Offline RL

1) Offline RL Introduction
2) Importance Sampling
3) Fitted-Q and BRM ☆ → { Bellman Completeness ☆ / Coverage Condition
4) Pessimism in Face of Uncertainty
5) Learning under Realizability
6) Marginalized Importance Sampling

Ref: Offline Reinforcement Learning in Large State Spaces: Algorithms and Guarantees by N. Jiang & T. Xie

---

Reinforcement Learning: learning from interactions (experiences)

↓

Online △

Empirical Deep RL

↓

simulation

In real scenarios: simulator ? ← medical / education / recommendation

when human are part of "enviroment"

⇒ offline RL: learning from pre-collected data

Data : $\mathcal{D}$ $\quad Q^{\pi}(Q^*) \in \mathcal{F}$? realizability

logging:

Function Approximation (RL in Large State Space)

e.g. state abstraction → piecewise-constant function approximation.

Supervised Learning

classification $\{(X_i, Y_i)\}_{i=1}^{n}$ i.i.d dataset.

$X_i \in \mathcal{X} \quad Y_i \in \{-1, +1\}$

hypothesis: $h: \mathcal{X} \to \{-1, +1\}$

to minimize risk (generalization) $\quad R(h) = \mathbb{E}_{(X,Y)\sim\mu}\left[\mathbb{1}\{h(X) \neq Y\}\right]$

$= \Pr(h(X) \neq Y)$

empirical risk: $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}$

L.L.N. $n \to \infty$ (asymptotic)

Non-asymptotic: by Hoeffding,

for any $h \in \mathcal{H}$, $|\hat{R}(h) - R(h)| \leq \sqrt{\frac{1}{2n} \cdot \log\frac{2}{\delta}}$. w.p. $\geq 1-\delta$.

$(\leq \delta / |\mathcal{H}|)$

ERM : $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{R}(h)$ — hypothesis class $\mathcal{H}$

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; R(h)$$

uniform convergence

$$\Rightarrow \; R(\hat{h}) - R(h^*) \leq 2 \cdot \max_{h \in \mathcal{H}} \left| \hat{R}(h) - R(h) \right|$$

$$\leq 2 \cdot \sqrt{\frac{1}{2n} \cdot \log \frac{2|\mathcal{H}|}{\delta}} \qquad (*)$$

Union Bound.

ERM-then-evaluate

Question: how to estimate the performance of a candidate function and establish a guarantee similar to $(*)$?

policy π

- behavior / logging policy π_b
- target / evaluation policy π (π_t)

$\Downarrow$

off-policy evaluation OPE 异策略评估

Model: MDP $(S, A, P, R, \gamma, d_0)$

large state space ($|S|$: arbitrarily large)

$P: S \times A \to \Delta(S)$ $\quad R: S \times A \to [0, R_{max}]$

policy $\pi : S \to \Delta(A)$

sampling $(t \to \infty)$

trajectory $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_{H-1}, a_{H-1}, r_{H-1})$

$\leq H$ steps

$s_0 \sim d_0$.

$a_t \sim \pi(\cdot \mid s_t)$

$r_t = R(s_t, a_t)$

$s_{t+1} \sim P(\cdot \mid s_t, a_t)$

$r = 0$.

(absorbing state)

Estimand: expected discounted return

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

$$= \mathbb{E}_\pi\left[\sum_{t=0}^{H-1} \gamma^t r_t\right]$$

target / evaluation policy

H: horizon

effective horizon $H = O\left(\frac{1}{1-\gamma}\right)$

Data $\mathcal{D} = \{\tau^{(i)} := (s_0^{(i)}, a_0^{(i)}, r_0^{(i)}, s_1^{(i)}, a_1^{(i)}, r_1^{(i)}, \cdots, r_{H-1}^{(i)})\}_i$.

$a_t^{(i)} \sim \pi_b(\cdot \mid s_t^{(i)})$

behavior / logging policy

Recall: OPE: based on data collected by π_b, accurately evaluate expected return of π

Estimating expectation using samples from a different distribution

estimand: $\mathbb{E}_{x\sim p}[f(x)]$.

↳ Monte Carlo: $X_1, \cdots, X_n \overset{i.i.d}{\sim} p \Rightarrow \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

However, we cannot access p. but we can access q.

If p is fully supported on q:

(informally) $\frac{p(x)}{q(x)} < \infty$ $\forall x \in \mathcal{X}$

$\forall x \in \mathcal{X}$: if $p(x) > 0 \Rightarrow q(x) > 0$.

$\Rightarrow$ Importance Weighted Estimator : $\frac{p(x)}{q(x)} \cdot f(x)$.

i.e. $X_1, \cdots, X_n \overset{i.i.d}{\sim} q \Rightarrow \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)} \cdot f(X_i)$

importance weight

/ importance ratio

/ inverse propensity score (IPS)

Unbiasedness :

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)} \cdot f(x)\right]$$

$$= \sum_{x\in\mathcal{X}} q(x) \cdot \frac{p(x)}{q(x)} \cdot f(x)$$

$$= \sum_{x\in\mathcal{X}} p(x) \cdot f(x) = \mathbb{E}_{x\sim p}[f(x)] \quad ☆$$

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\right] = 1$$

Importance Sampling in Contextual Bandits

(one-step RL)

- context space $\mathcal{X}$
- action space $\mathcal{A}$,
- reward $R : \mathcal{X} \times \mathcal{A} \to \Delta([0,1])$,
- policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$
- estimand $J(\pi) = \mathbb{E}_\pi[r]$

$\Rightarrow \{(x,a,r)^2\}$ by π_b

$x \sim d_0$.

$a \sim \pi_b(x)$

$r \sim R(x,a)$

If π is fully supported on π_b. ε-greedy / Uniform.

for a single sample (x,a,r),

$\to \rho \cdot r$ where $\rho = \dfrac{\pi(a|x)}{\pi_b(a|x)}$,

$$\mathbb{E}[r \mid a \sim \pi] = \mathbb{E}_{(x,a,r)\sim p}[r] = \mathbb{E}_{(x,a,r)\sim q}\left[\frac{p(x,a,r)}{q(x,a,r)} \cdot r\right]$$

generated by π generated by π_b

$$\Rightarrow \frac{p(x,a,r)}{q(x,a,r)} = \frac{\cancel{d_0(x)} \cdot \pi(a|x) \cdot \cancel{R(r|x,a)}}{\cancel{d_0(x)} \cdot \pi_b(a|x) \cdot \cancel{R(r|x,a)}} = \frac{\pi(a|x)}{\pi_b(a|x)}$$

Caveat: While the estimator is unbiased, the variance can be quite large when π and π_b are very different from each other.

π_b: uniform random

π: deterministic

$K = |A|$

$$\Rightarrow \rho = \frac{\pi(a|x)}{\pi_b(a|x)} = \frac{\mathbb{1}\{a=\pi(x)\}}{1/K}$$

Assume r is constant. $\rightarrow$ MC estimator

$$\mathrm{Var}_{\pi_b}[\rho r] = r^2 \cdot \mathrm{Var}_{\pi_b}[\rho]$$

importance weight.

$\mathbb{E}_{\pi_b}[\rho] = 1$

$$= r^2 \cdot \left(\mathbb{E}_{\pi_b}[\rho^2] - \mathbb{E}_{\pi_b}[\rho]^2 \right)$$

$$= r^2 \cdot \left(\mathbb{E}\left[\frac{\mathbb{1}\{a=\pi(x)\}}{1/K^2} \right] - 1 \right)$$

$$= r^2 \cdot \left(K \cdot \mathbb{E}_{\pi_b}[\rho] - 1 \right) = r^2 (K-1)$$

Intuition: only $1/K$ data points are "valid".

$\rightarrow$ effective sample size.

Generally: $\mathrm{Var}[\rho r] \leq \mathbb{E}[\rho^2 r^2] \leq \mathbb{E}[\rho^2] = K$.

roughly tight.

