17-Nov-2018

Chapter 6 Q-Learning

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

greedy-ϵ policy wp $1-\epsilon$ max, ϵ rando

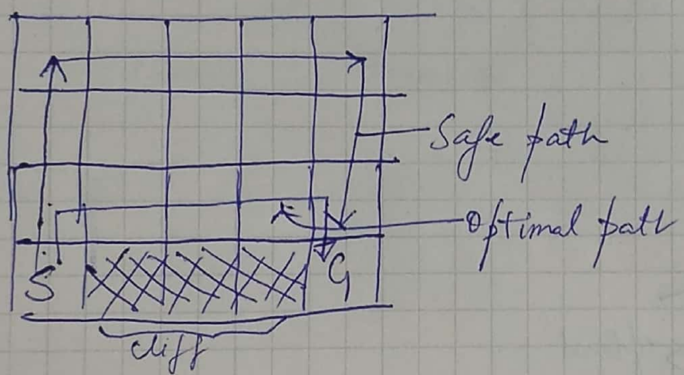
$Q(S_t, A_t)$ = Q-value associated with tuple (S_t, A_t)

will give optimal Q-value.

$$Q(S_t, A_t) \longrightarrow Q^*(S_t, A_t)$$

maximising action (optimal action) in state S_t $\underset{a}{\operatorname{argmax}}\, Q^*(S_t, a)$

Cliff Walking:



reward = -1 for any action that takes you to "non cliff" position.

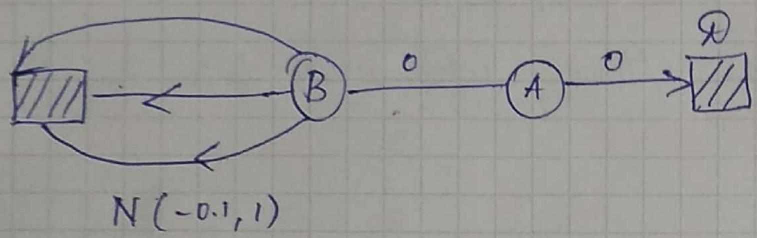
reward = -100 for any action that takes to the cliff

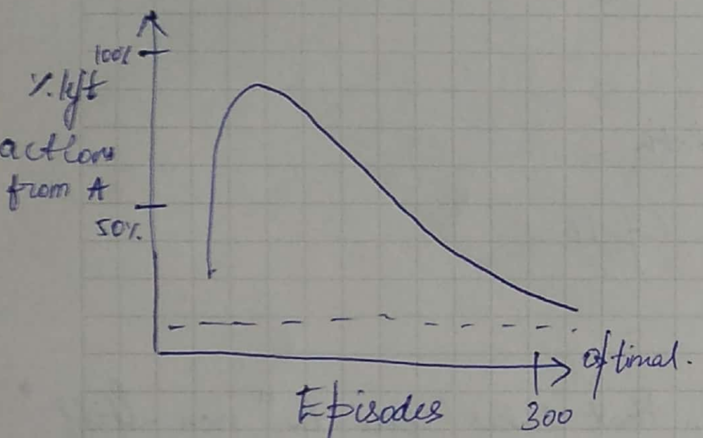
Expected SARSA:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma E_\pi [Q(S_{t+1}, A_{t+1}) | S_t] - Q(S_t, A_t) \right]$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

Q-learning suffers from maximization bias:





Double Q-learning (2010-11 algorithm)

- Have 2 estimates of Q-value, Q_1 & Q_2.

Double Q-learning for estimating $Q_1 \approx Q_2 \approx q^*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q_1(s,a)$ and $Q_2(s,a)$ $\forall s \in S^+, a \in A(s)$
st $Q(\text{terminal}) = 0$

Loop for each episode:

Intialize S

Loop for each episode:

choose action A in S using the policy ϵ-greedy in Q_1 + Q_2

Take action A, observe R, S'

with prob 0.5,

$$Q_1(S,a) \leftarrow Q_1(S,A) + \alpha(R + \gamma Q_2(S', \arg\max_a Q_1(S',a)) - Q_1(S,A))$$

else,

$$Q_2(S,a) \leftarrow Q_2(S,A) + \alpha(R + \gamma Q_1(S', \arg\max_a Q_2(S',a)) - Q_2(S,A))$$

Untill S is terminal.

