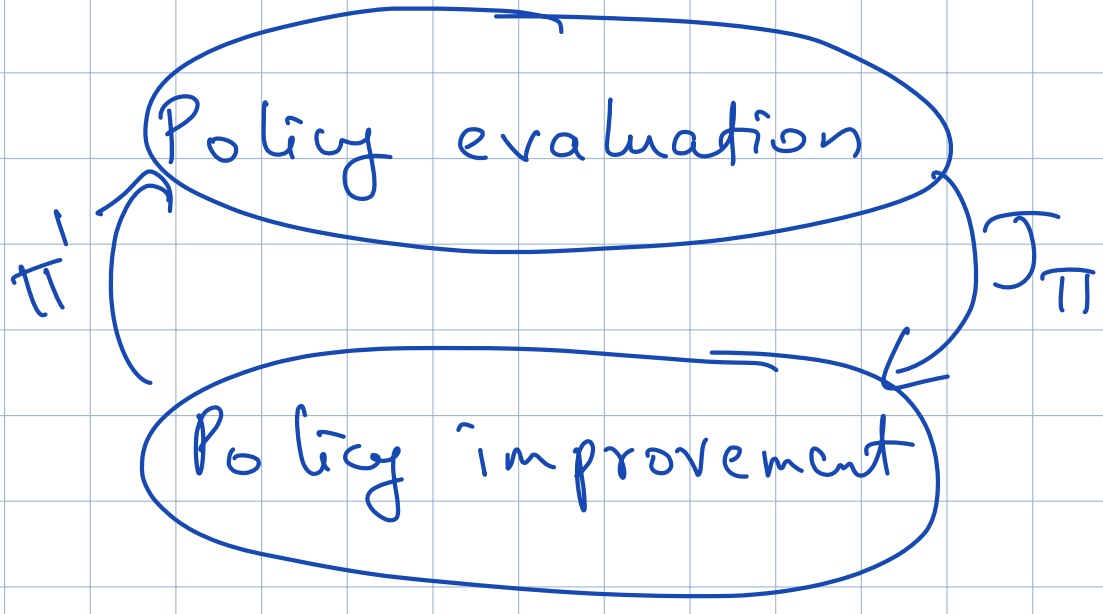
$$TJ(i) = \min\left(g(i,a) + \alpha \sum_{j=1}^{2} P_{ij}(a) J(j), \right.$$

$$g(i,b) + \alpha \sum_{j=1}^{2} p_{ij}(b) J(j)$$

$$J_2 \simeq (1.287, 1.562)$$

Lecture-22

Policy iteration



$\pi_0 \xrightarrow{P.E.} J_{\pi_0} \xrightarrow{P.I.} \pi_1$ $\left(J_{\pi_1}(i) \leq J_{\pi_0}(i) \right.$ with strict ineq for at least one state. $\left.\right)$

P.E.

$$\pi^* \cdots \leftarrow \pi_2 \xleftarrow{P.I} J_{\pi_1}$$

Claim:- A policy π is optimal if and only if $\pi(i)$ attains the minimum in the Bellman equation

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

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

(or)

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

Proof: Assume $TJ^* = T_\pi J^*$ for some π

We know $J^* = TJ^*$

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

$\Rightarrow J^* = J_\pi$ & hence, π is optimal.

Converse: π is optimal

$\Rightarrow J^* = J_\pi$

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

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

Policy iteration algorithm

Step 1: Fix π_0

Step 2: Find J_{π_k} corresponding to π_k

Policy evaluation

Step 3: Obtain π_{k+1} as Policy improvement

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

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

Claim: Let π, π' satisfy

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

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

& if π is not optimal, then

$J_{\pi'}(k) < J_\pi(k)$ for at least one state k.

