Lecture-18*

A high level idea about the RL setting:

Task: Policy evaluation, given a policy π in your favorite MDP (SSP, discounted, etc)

If we know the transition probabilities, then we know T_π & hence, can solve

$$J_\pi = T_\pi J_\pi \quad \text{directly}$$

or start with some J & keep applying T_π.

In RL setting: $p_{ij}(a)$ is not known.

Instead, we get to see a sample path.

$$s_0 \xrightarrow{\pi(s_0)} s_1 \xrightarrow{\pi(s_1)} s_2 \longrightarrow \cdots \quad \text{so on}$$

s_1: Sampled from $P_{s_0 s_1}(\pi(s_0))$

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

p_{ij}: this ain't known, but we get to see a sample path that follows $p_{ij}(a)$.

Assume cost function is known & is a deterministic function.

For simplicity, we may assume $g(\cdot,\cdot)$ is a function of the current state & current action.

The goal is to solve MDP as before, but the information available is only through the sample path.

So, an element of "learning" involved.

Policy evaluation: Temporal difference (TD) learning

Control: Q-learning

Background on Stochastic iterative algorithms

Ref: Chapter 4 of NDP book.

Aim: Solve the system of equations

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

in RL applications, think of H as the Bellman operator T or T^π & r as the value function

A direct approach: Start with r_0 & do $r_{n+1} = Hr_n$

a variation to the above!

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

Iterative scheme →

β: step size, $0 < \beta < 1$

An interesting special case!

$$Hr = r - \nabla f(r) \quad \text{for some } f : \mathbb{R}^n \to \mathbb{R}^n$$

Then $Hr = r \iff \nabla f(r) = 0$

problem of finding the minima of f.

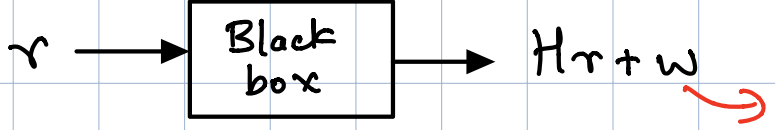
$$r_{n+1} = (1-\beta) r_n + \beta (r_n - \nabla f(r_n))$$

(or) $$r_{n+1} = r_n - \beta \nabla f(r_n)$$ ← Gradient descent algorithm

Remark: If $r_n \to r^*$ and H is continuous at r^*, then $Hr^* = r^*$.

So far, we assumed Hr is perfectly observable $\forall r$

Now, we consider a setting where H is not precisely known.



w: zero-mean noise e.g. $w \sim N(0,1)$

$$r_{m+1} = (1-\beta) r_m + \beta (H r_m + w_m) \quad \text{——— (*)}$$

noisy observation of Hr_m

Simplicity: assume $E(w_i) = 0$, $E w_i^2 < \infty$, w_i i.i.d.

or independent & identically distributed.

(*) is a stochastic iterative scheme.

Want: $r_m \to r^*$ with probability (w.p.) 1 as $m \to \infty$

→ same as almost surely (a.s.)

We shall identify conditions on

(i) H — e.g. contraction

(ii) step sizes — e.g. constant β or diminishing $\beta_m = \frac{1}{m}$

(iii) noise w_m — e.g. $E w_i = 0$, $E w_i^2 < \infty$ & $\{w_i\}$ i.i.d.

so that we can infer $r_m \to r^*$ w.p. 1 as $m \to \infty$

An application: Mean estimation

Consider a random variable (r.v.) X with mean μ & finite variance, say σ^2.