Function Approx Based Methods:

Neural N/w based architecture

No of states — 10^{1000} states i.e, large

1. State i is encoded as $x = (x_1(i), \dots x_L(i))$
2. x is transformed ~~to~~ linearly through a layer as

$$\sum_{l=1}^{k} r(k,l)\, x_l(i), \quad k = 1 \dots k$$

3. Non-linear transformation via a sigmoid $\sigma(\cdot)$

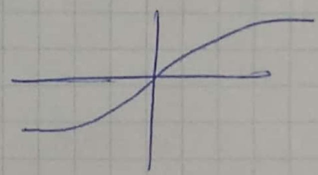
we obtain $\sigma\left(\sum_{l=1}^{k} r(k,l)\, x_l(i)\right)$

$\sigma(\cdot)$ are differentiable f^n satisfying

$$-\infty < \lim_{z \to -\infty} \sigma(z) < \lim_{z \to \infty} \sigma(z) < \infty$$

$\sigma(\cdot)$ are non-decreasing

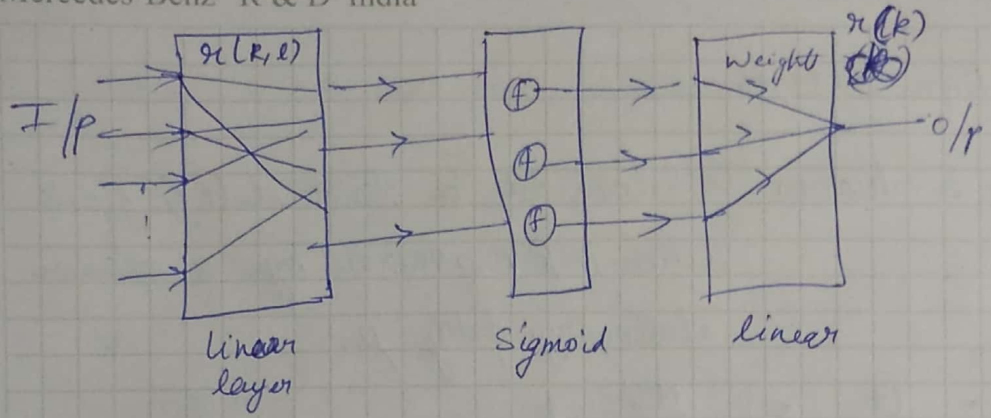
eg: $\sigma(z) = \tanh(z) = \dfrac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$



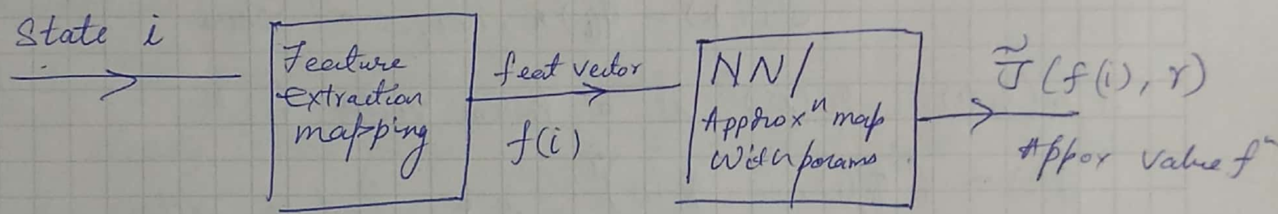
$\sigma(z) = \dfrac{1}{1+e^{-z}}$

4) Final o/p

$$\tilde{J}(i,r) = \sum_{k=1}^{k} r(k)\, \sigma\left(\sum_{l=1}^{L} r(k,l)\, x_L(i)\right)$$

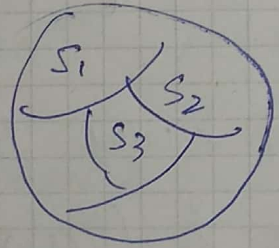


Features: - Capture the most important aspects of a state
- When used J is a complicated f^n [non linear]



Feature: A mapping $f_k : S \to \mathbb{R}$. Suppose $f_1, f_2 \ldots f_k$ are 'k' features

Feature vector $f(i) \equiv (f_1(i) \ldots f_k(i))$



$$r^* = \underset{r}{\operatorname{argmin}} \sum_i \Big(J(i) - \sum_k r(k)\,\phi(f(i))\Big)^2$$