$\rho \in [0, K]$

$r \in [0, 1]$

$\Rightarrow \rho r \in [0, K]$

① Concentration: "low-variance regime"

w.r.t. its range

Hoeffding's inequality: $K \cdot \sqrt{\frac{1}{2n} \cdot \log \frac{2}{\delta}}$.

Bernstein's inequality: $\sqrt{\frac{2K}{n} \log \frac{2}{\delta}} + \frac{2K}{3n} \log \frac{2}{\delta}$.

② Variance Reduction

1) Weighted Importance Sampling

- no variance
- biased
- consistent.

IS: $\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{a_i = \pi(x_i)\}}{1/K} \cdot r_i = \frac{1}{n/K} \cdot \sum_{i: a_i = \pi(x_i)} r_i$

$\Downarrow$

$|\{i : a_i = \pi(x_i)\}|$

"normalization" by expected size $\frac{n}{K}$.

WIS: $\frac{1}{\sum_{i=1}^{n} \rho_i} \cdot \sum_{i=1}^{n} \rho_i r_i$

→ biased

→ asymptotically unbiased.

(when n large: $\sum_{i=1}^{n} \rho_i \doteq n/K$)

2) Control Variate

$Var[\rho r] = r^2 (K-1)$

→ $Var[\rho(r-c)] = (r-c)^2 \cdot (K-1)$

$\Downarrow$

If we set $\underline{\underline{C}} = r$: 0 variance !

Doubly Robust: Suppose $\hat{R} : \mathcal{X} \times A \to \mathbb{R}$

s.t. $\hat{R}(x,a) \approx \mathbb{E}[r \mid x, a]$

estimand

$\Rightarrow$ for any (x, a, r)

$$\boxed{\mathbb{E}_{a' \sim \pi}\left[\hat{R}(x, a')\right]} + \rho \cdot \left(r - \hat{R}(x,a)\right)$$

1) $\hat{R}$ is deterministic $\Rightarrow \rho r \Rightarrow$ IS estimator.

2) unbiased. but reduced variance.

3) $\hat{R}(x,a)$ ✓ ρ ✗ $\Rightarrow$ DR ✓

ρ ✓ $\hat{R}(x,a)$ ✗ $\Rightarrow$ DR ✓

C.f. Doubly Robust Off-Policy Value Evaluation for RL.

Importance Sampling in multi-step RL

$M = (S, A, P, R, \gamma, d_0)$ $\quad \tau = (s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_{H-1}, a_{H-1}, r_{H-1})$

trajectory induced by π_b.

$$J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{H-1} \gamma^t r_t\right] = \mathbb{E}_{\pi_b}\left[\frac{\Pr(\tau \sim \pi)}{\Pr(\tau \sim \pi_b)} \cdot \sum_{t=0}^{H-1} \gamma^t r_t\right].$$

$$\rho = \frac{\Pr(\tau\sim\pi)}{\Pr(\tau\sim\pi_b)} = \frac{\cancel{d_0(s_0)}\cdot\pi(a_0|s_0)\cdot\cancel{P(s_1|s_0,a_0)}\cdots\pi(a_{H-1}|s_{H-1})}{\cancel{d_0(s_0)}\cdot\pi_b(a_0|s_0)\cdot\cancel{P(s_1|s_0,a_0)}\cdots\pi_b(a_{H-1}|s_{H-1})}$$

$$= \prod_{t=0}^{H-1}\frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} := \prod_{t=0}^{H-1}\rho_t := \rho_{0:H-1}$$

$\Rightarrow$ Per-trajectory IS estimator: $\rho_{0:H-1}\cdot\sum_{t=0}^{H-1}\gamma^t r_t = IS(\tau)$

Consider: π deterministic

π_b uniform.

$\Rightarrow$ Variance: $O(K^H)$ $\quad C_A = K$

K: # of actions $(|A|)$

H: horizon.

Concentration: $\{\tau^{(i)}\}_{i=1}^n$

$\Rightarrow \hat{J}_{IS}(\pi) = \frac{1}{n}\sum_{i=1}^n IS(\tau^{(i)})$

$$\sum_{t=0}^{H-1}\gamma^t r_t \in [0, V_{max}]$$

where $V_{max} = \frac{R_{max}}{1-\gamma}$.

Coverage: $\frac{\pi(a|s)}{\pi_b(a|s)} < \infty$

$\Rightarrow$ assume $\exists C_A$: $\max_{s,a}\frac{\pi(a|s)}{\pi_b(a|s)} \le C_A$.

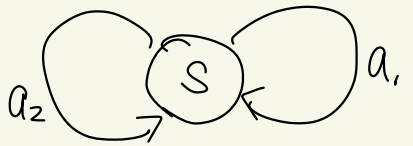
by Bernstein's inequality: $\left|\hat{J}_{IS}(\pi) - J(\pi)\right| \lesssim V_{max}\sqrt{\frac{(C_A)^H}{n}\log\frac{1}{\delta}}$

$\Rightarrow$ The Curse of Horizon. 时域诅咒

If $C_A = 1 + O\left(\frac{1}{H}\right) \Rightarrow (C_A)^H = O(1)$

$\Rightarrow$ π should be closed to π_b. ✗

The Curse of Horizon even exists in simple MDP!



$r = 0$.

π_b: uniform

π: deterministic.

$\rightarrow \rho \sim 2^H$

$\rightarrow Var \sim 4^H$.

By $(\Delta) \Rightarrow$ Sample complexity: $|\hat{J}_{IS}(\pi) - J(\pi)| \leq \varepsilon$

$$\Rightarrow n = O\left(\frac{(C_A)^H V_{max}^2}{\varepsilon^2} \log\frac{1}{\delta}\right) \quad \text{w.p.} \geq 1-\delta$$

☆ implicit constant factor (problem-dependent)

Per-trajectory IS estimator:

$$IS(\tau) = \rho_{0:H-1} \cdot \sum_{t=0}^{H-1} \gamma^t r_t \qquad \rho_{0:H-1} = \prod_{t=0}^{H-1} \rho_t.$$

r_t only depends on $a_{0:t}$

$\Downarrow$

Per-step IS estimator: $\sum_{t=0}^{H-1} \gamma^t \rho_{0:t} r_t$ unbiased

Define $\quad v_0 = 0$

$$v_{H-h+1} = \rho_h (r_h + \gamma \cdot v_{H-h})$$ ← "Bellman equation"

$$\Rightarrow \quad v_{H-1} = \sum_{t=0}^{H-1} \gamma^t \rho_{0:t} r_t \quad \text{(per-step IS)}$$

Assume v_{H-h} is an unbiased estimator of $V^\pi(s_{h+1})$

$\Rightarrow$ $r_h + \gamma \cdot v_{H-h}$ is an unbiased estimator of $Q^\pi(s_h, a_h)$

$a_h \sim \pi_b(\cdot | s_h)$

At step h: bandit $\quad s_h$ context

a_h action

$r_h + \gamma \cdot v_{H-h}$ reward (mean $Q^\pi(s_h, a_h)$)

$\Rightarrow$ $\rho_h (r_h + \gamma \cdot v_{H-h})$ is an unbiased estimator for $Q^\pi(s_h, \pi) = V^\pi(s_h)$

A Doubly Robust $\rightarrow$ $\hat{Q}^\pi$: estimated Q-value.

$$v^{DR}_{H-h+1} = \mathbb{E}_{a \sim \pi}[\hat{Q}^\pi(s_h, a)] + \rho_h (r_h + \gamma \cdot v^{DR}_{H-h} - \hat{Q}^\pi(s_h, a_h))$$

* WIS estimator

Variance of Per-step IS

$$V^\pi(s_h) = \mathbb{E}_h[V_{H-h+1} \mid s_h]$$

$$\mathrm{Var}_h[V_{H-h+1}] = \mathbb{E}_h[V_{H-h+1}^2] - \mathbb{E}_h[V_{H-h+1}]^2.$$

$$= \mathbb{E}_h[V_{H-h+1}^2] - \left(\mathbb{E}_h[V^\pi(s_h)]\right)^2$$

$$= \mathbb{E}_h\Big[\big(\rho_h Q^\pi(s_h,a_h) + \rho_h(r_h + \gamma \cdot V_{H-h} - Q^\pi(s_h,a_h))$$

$$+ V^\pi(s_h) - V^\pi(s_h)\big)^2\Big] - \left(\mathbb{E}_h[V^\pi(s_h)]\right)^2.$$

$$= \mathrm{Var}_h[V^\pi(s_h)] + \underbrace{\mathbb{E}_h[\mathrm{Var}_h[\rho_h Q^\pi(s_h,a_h)] \mid s_h]}_{0}$$

$$+ \gamma^2 \cdot \mathbb{E}_h[\rho_h^2 \cdot \mathrm{Var}_{h+1}[V_{H-h}]].$$

Special case: $\pi_b = \pi$ (on-policy).