Suppose we are given iid samples $X_1, \ldots, X_m$

Let r_m be the estimate of μ.

$$r_m = \frac{1}{m} \sum_{k=1}^{m} x_k$$

$$r_{m+1} = \frac{1}{m+1} \sum_{k=1}^{m+1} x_k$$

$$= \frac{m}{m+1} \left(\frac{1}{m} \sum_{k=1}^{m} x_n \right) + \frac{1}{m+1} x_{m+1}$$

$$r_{m+1} = \frac{m}{m+1} r_m + \frac{1}{m+1} x_{m+1}$$ ← iterative scheme for updating sample mean

$$r_{m+1} = r_m + \frac{1}{m+1} (x_{m+1} - r_m)$$

$$r_{m+1} = r_m + \beta_m (x_{m+1} - r_m) \quad — \quad (**)$$

where $\beta_m = \frac{1}{m+1}$

(**) resembles the sto. iter. algo (*)

What does Strong Law of large numbers say about r_m?

$$r_m \longrightarrow \mu \text{ a.s. as } m \to \infty$$

with stepsize $\beta_m = \frac{1}{m+1}$

$$r_{m+1} = r_m + \beta_m \left(X_{m+1} - r_m \right)$$

$$= r_m + \beta_m \left((\mu - r_m) + (X_{m+1} - \mu) \right)$$

$$r_{m+1} = (1 - \beta_m) r_m + \beta_m \left(\underbrace{\mu}_{Hr_m} + \underbrace{(X_{m+1} - \mu)}_{w_m} \right)$$

$$E w_m = 0, \quad E w_m^2 < \infty, \quad Hr^* = r^* \Leftrightarrow r^* = \mu.$$

So, (**) is really a sto. iterative algorithm.

And, we get $r_m \to r^* (= \mu)$ a.s. as $m \to \infty$.

The theory of sto. itr. algo that we shall develop ensures $r_m \to \mu$ a.s. as $m \to \infty$ for more general step sizes that satisfy

$$\sum_m \beta_m = \infty, \quad \sum_m \beta_m^2 < \infty$$

Question: Why do we need these conditions?

On step-size requirements:

$$r_{m+1} = (1-\beta_m) r_m + \beta_m (H r_m + w_m)$$

$$r_{m+1} = r_m + \beta_m \underbrace{(H r_m + w_m - r_m)}_{\text{increment}}$$

Suppose $\{w_i\}$ is independent of $r_m, \forall m$ & has variance σ^2

Variance of r_{m+1}

$$= \text{Var}(r_{m+1})$$

$$= \text{Var}\left[(1-\beta_m) r_m + \beta_m H r_m \right] + \beta_m^2 \text{Var}(w_m)$$

$$= \text{Var}\left[(1-\beta_m) r_m + \beta_m H r_m \right] + \beta_m^2 \sigma^2$$

$$\geq \beta_m^2 \sigma^2$$

Now, if $\beta_m = \beta \ \forall m$, then

$$\text{Var}(r_{m+1}) \geq \beta^2 \sigma^2$$

So, fixing $\beta > 0$, $r_m \not\to r^*$ for any r^*.

So, the step-size has to vanish asymptotically.

But, β_m cannot go down too fast.

$$r_{m+1} = r_m + \beta_m (H r_m + w_m - r_m)$$

$$| r_m - r_0 | \leq \sum_{\tau=0}^{m-1} \beta_\tau | H r_\tau - r_\tau + w_\tau |$$

So, if $| H r_\tau - r_\tau + w_\tau | \leq C_1$ and

if $\sum_{\tau=0}^{\infty} \beta_\tau \leq C_2 < \infty$, then

$| r_m - r_0 |$ is bounded above

$(\Rightarrow)$ r_m is within a certain radius of r_0

problematic if r^* lies outside the radius.

So, we need $\sum_{\tau} \beta_\tau = \infty$

Lecture-19*

Notions of convergence of r.v.s:

I Almost sure or w.p. 1 Convergence!

$\{X_n\}$ r.v.s.

$X_n \to X$ a.s. or $X_n \to X$ w.p.1 as $n \to \infty$

if $$P\left(\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1$$

e.g. SLLN

II Convergence in probability

$X_n \xrightarrow{P} X$ if $$\lim_{n\to\infty} P(\omega \mid |X_n(\omega) - X(\omega)| > \epsilon) = 0, \quad \forall \epsilon > 0.$$

usually written as

$$P(|X_n - X| > \epsilon)$$

e.g. WLLN

III L^2- Convergence

$X_n \xrightarrow{L^2} X$ if $E|X_n - X|^2 \to 0$ as $n \to \infty$.

