Lecture-28* (continued)

Reinforcement learning with function approximation

Why approximation?

TD/Q-learning: use look-up table representation
i.e., need an entry $Q(i,a)$ or $J(i)$
for every state i & action a.

On MDPs with large state spaces, these algorithms may not even be implementable.

e.g. Go: 10^{170} states.
other practical applications have large state spaces.

Practical alternative: Approximate J^π or Q^π.

Let's look at approximation in value space i.e.,

$$J^\pi(i) \approx \tilde{J}(i, r)$$

(r: parameter.)

E.g., Linear function approximation

$$\tilde{J}(i, r) = \phi(i)^T r$$

($\phi(i)$: feature vector; r: parameter)

$\phi(i) \in \mathbb{R}^d$, $r \in \mathbb{R}^d$, $|X| >> d$

So, no lookup table. In place of $J^{\pi}(i)$, we use $\tilde{J}(i,r)$

Question: What "r" to use in the approximation?
What features $\phi(\cdot)$ to employ → out of scope of this course.

$\tilde{J}$ → non-linear function of features.

The question of feature selection is orthogonal to choosing the best parameter tuning using these approximations. We focus on the latter.

Suppose we use $\tilde{J}$ in place of J^* & pick actions using a greedy policy:

$$\pi(i) = \arg\min_{a} \sum_{j} p_{ij}(a)\left(g(i,a,j) + \alpha \tilde{J}(j,r)\right)$$

If $\tilde{J}$ is close to J^*, then is π close to π^*?

Prop 1: (Ref: Chap. 6 of NDP book)

α → discount factor Assume finite state & finite action space.

$\|\cdot\|_\infty$ → max-norm. We have a discounted MDP.

Suppose we have a vector J s.t. $\|J - J^*\|_\infty = \epsilon$, $\epsilon \geq 0$

If π is a greedy policy based on J, then

$$\| J_\pi - J^* \| \leq \frac{\alpha \epsilon}{1-\alpha}.$$

Further, one can choose an ϵ_0 s.t. $\forall \epsilon < \epsilon_0$,

π is an optimal policy.

Pf:

$$\| J_\pi - J^* \|_\infty \overset{\text{since } J_\pi \text{ is the fixed pt of } T_\pi}{=} \| T_\pi J_\pi - J^* \|_\infty$$

$$\overset{\Delta^{\text{le}} \text{ ineq.}}{\leq} \| T_\pi J_\pi - T_\pi J \|_\infty + \| T_\pi J - J^* \|_\infty$$

(because π is greedy wrt J, we have $T_\pi J = TJ$)

$$\overset{T_\pi \text{ is } \alpha\text{-contraction in } \|\cdot\|_\infty}{\leq} \alpha \| J_\pi - J \|_\infty + \| TJ - J^* \|_\infty$$

$$\overset{J^* = TJ^* \text{ \& } T \text{ is } \alpha\text{-contraction in } \|\cdot\|_\infty}{\leq} \alpha \| J_\pi - J \|_\infty + \alpha \| J - J^* \|_\infty$$

$$\| J_\pi - J \|_\infty \leq \alpha \| J_\pi - J \|_\infty + \alpha \epsilon$$

$$\text{So, } \| J_\pi - J \|_\infty \leq \frac{\alpha \epsilon}{1-\alpha}$$

Let $\delta := \min_{\pi'} \| J^{\pi'} - J^* \|_\infty$

min attained since # of policies finite. So $\delta > 0$.

Choose ϵ s.t. $\frac{\alpha\epsilon}{1-\alpha} < \delta$ & let $\bar{\pi}$ be the greedy policy with this ϵ.

Then $\| J_{\bar{\pi}} - J^* \| < \delta \Rightarrow \bar{\pi} = \pi^*$. ■

Approximate policy evaluation using TD-type algorithms

Ref: DPOC-vol. II, 4th edition
Section 6.3

Consider a finite-state MDP.

Fix policy $\pi \in$ we shall not attach "π" to the symbols used & keep the policy implicit.

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

Transition probabilities $= p_{ij}$ (skipping π in notation here)

Fix $\pi \Rightarrow$ we have a Markov chain at hand.

Aim: Estimate $J_\pi(i) = E\left(\sum_{k=0}^{\infty} \alpha^k g(i_k, i_{k+1}) \mid i_0 = i \right)$

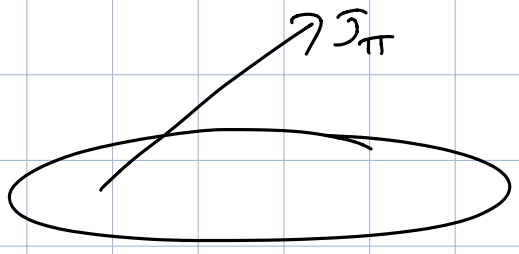
no π here, it is implicit

in Reaction chosen.