$\rho_h \equiv 1$ (deterministic)

$$\Rightarrow \mathrm{Var}_h[V_{H-h+1}] = \mathrm{Var}_h[V^\pi(s_h)] + \gamma^2 \cdot \mathbb{E}_h[\mathrm{Var}_{h+1}[V_{H-h}]]$$

Bellman equation for variance!

"reward" "discount factor"

$$\Rightarrow \mathrm{Var}[V_{H-1}] = \sum_{t=0}^{H-1} \gamma^{2(H-1)} \cdot \mathbb{E}_{s_t \sim d_t^\pi,\, a_t \sim \pi}\left[\mathrm{Var}_{s_t \sim P(\cdot \mid s_{t-1}, a_{t-1})}[V^\pi(s_t)]\right]$$

Recall: OPE $\left\langle \begin{array}{l} \text{target policy } \pi \rightarrow \text{to evaluate} \\ \text{behavior policy } \pi_b \rightarrow \text{to collect data.} \end{array} \right.$

IS estimator $\Rightarrow$ the curse of horizon

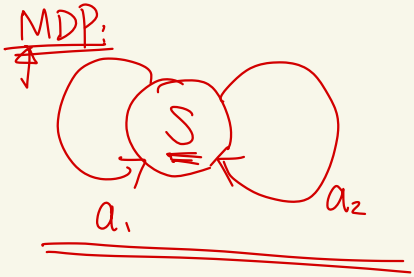
(sample complexity $\propto (C_A)^H$)

$$\left(\prod_{t=0}^{H-1} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}\right) \cdot \left(\sum_{t=0}^{H-1} \gamma^t r_t\right)$$

"trajectory-level"

State $\rightarrow$ Markovian

MDP:



Value Function

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right],$$

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s,\ a_0 = a\right].$$

$$\rightarrow J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, \pi\right]$$

$$= \mathbb{E}_{s\sim d_0}\left[V^\pi(s)\right]$$

$$= \mathbb{E}_{s\sim d_0}\left[Q^\pi(s,\pi)\right]$$

$$\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

Q1: How to estimate Q^π from data ?

Q2: What guarantees can we say about estimated $J(\pi)$?

Q^π: Bellman equation $Q^\pi = \mathcal{T}^\pi Q^\pi$

(policy-specific) Bellman operator

$$\forall f \in \mathbb{R}^{S \times A} : (\mathcal{T}^\pi f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(s,a)} [f(s', \pi)]$$

γ-contraction: $\| \mathcal{T}^\pi f - \mathcal{T}^\pi f' \|_\infty \leq \gamma \cdot \| f - f' \|_\infty$

strict

Tabular: dynamic programming

e.g. Value Iteration (VI) $f_{k+1} \leftarrow \mathcal{T}^\pi f_k$

$$\| f_{k+1} - Q^\pi \|_\infty = \| \mathcal{T}^\pi f_k - \mathcal{T}^\pi Q^\pi \|_\infty$$

$$\leq \gamma \cdot \| f_k - Q^\pi \|_\infty$$

$$\Rightarrow \| f_n - Q^\pi \|_\infty \leq \gamma^n \cdot \| f_0 - Q^\pi \|_\infty$$

$\leq V_{max}$

Q^π: $Q^\pi = \mathcal{T}^\pi Q^\pi$.

Bellman Operator

在大状态空间中，
不能遍历所有状态
⇒ 不能直接"作用" $\mathcal{T}^\pi$.

$$(\mathcal{T}^\pi f)(s,a) = \mathbb{E}\left[r + \gamma \cdot f(s', \pi) \right]$$

$r = R(s,a)$

$s' \sim p(s,a)$

$\rightarrow$ $r, s' \mid s, a$

$f' = \mathbb{E}_r[\, Y \mid X \,]$

$\Uparrow$

$\min_{f'} \mathbb{E}\left[\left(f'(X) - Y \right)^2 \right]$

Least-Square Regression

$$\mathcal{T}^\pi f \in \operatorname*{argmin}_{f' \in \mathbb{R}^{S \times A}} \mathcal{L}(f'; f, \pi)$$

☆ MSE: $\mathcal{L}(f'; f, \pi) = \mathbb{E}_D\left[\left(f'(s,a) - r - \gamma \cdot f(s', \pi) \right)^2 \right]$.

Goal: approximate $\mathcal{T}^\pi$

by least-square regression

$(s,a) \sim d^D$

$r = R(s,a)$

$s' \sim P(\cdot \mid s,a)$

Data:

$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$

$\rightarrow (s_0, a_0, r_0, s_1)$

(s_1, a_1, r_1, s_2) ....

⇒ Data: i.i.d

$\rightarrow (s, a, r, s') \overset{i.i.d}{\sim} D$

Fitted-Q [Ernst et al., 2005].

Goal: estimate Q^π

→ Fitted-Q Evaluation (FQE)

$$f_{k+1} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \hat{\mathcal{L}}(f'; f_k, \pi)$$

$\mathcal{T}^\pi f_k$

① Function class: $\mathcal{F}$.

② $\hat{\mathcal{L}}(f'; f, \pi) = \frac{1}{|D|} \sum_{(s,a,r,s') \in D} \left(f'(s,a) - r - \gamma \cdot f(s', \pi) \right)^2$.

FQE: $f_{k+1} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \hat{\mathcal{L}}(f'; f_k, \pi)$

→ to approximate $\mathcal{T}^\pi f_k$ by data and function class

Function class: $\mathcal{F}$.

$\downarrow$

$|\mathcal{F}|$ finite (but exponentially large)

Function Approximation

$\mathcal{F} = \{ f_\theta : \theta \in \Theta \}$

$$f_{k+1} \leftarrow \underset{f' \in \mathcal{F}}{\operatorname{argmin}} \sum_{(s,a,r,s') \in D} \left(f'(s,a) - r - \gamma \cdot f_k(s', \pi) \right)^2 .$$

Q_θ^π

Stochastic optimization (Semi-SGD)

"online regression"

$$\theta \leftarrow \theta - \frac{\alpha}{2} \cdot \nabla_\theta \Big(f_\theta(s,a) - \underbrace{r - \gamma f_\theta(s', \pi)}_{\text{fixed (don't pass gradient)}} \Big)^2 .$$

$$= \theta - \alpha \cdot \left(f_\theta(s,a) - r - \gamma \cdot f_\theta(s', \pi) \right) \cdot \nabla_\theta f_\theta(s,a)$$

Proposition: FQE can diverge even with all of the following:

$\theta \to \infty$

1) $|D| = \infty$
2) minimization is exact.
3) $\mathcal{F}$ is a 1-dim linear function class.

$$f_\theta(s) = \theta^\top \cdot \phi(s)$$

feature.

