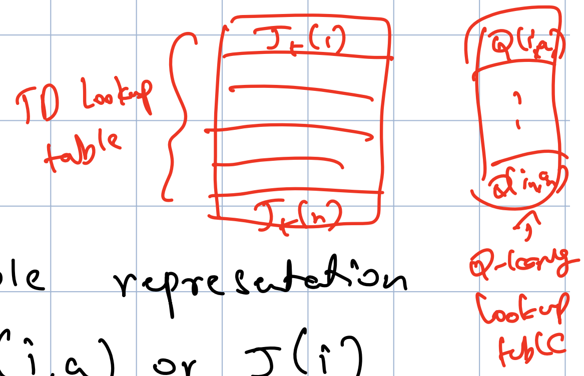
Lecture-28*

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.

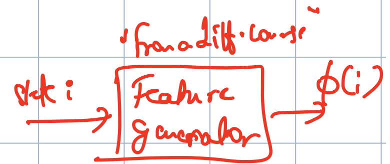
e.g. Go: 10^{170} states, Chess $10^{??}$ "large number" states
other practical applications have large state spaces.

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

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

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

parameter.



E.g., Linear function approximation

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

feature vector

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 $\rightarrow$ out of scope of this course.

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

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

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

$$\pi(i) = \arg\min_a \sum_j P_{ij}(a)\left(g(i,a,j) + \alpha \tilde{J}(j,r)\right) \quad (*)$$

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

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

$\alpha \rightarrow$ discount factor. Assume finite state & finite action space.

$\|\cdot\|_\infty \rightarrow$ 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 $\tilde{J}$, i.e., using (*) with "$\tilde{J}$" instead of J

$$\boxed{\|J_\pi - J^*\|_\infty \le \frac{\alpha \epsilon}{1-\alpha}}.$$

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

π is an optimal policy.

Pf:

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

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

because π is greedy wrt $\tilde{J}$, we have $T_\pi \tilde{J} = T\tilde{J}$

$$\overset{T_\pi \text{ is a } \alpha\text{-contraction in } \|\cdot\|_\infty}{\le} \alpha \|J_\pi - \tilde{J}\|_\infty + \|T\tilde{J} - J^*\|_\infty$$

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

$$\le \alpha \|J_\pi - J^*\|_\infty + \alpha\|J^* - \tilde{J}\|_\infty + \alpha\|\tilde{J} - J^*\|_\infty$$

$$= \alpha \|J_\pi - J^*\|_\infty + 2\alpha\epsilon$$

$$\|J_\pi - J^*\|_\infty \le \alpha \|J_\pi - J^*\|_\infty + 2\alpha\epsilon$$

So, $$\boxed{\|J_\pi - J^*\|_\infty \le \frac{2\alpha\epsilon}{1-\alpha}}$$

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

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

Choose ϵ s.t. $\frac{2\alpha\epsilon}{1-\alpha} < \delta$ & 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 π & we shall not attach "π" to the symbols used & keep the policy implicit.

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

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

"implicit in notation"

$$i_k \xrightarrow{\pi(i_k)} i_{k+1}$$

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

Aim: Estimate $J_\pi(i) = E\left(\sum_{k=0}^{\infty} \alpha^k g(\underbrace{i_k, i_{k+1}}_{\text{no } \pi \text{ here. it is implicit}}) \,\middle|\, i_0 = i, \pi \right)$

in the action closer.

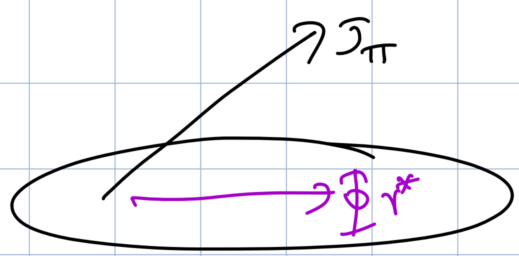
Linear function approximation:

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

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

State $i \rightarrow \boxed{\text{Value approximator}} \rightarrow \phi(i)^T r$

$\phi(i)$: on the fly; r: fixed / for d



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

linear space

$\Phi \rightarrow$ big Phi (matrix)

$\phi(i) \rightarrow$ small phi "feature vector"

Let

$$\Phi = \begin{bmatrix} - & \phi(1)^T & - \\ - & \phi(2)^T & - \\ & \vdots & \\ - & \phi(n)^T & - \end{bmatrix}$$

Feature matrix

"Tall matrix"

$n \times d$, $n >> d$

For analysis

"computing algo will not work with Φ" explicitly

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

$n \times d$ matrix $\rightarrow$ a tall one.