Projected Eqn Methods:

Policy Evaluation: Consider a finite state & finite action MDP governed by a ~~stationa~~ stationary policy μ.

$$J_\mu(i) = \lim_{N \to \infty} E\left[\sum_{k=0}^{N-1} \alpha^k g(i, k_{+1}) \,\middle|\, i_0 = i \right] \quad i = 1, \ldots, n$$

Approximate $J_\mu(i)$ according to

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

$r = (r(1) \ldots r(s))^T$ is a param vector &

$\phi(i) = (\phi_1(i), \ldots \phi_s(i))^T$ is a feature vector associated with state i. let ϕ =

$$\begin{bmatrix} -\phi(1)^T- \\ -\phi(2)^T- \\ \vdots \\ -\phi(n)^T- \end{bmatrix}_{n \times s}$$

$$\phi = \begin{bmatrix} \phi_1(1) & \cdots & \phi_s(1) \\ \phi_1(2) & & \phi_s(2) \\ \vdots & & \\ \phi(n) & \cdots & \phi_s(n) \end{bmatrix}_{n \times s}$$

ϕ is called feat matrix

Let $\tilde{J}_r = (\tilde{J}(i, r), i \in S)^T$ where $= \{1, 2 \ldots n\}$ is the state space

Then $\tilde{J}_r = \phi r$

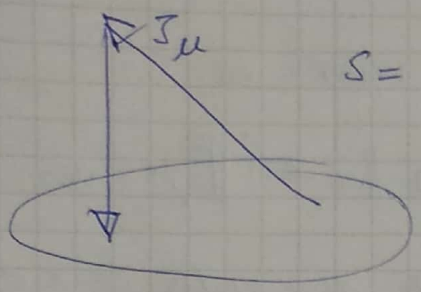
$$\tilde{J}_r = \begin{pmatrix} \tilde{J}(1, r) \\ \tilde{J}(2, r) \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} \phi(r)_1 \\ \phi(r)_2 \\ \vdots \\ \phi(r)_n \end{pmatrix}.$$

eg: $\phi_1(i) = 1$

$\phi_2(i) = i$

$\phi_3(i) = i^2$

Goal: Find the best approx of J_μ within



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

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

Note that $\Phi r = \sum_{i=1}^{s} \phi_i r_i \qquad \phi_i = \begin{pmatrix} \phi_i(1) \\ \vdots \\ \phi_i(n) \end{pmatrix}$

Assumptions:

1. The markov chain $\{X_n\}$ has steady state prob $\xi_1 \ldots \xi_n$ that are all positive, i.e., $\forall i = 1, \ldots n$,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} P(i_k = j \mid i_0 = i) = \xi_j > 0 \quad \forall j = 1, \ldots n$$

The matrix Φ has rank s $[s \leq n]$

$n \approx 10^{1000}$, $s \approx 10-100$

Define a weighted Euclidean norm as $\mathbb{R}^n$ as

$$\|J\|_v = \sqrt{J^T V J} = \sqrt{\sum_{i=1}^{n} v_i (J(i))^2}$$

Here $V = \begin{bmatrix} v_1 & & 0 \\ & v_2 & \\ & & \ddots \\ 0 & & v_n \end{bmatrix}$ $\quad J = (J(i), i \in S)^T$

Assume $v_1, v_2 \ldots v_n > 0$

Let π be the projection operator into S wrt this norm

For any $J \in \mathbb{R}^n$, π^J is the unique vector in

$S = \{ \phi_r \mid r \in \mathbb{R}^s \}$ that minimizes $\|J - \hat{J}\|_V^2$ over

all $\hat{J} \in S$. Since ϕ has rank s, any vector $\hat{J} \in S$

is uniquely written as $\hat{J} = \phi r$ for some $r \in \mathbb{R}^n$.