Linear function approximation:

$$\tilde{J}(i,r) = \phi(i)^T r \quad , \quad i = 1 \dots n$$

$$\phi(i) \in \mathbb{R}^d, \quad r \in \mathbb{R}^d, \quad \text{"}n >> d\text{"}.$$



$$S = \{ \Phi r \mid r \in \mathbb{R}^d \}$$

Linear space

Let $$\Phi = \begin{bmatrix} — \ \phi(1)^T \ — \\ — \ \phi(2)^T \ — \\ \vdots \\ — \ \phi(n)^T \ — \end{bmatrix}$$ Feature matrix

$$\Phi = \begin{bmatrix} \phi_1(1) & \cdots & \phi_d(1) \\ \vdots & & \\ \phi_1(n) & \cdots & \phi_d(n) \end{bmatrix}$$

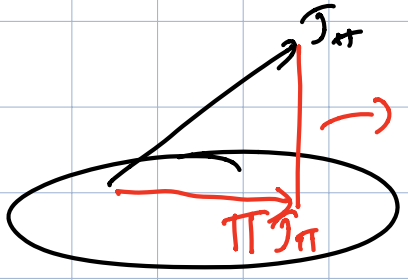
nxd matrix → a tall one.

$$\tilde{J}_r = \begin{bmatrix} \tilde{J}(1,r) & \cdots & \tilde{J}(n,r) \end{bmatrix}$$

$$\tilde{J}_r = \Phi r$$

Aim: Find the best approximation to J_π in the space $S = \{\Phi r \mid r \in \mathbb{R}^d\}$

Lecture-30 (continued after Markov chains review)



Π: Projection operation

$S = \{\Phi r \mid r \in \mathbb{R}^d\}$

For regular policy evaluation, one solves

$$J_\pi = T_\pi J_\pi \leftarrow \text{fixed point relation}$$

$$J_\pi \approx \Phi r$$

$$\Phi r = T_\pi(\Phi r)$$

doesn't make sense since T_π could map any Φr outside S.

Projected fixed point equation

$$\Phi r = \underbrace{\Pi T_\pi}_{\Pi}(\Phi r) \quad \text{---(*)}$$

Π: projection onto linear space S

projected fixed point equation

$\Pi \Rightarrow$ definition requires stationary distribution of the markov chain underlying policy π.

(can solve (*) if "ΠT_π" is a contraction.
& we will show it is the case.

Lecture-31*

TD with linear function approximation

Assumptions:

(C1) The Markov chain underlying policy π is irreducible, positive recurrent
i.e., $\exists$ a stationary distribution $\{\xi_1, \ldots \xi_n\}$ for this chain [$\xi = \xi P$]

(C2) The matrix Φ has full column rank
or rank $(\Phi) = d$ [Note: we assume $n \gg d$]

Towards a projected fixed point equation: (Policy π fixed throughout)

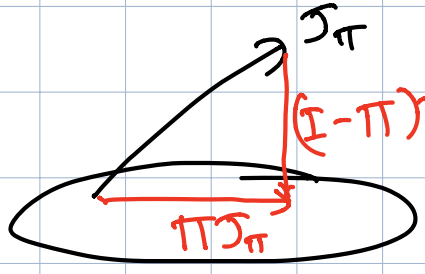
for a n-vector J, define → Stationary distribution

$$\|J\|_\xi^2 = \sum_{i=1}^{n} \xi_i (J(i))^2$$

→ weighted norm

$$D = \begin{bmatrix} \xi_1 & & \\ & \xi_2 & \\ & & \ddots \\ & & & \xi_n \end{bmatrix}$$

$$\|J\|_\xi^2 = J^T D J$$



$S = \{\Phi r \mid r \in \mathbb{R}^d\}$

Π: Projection operator

"Projection is orthogonal & performed using $\|\cdot\|_\xi$"

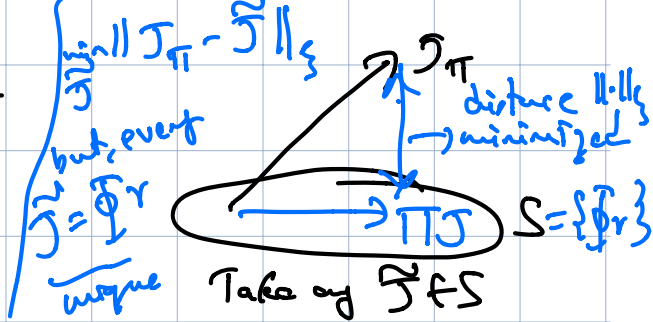
ΠJ_π is the "unique" vector in S that minimizes $\|J_\pi - \tilde{J}\|_\xi$, over all $\tilde{J} \in S$.