Illustration of policy iteration:

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

$$P_a = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix} \qquad P_b = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}$$

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

$\alpha = 0.9$

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

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

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

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

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

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

Policy Improvement:

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

$$(T J_{\pi_0})(1) = \min \text{ of}$$

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

$$0.5 + 0.9\left(\frac{1}{4} \times 24.12 + \frac{3}{4} \times 25.96\right) \Big)$$

$$= \min(24.12, 23.45) = 23.45 \; (= \text{action b})$$

$$(T J_{\pi_0})(2) = \min\left(\underline{23.12}, \underline{25.95} \right)$$

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

H.W:- Do one more step of policy iteration

i.e., Find $J_{\pi_1}(1), J_{\pi_1}(2)$

& we that to perform policy improvement

You should get $\boxed{\pi_2 = \pi_1}$

Sample path :-

$$S_0 \xrightarrow{\pi(S_0)} S_1$$

$$S_1 \sim P_{S_0 S_1}(\pi(S_0))$$

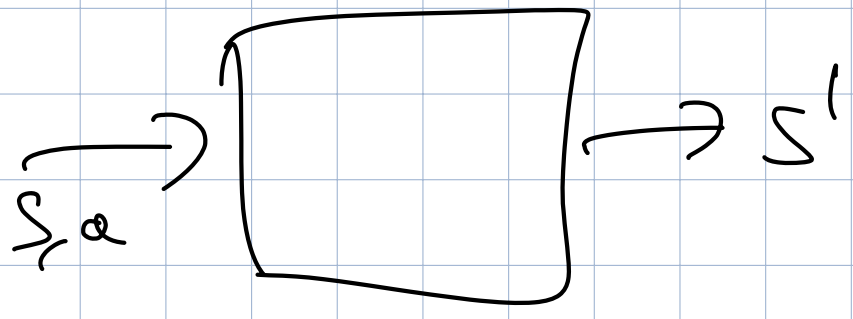
for e.g. $S = \{1, 2, 3\}$

$$S_0 = 1$$

P_{11} (action) = 0

P_{12} (action) = 0.4

P_{13} (— u —) = 0.6



$(S_0, a_0, S_1, a_1, S_2, \text{---})$

Sample path

Policy evaluation (prediction)

$(S_0, \pi(S_0), S_1, \pi(S_1), \text{----})$

Goal: Find J_π

Policy control :-

$(S_0, a_0, S_1, \text{----})$

Reinforcement Learning

Mean estimation:

R.v. v with mean μ and finite variance.

$\{v_1, \cdots, v_n\} \leftarrow$ iid samples from the distribution of v.

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

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

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

$$= \frac{n}{n+1} r_n + \frac{1}{n+1} v_{n+1}$$

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

More generally,

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

η_n: step-size

If $\sum_{n=1}^{\infty} \eta_n = \infty$, $\sum_{n=1}^{\infty} \eta_n^2 < \infty$, then

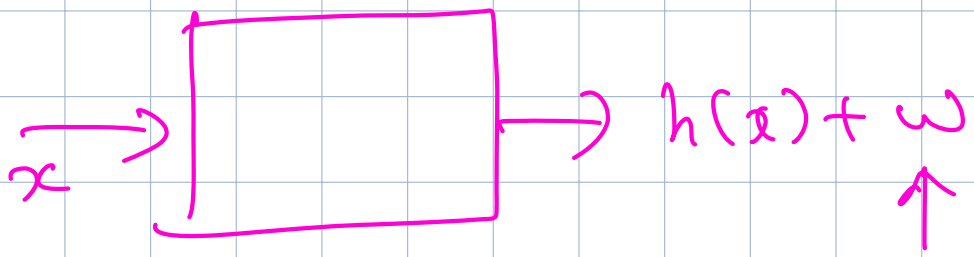
$r_n \longrightarrow \mu$ w.p. 1 as $n \to \infty$.

Stochastic approximation

(Robbins Monro algorithm)

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

h is Lipschitz



w: zero-mean noise

e.g. $w \sim N(0,1)$

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

Under some conditions, it can be shown that $x_n \to x^*$ w.p. 1 as $n \to \infty$,

where $h(x^*) = 0$

Two special cases:-

(1) $\min_x f(x)$ Then $h = f'$

or in higher dimensions

$h = \nabla f$

"Gradient descent"

(2) Want to solve $f(x) = x$ } Fixed point Iteration.

