DATE :

~~Two danger in RL~~

During learning (training): we update $\hat{w}$ as (current estimate)

via gradient descent method

$$\hat{w} \leftarrow \hat{w} - \alpha \underbrace{\nabla e_{ms}(w)}_{\text{grad of err}}$$

here min

Where the grad is $\nabla \overset{\text{def}}{=} \frac{d}{dw}$

$$\underbrace{\nabla E_{p} \left[v(s) - \hat{v}(s;w)^2 \right]}_{\text{def of } e_{ms}} = E\left[\nabla (v(s) - \hat{v}(s;w))^2 \right]$$

exchange ∇ and E

$$= \frac{1}{2} E \left[2(\underbrace{v(s)}_{\text{true}} - \underbrace{\hat{v}(s;w)}_{\text{approximate}})' \underbrace{(+\nabla \hat{v}(s;w))}_{f(s) \text{ in linear case}} \right]$$

SL vs RL → Approximate true value (via TD, MC)

① training data exist — $v(s)$ is unknown (no training data)

② iid datapoints (samples) — Markovian data (not iid) collected by the agent

$$\approx E\left[\{ \underbrace{r(s,a) - g + \hat{v}(s';w)}_{\text{approx } v(s) \text{ based BEE}} \} - \hat{v}(s;w) \right] f(s)$$

DATE :

$= \{ r(s,a) - \hat{g} + \hat{v}(s';w) - \hat{v}(s;w) \} f(s)$

↳ Sampling approx using a sample of (s,a,r',s') to the expectation

RL Scheme
- approx value iter
- approx poly iter (GPI)
  - approx pol eval → parametric func approx → linear func approx param
    - $w \in W = \mathbb{R}^{dim(w)}$
    - optimal $w^* = \arg e(w)$, $w \in W$ [illegible]
    - error function
      - e_{ms} ① → update w/ SGD
      - e_{pb}
  - approx improvement

e_{PB} ① doesn't include true value → stable

PB : Projected Bellman error

↳ Bellman operator

From BEE: Immediate

Scalar: $v(s) = r(s,a) - g + E[v(s')]$ ~ BEE

with deterministic policy — equality — ↓ (value next state)

$\underbrace{\text{Value(curr)} = \text{Func}[\text{value (next state}_{t+1})]}$

TD = RHS − LHS = 0 ... theory

DATE :

Matrix: temporal

(vector) $\vec{v} = \vec{r} - g\,\underline{1} + P \cdot \vec{v} \Longleftrightarrow \overbrace{(I-P)}^{\text{singular}} v = r - g\,\mathbb{1}$

↳ ~~State-transition~~ one step State-transition $|S|$ by $|S|$

$\vec{v} = B[v]$

↳ \mathbb(B)

↳ Bellman (evaluator) operator on $\vec{v}$

$e_{PB} \overset{\text{def}}{=} \| \hat{v}(w) - \Pi B\, \hat{v}(w) \|^2_{p^*}$ -- norm vector

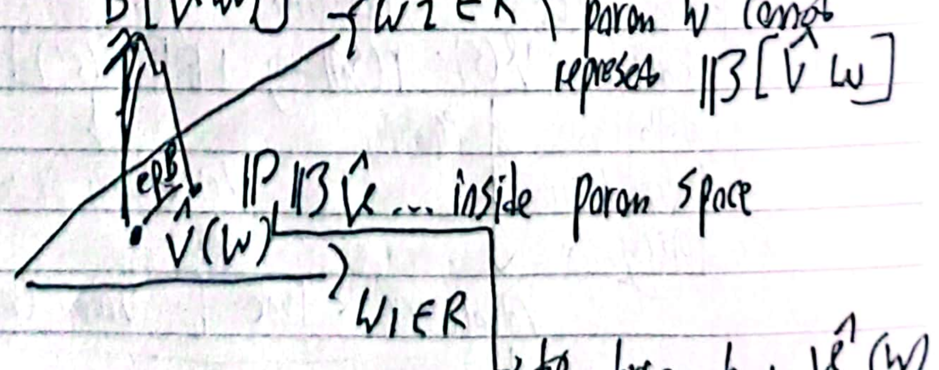
$= (\Delta v)^T D_{p^*} \cdot \Delta v$ where $\Delta v = \hat{v}(w) - \Pi B \hat{v}(w)$

$= \sum p^*(s)\, (\Delta v(s))^2$

$= \mathbb{E}_{p^*} [\Delta v(s)]^2$

Let $\underbrace{\vec{w} = [w_1 \;\; w_2]^T}_{\text{2-dim param vector}}$, $w_1 \in R$

is two in Parameter space



$$\mathbb{P}v = F\tilde{w}$$ — The projection of subspace (feature matrix, true)

where $\tilde{w} = \underset{w \in W}{\operatorname{argmin}} \left[\| \underbrace{Fw}_{\hat{v}} - \underbrace{v}_{\text{true}} \|^2_{p^*} \right]$

exceed derive P!

$$\boxed{\mathbb{P} = F(F^T D_{p^*} F)^{-1} F^T D_{p^*}}$$

determine $W^* = \underset{w}{\operatorname{argmin}} ~~\text{[illegible]}~~ e_{PB}(w)$

How to get w^*

① take the $\underbrace{\text{grad with respect to}}_{\nabla e_{PB}(w)}$ w

② set the grad to zero

Then we can show that
$w^* = X^{-1} Y$

Where
$X = \sum p^*(s) \sum p(s'|s) \left[f(s) \{ f(s) - f(s') \}^T \right]$

$= F^T D_{p^*} (I - P) F$

feature matrix (F), trans matrix (P)

DATE :

$$y = \sum p^*(s) \left[r(s) - g \mathbf{1} \right] f(s)]$$
$$= F^T D p^* (r - g\mathbf{1})$$

$$\underbrace{\hat{X} = \frac{1}{n} \sum_{i=1}^{n} f(s_i) \left[f(s_i) - f(s_i') \right]^T}_{\text{Sampling approx of } X}$$

where n is the number of sample

$$\underbrace{\hat{y} = \frac{1}{n} \sum_{i=1}^{n} (r(s_i) - g) f(s_i)}_{\text{Sampling approx of } X}$$

$$\boxed{\hat{w} = \hat{X}^{-1} \hat{y}} \longrightarrow \text{LSTD}$$

$$\hat{X} \longleftarrow \hat{X} + \varepsilon I$$

In RL, no separation between $\underbrace{\underbrace{\text{training}}_{\text{before convergence}} \text{ vs } \underbrace{\text{testing}}_{\text{after convergence}}}_{\text{If we want to distinguish training \& testing}}$

$\underbrace{\text{If we want to distinguish training \& testing}}_{\text{in the same environment}}$

More challenges ② → non stationary env
① different but similar env