4) $\mathcal{F}$ can exactly capture Q^π.

i.e. $Q^\pi \in \mathcal{F}$. (realizability)

Instability

Empirically. Deep RL. → instability.

Deadly Triad

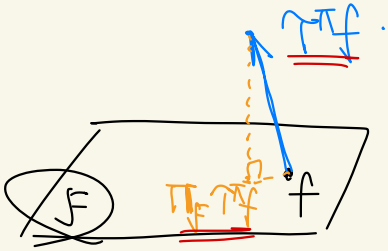
致命三要素

- function approximation ✓✓✓
- off-policy. ✓✓✓
- bootstrapping ✓

min $f(x)$

$\mathcal{T}^{\pi} f$ may not have succint representation

⇒ find the closest approximation in $\mathcal{F}$. (projection)

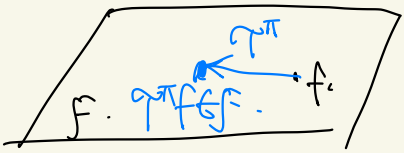


→ minimize PBE

(Projected Bellman Error)

→ Bellman Completeness: $\forall f \in \mathcal{F}.\ \mathcal{T}^{\pi} f \in \mathcal{F}.$

(a.k.a. Bellman closedness / closure)



- For finite $|\mathcal{F}|$:

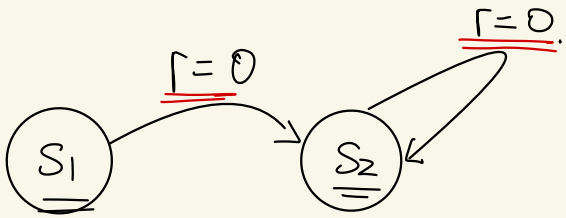
Completeness ⇒ realizability

$$(\gamma^\pi)^\infty f = Q^\pi \in \mathcal{F}.$$

Divergence of FQE

[Tsitsiklis and van Roy, 1996]

MDP:



$\theta^* = 0$

$$V_\theta(s) = \theta^* \cdot \phi(s)$$

$$\phi(s_1) = 1$$

$$\phi(s_2) = 2.$$

$r + \gamma \cdot f_k(s, \pi)$

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\left(\theta - \text{target}_1\right)^2 + \left(2\theta - \text{target}_2\right)^2$$

$$= \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\left(\theta - \gamma \cdot \underbrace{2\theta_k}_{s_2}\right)^2 + \left(2\theta - \gamma \cdot 2\theta_k\right)^2.$$

$$\Rightarrow \left(\theta - \gamma \cdot 2\theta_k\right) + 2 \cdot \left(2\theta - \gamma \cdot 2\theta_k\right) = 0.$$

$$\Rightarrow 5\theta = 6\gamma\theta_k \Rightarrow \boxed{\theta_{k+1} = \frac{6}{5}\gamma \cdot \theta_k}$$

(semi-SGD update)

This diverges if $\gamma \geq \frac{5}{6}$.

$$(s, a) \mapsto r + \gamma \cdot f_k(s', \pi).$$

$\mathcal{T}^\pi f_k \in \mathcal{F}$?

projection is well-specified

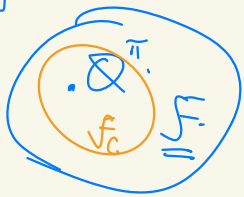
Bellman Completeness: $\forall f \in \mathcal{F}.\ \mathcal{T}^\pi f \in \mathcal{F}$

① Finite $\mathcal{F}$: completeness $\Rightarrow$ realizability

SL

non-monotone

add $f' \in \mathcal{F}$. $\mathcal{T}^\pi f' \notin \mathcal{F}$.



② If true MDP is in some MDP class $\mathcal{M}$.

bounded log-cardinality

$$\mathcal{F} \leftarrow \mathcal{F} \cup \{ \mathcal{T}_M^\pi f : f \in \mathcal{F},\ M \in \mathcal{M} \}$$

$O(H)$

$\rightarrow$ (approximate) Bellman-complete $\mathcal{F}$.

③ Bellman completeness is often satisfied in structured MDPs.

e.g.1. $\boxed{\text{low-rank MDP}}$

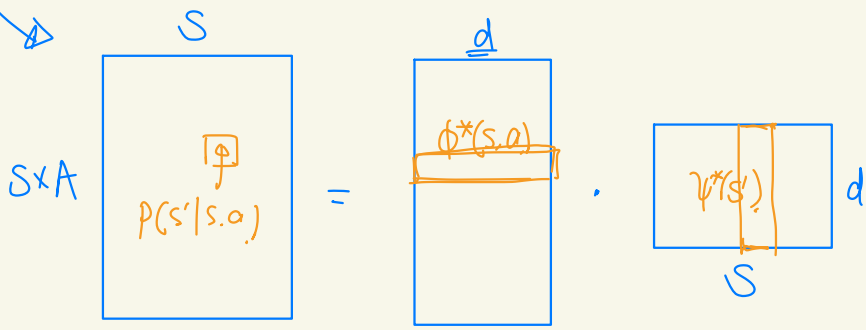
e.g.2. bisimulation abstraction (see "state abstraction")

def (low-rank MDP). MDP with rank d. if

$\exists\ \phi^*: S\times A \to \mathbb{R}^d,\quad \psi^*: S \to \mathbb{R}^d.$

s.t. $\boxed{P(s'|s,a) = \langle \phi^*(s,a), \psi^*(s') \rangle}$

$R(s,a) = \langle \phi^*(s,a), \theta_R \rangle$ for some $\theta_R \in \mathbb{R}^d$.



def (linear MDP) ϕ^* is known.

features.

linear function approximation: $\mathcal{F} = \{ f_\theta = \theta^T x : \theta \in \Theta \}$

linear MDP: P. R

Q^π / Q^* is linear

In linear MDP: $\forall f \in \mathbb{R}^{S \times A}$, π:

$$(\mathcal{T}^{\pi} f)(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a)} [f(s', \pi)],$$

$$= \phi^*(s,a)^T \theta_R + \gamma \cdot \sum_{s'} \phi^*(s,a)^T \psi^*(s') \cdot f(s', \pi),$$

$$= \phi^*(s,a)^T \left(\theta_R + \gamma \cdot \sum_{s'} \psi^*(s') \cdot f(s', \pi) \right)$$

$$= \langle \phi^*(s,a), \quad \rangle$$ is linear in ϕ^*.

$Q^{\pi} = (\mathcal{T}^{\pi})^{\infty} f$ is linear in ϕ^*.

Linear Function Approximation

$$\mathcal{F}_{\phi^*} = \{ \langle \phi^*, \theta \rangle : \theta \in \mathbb{R}^d \}$$

where $f_\theta \in \mathcal{F}_{\phi^*}$ satisfies $f_\theta(s,a) = \phi^*(s,a)^T \theta$.

$\Rightarrow \mathcal{F}_{\phi^*}$ is Bellman complete!

* If ϕ^* unknown, but $\phi^* \in \Phi$, $\bigcup_{\phi \in \Phi} \mathcal{F}_\phi$ is Bellman complete!

Recall: FQE: $f_{k+1} = \underset{f' \in \mathcal{F}}{\operatorname{argmin}} \; \hat{\mathcal{L}}(f'; f_k, \pi)$

iterative

$$\hat{\mathcal{L}}(f'; f, \pi) = \sum_{(s,a,r,s') \in D} \left(f'(s,a) - r - \gamma \cdot f(s', \pi) \right)^2$$

Deadly Triad
- function approximation
- off-policy
- bootstrapping

ERM

- "Global" loss function ?