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

In a MDP context, one usually wants to solve $Hx = x$

$H : \mathbb{R}^n \to \mathbb{R}^n$, $r \in \mathbb{R}^n$

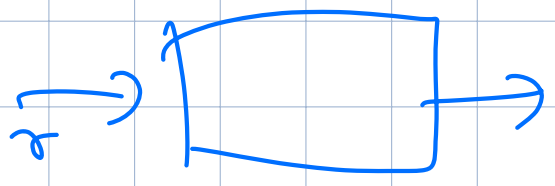
e.g. $H = T_\pi$

A direct method:

$$r_{n+1} = H r_n$$

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

In a RL setting,



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

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

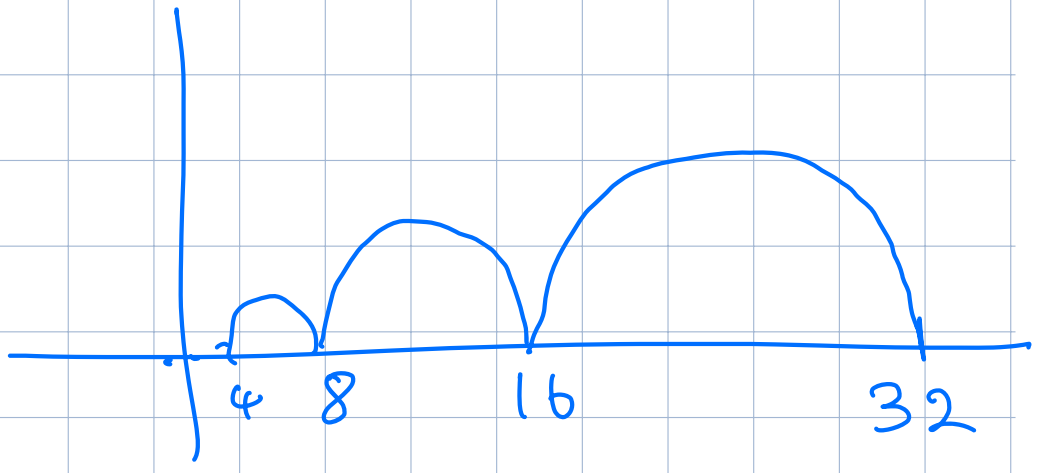
One can show that

$$r_n \to r^* \text{ a.s. as } n \to \infty$$

where $\eta_{r^*} = r^*$

A brief tour of contraction mappings.

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



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

$$f(0) = 0$$

Def: $f: \mathbb{R} \to \mathbb{R}$ is a contraction mapping with modulus $\beta \in (0,1)$ if

$$|f(x) - f(y)| \leq \beta |x - y|, \quad \forall x, y \in \mathbb{R}$$

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

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

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

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

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

↑ discount factor

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

for any stationary π.

Corollary:

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

Corollary :-

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

Lecture-23

Stochastic iterative algorithm for finding fixed point of a contraction mapping

Goal: Solve $H x^* = x^*$

$H : \mathbb{R}^n \to \mathbb{R}^n$, $x \in \mathbb{R}^n$

Update rule:

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

$w_t(i)$: noise

Assumptions:

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

(history)

(A1) $$\mathbb{E}\left(w_t(i) \mid \mathcal{F}_t \right) = 0$$

$$\mathbb{E}\left(w_t^2(i) \mid \mathcal{F}_t \right) \le A + B \|r_t\|^2$$

(A2) $$\sum_{t=1}^{\infty} \eta_t = \infty \,, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty .$$

(A3) H is a max-norm pseudo-contraction

$$\| Hr - r^* \|_\infty \leq \alpha \| r - r^* \|, \quad \forall r$$

Full contraction
$\| Hr - Hr' \|_\infty \leq \alpha \| r - r' \|_\infty$
$\forall r, r'$