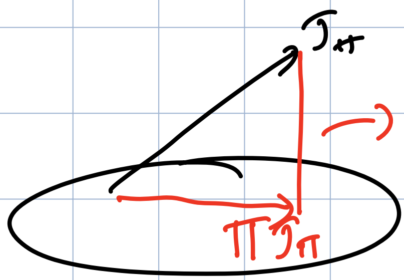
$$\tilde{J}_r = [\tilde{J}(1,r) \ \dots \ \tilde{J}(n,r)]^T \rightarrow \text{approx. value function}$$

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

" Want to bound $\| \Phi r - J^\pi \|_{\rightarrow \text{some norm}}$ "

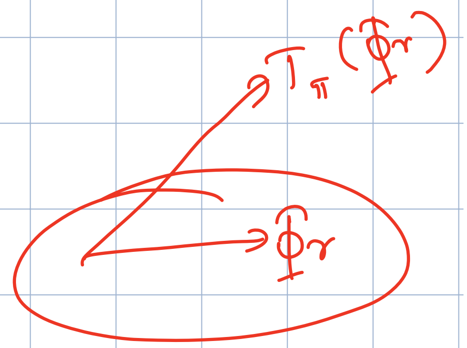
Aim: Find the best approximation to J_π in the space $S = \{ \Phi r \mid r \in \mathbb{R}^d \}$

Lecture-30 (continued after Markov chains review)



Π: Projection operation

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



Q: In $J_\pi \approx \Phi r$, what "r" to choose?

For regular policy evaluation, one solves

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

$$J_\pi \approx \Phi r^*$$

$$\boxed{\Phi r^* = T_\pi (\Phi r^*)}$$

doesn't make sense

since $\Pi(\Phi r^*)$ need not be the same as Φr^* (unless $J_\pi = \Phi r^*$)

see

This fixed point relation does not necessarily hold

Projected fixed point equation

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

projection onto linear space S

projected fixed point equation

$\tilde{J} = \underbrace{\Pi T_\pi}_{\text{contraction}} \tilde{J}$

$\tilde{J}$ is unique

$\Downarrow$

r is also unique? where $\tilde{J} = \Phi r$

$\Pi \rightarrow$ 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.

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 $\text{rank}(\Phi) = d$ [Note: we assume $n \gg d$]

Lecture-31*

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

For a n-vector J, define
# of states
Stationary distribution (assumed to exist)

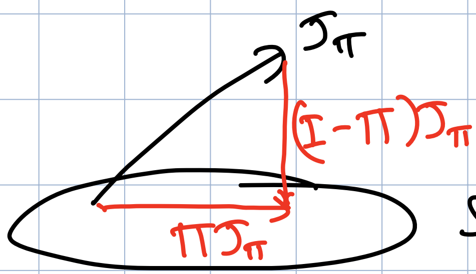
$$\|J\|_\xi^2 = \sum_{i=1}^{n} \xi_i (J(i))^2$$

weighted l_2 norm

$$D = \begin{bmatrix} \xi_1 & & & 0 \\ & \xi_2 & & \\ & & \ddots & \\ 0 & & & \xi_n \end{bmatrix}$$

$\xi_i > 0, \forall i$

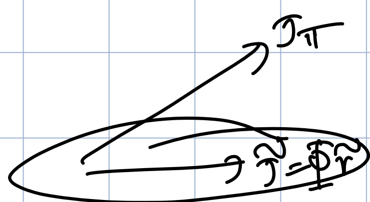
$$\|J\|_\xi^2 = J^T D J, \quad \forall J \in \mathbb{R}^n$$