Thus $\|J - \hat{J}\|_V^2 = \|J - \phi r\|_V^2 = (J - \phi r)^T V (J - \phi r)$

Thus,

$$\pi J = \phi r_J \quad \text{where } r_J = \underset{r \in \mathbb{R}^c}{\arg\min} \|J - \phi r\|_V^2, \; J \in \mathbb{R}^n$$

In orde to fing r_J,

$$\nabla_r \left(\|J - \phi r\|^2 \right) = 0$$

or $\nabla_r \left((J - \phi r)^T V (J - \phi r) \right) = 0$

or $\phi^T V (J - \phi r_J) = 0$

or $\phi^T V J - \phi^T V \phi r_J = 0$

$\Rightarrow \phi V \phi r_J = \phi^T V J$

$\Rightarrow r_J = (\phi^T V \phi)^{-1} \phi^T V J$

$\phi r_J = \pi J$ Thus,

$$\pi J = \phi(\phi^T V \phi)^{-1} \phi V J$$

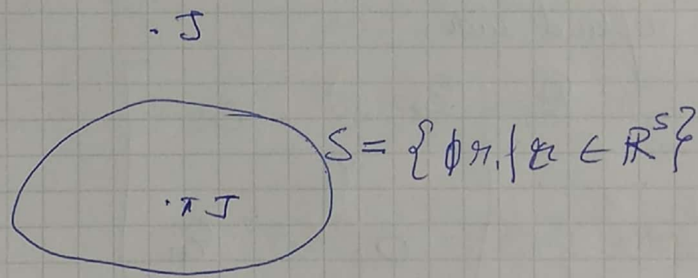
Thus $\pi = \phi(\phi^T V \phi)^{-1} \phi^T V$

Thus π is also called the projection matrix. The optimality condition (*) can be re-written as,

$$r^T \phi^T V (J - \phi r_J) = 0 \quad \text{or}$$

$$\bar{J}^T V (J - \phi r_J) = 0 \quad \text{where } \bar{J} = \phi r \in S$$

The difference / approx error $(J - \phi r_J)$ is othogonal to subspace S.



The vectors x & y are orthogonal if $x^T V y = 0$

$$\text{or } \sum_{i=1}^{n} v_i x_i y_i = 0$$

Consider now the mapping,

$T: V \to V$ where V = space of bounded f^n from S to $\mathbb{R}$ defined by

$$(TJ)(i) = \sum_{j=1}^{n} p_{ij} (g(ij) + \alpha J(j)), \quad i = 1 \ldots n)$$

or $TJ = g + \alpha PJ$

where $g = (g_1 \ldots g_n)^T$ with

$$g_i = \sum_{i=1}^{n} p_{ij}\, g(ij), \quad i = 1 \ldots n$$

$$P = [[p_{ij}]]_{i,j=1}^{n}$$

Bellaman Eqn:

$$J = TJ$$

$$\boxed{\phi r = \Pi T(\phi r)} \qquad T(\phi r) = g + \alpha P(\phi r)$$

↓

projected Bellman Eqn

V_i replaced with

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

$v_i = \xi_i$ = steady state prob of markov chain in state i

Lemma 1: $\|Pz\|_\xi \leq \|z\|_\xi \;\; \forall \; z \in \mathbb{R}^n$ where

P is the transition prob matrix for the markov chain for which $\xi = \{\xi_1, \dots \xi_n\}$ is the stationary distribution

Proof $$\|Pz\|_\xi^2 = \sum_{i=1}^{n} \xi_i \left(\sum_{j=1}^{n} p_{ij} z_j\right)^2 \quad \text{—①}$$

according to Jensons inequality,

$$E[z|i]^2 \leq E[z^2|i]$$

∴ ① becomes

$$\leq \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} p_{ij} z_j^2$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} \xi_i p_{ij} z_j^2$$

$$= \sum_{j=1}^{n} \xi_j z_j^2 = \|z\|_\xi^2$$

or $\|Pz\|_\xi \leq \|z\|_\xi \;\; \forall \; z \in \mathbb{R}^n$

Lemma 2: The projection map π is non ~~expasive~~ expansive i.e.,

$$\|\pi J - \pi \bar{J}\|_v \leq \|J - \bar{J}\|_v \;\; \forall \; J, \bar{J} \in \mathbb{R}^n$$

Proof: Note that,

$$\|\pi(J-\bar{J})\|_v^2 \leq \|\pi(J-\bar{J})\|_v^2 + \|(I-\pi)(J-\bar{J})\|_v^2$$

Note that $\underset{\in S}{\pi(J-\bar{J})} \perp ((J-\bar{J}) - \pi(J-\bar{J}))$

by pythogorean theorem, we have

$$\|\pi(J-\bar{J})\|_v^2 + \|(I-\pi)(J-\bar{J})\|_v^2 = \|\pi\cancel{(J-\bar{J})} + (J-\bar{J}) - \pi\cancel{(J-\bar{J})}\|_v^2$$
$$= \|J-\bar{J}\|_v^2$$

$$\Rightarrow \|\pi(J-\bar{J})\|_v \leq \|J-\bar{J}\|_v \quad \forall J, \bar{J} \in \mathbb{R}^n$$

<u>Proposition</u>: The mappings T & πT are contractions of modulus α with respect to $\|\cdot\|_\xi$ where $\xi = (\xi_1 \cdots \xi_n)^T$ is the steady state distribution of the Markov Chain.