Q^π is the unique solution for $f = \mathcal{T}^\pi f$.

def (Bellman error / residual) $f - \mathcal{T}^\pi f$

$$\Rightarrow \mathcal{E}(f; \pi) = \mathbb{E}_{d^D}\left[\left(f - \mathcal{T}^\pi f \right)^2 \right].$$

$$\mathbb{E}_{(s,a) \sim d^D}\left[\left(f(s,a) - (\mathcal{T}^\pi f)(s,a) \right)^2 \right]$$

minimize $\mathcal{E}(f; \pi)$

$\rightarrow \mathcal{E}(Q^\pi; \pi) = 0$ is a minimizer!

Bellman Residual Minimization (BRM)

[Antos et al. 2008]

minimize $\mathcal{E}(f;\pi) = \mathbb{E}_{d^D}\left[(f - \mathcal{T}^\pi f)^2\right]$

$(s,a)\sim d^D$

Non-trivial !

$$\mathcal{E}(f;\pi) = \mathbb{E}_{(s,a)\sim d^D}\left[\left(f(s,a) - \boxed{\mathbb{E}\left[r+\gamma\cdot f(s',\pi)\right]}\right)^2\right]$$

$r = R(s,a)$

$s'\sim p(s,a)$ $\rightarrow \mathbb{E}[Y|X]$

(with infinite data) $\#$

$$\mathcal{L}(f;f,\pi) = \mathbb{E}_{(s,a)\sim d^D}\left[\left(f(s,a) - r - \gamma\cdot f(s',\pi)\right)^2\right]$$

$r = R(s,a)$

$s'\sim p(s,a)$

Y

$\hat{\mathcal{L}}(f;f,\pi)$

from data $\{(s,a,r,s')\}$.

$$\underbrace{\mathbb{E}_{X,Y}\left[(Y - h(X))^2\right]}_{\text{(MSE loss)}} = \mathbb{E}_X\left[\left(\mathbb{E}_Y[Y|X] - h(X)\right)^2\right] \text{ (excess risk)}$$

$$+ \mathbb{E}_{X,Y}\left[\left(\mathbb{E}_Y[Y|X] - Y\right)^2\right] \text{ (label noise)}$$

bias-variance decomposition

$Y = r + \gamma\cdot f(s',\pi)$

$X = s, a.$

$$\mathcal{L}(f; f, \pi) = \mathcal{E}(f; \pi) + \mathcal{L}(\mathcal{T}^{\pi} f; f, \pi)$$

can be approximated by data

$$\hat{\mathcal{L}}(\cdot) \to \mathcal{L}(\cdot)$$

$$\mathcal{L}(\cdot) \neq \mathcal{E}(\cdot)$$

"label noise"

In SL: minimize excess risk

$\Leftrightarrow$ minimize MSE loss

(label noise is independent to hypothesis h)

$\min_{h \in \mathcal{H}} \mathcal{L}(h)$

$\mathcal{L}(\mathcal{T}^{\pi} f; f, \pi)$ depends on candidate f !

Double Sampling Problem

$$\mathcal{E}(f; \pi) = \mathbb{E}_{d^{\pi}}\left[\left(f - \mathcal{T}^{\pi} f\right)^2\right]$$

$$(\mathcal{T}^{\pi} f)(s,a) = \mathbb{E}_{r, s' | s, a}\left[r + \gamma \cdot f(s', \pi)\right]$$

Workaround #1: (Double Sampling)

for each (s, a), we sample two independent (r_A, s'_A), (r_B, s'_B)

$\rightarrow$ form an unbiased estimator of $\mathcal{E}(f; \pi)$

$\Downarrow$

$$\left(f(s,a) - (r_A + \gamma \cdot f(s'_A, \pi))\right) \cdot \left(f(s,a) - (r_B + \gamma \cdot f(s'_B, \pi))\right)$$

is an unbiased estimator of $\mathcal{E}(f; \pi)$

$$\mathbb{E}\left[\left(h(x)-\mathbb{E}[Y|x]\right)^2\right] = \mathbb{E}\left[\left(h(x)-\underline{Y_A}\right)\left(h(x)-\underline{Y_B}\right)\right]$$

"label noise": $\mathbb{E}_{x,Y}\left[\left(\underline{\mathbb{E}[Y|x]-Y}\right)^2\right] \geq 0$

$\mathbb{E}[X^2] \geq \underline{\mathbb{E}[X]^2} \geq 0$

$\xrightarrow{\text{double sampling}}$ $\mathbb{E}_{x,Y}\left[\left(\underline{Y_A-Y}\right)\left(\underline{Y_B-Y}\right)\right]$

$= \mathbb{E}\left[\underline{Y_A-Y}\right]\cdot\mathbb{E}\left[\underline{Y_B-Y}\right] = 0$

independence

Caveat: only doable with simulator (with reset)!

<u>Workaround #2</u>: $\mathcal{L}(f;f,\pi) = \mathcal{E}(f;\pi) + \mathcal{L}(\mathcal{T}^\pi f;f,\pi)$

Can we estimate it from data?

$$\mathcal{L}(\mathcal{T}^\pi f;f,\pi) = \mathbb{E}_{\substack{(s,a)\sim d^\pi \\ r=R(s,a) \\ s'\sim P(\cdot|s,a)}}\left[\left((\mathcal{T}^\pi f)(s,a) - \left(r+\gamma\cdot f(s',\pi)\right)\right)^2\right]$$

is the expectation of (conditional expectation)

$$\boxed{\min_{g\in\mathbb{R}^{SA}} \mathcal{L}(g;f,\pi) = \mathcal{L}(\mathcal{T}^\pi f;f,\pi)}$$

$$\mathcal{E}(f;\pi) = \max_{g\in \underline{G}}\left(\mathcal{L}(f;f,\pi) - \mathcal{L}(g;f,\pi)\right)$$

use some function class G to model g.

$$\hat{\mathcal{E}}(f;\pi) = \max_{g\in\mathcal{G}} \left(\hat{\mathcal{L}}(f;f,\pi) - \hat{\mathcal{L}}(g;f,\pi) \right)$$

$\Rightarrow$ BRM algorithm : $\boxed{\hat{f}^{\pi} = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \hat{\mathcal{E}}(f;\pi)}$

① Computation : minimax optimization

(tractable in linear setting. with DNN ?)

LSTDQ

② function class: $f \in \mathcal{F}$.
$g \in \mathcal{G}$.

$$\min_{g\in\mathcal{G}} \mathcal{L}(g;f,\pi) \approx \mathcal{L}(\mathcal{T}^{\pi}f;f,\pi) = \min_{g\in\mathbb{R}^{SA}} \mathcal{L}(g;f,\pi)$$

expressiveness

Assumption: $\forall f\in\mathcal{F}.\ \mathcal{T}^{\pi}f \in \mathcal{G}$

If $\mathcal{F}=\mathcal{G}$ $\Rightarrow$ $\forall f\in\mathcal{F}\ \ \mathcal{T}^{\pi}f\in\mathcal{F}$

(Bellman Completeness)

Iterative alg: FQE
Minimax alg: BRM

$\rightarrow$ both need expressiveness assumption (whether bootstrapping or not)

Recall: BRM $\hat{f}^{\pi} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \hat{\mathcal{E}}(f;\pi)$

squared Bellman error (empirical)

By concentration (Hoeffding)

w.p. $\geq 1-\delta$, $\forall f \in \mathcal{F}$:

$$\left| \hat{\mathcal{E}}(f;\pi) - \mathcal{E}(f;\pi) \right| \lesssim \frac{V_{max}^2}{n} \cdot \log \frac{|\mathcal{F}|}{n}.$$

① $\mathcal{E}$ is quadratic

② union bound over $\mathcal{F}$.

$\mathcal{E}(f;\pi) = \mathbb{E}[(f - \mathcal{T}^{\pi} f)^2]$

Bellman completeness

$\Rightarrow f - \mathcal{T}^{\pi} f \in \mathcal{F}$.