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

Π: Projection operator

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



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

Any $J \in S$ is of the form Φr

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

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

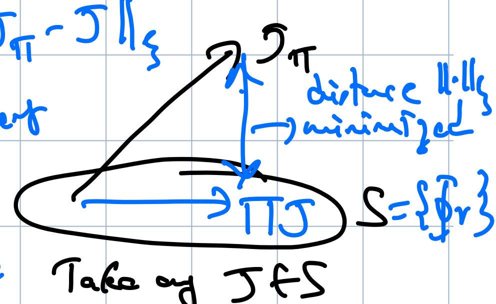
and $\Pi J_\pi = \Phi \tilde{r}$

projection operator

$\min_J \|J_\pi - J\|_\xi$

but, every $J = \Phi r$

unique



To find $\tilde{r}$:

$$\nabla \| J_\pi - \Phi \tilde{r} \|_\xi^2 = 0$$

$$\Leftrightarrow \nabla (J_\pi - \Phi \tilde{r})^T D (J_\pi - \Phi \tilde{r}) = 0$$

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

$$\Leftrightarrow \Phi^T D J_\pi - \Phi^T D \Phi \tilde{r} = 0$$

$$\Leftrightarrow \Phi^T D \Phi \tilde{r} = \Phi^T D J_\pi$$

$$\Leftrightarrow \tilde{r} = \underbrace{(\Phi^T D \Phi)^{-1}}_{\text{why is this invertible?}} \Phi^T D J_\pi$$

Why is this invertible?
$\Phi \to$ full rank and
diagonal elements of $D > 0$.

Cannot solve $\Phi r^* = \underbrace{T_\pi(\Phi r^*)}_{\text{Can be outside } S}$

So, project $T_\pi(\Phi r^*)$
i.e., $\Pi(T_\pi(\Phi r^*))$
& then solve
$\Phi r^* = \Pi(T_\pi(\Phi r^*))$

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

Now, $\Pi J_\pi = \Phi \tilde{r}$

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

projection operator

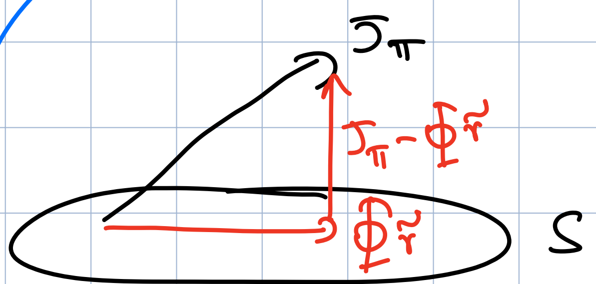
diagonal matrix with stationary distribution values

From (*),

$$\Phi^T D (J_\pi - \Phi \tilde{r}) = 0$$

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

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



This equation $\Leftrightarrow$

$$\Phi\tilde{r} \perp (J_\pi - \Phi\tilde{r})$$

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

$$\Leftrightarrow \quad \langle (\Phi\tilde{r}), J_\pi - \Phi\tilde{r} \rangle_\xi = 0$$

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

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

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

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

$g = (g_1, \ldots, g_n) \quad g_i = \sum_j P_{ij}\, g(i,j)$

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

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

t.p.m. of M.C. underlying policy π

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 ⊕ $p_{ij}^2 \le p_{ij}$

$$\le \sum_{i=1}^n \xi_i \sum_{j=1}^n p_{ij} J(j)^2 \quad \text{--- } (**)$$

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

this step requires stationarity

Using $\xi = \xi P$ $\Rightarrow$ $\boxed{\xi_j = \sum_{i=1}^n \xi_i p_{ij}}$

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

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

So, $$\|PJ\|_\xi^2 \le \|J\|_\xi^2$$

∎

"Projection is non-expansive"