Any $\tilde{J} \in S$ is of the form Φr

[Any $\tilde{J} \in S$ can be uniquely written as some Φr since Φ is full column rank]

So, $r^* = \underset{r \in \mathbb{R}^d}{\operatorname{argmin}} \|J_\pi - \Phi r\|_\xi^2$

and $\Pi J_\pi = \Phi r^*$



To find r^* :

$$\nabla_r \left\| J_\pi - \Phi r^* \right\|_\xi^2 = 0$$

$$\Leftrightarrow \nabla_r \left(J_\pi - \Phi r^* \right)^T D \left(J_\pi - \Phi r^* \right) = 0$$

$$\Leftrightarrow \Phi^T D \left(J_\pi - \Phi r^* \right) = 0 \quad \text{— (*)}$$

$$\Leftrightarrow \Phi^T D J_\pi - \Phi^T D \Phi r^* = 0$$

$$\Leftrightarrow \Phi^T D \Phi r^* = \Phi^T D J_\pi$$

$$\Leftrightarrow r^* = \left(\Phi^T D \Phi \right)^{-1} \Phi^T D J_\pi$$

Why is this invertible?
Φ → full rank and
diagonal elements of $D > 0$.

Cannot solve $\Phi r = T(\Phi r)$
$T(\Phi r)$ can be outside S
So, project $T(\Phi r)$
ie., $\Pi(T(\Phi r))$
& then solve
$\Phi r = \Pi(T(\Phi r))$

Φ is $n \times d$
D is $n \times n$
r is $d \times 1$

Now, $\Pi J_\pi = \Phi r^*$

$$\Rightarrow \Pi = \Phi \left(\Phi^T D \Phi \right)^{-1} \Phi^T D$$

Π : projection operator

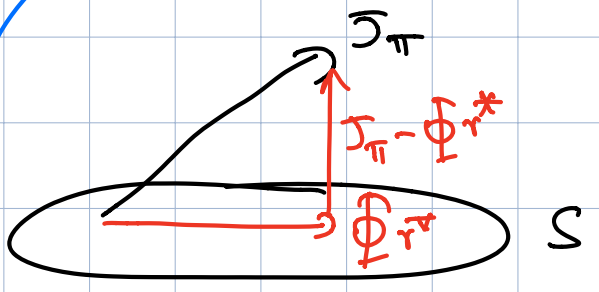
D : diagonal matrix with stationary distribution values

From (*),

$$\Phi^T D \left(J_\pi - \Phi r^* \right) = 0$$

$$\Leftrightarrow \quad r^{*T} \Phi^T D \left(J_\pi - \Phi r^* \right) = 0$$

$$\Leftrightarrow \quad \left(\Phi r^* \right)^T D \left(J_\pi - \Phi r^* \right) = 0$$



This equation $\Leftrightarrow$

$\Phi r^* \perp (J_\pi - \Phi r^*)$

$\perp$: orthogonal in $\|\cdot\|_\xi$ norm

$$\Leftrightarrow \quad \left\langle \left(\Phi r^* \right), J_\pi - \Phi r^* \right\rangle = 0$$

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i \xi_i y_i$$

Recall, for a bounded $J = (J(1), \ldots, J(n))$

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

In compact notation: $T_\pi J = g + \alpha P J$

$g = (g_1, \dots, g_n)$ $g_i = \sum_j p_{ij}\, g(i,j)$

$$P = \begin{bmatrix} p_{11} & & \\ & \ddots & \\ & & p_{nn} \end{bmatrix}$$

Projected fixed point equation:

$$\Phi r^* = \Pi T_\pi(\Phi r^*)$$

Here ΠT is the composition of Π with T.

IF: ΠT is a contraction wrt $\|\cdot\|_\xi$, then deriving VI or sto-iter-algo variations are straightforward.

"ΠT is contractive"

Letting $\tilde{J} = \Phi r^*$, we have

$$\tilde{J} = \Pi T_\pi(\tilde{J}) \quad \leftarrow \text{projected eqn}$$

Contrast with regular fixed point equation:

$$J_\pi = T_\pi J_\pi$$

Π: Come in because $\tilde{J} \in S$ "linear space".