Proof: Recall that,

$$TJ = g + \alpha PJ$$
$$T\bar{J} = g + \alpha P\bar{J}$$

$$\Rightarrow (TJ - T\bar{J}) = \alpha P(J-\bar{J})$$

$$\|TJ - T\bar{J}\|_\xi \leq \alpha \|P(J-\bar{J})\|_\xi$$
$$\leq \alpha \|J-\bar{J}\|_\xi \quad \text{by lemma 1}$$

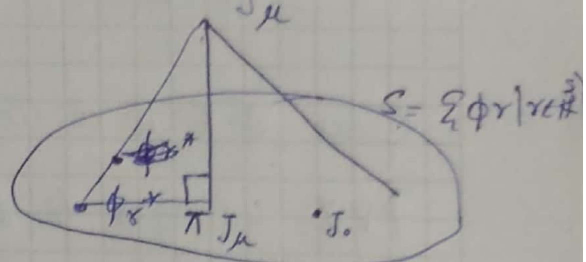
T is a contraction of moduls α wrt $\|\cdot\|_\xi$

Now $\|\pi TJ - \pi T\bar{J}\|_\xi \leq \|TJ - T\bar{J}\|_\xi$
$$\leq \alpha \|J-\bar{J}\|_\xi$$

πT is a contraction of modulus α wrt $\|\cdot\|_\xi$

Proposition 2: Let ϕr^* be the fixed point of πT. Then,

$$\| J_\mu - \phi r^* \|_\xi \leq \frac{1}{\sqrt{1-\alpha^2}} \| J_\mu - \pi J_\mu \|_\xi$$



Proof:

Note that,

$$\| J_\mu - \phi r^* \|_\xi^2 = \| J_\mu - \pi J_\mu \|_\xi^2 + \| \pi J_\mu - \phi r^* \|_\xi^2 \quad \text{by pythagoras theorem}$$

$\pi T \phi r^* = \phi r^*$

$$= \| J_\mu - \pi J_\mu \|_\xi^2 + \| \pi T J_\mu - \pi T (\phi r^*) \|_\xi^2$$

$$\leq \| J_\mu - \pi J_\mu \|_\xi^2 + \alpha^2 \| J_\mu - \phi r^* \|_\xi^2$$

$$\Rightarrow (1-\alpha^2) \| J_\mu - \phi r^* \|_\xi^2 \leq \| J_\mu - \pi J_\mu \|_\xi^2$$

or $$\| J_\mu - \phi r^* \|_\xi \leq \frac{1}{\sqrt{1-\alpha^2}} \| J_\mu - \pi J_\mu \|_\xi$$

The solution to the ~~projected~~ projected Bellman eqn is the vector $J = \phi r^*$ where $r^* = \underset{r \in \mathbb{R}^s}{\operatorname{argmin}} \| \phi r - g + \alpha P \phi r^* \|_\xi^2$

ϕr^* is the projection of $g + \alpha P \phi r^*$ to the subspace S

$$\Rightarrow \phi^T D (\phi r^* - (g + \alpha P \phi r^*)) = 0$$

or $$\phi^T D ((I - \alpha P) \phi r^* - g) = 0$$

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

or $$\underbrace{\phi^T D (I - \alpha P) \phi}_{C} r^* = \phi^T D g$$

$$C r^* = \phi^T D g \quad \text{or} \quad r^* = C^{-1} \phi^T D g$$

Numerical Sol^n to PBE:

Value iteration:

$$\phi r_{k+1} = \pi T(\phi r_k), \ k = 0, 1, 2 \ldots$$

Since πT is a contraction, $\{\phi r_k\}$ generated by PVI converges to the unique fixed point ϕr^* of πT