Under A1 – A3,

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

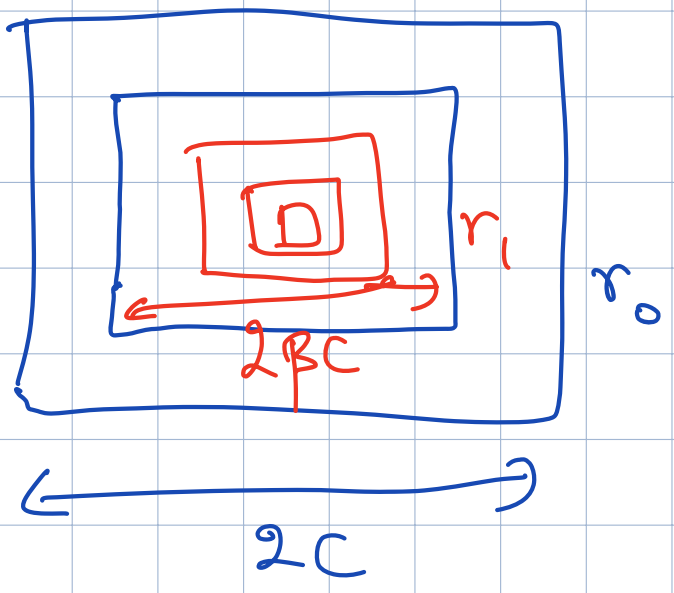
where $Hr^* = r^*$

An intuitive justification:-

$H0 = 0$, $r^* = 0$

$$|r_0(i)| \leq C$$

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



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

Similar behavior, but shrinkage is not immediate

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 :-

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

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

$$J_{t+1}(i) = J_t(i) + \eta_t \Big(\underbrace{g(i,\bar{i}) + \alpha J_t(\bar{i})}_{\text{Proxy for } E\left(g(i,\bar{i}) + \alpha J_t(\bar{i})\right) = (T_\pi J_t)(i)} - J_t(i) \Big)$$

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

Using result for r_t above, we can infer

$J_t \to J_\pi$ w.p.1 as $t \to \infty$

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

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

$\bar{i}$: next state; $\bar{\bar{i}}$: next-to-next state

TD(λ), $\lambda \in (0,1)$:

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

Fix $\lambda < 1$, and multiply $(1-\lambda)\lambda^l$ & 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 $\bar{i} = 1, \dots n$

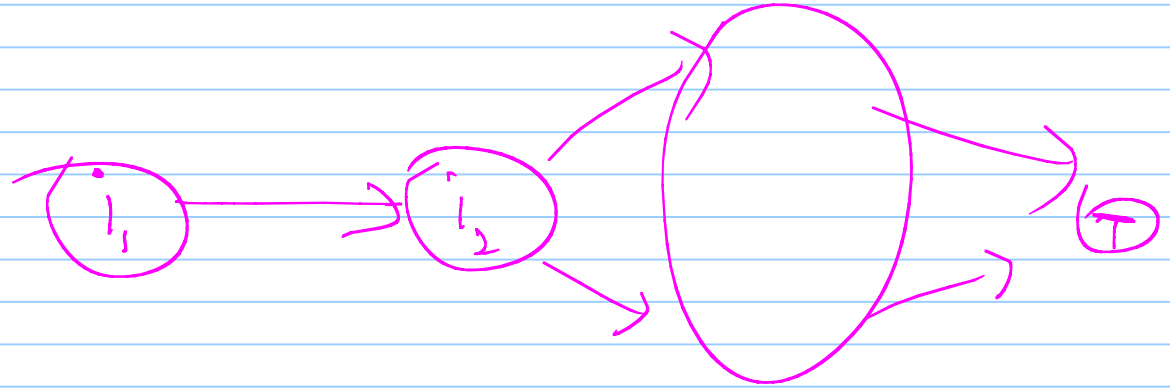
TD(1) :- Simulate episodes of the underlying SSP using π.

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

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

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

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



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

Lecture-24

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

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

Under some conditions, HQ is a contraction mapping.

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

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

$Q_t \to Q^*$ 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

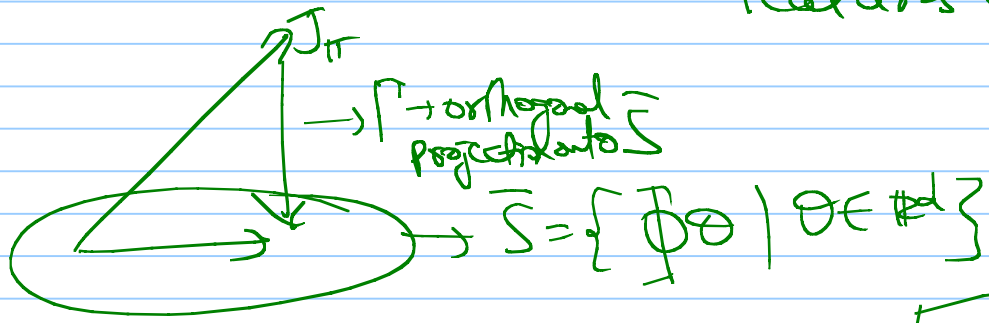
Fix a policy π.