IV Convergence in distribution

$X_n \xrightarrow{d} X$ if $Ef(X_n) \to Ef(X)$

$\forall$ bdd continuous f.

(Check textbook for equivalent definitions). e.g. CLT.

Note: a.s. $\Rightarrow$ (conv. in prob) $\Rightarrow$ (conv. in dist.)

In this course, we provide a.s. convergence guarantees for the popular RL algorithms, e.g., TD-learning & Q-learning.

A crash course in selected topics in prob:

Def: A collection $\mathcal{F}$ of subsets of Ω is a **σ-field** if

(a) $\phi \in \mathcal{F}$ (b) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

(c) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

Def: A probability measure P on $(\Omega, \mathcal{F})$ is a function $P: \mathcal{F} \to [0,1]$ satisfying

(a) $P(\Omega) = 1$

(b) $A_1, A_2, \ldots$ disjoint, i.e., $A_i \cap A_j = \phi$ $(i \neq j)$

then, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Conditional expectation: Check CS601S course notes, or any standard probability textbook.
(e.g. Grimmett-Stirzaker, Prob & random process)

σ-field generated by a r.v. X, denoted by $\sigma(X)$:

$\sigma(X)$ is the smallest σ-field containing all sets of the form $\{X \le \alpha\} = \{\omega \mid X(\omega) \le \alpha\}$

↓

Take all such sets, add complements & countable unions to get $\sigma(X)$.

Given $\sigma(X)$, $E(Y \mid \sigma(X))$ is denoted by $E(Y \mid X)$

Extend $\sigma(X)$ to the σ-field generated by a collection of r.v.s $\{X_1, \ldots X_m\}$ & denote it by $\sigma(X_1, \ldots X_m)$

Also, $E(Y \mid \sigma(X_1 \ldots X_m)) = E(Y \mid X_1 \ldots X_m)$

Filtration: increasing sequence of σ-fields.

e.g. $\mathcal{F}_1 = \sigma(X_1)$, $\mathcal{F}_2 = \sigma(X_1, X_2)$, and so on

$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \ldots$ $\{\mathcal{F}_k\}$ is a filtration

$\mathcal{F}_k = \sigma(X_1, \ldots X_k)$ ← filtration

$(X_k, \mathcal{F}_k)$ is a martingale if

$$E(X_{k+1} \mid \mathcal{F}_k) = X_k, \quad \forall k$$

Note: If $E(X_{k+1} \mid \mathcal{F}_k) = 0$ then $(X_k, \mathcal{F}_k)$ is a martingale difference sequence.

(*) $\{X_k\}$ are not necessarily i.i.d.

To see martingale difference, contraction mapping, etc in a stoc. iter. algo in RL context,
Let's look at TD(0) learning.

Fix policy π. Consider a discounted MDP.

Also, assume single stage cost $g(i, \pi(i))$ & is indep. of next state

$$(HJ)(i) = g(i, \pi(i)) + \alpha \sum_j p_{ij}(\pi(i)) J(j)$$

Want to solve $J(i) = (HJ)(i) \quad \forall i \quad - (*)$

Sto. iter. algo for solving (*)

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

$\bar{i}$ is a state observed along the sample path, i.e., in state i, take action $\pi(i)$, & we transition to $\bar{i}$ & $\bar{i}$ follows $P_{ij}(\pi(i))$

$$= J_t(i) + \beta_t\Bigg(g(i,\pi(i)) + \alpha \sum_j P_{ij}(\pi(i))\, J_t(j) - J_t(i)$$

$$+ \Big\{ g(i,\pi(i)) + \alpha J_t(\bar{i}) - g(i,\pi(i)) - \alpha \sum_j P_{ij}(\pi(i)) J_t(j) \Big\}\Bigg)$$

g added & subtracted this

$$J_{t+1}(i) = J_t(i) + \beta_t\left(\left((HJ_t)(i) - J_t(i)\right) + \omega_t(i)\right),$$

where $\omega_t(i)$ is the quantity in flower braces.