Bounding Bellman error:

$$\mathcal{E}(\hat{f}^{\pi};\pi) = \mathcal{E}(\hat{f}^{\pi};\pi) - \mathcal{E}(Q^{\pi};\pi)$$

$$= \mathcal{E}(\hat{f}^{\pi};\pi) - \hat{\mathcal{E}}(\hat{f}^{\pi};\pi)$$

$$+ \underbrace{\hat{\mathcal{E}}(\hat{f}^{\pi};\pi)}_{\leq \hat{\mathcal{E}}(Q^{\pi};\pi)} - \mathcal{E}(Q^{\pi},\pi)$$

uniform

$$\leq 2 \sup_{f \in \mathcal{F}} \left| \mathcal{E}(f;\pi) - \hat{\mathcal{E}}(f;\pi) \right|$$

$$\lesssim \frac{V_{max}^2}{n} \cdot \log \frac{|\mathcal{F}|}{n}.$$

for finite $\mathcal{F}$.

$\hat{f}^\pi \in \mathcal{F}$.

$Q^\pi \in \mathcal{F}$

(realizability)

def (distribution-weighted p-norm)

$$f \in \mathbb{R}^{\mathcal{X}}: \quad \|f\|_{p,\mu}^p = \mathbb{E}_{x\sim\mu}[(f(x))^p].$$

for $\mu \in \Delta(\mathcal{X})$

e.g. μ is uniform on (x_1, x_2, x_3)

$$\Rightarrow \|f\|_{2,\mu}^2 = \frac{x_1^2 + x_2^2 + x_3^2}{3}.$$

$\|f\|_2^2 = x_1^2 + x_2^2 + x_3^2$

$\|f\|_{1,\mu} \le \|f\|_{2,\mu} \le \|f\|_{\infty,\mu}$.

$\|f\|_1 \ge \|f\|_2 \ge \|f\|_\infty$

$\Rightarrow$ Bellman error: $\|\hat{f}^\pi - \mathcal{T}^\pi \hat{f}^\pi\|_{2,d^D} \lesssim V_{max} \sqrt{\frac{1}{n} \cdot \log \frac{|\mathcal{F}|}{n}}$

$\downarrow$

Value error ?

A classical result: $\|f - Q^\pi\|_\infty \le \frac{1}{1-\gamma} \cdot \|f - \mathcal{T}^\pi f\|_\infty$

$J_f(\pi) - J(\pi)$

$\mathbb{E}_{s\sim d_0}[f(s,\pi)]$

Norm mismatch !

Recall: Under Bellman Completeness,

$$\|\hat{f}^{\pi} - \mathcal{T}^{\pi}\hat{f}^{\pi}\|_{2,d^D} \lesssim V_{max}\sqrt{\frac{1}{n}\log\frac{|\mathcal{F}|}{n}}$$

$$\|\cdot\|_{p,\mu}^{p} = \mathbb{E}_{\mu}[|\cdot|^{p}]$$

Lemma (Bellman error telescoping)

$$\forall \pi.\ f \in \mathbb{R}^{S\times A}: \quad J_f(\pi) = \mathbb{E}_{s\sim d_0}[f(s,\pi)]$$

$$J_f(\pi) - J(\pi) = \frac{1}{1-\gamma}\cdot \mathbb{E}_{d^{\pi}}[f - \mathcal{T}^{\pi}f]$$

Bellman error.

state-action occupancy (visitation frequency)

$$d^{\pi} = (1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^{t}\cdot d_t^{\pi}$$

$$d_t^{\pi}(s,a) = \Pr_{\pi}(s_t = s, a_t = a)$$

Guarantee of BRM:

$$J_f(\pi) - J(\pi)$$

$$= \frac{1}{1-\gamma}\cdot \mathbb{E}_{d^{\pi}}[f - \mathcal{T}^{\pi}f]$$

$$\leq \frac{1}{1-\gamma}\cdot \|f - \mathcal{T}^{\pi}f\|_{1,d^{\pi}}$$

$$\leq \frac{1}{1-\gamma}\cdot \|f - \mathcal{T}^{\pi}f\|_{2,d^{\pi}} \quad \longleftrightarrow \quad \|f - \mathcal{T}^{\pi}f\|_{2,d^{D}}$$

target distribution

data distribution

$$\leq \frac{\sqrt{C_{\pi}}}{1-\gamma}\cdot \|f - \mathcal{T}^{\pi}f\|_{2,d^{D}} \lesssim \frac{V_{max}}{1-\gamma}\cdot\sqrt{\frac{C_{\pi}}{n}\cdot\log\frac{|\mathcal{F}|}{\delta}}$$

Lemma For any function ξ over space $\mathcal{X}$. $\mu, \nu \in \Delta(\mathcal{X})$

$1 \le p < \infty$:

$$\|\xi\|_{p,\mu}^p \le \left\|\frac{\nu}{\mu}\right\|_\infty \|\xi\|_{p,\nu}^p$$

norm transform

$$\left\|\frac{\nu}{\mu}\right\|_\infty = \max_x \frac{\nu(x)}{\mu(x)}$$

"Coverage"

Proof: $\|\xi\|_{p,\mu}^p = \mathbb{E}_{x\sim\mu}[\,|\xi(x)|^p\,]$

$$= \sum_{x\in\mathcal{X}} \mu(x)\cdot|\xi(x)|^p$$

$$= \sum_{x\in\mathcal{X}} \frac{\mu(x)}{\nu(x)}\,\nu(x)\cdot|\xi(x)|^p$$

$$\le \left\|\frac{\nu}{\mu}\right\|_\infty \cdot \sum_{x\in\mathcal{X}} \nu(x)\cdot|\xi(x)|^p = \left\|\frac{\nu}{\mu}\right\|_\infty \cdot \|\xi\|_{p,\nu}^p \quad \square$$

Assumption 2: (Coverage Condition) $\left\|\frac{d^\pi}{d^D}\right\|_\infty \le C_\pi < \infty$

$$\forall (s,a):\quad d^\pi(s,a) > 0 \Rightarrow d^D(s,a) > 0.$$

i.e. data "covers" all (s,a) pairs that target policy π will visit.

Two core assumptions in offline RL
- function approximation → Bellman Completeness
- Data → Coverage Condition

$$|J_{\hat{f}^\pi}(\pi) - J(\pi)| = \left| \frac{1}{1-\gamma} \mathbb{E}_{d^\pi}[\hat{f}^\pi - \mathcal{T}^\pi \hat{f}^\pi] \right|$$

$$\leq \frac{1}{1-\gamma} \cdot \| \hat{f}^\pi - \mathcal{T}^\pi \hat{f}^\pi \|_{1, d^\pi}$$

NOT TIGHT!

$$\leq \frac{1}{1-\gamma} \cdot \| \hat{f}^\pi - \mathcal{T}^\pi \hat{f}^\pi \|_{2, d^\pi}$$

coverage

$$\leq \frac{\sqrt{C_\pi}}{1-\gamma} \cdot \| \hat{f}^\pi - \mathcal{T}^\pi \hat{f}^\pi \|_{2, d^D}$$

$C_\pi = \left\| \frac{d^\pi}{d^D} \right\|_\infty$: concentrability coefficient.

Improve 1: $C_\pi^{sq} = \max_{f \in \mathcal{F}} \frac{\| f - \mathcal{T}^\pi f \|_{2, d^\pi}^2}{\| f - \mathcal{T}^\pi f \|_{2, d^D}^2}$

e.g. $\mathcal{F}$ is linear class induced by a feature map $\phi \in \mathbb{R}^d$.

$\forall f \in \mathcal{F}. \ \exists \theta$ s.t. $f(s,a) = \phi(s,a)^T \theta$.

Bellman completeness $\Rightarrow \mathcal{T}^\pi f \in \mathcal{F}$.

$\Rightarrow$ Bellman error: $f - \mathcal{T}^\pi f \in \mathcal{F}$.