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

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

Policy gradient methods (PG methods)

The main idea behind PG methods is the "likelihood ratio" trick (aka score function)

An illustration of this trick in a very simple setting.

Let X be a discrete r.v. with mass function $p(\theta, \cdot)$ (θ → parameter)

i.e., $p(X = x)$ is parameterized by θ.

$$J(\theta) = E f(X)$$ e.g. $\theta \in \mathbb{R}^d$

Goal: $\min_\theta J(\theta)$ → min among a class of parameterized r.v.s

Want to find best θ using a gradient method

$$\theta_{t+1} = \theta_t - \beta_t \nabla J(\theta_t)$$ ← Gradient descent

Need: "$\nabla J(\theta)$"

Use "likelihood ratio" trick, which is shown below.

$$J(\theta) = \sum_x f(x)\, p(\theta, x)$$ ← LOTUS

$$\nabla J(\theta) = \nabla_\theta \left(\sum_x f(x)\, p(\theta, x) \right)$$

$$= \sum_x f(x)\, \nabla p(\theta, x)$$

need a few regularity conditions to justify interchange of $\sum$ & ∇

mild regularity conditions → usually let one invoke "dominated

Convergence theorem

$$\nabla J(\theta) = \sum_x f(x) \nabla p(\theta, x)$$

$$\nabla J(\theta) = \sum_x \left(f(x) \frac{\nabla p(\theta,x)}{p(\theta,x)} \right) p(\theta,x)$$

→ RHS is an expectation

$$\nabla J(\theta) = E\left(f \frac{\nabla p}{p} \right) = E(f \nabla \log p)$$

To get an estimate of $\nabla J(\theta)$, I can sample from →

$\frac{\nabla p(\theta,x)}{p(\theta,x)}$ → likelihood ratio.

$\nabla \log p(\theta) = \frac{\nabla p(\theta)}{p(\theta)}$

Connecting likelihood ratio to RL:

Fix an SSP problem with state space $\mathcal{X}$, action space $\mathcal{A}$.

Π_{det} : class of admissible stationary deterministic policies

$\{ \pi : \pi : \mathcal{X} \to \mathcal{A}$ & it is time invariant$\}$

θ → [unparametrized policy] → $\pi(x)$ → an action

π: states → actions

for PG methods, we consider stationary randomized policies, i.e.,

$\pi : \mathcal{X} \to \Delta(\mathcal{A})$ → set of all distributions over the actions.

For simplicity, assume all actions available in all states.

e.g. $\pi_\theta(x,a) = \frac{\exp(h(\theta,x,a))}{\sum_b \exp(h(\theta,x,b))}$ → randomized policy parameterized by "θ"

"is the probability of choosing action a in state x"

$\pi_\theta(x) = [\ \pi_\theta(x,a),\ \forall a \in \mathcal{A}\ ]$

π_θ → distribution over $\mathcal{A}$.

$\pi_\theta(x,a) = \frac{\exp(\theta_{x,a})}{\sum_b \exp(\theta_{x,b})}$ θ → vector of size |states|×|actions|

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

Simple example for h: $h(\theta,x,a) = \theta^T \phi(x,a)$ $\in \mathbb{R}$, ϕ → state-action features

$\pi_\theta(x,a) = \frac{\exp(\theta^T \phi(x,a))}{\sum_b \exp(\theta^T \phi(x,b))}$ Boltzmann distribution aka soft-max

$\theta \in \mathbb{R}^d$, $\phi(x,\cdot) \in \mathbb{R}^d$

Assumption A1: Policy π_θ is a continuously differentiable function of θ a.k.a $\nabla \log \pi_\theta$ exists.

Note: Every π_θ is identified by its parameter $\theta \in \mathbb{R}^d$

Goal: $\min_{\theta \in \Theta} J_{\pi_\theta}(x^0)$ → Find an approximately optimal policy in the class of parameterized policies

& we want to find the best parameter in a class

$\{ \pi_\theta \mid \theta \in \Theta \}$ e.g. $\Theta \subset \mathbb{R}^d$

Want to find a $\theta^* \in \arg\min_{\theta \in \Theta} J_{\pi_\theta}(x^0)$ x^0 → fixed start state

"$\in$" is necessary as convex function over parameter θ

A formula for policy gradient in an SSP

SSP: 0 is a special cost-free absorbing state.