In the absence of noise term $\omega_t(i)$, it is easy to infer

$$J_t \to J_\pi \text{ , where } J_\pi = H J_\pi$$

What conditions of $w_t(i)$ would ensure the above convergence?

Let $\mathcal{F}_t = \sigma(J_0, \ldots, J_t, w_0, \ldots w_{t-1})$

$$E(w_t(i) \mid \mathcal{F}_t) = 0$$

We shall get back to TD(0) later.

Coming next:

Sto. iter. algo $\quad r_{m+1}(i) = r_m(i) + \beta_m (H r_m(i) + w_m(i) - r_m(i))$

Thm,

if ① H is a contraction mapping wrt max-norm

$Hr^* = r^*$

② $E(w_m(i) \mid \mathcal{F}_m) = 0$

$E(w_m^2(i) \mid \mathcal{F}_m) \le$ "(bounded)"

③ $\sum \beta_m = \infty, \quad \sum \beta_m^2 < \infty$

Thm, $r_m \to r^*$ a.s. as $m \to \infty$.

Lecture-20

Convergence of stochastic iterative algorithms

Ref: Sec 4.3 of NDP book

$$r_{m+1}(i) = (1-\beta_m)\, r_m(i) + \beta_m \left(H r_m(i) + w_m(i) \right), \quad i = 1, \ldots, n \qquad (1)$$

It helps to look at per-component update because RL applications would estimate the value function, say $J_\pi(i)$, with start state i using an iterate of the form $r_{m+1}(i)$

The underlying σ-field is

$$\mathcal{F}_m = \sigma\left(r_0(i), \ldots, r_m(i), w_0(i), \ldots, w_{m-1}(i), \; i = 1, \ldots, n \right)$$

all the random variables upto time m

Assumptions:

<(conditions on noise factors $\{w_m\}$>

(A1) $\forall i$ and $\forall m$, (i) $E\left(w_m(i) \mid \mathcal{F}_m \right) = 0$ (conditionally zero mean)

$$\text{(ii)}\; E\left(w_m^2(i) \mid \mathcal{F}_m \right) \leq A + B \| r_m \|^2$$

bound on the conditional variance

Note: For mean-estimation example,

$w_m \to$ iid, indep. of $\{r_m\}$, bounded variance

So, $E(w_m \mid \mathcal{F}_m) = E(w_m) = 0$

$E(w_m^2 \mid \mathcal{F}_m) = E(w_m^2) \le (\text{const})^2$

↓ since $\text{Var}(w_m)$ is bounded.

i.e., $B = 0$ in (A1) part (ii)

(A2) ⟨Condition on H⟩

Define weighted norm

$$\|r\|_\xi = \max_i \frac{|r(i)|}{\xi(i)}$$

→ we saw an example of such a norm in SSP

If $\xi(i) \equiv 1 \ \forall i$, then we get the max-norm. → standard discounted MDP

Assume H is a "weighted max-norm pseudo-contraction", i.e., $\exists$ a positive $\xi = (\xi(1), \dots, \xi(n))$ and a constant $\beta \in [0,1)$ s.t.

$$(*) \to \boxed{\|Hr - r^*\|_\xi \le \beta \|r - r^*\|_\xi}, \quad \forall r \in \mathbb{R}^n.$$

$$(\Leftrightarrow) \quad \frac{|Hr(i) - r^*(i)|}{\xi(i)} \le \beta \max_j \frac{|r(j) - r^*(j)|}{\xi(j)}, \quad \forall i, r.$$

[Recall → full contraction if
$\| Hr - Hr' \|_\xi \leq \beta \| r - r' \|_\xi$]

Claim w/o proof: (*) $\Rightarrow$ $\exists r^* = r^*$ & r^* is the unique fixed point.