Lemma 1: $\|PJ\|_\xi \le \|J\|_\xi \quad \forall J \in \mathbb{R}^n$

P: t.p.m. of M.C. underlying policy π

ξ: Stationary distribution vector

Pf:

$$\|PJ\|_\xi^2 = \sum_{i=1}^n \xi_i \left(\sum_{j=1}^n P_{ij} J(j) \right)^2$$

Jensen's inequality ⊛ $P_{ij}^2 \le P_{ij}$

$$\le \sum_{i=1}^n \xi_i \sum_{j=1}^n P_{ij} J(j)^2$$

$$= \sum_{j=1}^n \left(\sum_{i=1}^n \xi_i P_{ij} \right) J(j)^2$$

Using $\xi = \xi P$ → $\xi_j = \sum_{i=1}^n \xi_i P_{ij}$

$$= \sum_{j=1}^n \xi_j J(j)^2$$

$$= \|J\|_\xi^2$$

So, $\|PJ\|_\xi^2 \le \|J\|_\xi^2$ ■

"Projection is non-expansive"

Lemma 2: $\|\Pi J - \Pi J'\|_\xi \le \|J - J'\|_\xi, \quad \forall J, J' \in \mathbb{R}^n$

Pf: PTO

$$\|\Pi J - \Pi J'\|_\xi^2 = \|\Pi(J - J')\|_\xi^2$$

Pythagoras theorem: ΠJ & $(I-\Pi)J$ are orthogonal

$$\leq \|\Pi(J - J')\|_\xi^2 + \|(I - \Pi)(J - J')\|_\xi^2 \quad \text{(*)}$$

Note:

$$\underbrace{\Pi(J - J')}_{\in S} \perp \underbrace{\left((J - J') - \Pi(J - J')\right)}_{\text{orthogonal to } S}$$

$$\|\Pi(J - J')\|_\xi^2 + \|(I - \Pi)(J - J')\|_\xi^2$$

$$= \|\Pi(J - J') + (I - \Pi)(J - J')\|_\xi^2$$

$$= \|(J - J')\|_\xi^2$$

Hence, from (*), we obtain

$$\|\Pi J - \Pi J'\|_\xi^2 \leq \|J - J'\|_\xi^2$$

■

Main claim:

T_π and ΠT_π are contraction mappings w.r.t. $\|\cdot\|_\xi$, and have modulus α ($\rightarrow$ discount factor)

Pf:

Recall $T_\pi J = g + \alpha P J$

For any $J, J' \in \mathbb{R}^n$,

$$\| T_\pi J - T_\pi J' \|_\xi = \alpha \| P(J - J') \|_\xi$$

Using Lemma 1 ($\|PJ\|_\xi \le \|J\|_\xi$) $\le \alpha \| J - J' \|_\xi$ — (**)

So, T_π is contractive with modulus α.

(Side note! We showed earlier that T_π is a contraction w.r.t. max-norm. Here, we showed T_π to be contractive wrt $\|\cdot\|_\xi$ as well)

Next

$$\| \Pi T_\pi J - \Pi T_\pi J' \|_\xi$$
$$= \| \Pi (T_\pi J - T_\pi J') \|_\xi$$

Π is non-expansive (Lemma 2) $\le \| T_\pi J - T_\pi J' \|_\xi$

From (**) $\le \alpha \| J - J' \|_\xi$

So, ΠT is contractive wrt $\|\cdot\|_\xi$ with modulus α.

■

Implication:

$$\Phi r^* = \Pi T_\pi (\Phi r^*)$$

→ Cannot solve $T_\pi(T_\pi J_\theta) = J_\theta$ in func. approx case. So, solve this eqn instead.

Projected equation has a "unique" solution

& we can do value iteration to get the solution

(i.e., $\Phi r_0 \xrightarrow{\Pi T} \Phi r_1 \to \cdots \to$ asymptotically converges to Φr^*)

Lecture-32*

Not-so- main claim: Recall r^* is the fixed point of ΠT_π, i.e., $\Phi r^* = \Pi T_\pi (\Phi r^*)$

$$\| J_\pi - \Phi r^* \|_\xi^2$$

because $(J_\pi - \Pi J_\pi) \perp (\Pi J_\pi - \Phi r^*)$ (orthogonal to; $\Pi J_\pi - \Phi r^* \in S$)

$$= \| J_\pi - \Pi J_\pi \|_\xi^2 + \| \Pi J_\pi - \Phi r^* \|_\xi^2$$

$$= \| J_\pi - \Pi J_\pi \|_\xi^2 + \| \Pi T_\pi J_\pi - \Pi T_\pi \Phi r^* \|_\xi^2$$

$$\leq \| J_\pi - \Pi J_\pi \|_\xi^2 + \alpha^2 \| J_\pi - \Phi r^* \|_\xi^2$$

Rearranging:

$$\| J_\pi - \Phi r^* \|_\xi^2 \leq \frac{1}{1-\alpha^2} \| J_\pi - \Pi J_\pi \|_\xi^2$$