Total cost objective:

$$J_{\pi_\theta}(x^0) := \mathbb{E}\left(\sum_{k=0}^{\infty} g(x_k, a_k, x_{k+1}) \,\middle|\, x_0 = x^0 \right)$$
$$a_k \sim \pi_\theta(x_k, \cdot)$$

A trajectory (aka episode) is

$$\tau = x_0\, a_0\, x_1\, a_1 \cdots x_T, \qquad x_T = 0, \text{ the terminal state.}$$

(x_0: fixed)

Probability of seeing a trajectory τ is

$$\pi(x_0, a_0)\, P(x_1 | x_0, a_0) \cdots P(x_T | x_{T-1}, a_{T-1})$$

J_{π_θ} → Can be seen as an average of the total cost over episodes.

Let $D(\tau)$ denote the total cost r.v.

$$D(\tau) = \sum_{k=0}^{T-1} g(x_k, a_k)$$

Here $\tau = (x_0, a_0, \ldots x_{T-1}, a_{T-1}, x_T)$ is the episode

$$J_{\pi_\theta}(x^0) = E(D(\tau))$$

E: expectation over episodes.

$$\nabla J_{\pi_\theta}(x^0) = \nabla_\theta \sum_\tau D(\tau)\, \text{Prob}(\tau \text{ under } \pi_\theta)$$

$D(\tau)$ is not a function of θ ★ interchange of ∇ & $\sum$ justified for finite state action spaces.

$$= \sum_\tau D(\tau)\, \nabla_\theta \text{Prob}(\tau \text{ under } \pi_\theta)$$

$$= \sum_\tau D(\tau)\, \text{Prob}(\tau \text{ under } \pi_\theta)\, \nabla \log\left(\text{Prob}(\tau \text{ under } \pi_\theta)\right)$$

$$= \sum_\tau D(\tau)\, \text{Prob}(\tau \text{ under } \pi_\theta)$$

$$\times \nabla \log\left\{ \pi_\theta(x_0, a_0)\, P(x_1 | x_0, a_0) \cdots \pi_\theta(x_{T-1}, a_{T-1})\, P(x_T | x_{T-1}, a_{T-1}) \right\}$$

why!?

$$= \sum_\tau D(\tau)\, \text{Prob}(\tau \text{ under } \pi_\theta) \sum_{k=0}^{T-1} \nabla \log \pi_\theta(x_k, a_k)$$

Main claim! pTO

$$\nabla J_{\pi_\theta}(x^0) = \underbrace{E}_{\text{expectation over trajectories.}} \left[D(\tau) \sum_{t=0}^{T-1} \nabla \log \pi_\theta(x_t, a_t) \right]$$

An unbiased estimate of $\nabla J_{\pi_\theta}(x^0)$

Given an episode $(x^0, a_0, \ldots, x_T)$

Calculate total cost of this episode, say $\hat{D}$

Calculate likelihood ratio $\hat{\psi} = \sum_{t=0}^{T-1} \nabla \log \pi_\theta(x_t, a_t)$

$$\hat{\nabla} J = \hat{D} \times \hat{\psi}.$$

$$\theta_{t+1} = \theta_t - a(t) \hat{\nabla} J$$

policy gradient update

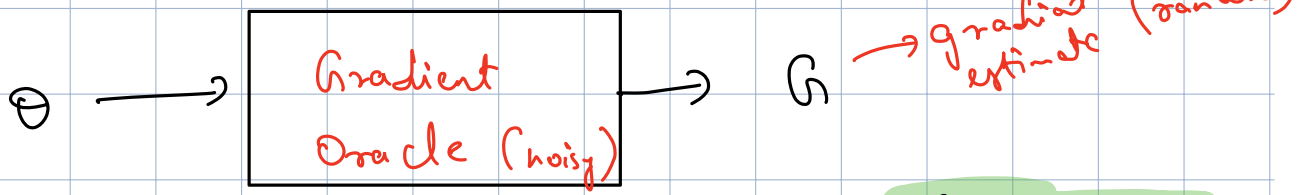
"REINFORCE". (Williams, 1992)

Lecture - 18

A detour → Non-asymptotic analysis of stochastic gradient algorithms

Aim: Solve $\min_{\theta} f(\theta)$ → smooth

Setting: Unbiased gradient information



$$\mathbb{E}(G) = \nabla f(\theta)$$

SG: $\theta_{k+1} = \theta_k - a(k)\, G(\theta_k, \xi_k)$ ①

algorithm

ξ_k ← noise

Assumption:

(A1) $$\mathbb{E}_{\xi_k}\left(G(\theta_k, \xi_k)\right) = \nabla f(\theta_k), \quad \forall k \geq 1$$

$$\mathbb{E}_{\xi_k} \| G(\theta_k, \xi_k) - \nabla f(\theta_k) \|^2 \leq \sigma^2$$

for some $\sigma > 0$.