A full-contraction is trivially a pseudo-contraction.

Connection to MDPs:

① In a SSP with all proper policies, the Bellman operator is a contraction.

② In a discounted MDP with bounded single stage cost, the Bellman operator is a contraction.

(A3) "Step size conditions"

$$\sum_m \beta_m = \infty, \quad \sum_m \beta_m^2 < \infty$$

Theorem 1 "Convergence of Sto. itr. algo"

Assume (A1), (A2), (A3). For r_m updated w/ Eq. (1), we have $r_m \to r^*$ a.s. as $m \to \infty$

A rough intuitive argument:

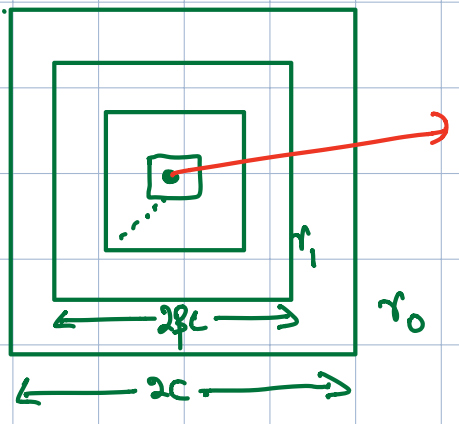
Take a very special case where

$H0 = 0$ i.e., $r^* = 0$, H is a contraction with modulus β

$w_m = 0 \quad \forall m$

$$r_{m+1}(i) = H r_m(i)$$

Suppose the initial point $|r_0(i)| \le c$



In the sto. iter. algo (Eq ①),

$$r_{m+1}(i) = (1-\beta_m) r_m(i) + \beta_m \left(H r_m(i) + w_m(i) \right)$$

noise

Because of noise, the shrinkage is not immediate, but the iterate r_m gets into smaller & smaller box as the algorithm proceeds.

A convergence result under monotonicity

Motivation: SSP where $\exists$ at least one proper policy & improper policies have infinite cost. In such SSPs, the Bellman operator is not "contractive", but we still have "monotonicity".

Instead of (A2), assume the following:

(A2') (i) H is monotone, i.e., $r \le r' \Rightarrow Hr \le Hr'$, $\forall r, r'$.

(ii) $\exists\, r^*$ s.t. $Hr^* = r^*$ & r^* is unique.

(iii) e = vector of all ones, $e \in \mathbb{R}^n$.

$\delta > 0$

$$Hr - \delta e \le H(r - \delta e) \le H(r + \delta e) \le Hr + \delta e$$

Theorem 2: Assume (A1), (A2') & (A3).

Consider the sto. iter. algo that updates r_m using Eq (1).

Suppose r_m is bounded w.p. 1 ie., $\sup_{m,i} |r_m(i)| < \infty$

Then, $r_m \to r^*$ a.s. as $m \to \infty$

Note: Unlike Thm 1, for the monotone case, we require the iterate r_m to satisfy a "boundedness" requirement.

A sufficient condition to ensure boundedness of the iterates:

The conditions are

(i) $\{\omega_m(i)\}$ satisfies (A1)

(ii) $\{\beta_m\}$ satisfies (A3)

(iii) $\exists$ a positive vector ξ, a $\beta \in [0,1)$ and a scalar $D > 0$ s.t.

$$\|H r_m\|_\xi \le \beta \|r_m\|_\xi + D, \quad \forall m$$

Under (i), (ii), (iii), r_m is bounded w.p. 1.