■

Matrix form of $\Phi r^* = \Pi T(\Phi r^*)$

$$\Pi = \Phi (\Phi^T D \Phi)^{-1} \Phi^T D$$

$$T_\pi J = g + \alpha P J$$

$\Pi J \rightarrow$ linear

$T_\pi J \rightarrow$ linear

$$\Phi r^* = \Pi T_\pi (\Phi r^*) \quad \text{---} \quad (*)$$

↓

"linear" system of equations

Want to write $(*)$ as "$Cr^* = d$"

ΠJ: minimize distance between J & $\tilde{J} \in S$ in $\|\cdot\|_\xi$

$$r^* = \arg\min_{r \in \mathbb{R}^d} \| \Phi r - T_\pi (\Phi r^*) \|_\xi^2$$

↑ arg $(*)$

$$= \arg\min_{r \in \mathbb{R}^d} \| \Phi r - (g + \alpha P \Phi r^*) \|_\xi^2$$

Weighted →

$$= \arg\min_{r \in \mathbb{R}^d} (\Phi r - (g + \alpha P \Phi r^*))^T D (\Phi r - (g + \alpha P \Phi r^*))$$

$$D = \begin{bmatrix} \xi_1 & & 0 \\ & \ddots & \\ 0 & & \xi_n \end{bmatrix}$$

Differentiating the expression being minimized, we obtain

$$\underbrace{\Phi^T D \left(\Phi r^* - (g + \alpha P \Phi r^*) \right) = 0}_{\text{Matrix form of } \Phi r^* = \Pi T(\Phi r^*)} \quad \text{— (*)}$$

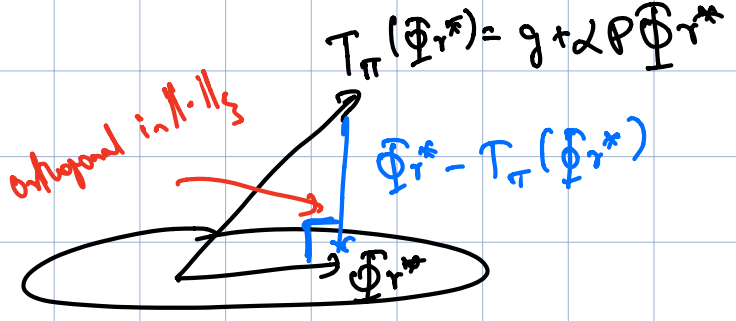
Intuitively: why (*), we have

$$r^{*T} \Phi^T D \left(\Phi r^* - (g + \alpha P \Phi r^*) \right) = 0$$

$$(\Phi r^*)^T D \left(\Phi r^* - (g + \alpha P \Phi r^*) \right) = 0$$

$$\left\langle \Phi r^*, \Phi r^* - (g + \alpha P \Phi r^*) \right\rangle = 0$$

where $\langle \cdot, \cdot \rangle$ leads to $\| \cdot \|_\xi$



$$\Phi^T D \left(\Phi r^* - (g + \alpha P \Phi r^*) \right) = 0$$

$$\Phi^T D g = \Phi^T D \Phi r^* - \alpha \Phi^T D P \Phi r^* = \Phi^T D (I - \alpha P) \Phi r^*$$

$$\Phi^T D g = \Phi^T D (I - \alpha P) \Phi r^* \quad \text{or, equivalently}$$

$$C r^* = d, \text{ where } C = \Phi^T D (I - \alpha P) \Phi, \; d = \Phi^T D g$$

↓ this is the same as $\Phi r^* = \Pi T_\pi (\Phi r^*)$

Explicit solution: $r^* = C^{-1} d$

(C invertible since Φ full col. rank & D has pve diagonals & $(I - \alpha P)$ is invertible)

Projected value iteration:

We know ΠT_π is a α-contraction

Start with r_0 & repeatedly apply ΠT_π

$$\Phi r_0 \xrightarrow{\Pi T_\pi} \Phi r_1 \to \cdots\cdots \longrightarrow \Phi r^*$$

Converge to

$$\Phi r_{k+1} = \Pi T_\pi (\Phi r_k), \quad k = 0, 1, \ldots \quad -(**)$$

↑ Value iteration + projection

PVI update $(**)$ in terms of C, d:

$$r_{k+1} = \arg\min_{r \in \mathbb{R}^d} \| \Phi r - (g + \alpha P \Phi r_k) \|_\xi^2$$

(Φr: variable; $g + \alpha P \Phi r_k$: fixed)

Differentiating,

$$\Phi^T D (\Phi r_{k+1} - (g + \alpha P \Phi r_k)) = 0 \quad (**)$$

$$r_{k+1} = r_k - (\Phi^T D \Phi)^{-1} (C r_k - d) \quad (*)$$

Check this by substituting expressions C, d in $(*)$.