(A0) f is L-smooth

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \le L\|\theta_1 - \theta_2\|, \quad \forall \theta_1, \theta_2 \in \mathbb{R}^d$$

Asymptotic convergence:

(A2) $$\sum_k a(k) = \infty, \qquad \sum_k a(k)^2 < \infty$$

e.g. $a(k) = \frac{1}{k}$ → standard stochastic approximation conditions.

Under (A0) - (A2), the parameter θ_k governed by ① converges almost surely to the set $\{\theta^* \mid \nabla f(\theta^*) = 0\}$

Tool used for proving this claim: Kushner-Clark Lemma

Eq ① (=)

$$\theta_{k+1} = \theta_k - a(k)\left(\nabla f(\theta_k) + \xi_k\right)$$

$$\xi_k \doteq G(\theta_k, \xi_k) - \nabla f(\theta_k)$$

We want to understand the non-asymptotic performance of the SG algorithm above.

Stationary point, say θ^*, has $\nabla f(\theta^*) = 0$

ϵ-stationary point: $\bar{\theta}$ is an ϵ-stationary point if $E\|\nabla f(\bar{\theta})\| \leq \epsilon$, where the expectation is over the randomness in the SG algorithm.

[Ghadimi-Lan, 2014, SIAM. J. Opt, "Stochastic first/zeroth order]

"Randomized stochastic gradient" (RSG)

Suppose we run (1) for N iterations.

$\{\theta_1, \theta_2, \theta_3, \ldots \theta_N\}$ → (random) set of iterates visited by SG algorithm.

RSG will pick a random iterate as follows:

θ_R is picked uniformly at random from $\{\theta_1, \ldots \theta_N\}$

i.e., $\mathbb{P}(\theta_R = \theta_i) = \frac{1}{N} \quad \forall i$

$$\mathbb{E}\,\theta_R = \frac{1}{N}\sum_{i=1}^{N}\theta_i$$ } average iterate.

The non-asymptotic bound for RSG is of the form

$$\mathbb{E}\,\|\nabla f(\theta_R)\|^2 \leq \frac{\text{Const}}{\sqrt{N}}$$

under suitable choice of step-size.

RL connection: $\theta \rightarrow$ policy parameter

Objective $f \rightarrow$ $J_{\pi_\theta}(x^0)$ or $J(\theta)$

$J_{\pi_\theta}(x^0)$: Value function with start state x^0 in a discounted/SSP

$J(\theta)$: Average cost

Proof of the non-asymptotic bound for RSG:

First $\rightarrow$ derive a bound for a general step-size

Second $\rightarrow$ Specialize the bound above with a particular step size.

$a_i \sim \pi_\theta(\cdot \mid s_i)$ terminal

$\tau = (x_0, a_0, x_1, a_1, \ldots, x_T)$

$$\nabla J_{\pi_\theta}(x^0) = E\left[D(\tau) \sum_{k=0}^{T-1} \nabla \log \pi_\theta(x_k, a_k) \right]$$

expectation over trajectories τ

"Each τ ends in a terminal state.

Random length trajectories.

but length bounded w.p.1 under policy θ is proper".

$\theta \to$ parameter $\quad J(\theta) = J_{\pi_\theta}(x^0)$ ← fixed start state

For RSG analysis/bound to be applicable, we need to show

① $\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \le 2 \|\theta_1 - \theta_2\|$

"J is 2-smooth".

② Gradient estimate from a sample trajectory τ.

$$\hat{G} = D(\tau) \sum_{k=0}^{T-1} \nabla \log \pi_\theta(x_k, a_k)$$

$E(\hat{G}) = \nabla J(\theta)$ ← unbiased gradient estimate.

To verify ① & ② in an RL context, we make the following assumptions:

(B1) State & action spaces are finite

(B2) $\forall \theta$, π_θ is proper.