Remark: If H is a pseudo-contraction, then

$$\|H r_t\|_\xi = \|H r_t - r^* + r^*\|_\xi \le \|H r_t - r^*\|_\xi + \|r^*\|_\xi$$

$$\le \beta \|r_t - r^*\|_\xi + \|r^*\|_\xi$$

$$\le \beta \|r_t\|_\xi + (1+\beta)\|r^*\|_\xi$$

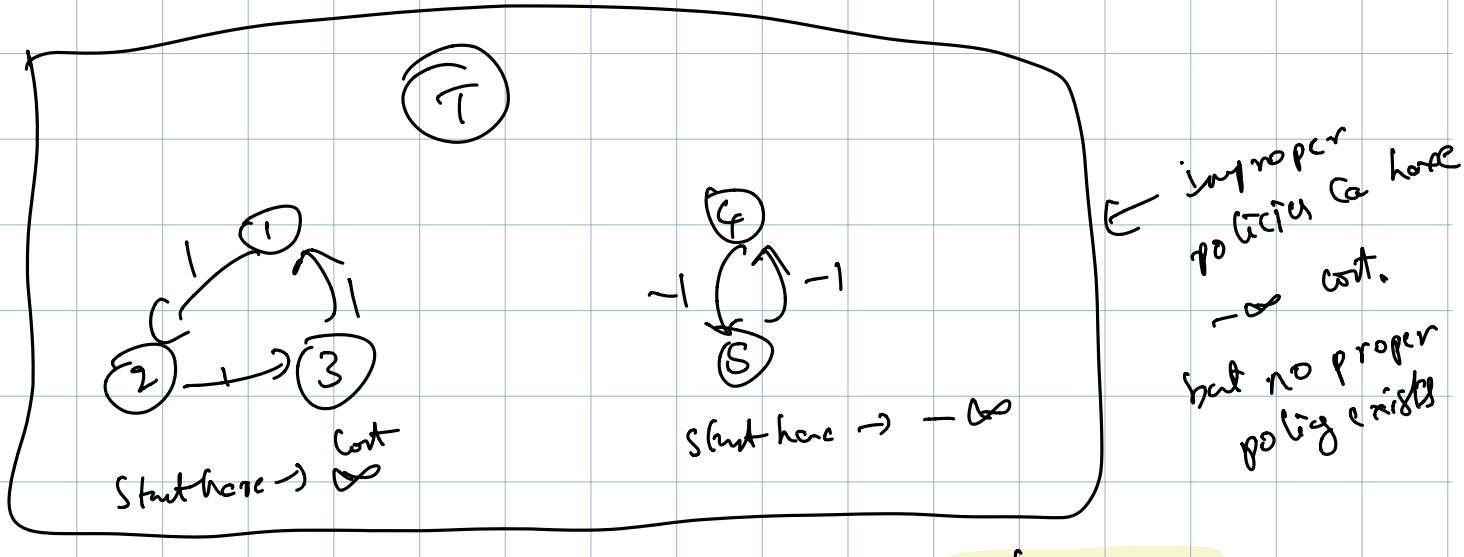
So, a pseudo-contractive $\mathcal{H}$ satisfies condition (iii) & hence, the iterate is bounded in the pseudo-contractive case.

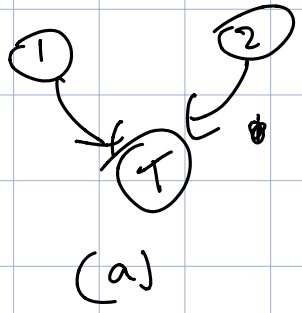
Summary

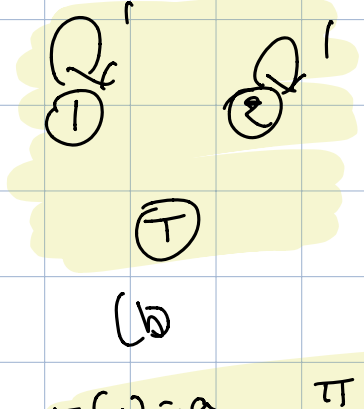
Eqn (1) converges

$\mathcal{H}$ is Contractive

$\mathcal{H}$ is Monotone + iterate bounded.







$J(\pi) = (\quad , -\infty)$

↑

improper policy

but no state without cost.