It is the same as $\Phi r_{k+1} = \Pi T_\pi (\Phi r_k)$

Remark: For PVI, need knowledge of t.p.m. P & stationary distribution values (through D) to form C & d, which are used in update iteration $(**)$

Solving $Cr^* = d$ using a sample path:

Recall $C = \Phi^T D (I - \alpha P) \Phi$, $d = \Phi D g$

have to estimate: $\Phi^T D \Phi$, $\Phi^T D P \Phi$, $\Phi^T D g$

to form estimates of C, d.

$$\Phi = \begin{bmatrix} \phi(1)^T \\ \vdots \\ \phi(n)^T \end{bmatrix} \qquad \Phi^T = \begin{bmatrix} \phi(1) & \cdots & \phi(n) \end{bmatrix}$$

$$\Phi^T D \Phi = \begin{bmatrix} \phi(1) & \cdots & \phi(n) \end{bmatrix} \begin{bmatrix} \xi_1 & & 0 \\ & \ddots & \\ 0 & & \xi_n \end{bmatrix} \begin{bmatrix} \phi(1)^T \\ \vdots \\ \phi(n)^T \end{bmatrix}$$

$$\Phi^T D \Phi = \sum_{i=1}^{n} \xi_i \, \phi(i) \, \phi(i)^T$$

Recall $P = \begin{bmatrix} p_{11} & & \\ & \ddots & \\ & & p_{nn} \end{bmatrix}$ $\quad g = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}$ $\quad g_i = \sum_{j=1}^{n} p_{ij} g(i,j)$

$$\Phi^T D P \Phi = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \, p_{ij} \, \phi(i) \, \phi(j)^T$$

$$\Phi^T D g = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \, p_{ij} \, \phi(i) \, g(i,j)$$

→ unknown in an RL setting

Using π, & some i_0, generate a sample path $(i_0, i_1, \ldots, i_k)$

Observe $g(i_t, i_{t+1})$, $\forall t$

Form sample-based estimate of $\Phi^T D \Phi$, $\Phi^T D P \Phi$ & $\Phi^T D g$.

approximate

$$\Phi^T D \Phi \approx \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\,\phi(i_t)^T$$

$$\Phi^T D P \Phi \approx \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\,\phi(i_{t+1})^T$$

$$\Phi^T D g \approx \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, g(i_t, i_{t+1})$$

Let $$C_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\,(\phi(i_t) - \alpha\,\phi(i_{t+1}))^T$$

$$d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, g(i_t, i_{t+1})$$

$$C_k \approx C \;\left(C = \Phi^T D (I - \alpha P) \Phi\right) \qquad d_k \approx d \;(= \Phi^T D g)$$

LSTD: Solve $\boxed{C_k r_k = d_k}$

Least-squared temporal difference.

$$C r_k - d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\Big(\underbrace{\phi(i_t)^T r_k - \alpha\,\phi(i_{t+1})^T r_k - g(i_t, i_{t+1})}_{\text{Temporal difference.}}\Big)$$

$\phi(i_t)^T r_k \approx J_\pi(i_t)$, $\phi(i_{t+1})^T r_k \approx J_\pi(i_{t+1})$

$$\text{TD-term} = \tilde{J}(i_t) - \left(g(i_t, i_{t+1}) + \alpha \tilde{J}(i_{t+1})\right)$$

Lecture-33

Recall projected system of equations:

$$\left(r^* = d \iff \Phi r^* = \Pi T_\pi(\Phi r^*)\right.$$

$$C = \Phi^T D(I - \alpha P)\Phi, \quad d = \Phi^T D g$$

$$C_1 := \Phi^T D \Phi = \sum_{i=1}^{n} \xi_i \, \phi(i)\phi(i)^T$$

$$C_2 := \Phi^T D P \Phi = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i p_{ij} \, \phi(i)\phi(j)^T$$

$$d := \Phi^T D g = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i p_{ij} \phi(i) g(i,j)$$

} → Sample from $\xi = \{\xi_i, i = 1 \dots n\}$

} $\xi P = \{\xi_i p_{ij}, \; i = 1 \dots n, \; j = 1 \dots n\}$

} ——u——

Suppose we obtain a sample path $\{i_0, i_1, \dots, i_k\}$ simulated using policy π & states picked according to the distribution ξP ⟹ pick an i_0 from ξ (starting distribution) & then pick a next state using p_{ij} & repeat.