($\Leftarrow$) "in a finite, say M steps, there is a positive prob. of reaching the terminal state".

(B3) $\| \nabla \log \pi_\theta(x,a) \| \leq G$

$\| \nabla^2 \log \pi_\theta(x,a) \| \leq H$

$\left[\nabla_i \log \pi_\theta(x,a) \right| \leq G$

$\left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \right| \leq H$

"Showing J is L-smooth as a function of θ"

Use this identity: for any twice diff'ble fn f:

$$\nabla^2 f(\theta) = f(\theta) \left(\nabla^2 \log f(\theta) + \nabla \log f(\theta) \nabla \log f(\theta)^T \right)$$

(Scalar Case: $f''(\theta) = f(\theta) \left(\frac{d^2}{d\theta^2} (\log f(\theta)) + \left(\frac{d}{d\theta} \log f(\theta) \right)^2 \right)$

$$J(\theta) = \sum_{\tau} \text{Prob}(\tau \text{ under } \pi_\theta)\, D(\tau)$$

Total cost from τ.

$$= \sum_{\tau} p_\theta(\tau)\, D(\tau)$$

$$\nabla^2 J(\theta) = \sum_{\tau} p_\theta(\tau)\, \nabla \log p_\theta(\tau)\, \nabla \log p_\theta(\tau)^T D(\tau)$$

$$+ \sum_{\tau} p_\theta(\tau)\, \nabla^2 \log p_\theta(\tau)\, D(\tau)$$

(*)

Proper policies (see (B2))

Assume: Trajectory length is finite w.p. 1

Let T be a r.v. denoting the length of a random trajectory τ.

Then $|T| < C_1 < \infty$ w.p. 1.

Fact 2: Finite state-action spaces ((B1))

$\Rightarrow \sup_{x,a} |g(x,a)| < C_2 < \infty.$

Fact 1 & 2 $\Rightarrow$

$$D(\tau) = \sum_{k=0}^{T-1} g(x_k, a_k)$$

$$|D(\tau)| \leq C_2 C_1$$

Using (B3) ($\|\nabla \log \pi\| \leq G$, $\|\nabla^2 \log \pi(\cdot|\cdot)\| \leq H$)

in (*),

$$\|\nabla^2 J(\theta)\| \leq C_1 C_2 \left(\sum_\tau P_\theta(\tau) \|\nabla \log P_\theta(\tau) \nabla \log P_\theta(\tau)^T\| + \sum_\tau P_\theta(\tau) \|\nabla^2 \log P_\theta(\tau)\| \right)$$

$$\leq C_1 C_2 \left(G^2 C_1 + H C_1 \right),$$

Since

$$\tau = (x_0, a_0, \dots x_T)$$

$$P_\theta(\tau) = \pi_\theta(x_0, a_0)\, P(x_1|x_0 a_0) \cdots P(x_T|x_{T-1}, a_{T-1})$$

$$\nabla \log P_\theta(\tau) = \sum_{k=0}^{T-1} \nabla \log \pi_\theta(x_k, a_k) \leq G C_1$$

So, $J(\theta)$ is L-smooth, where

$$L = C_1 C_2 \left(G^2 C_1 + H C_1 \right)$$

Homework:

$$E \| \hat{G} - E\hat{G} \|^2$$

$$= E \| \hat{G} - \nabla J(\theta) \|^2$$

$$\leq E \| \hat{G} \|^2 \qquad \hat{G} = D(\tau) \sum_{k=0}^{T-1} \nabla \log \pi_\theta(s)$$

$$\leq C_1^2 C_2^2 C_1 E \sum_{k=0}^{T-1} \| \nabla \log \pi(x_k, a_k) \|^2$$

$$\leq C_1^3 C_2^2 G^2 C_1$$

So, variance of the gradient estimate is bounded.

$\Rightarrow$ RSG analysis is applicable, leading to

$$E \| \nabla J(\theta_R) \|^2 \leq \frac{\text{const}}{\sqrt{N}},$$

where θ_k chosen unif. at. random from

$$\{\theta_1, -- \theta_N\}$$

& $\theta_{k+1} = \theta_k - a(k)\left(\hat{G}_k\right)$

↓

single trajectory estimate of $\nabla J(\theta_k)$