$\phi r_k \longrightarrow \phi r^*$ where $\phi r^* = \pi T(\phi r^*)$

Note that

$$r_{k+1} = \underset{r \in \mathbb{R}^s}{\arg\min} \|\phi r - (g + \alpha P \phi r_k)\|_\xi^2$$

Consider $\nabla_r (\phi r - g + \alpha P \phi r)^T D (\phi r - g - \alpha P \phi r_k)$

$$= 2\phi^T D (\phi r_{k+1} - (g + \alpha P \phi r_k)) = 0$$

$$= \phi^T D \phi r_{k+1} = \phi_D D(g + \alpha P \phi r_k)$$

$$\therefore r_{k+1} = (\phi^T D \phi)^{-1} \phi^T D g + \alpha (\phi^T D \phi)^{-1} (\phi^T D P \phi r_k), \ k \geq 0$$

$$d = \phi D g.$$

Thus,

$$r_{k+1} = r_k + (\phi^T D \phi)^{-1} d - (\phi^T D \phi)^{-1} \phi^T D \phi r_k + \alpha (\phi^T D \phi)^{-1} \phi^T D P \phi r_k$$

$$= r_k - (\phi^T D \phi)^{-1} \phi^T D (I - \alpha P) \phi r_k + (\phi^T D \phi)^{-1} d$$

$$= r_k - (\phi^T D \phi)^{-1} (C r_k - d) \qquad \text{where } C = \phi^T D (I - \alpha P) \phi$$

Proposition 3: The matrix $C = \phi^T D(I-\alpha P)\phi$ is positive definite.

Proof from pythogorean theorem,

$$\|P\phi r\|_\xi^2 = \|\pi P\phi r_*\|_\xi^2 + \|(I-\pi)P\phi r\|_\xi^2$$

$$\Rightarrow \|\pi P r\|_\xi^2 \le \|P\phi x\|_\xi^2$$

$$\Rightarrow \|\pi P\phi r\|_\xi \le \|P\phi r\|_\xi \le \|\phi r\|_\xi$$

Note that all vectors from ϕr are orthogonal to all vectors of the form $x - \pi x$ i.e.,

$$r^T \phi^T D(x-\pi x) = 0 \quad \forall\, r \in \mathbb{R}^S,\ x \in \mathbb{R}^n$$

Thus $\forall\ r \neq 0$, we have

$$r^T C r = r^T \phi^T D(I-\alpha P)\phi r$$

$$= r^T \phi^T D(I - \alpha\pi P + \alpha(\pi - I)P)\phi r$$

$$= \underbrace{r^T \phi^T D(I-\alpha\pi P)\phi r}_{①} + \underbrace{\alpha r^T \phi^T D(\pi - I)P\phi v}_{② \;=\; 0}$$

② $= 0$ Since ϕr is orthogonal to $(\pi - I)P\phi r$

Then ① $= r^T\phi^T D\phi r - \alpha r^T \phi^T D \pi P \phi r$

$$\ge \|\phi r\|_\xi^2 - \underbrace{\alpha\|\phi r\|_\xi \|\pi P\phi r\|_\xi}_{\text{Cauchy's inequality}}$$

weighted inner product

$$\langle \phi r, \pi P\phi r\rangle_D \le \|\phi r\|_\xi \|\pi P\phi r\|_\xi$$

$$\ge \|\phi r\|_\xi^2 - \alpha\|\phi r\|_\xi^2 = (1-\alpha)\|\phi r\|_\xi^2 > 0 \text{ whenever } r \neq 0$$

Thus $C = [\phi^T D (I - \alpha P) \phi]$ is a +'ve definite matrix

Material: D Bertsekas Optimal Control & Dynamic Program chap 6 [Approx Dynamic Programming]

Chapter 9 : On-Policy prediction with Approx

Mean Square Value Error: $\overline{VE}(w) \triangleq \sum_{s \in S} \mu(s) \left[v_\pi(s) - \hat{v}(s, w) \right]^2$

$\mu = (\mu(1), -- \mu(n))^T$ [steady state dist. of the markov chain]

$\mu(s) \equiv$ fraction of markov chain spends in state s.

On-policy distribution for episodic tasks : let $h(s)$ = prob that an episode begins in state s. Let $\eta(s)$ = no of times step spent in state 's' on avg

Then $\eta(s) = h(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a \mid \bar{s}) p(s \mid \bar{s}, a)$, $\forall s \in S$

$\bar{s}$ - avg.

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}$$