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

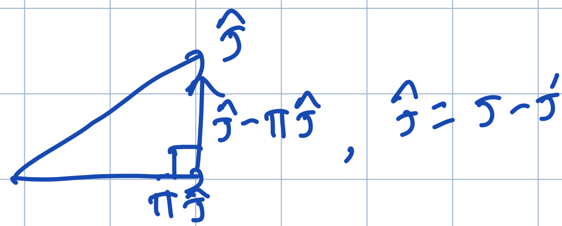
Pf: PTO

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

Pythagoras theorem

ΠJ & $(I-\Pi)J$ are orthogonal

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



Note:

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

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

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

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

Hence, from $(*)$, we obtain

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

$\blacksquare$

Main claim:

T_π and ΠT_π are contraction mappings w.r.t. $\|\cdot\|_\xi$, and have modulus α ($\to$ 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 $\|PJ\|_\xi \leq \|J\|_\xi$

$$\leq \alpha \| J - J' \|_\xi \qquad (**)$$

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 nonexpansive (Lemma 2)

$$\leq \| T_\pi J - T_\pi J' \|_\xi$$

From (**)

$$\leq \alpha \| J - J' \|_\xi$$

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

Implication:

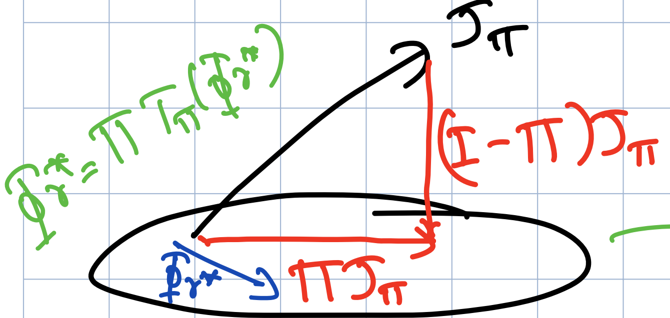
r^* is unique because (i) ΠT_π is contractive, (ii) Φ is full col. rank

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

Cannot solve $J_\pi = T_\pi J_\pi$ 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., $\Phi r_0 \xrightarrow{\Pi T_\pi} \Phi r_1 \to - - - - \to$ asymptotically converge to Φr^*)

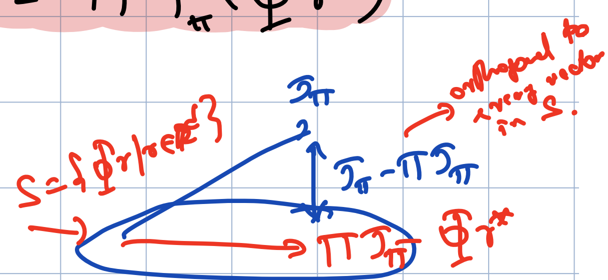


$\to S = \{\Phi r \mid r \in \mathbb{R}^d\}$

$\Pi J_\pi = \Phi \tilde{r}$ (Don't confuse $\tilde{r}$ with r^*)

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

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

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

ΠT_π is a α-contraction

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

Re-arranging:

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

Lecture-32

Matrix form of $\Phi r^* = \Pi T_\pi(\Phi r^*)$ i.e., $C r^* = d$

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

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

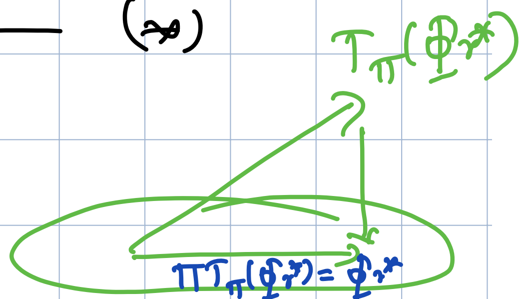
$\Pi J \rightarrow$ linear

$T_\pi J \rightarrow$ linear

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

$\downarrow$

"linear" system of equations



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

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

$\tilde{r} = \arg\min_r \|\Phi r - J\|_\xi^2$

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

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

using $(*)$

$T_\pi = g + \alpha P$

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

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

min is over "r"   non-variable

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

Differentiating the expression being minimized, we obtain

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

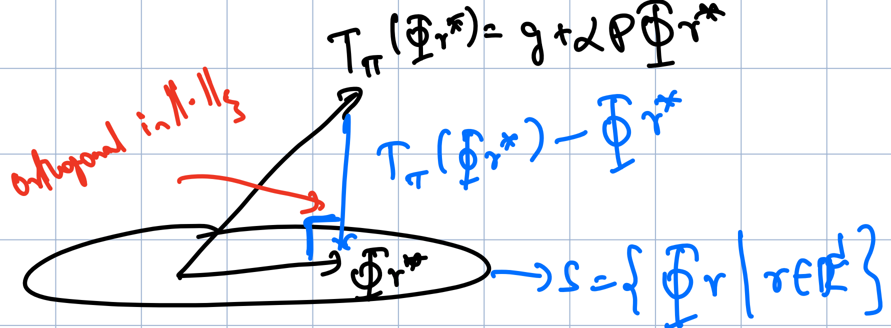
Intuitively: using $(**)$, we have

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

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

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

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



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

$$\tilde{J} = \Pi(T_\pi(\tilde{J}))$$

$$\tilde{J} = \Phi r^*$$

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

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

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

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

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

Explicit solution! $\quad r^* = C^{-1} d$

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

$$J_{t+1}(i) = J_t(i) + \beta \left(\overset{\text{Incremental}}{\text{TD-error}} \right)$$

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

Projected value iteration:

We know ΠT_π is a α-contraction

Start with r_0 & repeatedly apply ΠT_π

$$\Phi r_0 \xrightarrow{\Pi T_\pi} \Phi r_1 \rightarrow \cdots\cdots \xrightarrow[\text{Converge to}]{} \Phi r^*$$

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

Value iteration + projection

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

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

variable

Differentiating,

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

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

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

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 $(**)$

Lecture-34*

$x_0 \xrightarrow{\pi(x_0)} x_1 \xrightarrow{\pi(x_1)} x_2 \longrightarrow - - - -$ observe single step costs

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.

$\hat{\Phi} = \begin{bmatrix} \phi(i_1)^T \\ \vdots \\ \phi(\cdot)^T \end{bmatrix}$ $\Phi = \begin{bmatrix} \phi(1)^T \\ \vdots \\ \phi(\cdot)^T \end{bmatrix}$

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

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

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

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

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

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

approximate

$(i_0, i_1 - - i_k)$

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

$\frac{1}{k+1} \phi(i)\phi(i)$ $+ \frac{1}{k+1} \phi(i)\phi(i)^T$

$(i_0, i_1 - - - - - - -)$

$\frac{\# N(i)}{k+1} \xrightarrow{k \to \infty} \xi_i$

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

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

Let

$$\boxed{\begin{aligned} C_k &= \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\left(\phi(i_t) - \alpha\,\phi(i_{t+1})\right)^T \\ d_k &= \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, g(i_t, i_{t+1}) \end{aligned}}$$

Simple-based approx of C & d

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

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

Least-squares temporal difference.

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

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

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

Recall projected system of equations:

$$C r^* = d \quad \Longleftrightarrow \quad \Phi r^* = \Pi T_\pi(\Phi r^*)$$

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

uses stationary dist^n (beca $\|\cdot\|_\xi$)

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

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

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

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

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

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

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

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

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

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

replacing $\hat{\xi}_i$ & $\hat{p}_{ij}$ by ξ_i & p_{ij} would get us to

$C_1 = \Phi^T D \Phi$

$C_2 = \Phi^T D P \Phi$ &

$d = \Phi^T D g$.

How to approximate $C r^* = d$?

$$\hat{A}\theta = \hat{b}$$

$$\hat{A} = \frac{1}{k+1} \sum_{m=0}^{k} \phi(i_m) \phi(i_m)^T$$

$$\hat{b} = \frac{1}{k+1} \sum_{m=0}^{k} \phi(i_m) y_m$$

"Least squares regression"

(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: $\rightarrow$ batch algorithm uses a dataset (not incremental)

Some analysis:-

initial dist^n μ_0 — transitions using t.p.m. P — stationary dist^n from M ...