Assume: $f - \mathcal{T}^\pi f = \phi^T u$.

$$\mathbb{E}_{(s,a) \sim d^\pi}[(f - \mathcal{T}^\pi f)(s,a)] = \sum_{(s,a)} d^\pi(s,a) \cdot (f - \mathcal{T}^\pi f)(s,a)$$

$$= \sum_{(s,a)} d^\pi(s,a) \cdot \phi(s,a)^T u.$$

$$\Rightarrow \|f - \mathcal{T}^\pi f\|_{2,d^\pi}^2 = u^T \left(\sum_{(s,a)} d^\pi(s,a)\, \phi(s,a)\, \phi(s,a)^T \right) u.$$

$$= u^T \cdot \mathbb{E}_{d^\pi}[\phi \phi^T] \cdot u$$

$$\|f - \mathcal{T}^\pi f\|_{2,d^\pi}^2 = u^T \Sigma_\pi u.$$

$$\|f - \mathcal{T}^\pi f\|_{2,d^D}^2 = u^T \Sigma_D u.$$

Generalized Rayleigh Quotient

$$\Rightarrow C_\pi^{sq} \leq \max_{f \in \mathcal{F}} \frac{\|f - \mathcal{T}^\pi f\|_{2,d^\pi}^2}{\|f - \mathcal{T}^\pi f\|_{2,d^D}^2} = \max_{u \in \mathbb{R}^d} \frac{u^T \Sigma_\pi u}{u^T \Sigma_D u.}$$

$$= \sigma_{max}\left(\Sigma_D^{-\frac{1}{2}} \Sigma_\pi \Sigma_D^{-\frac{1}{2}} \right)$$

$$\Sigma_\pi = \mathbb{E}_{d^\pi}[\phi \phi^T]$$

$$\Sigma_D = \mathbb{E}_{d^D}[\phi \phi^T].$$

Rayleigh Quotient:

$$\max_{u \in \mathbb{R}^d} \frac{u^T A u}{u^T u} = \sigma_{max}(A)$$

(A is PSD)

$\Downarrow$

Generalized Rayleigh Quotient:

$$\max_{u \in \mathbb{R}^d} \frac{u^T A u}{u^T B u} = \frac{(u B^{\frac{1}{2}})^T B^{-\frac{1}{2}} A \cdot B^{-\frac{1}{2}} (u B^{\frac{1}{2}})}{(u B^{\frac{1}{2}})^T (u B^{\frac{1}{2}})}$$

$$= \max_{v \in \mathbb{R}^d} \frac{v^T B^{-\frac{1}{2}} A B^{-\frac{1}{2}} v}{v^T v.}$$

$$= \sigma_{max}(B^{-\frac{1}{2}} A B^{-\frac{1}{2}})$$

PCA. LDA

Coverage

C_π^{sq} : $(s,a) \sim d^D$ to cover (hit)

all feature directions activated by d^π

Remark: $\exists \pi$: $\|d^\pi / d^D\| \to \infty$

with bounded C_π^{sq}.

Improve 2 : $C_\pi^{avg} = \max_{f \in \mathcal{F}} \frac{\left(\mathbb{E}_{d^\pi}[f - \mathcal{T}^\pi f]\right)^2}{\mathbb{E}_{d^D}[(f - \mathcal{T}^\pi f)^2]}$.

e.g. linear setting: $C_\pi^{avg} \le \mathbb{E}_{d^\pi}[\phi]^T \Sigma_D^{-1} \mathbb{E}_{d^\pi}[\phi]$.

↳ data only need to hit the average feature direction under d^π.

BRM : $\hat{f}^\pi = \arg\min_{f \in \mathcal{F}} \hat{\mathcal{E}}(f;\pi)$. → bounding value error by Bellman error

$J_f(\pi) - J(\pi) = \mathbb{E}_{d^\pi}[f - \mathcal{T}^\pi f]$

FQE: $f_{k+1} = \arg\min_{f \in \mathcal{F}} \hat{\mathcal{L}}(f; f_k, \pi)$.

least square regression

$\mathcal{L}(f; f_k, \pi) = \mathbb{E}_D\left[\left(f(s,a) - r - \gamma \cdot f_k(s', \pi)\right)^2\right]$.

Claim: $\|f_{k+1} - \mathcal{T}^\pi f_k\|_{2,d^D}^2 = \mathcal{L}(f_{k+1}; f_k, \pi) - \mathcal{L}(\mathcal{T}^\pi f_k; f_k, \pi)$

"Bellman error"

Concentration : w.p. $\ge 1-\delta$. $\forall f, f' \in \mathcal{F}$.

$$\left| \hat{\mathcal{L}}(f'; f, \pi) - \mathcal{L}(f'; f, \pi) \right| \le \varepsilon = O\left(\frac{V_{max}^2}{n} \cdot \log \frac{|\mathcal{F}|}{\delta}\right)$$

minimizer of $\hat{\mathcal{L}}$

$$\Rightarrow \underline{\underline{\| f_{k+1} - \mathcal{T}^\pi f_k \|_{2,d^D}^2}} \le \underbrace{\hat{\mathcal{L}}(f_{k+1}; f_k, \pi) - \hat{\mathcal{L}}(\mathcal{T}^\pi f_k; f_k, \pi)}_{\le 0} + 2\varepsilon$$

$\in \mathcal{F}$.

(Bellman Completeness)

$$\Rightarrow \boxed{\| f_{k+1} - \mathcal{T}^\pi f_k \|_{2,d^D} \lesssim V_{max} \cdot \sqrt{\frac{1}{n} \cdot \log \frac{|\mathcal{F}|}{\delta}}}$$

<u>Proof of Claim</u>: $\mathcal{L}(f_{k+1}; \underline{f_k}, \pi) - \mathcal{L}(\mathcal{T}^\pi f_k; f_k, \pi)$

$$= \mathbb{E}_D \left[(f_{k+1} - y)^2 - (\mathcal{T}^\pi f_k - y)^2 \right]$$

$y = r + \gamma \cdot f_k(s', \pi)$

$$= \mathbb{E}_D \left[f_{k+1}^2 - (\mathcal{T}^\pi f_k)^2 - 2y \cdot (f_{k+1} - \mathcal{T}^\pi f_k) \right]$$

condition expectation: $\mathbb{E}[y \mid s, a] = (\mathcal{T}^\pi f_k)(s, a)$

law of total expectation

$$= \mathbb{E}_D \left[f_{k+1}^2 - (\mathcal{T}^\pi f_k)^2 - 2 \cdot (\mathcal{T}^\pi f_k)(f_{k+1} - \mathcal{T}^\pi f_k) \right]$$

$$= \mathbb{E}_D \left[f_{k+1}^2 - 2 \cdot f_{k+1} \cdot (\mathcal{T}^\pi f_k) + (\mathcal{T}^\pi f_k)^2 \right]$$

$$= \mathbb{E}_D \left[(f_{k+1} - \mathcal{T}^\pi f_k)^2 \right] = \underline{\underline{\| f_{k+1} - \mathcal{T}^\pi f_k \|_{2,d^D}^2}} \quad \square$$

per-iteration Bellman error.