Empirical frequencies: $\hat{\xi}_i = \dfrac{\sum_{t=0}^{k} I(i_t = i)}{k+1}$

$$\hat{P}_{ij} = \frac{\sum_{t=0}^{k} I(i_t = i, i_{t+1} = j)}{k+1}$$

Using $\hat{\xi}_i$, $\hat{P}_{ij}$, we estimate C_1, C_2, d as follows:

$$\hat{C}_1 = \sum_{i=1}^{n} \hat{\xi}_i \, \phi(i) \, \phi(i)^T$$

$$\hat{C}_2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\xi}_i \, \hat{P}_{ij} \, \phi(i) \, \phi(j)^T$$

$$\hat{d} = \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\xi}_i \, \hat{P}_{ij} \, \phi(i) \, g(i,j)$$

How to approximate $Cr^* = d$?

① Estimate C by C_k as follows:

$$C_k = \hat{C}_1 - \alpha \hat{C}_2$$

$$d_k = \hat{d}$$

② Solve $C_k \hat{r}_k = d_k$

$\downarrow$

$\hat{r}_k \rightarrow$ LSTD solution.

Some analysis:-

As the trajectory length k in $(i_0, \ldots, i_k)$ goes to infinity, do the estimates $\hat{\xi}_i$, $\hat{p}_{ij}$ converge?

As $k \to \infty$, $\hat{\xi}_i \longrightarrow \xi_i$ w.p. 1 (Version of SLLN)

$\hat{p}_{ij} \longrightarrow p_{ij}$ w.p. 1

$$C_k \xrightarrow[k\to\infty]{} C, \quad d_k \to d$$

Hence, $\hat{r}_k \to r^*$ w.p. 1 as $k \to \infty$

Remark: C_k and d_k can be written alternatively as

$$\hat{C}_1 = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, \phi(i_t)^T$$

$$\hat{C}_2 = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, \phi(i_{t+1})^T$$

$$\hat{d} = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, g(i_t, i_{t+1})$$

Another remark:

LSTD solution: $C_k \hat{r}_k = d_k$

Can we write $r_k = C_k^{-1} d_k$? NO.

Trivial example: $(i_0, i_1, \dots, i_k)$

Suppose $i_0 = i_1 = \dots = i_k = i$ (some state)

$$\hat{C}_1 = \frac{1}{k+1} \sum \phi(i_t)\phi(i_t)^T$$

$\hat{C}_1 \to$ with d columns, but each row is identical

So $\operatorname{rank}(\hat{C}_1) < d$

As an alternative, solve $\left(C_k + \beta I\right) \hat{r}_k = d_k$

$$\Rightarrow \quad \hat{r}_k = \underbrace{\left(C_k + \beta I\right)^{-1}}_{\text{this is invertible if } \beta \text{ is large enough}} d_k$$

Where is "temporal difference" in "least squares temporal difference (LSTD)"?

LSTD: $C_k r_k - d_k = 0$

$$C_k r_k - d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\left[\phi(i_t)^T r_k - \alpha\, \phi(i_{t+1})^T r_k - g(i_t, i_{t+1})\right]$$

Since $C_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\left(\phi(i_t) - \alpha\, \phi(i_{t+1})\right)^T$,

$$d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, g(i_t, i_{t+1})$$

the term $\underbrace{\left[\phi(i_t)^T r_k - \left[g(i_t, i_{t+1}) + \alpha\, \phi(i_{t+1})^T r_k\right]\right]}_{(*)}$

is a temporal difference term because

$$\phi(i_t)^T r_k \approx J_\pi(i_t)$$

$$\phi(i_{t+1})^T r_k \approx J_\pi(i_{t+1})$$

So $(*) \approx J_\pi(i_t) - \left(g(i_t, i_{t+1}) + \alpha\, J_\pi(i_{t+1})\right)$

So $(*) \Leftrightarrow$ TD-term.

Lecture-34*

TD(0) with linear function approximation

Recall the story of TD(0) without function approximation!