$i_0 \rightarrow i_1 \rightarrow i_2 \rightarrow \cdots$

Generative model (for analysis)

$\mu_0 P^n \xrightarrow[\text{dist.}]{} \xi$

Stationary distribution $\{\xi_1, \ldots, \xi_n\}$ → state i sampled from stationary distribution

i → t.p.m P → i' that is sampled from $P_{ij}(\pi(i)) = P_{ij}$

sample transition from the MDP

from stationary distbn

from t.p.m

Batch dataset

$$\{(i_0, i_0', g(i_0, i_0')), (i_1, i_1', g(i_1, i_1')), \ldots, (i_k, i_k', g(i_k, i_k'))\}$$

$i_0, i_1, \ldots$ sampled from $\{\xi_1, \ldots, \xi_n\}$

$i_0', i_1', \ldots$ sampled using t.p.m. P.

form $\hat{C}_1$, $\hat{C}_2$, $\hat{d}$

$$C_k = \hat{C}_1 - \alpha \hat{C}_2, \quad d_k = \hat{d}$$

Solve $C_k \hat{r}_k = d_k$

LSTD solution

$$\hat{\xi}_i = \frac{\#\text{ visits to } i \text{ in } (i_0, \ldots, i_k)}{k+1} \xrightarrow[k \to \infty]{} \xi_i$$

Suppose mixing time is $\tilde{\tau}$

$$\hat{\xi}_i = \frac{\#\text{visits in } (i_0 \ldots i_{\tilde{\tau}})}{k+1} + \frac{\#\text{visits in } (i_{\tilde{\tau}+1}, \ldots)}{k+1}$$

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 $k \to \infty$,

$$\hat{\xi}_i \longrightarrow \xi_i \quad \text{w.p. } 1$$

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

(Version of SLLN)

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

→ LLN-type result for "LSTD"

Hence, $\hat{r}_k \longrightarrow r^*$ 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! $(i_0, i_1, \ldots, i_k)$

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

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

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

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

So, C_k is not necessarily invertible

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

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

Lecture-35*

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

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.

"LSTD(λ) can be worked out using TD(λ) + LSTD idea"

Lecture-34*

TD(0) with linear function approximation

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

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

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

$\downarrow$

sto-iter-algo for solving this eqn

Sampled from RMS.

Tabular TD(0) or TD(0) with full state representations

$\downarrow$

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

$g(i,\bar{i}) + \alpha\, J_t(\bar{i})$ 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

$$r_{t+1} \cancel{=} r_t + \beta_t \big(\underbrace{g(i,\bar{i}) + \alpha\, r_t^T \phi(\bar{i}) - r_t^T \phi(i)}_{\in \mathbb{R}} \big)$$

$r_t \in \mathbb{R}^d$ $\quad \beta_t \in \mathbb{R}$

TD(0) with linear function approximation

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

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

On line algorithm

Go back to projected fixed point:

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

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

$$d = \Phi D g$$

Want to use a sample path $(i_0, i_1, \ldots)$ to find r^* s.t. $C r^* = d$.

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

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

$$r_t \to r^*$$

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

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

Picked from ξ picked from P_{ij}

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

taken with ξP distribution

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

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

i picked from $\xi = (\xi_1, \dots, \xi_n)$
j — " — $((P_{ij}))$

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

Then, do the following update

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

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.

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

initial state picked using some distribution "μ"

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

Then, after k steps $\rightarrow$ the distribution μP^k

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

$$(\underbrace{i_0, i_1, \ldots, \quad}_{\text{transient phase}} i_N, i_{N+1}, \ldots)$$

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

$\Leftarrow$. it can be shown that $(*)$ (TD with LFA)
would converge to the same fixed point,
i.e., r^* satisfying $Cr^* = d \iff \Phi r^* = \Pi T_\pi(\Phi r^*)$

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"