Want to solve $J_\pi = T_\pi J_\pi$

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

sto-iter-algo for solving this eqn

Sampled from RHS.

$$J_{t+1}(i) = J_t(i) + \beta_t \underbrace{\left(g(i, \bar{i}) + \alpha\, J_t(\bar{i}) - J_t(i)\right)}_{\text{TD-term}}$$

$g(i, \bar{i}) + \alpha J_t(i)$ is a proxy for $E_{\bar{i}}\left(g(i, \bar{i}) + \alpha J_\pi(\bar{i})\right)$

Onto linear function approximation case:

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

Cannot do this

$$r_{t+1} \not\leftarrow r_t + \beta_t \left(g(i,\tilde{i}) + \alpha\, r_t^T \phi(\tilde{i}) - r_t^T \phi(i) \right)$$

$r_t \in \mathbb{R}^d$ $\quad \beta_t \in \mathbb{R}$ $\quad g(i,\tilde{i}) + \alpha\, r_t^T \phi(\tilde{i}) - r_t^T \phi(i) \in \mathbb{R}$

TD(0) with linear function approximation

On a transition (i_t, i_{t+1}) in a sample path $(i_0, \ldots)$

$$r_{t+1} = r_t + \beta_t \phi(i_t)\left(g(i_t, i_{t+1}) + \alpha r_t^T \phi(i_{t+1}) - r_t^T \phi(i_t) \right)$$

$r_{t+1}, r_t \in \mathbb{R}^d$ $\quad \phi(i_t) \in \mathbb{R}^d$ $\quad g(i_t, i_{t+1}) + \alpha r_t^T \phi(i_{t+1}) - r_t^T \phi(i_t) \in \mathbb{R}$

Go back to projected fixed point:

$$\Phi r^* = \Pi T_\pi(\Phi r^*)$$

$$\Leftrightarrow \quad C r^* = d, \quad C = \Phi^T D (I - \alpha P) \Phi$$

$$d = \Phi^T D g$$

Want to use a sample path $(i_0, i, \ldots)$ to find r^* s.t. $C r^* = d$.

Sto-iter-algo: $r_{t+1} = r_t - \beta_t \left((r_t - d) \right)$ — (*)

Where would r_t converge? Ans: wherever $Cr^* = d$

$$r_t \to r^*$$

$$Cr^* - d = \Phi^T D (I - \alpha P) \Phi r^* - \Phi D g$$

$$= \sum_{i,j} \xi_i P_{ij} \phi(i) \left(\phi(i)^T r^* - \alpha \phi(j)^T r^* - g(i,j) \right)$$

$$= E \left(\phi(i) \left(\phi(i)^T r^* - \alpha \phi(j)^T r^* - g(i,j) \right) \right)$$

↑ taken with ξP distribution

Q: Want to find an r^* s.t.

$$E \left(\phi(i) \left(\phi(i)^T r^* - \alpha \phi(j)^T r^* - g(i,j) \right) \right) = 0$$

↑ i picked from $\xi = (\xi_1 \cdots \xi_n)$

j — " — $((P_{ij}))$

Let (i_t, i_{t+1}) be a sample transition

Then, do the following update

(*)-) $r_{t+1} = r_t - \beta_t \phi(i_t) \left(\phi(i_t)^T r_t - \alpha \phi(i_{t+1})^T r_t - g(i_t, i_{t+1}) \right)$

↑ TD(0) with linear function approximation.

Remark:

① In the above, we assumed sampling from the stationary distribution. Under this, it is straightforward to invoke "Sto-iter-algo general convergence result" under contraction case.

② TD(0) with linear function approximation would converge even if sampling is not from the stationary distribution.

$(i_0, i_1, \ldots, \ldots)$

i_0: initial state picked using some distribution "μ"

Let the Markov chain have t.p.m. P.

Then, after k steps → the distribution μP^k

$\mu P^k \longrightarrow \xi$ as $k \to \infty$ (assuming irreducible + positive recurrence)

$(i_0, i_1, \ldots, i_N, i_{N+1}, \ldots)$

transient phase

i_N: after a large # N of iterations, the Markov chain is in steady state i.e., the state distribution is ξ.

4. It can be shown that (*) (TD with LFA) would converge to the same fixed point, i.e., γ^* satisfying $C\gamma^* = d$ ($\Leftrightarrow \Phi\gamma^* = \Pi T_{\pi,\alpha}(\Phi\gamma^*)$)

Ref: J. N. Tsitsiklis & B. V. Roy "Analysis of TD with LFA", IEEE trans. auto. control, 1997.

(3) Can extend TD(0) to TD(λ) with LFA.

"Read it from NDP book or DPOC Vol II Chapter 6"