Suppose run FQE for K rounds.

$$\left| J_{f_K}(\pi) - J(\pi) \right| = \underline{\underline{\left| \mathbb{E}_{s \sim d_0} [f_K(s, \pi)] - \mathbb{E}_{s \sim d_0} [Q^\pi(s, \pi)] \right|}}$$

$$= \left| \mathbb{E}_{s\sim d_0} \Big[\underbrace{f_k(s,\pi) - \mathcal{T}^\pi f_{k-1}(s,\pi)}_{\text{(I) per-iteration Bellman error}} + \underbrace{\mathcal{T}^\pi f_{k-1}(s,\pi) - \overset{\mathcal{T}^\pi Q^\pi}{Q^\pi(s,\pi)}}_{\text{(II) "}\gamma\text{-contraction"}} \Big] \right|$$

$$(\mathrm{I}): \le \| f_k - \mathcal{T}^\pi f_{k-1} \|_{1,\, d_0 \times \pi}$$

s a

$$\le \| f_k - \mathcal{T}^\pi f_{k-1} \|_{2,\, d_0 \times \pi}$$

d_1^π

$$\le \sqrt{C_\pi} \cdot \| f_k - \mathcal{T}^\pi f_{k-1} \|_{2,\, d^D}$$

$$\lesssim V_{max} \cdot \sqrt{\frac{C_\pi}{n} \cdot \log \frac{|\mathcal{F}|}{\delta}} \triangleq \varepsilon_c$$

Coverage Condition of FQE:

BRM: $\left\| \frac{d^\pi}{d^D} \right\|_\infty \le C_\pi < \infty$

FQE: $\left\| \frac{d_t^\pi}{d^D} \right\|_\infty \le C_{\pi,t} < \infty$ $\forall t$.

occupancy at time t

$$(\mathrm{II}): \left| \mathbb{E}_{s\sim d_0} \Big[(\mathcal{T}^\pi f_{k-1})(s,\pi) - (\mathcal{T}^\pi Q^\pi)(s,\pi) \Big] \right|$$

$$\le \left| \cancel{\mathbb{E}_{s\sim d_0} \big[R(s,\pi)} + \gamma \cdot \mathbb{E}_{s'\sim P(s,\pi)} \big[f_{k-1}(s',\pi) \big] \right.$$

$$\left. - \cancel{\mathbb{E}_{s\sim d_0} \big[R(s,\pi)} - \gamma \cdot \mathbb{E}_{s'\sim P(s,\pi)} \big[Q^\pi(s',\pi) \big] \right|$$

$$\le \gamma \cdot \mathbb{E}_{\substack{s\sim d_0,\, a\sim\pi \\ s'\sim P(s,a)}} \big[f_{k-1}(s',\pi) - Q^\pi(s',\pi) \big]$$

$\Leftrightarrow s' \sim d_1^\pi$

c.f. Goal: $\mathbb{E}_{s\sim d_0} \big[f_k(s,\pi) - Q^\pi(s,\pi) \big]$.

$$\Rightarrow \left| \mathbb{E}_{s\sim d_0}\left[f_k(s,\pi) - Q^\pi(s,\pi)\right] \right|$$

$$\leq \varepsilon_c + \gamma \cdot \left| \mathbb{E}_{s\sim d_1^\pi}\left[f_{k-1}(s,\pi) - Q^\pi(s,\pi)\right] \right|$$

$$\leq \varepsilon_c + \gamma \cdot \underline{\varepsilon_c} + \gamma^2 \left| \mathbb{E}_{s\sim d_2^\pi}\left[f_{k-2}(s,\pi) - Q^\pi(s,\pi)\right] \right|$$

$$\leq \cdots$$

$$\leq \underline{\varepsilon_c(1+\gamma+\gamma^2+\cdots\gamma^{k-1})} + \gamma^k \left| \mathbb{E}_{s\sim d_k^\pi}\left[\underline{Q^\pi(s,\pi)}\right] \right|$$

$$\leq \frac{\varepsilon_c}{1-\gamma} + \gamma^k \cdot V_{max} \qquad (f_0 \equiv 0)$$

dominate $= O\left(\frac{V_{max}}{1-\gamma} \cdot \sqrt{\frac{C_\pi}{n} \log \frac{|F|}{\delta}} \right)$

A "modern" way of analyzing FQE.

Key observation: running FQE for K rounds can be viewed as learning the non-stationary value function for a truncated, finite horizon MDP.

$$J_K(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{K-1} \gamma^t r_t \right].$$

$J_K(\pi)$ well-approximates $J(\pi)$

(approximation error γ^K)

Value function (time-dependent)

$$Q_k^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{k-1}\gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

- $J_{Q_k^\pi}(\pi) = J_k(\pi)$
- Q_k^π satisfies Bellman equation: $Q_k^\pi = \mathcal{T}^\pi Q_{k-1}^\pi$

Q^π

$Q_k^\pi \quad Q_{k-1}^\pi \quad \cdots \quad Q_1^\pi \quad Q_0^\pi \equiv 0$

$f_k \quad f_{k-1} \quad \cdots \quad f_1$

iterative nature

Bellman error: $\|f_k - \mathcal{T}^\pi f_{k-1}\|_{2,d^\pi}$

Lemma
- policy sequence $\pi_{k:1}$ (non-stationary)
- function sequence: $f_{k:1}$

$$\Rightarrow J_{f_k}(\pi_k) - J_k(\pi_{k:1}) = \sum_{t=0}^{k-1}\gamma^t \mathbb{E}_{d_t^{\pi_{k:1}}}\left[f_{k-t} - \mathcal{T}^{\pi_{t+1}} f_{k-t-1}\right]$$

Here, $\pi_k = \pi_{k-1} = \cdots = \pi_1 = \pi$:

$$J_{f_k}(\pi) - J_k(\pi) = \sum_{t=0}^{k-1}\gamma^t \mathbb{E}_{d_t^\pi}\left[f_{k-t} - \mathcal{T}^\pi f_{k-t-1}\right]$$

suffer an approx. error.

c.f. (BRM) $J_f(\pi) - J(\pi) = \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{d_t^\pi}\left[f - \mathcal{T}^\pi f\right]$

$$\mathbb{E}_{d_t^\pi}\left[f_{k-t} - \mathcal{T}^\pi f_{k-t-1}\right] \le \| f_{k-t} - \mathcal{T}^\pi f_{k-t-1} \|_{1, d_t^\pi}$$

$$\le \| f_{k-t} - \mathcal{T}^\pi f_{k-t-1} \|_{2, d_t^\pi}$$

Coverage:
need to cover all d_t^π

$$\le \sqrt{C_\pi} \cdot \underbrace{\| f_{k-t} - \mathcal{T}^\pi f_{k-t-1} \|_{2, d^D}}_{\text{well-controlled}}$$

Proof of Lemma:

$$LHS = \mathbb{E}_{s \sim d_0}\left[f_k(s, \pi_k)\right] - \mathbb{E}_{\pi_{k:1}}\left[\sum_{t=0}^{k-1} \gamma^t r_t\right]$$

$$\left(f_{k-t} - \mathcal{T}^{\pi_{t+1}} f_{k-t-1}\right)(s,a)$$

$$= f_{k-t}(s,a) - r_{k-t} - \gamma \cdot \mathbb{E}_{\substack{s' \sim P(s,a) \\ a' \sim \pi_{t+1}(s')}}\left[f_{k-t-1}(s', a')\right]$$

$$(s,a) \sim d_t^{\pi_{k:1}}$$

$$\Rightarrow (s', a') \sim d_{t+1}^{\pi_{k:1}}$$

$$RHS = \mathbb{E}_{d_0^\pi}\left[f_k - \underline{\mathcal{T}^{\pi_1} f_{k-1}}\right]$$

$$+ \gamma \cdot \mathbb{E}_{d_1^\pi}\left[f_{k-1} - \mathcal{T}^{\pi_2} f_{k-2}\right]$$

$$+ \gamma^2 \cdot \mathbb{E}_{d_2^\pi}\left[f_{k-2} - \mathcal{T}^{\pi_3} f_{k-3}\right]$$

$$+ \dots$$

$$+ \gamma^{k-1} \cdot \mathbb{E}_{d_{k-1}^\pi}\left[f_1 - \mathcal{T}^{\pi_k} f_0\right]$$

$$f_0 \equiv 0$$

$$= \mathbb{E}_{d_0^\pi}[f_k] - \gamma \cdot \mathbb{E}_{d_1^\pi}[f_{k-1}] - r_0$$

$$+ \gamma \cdot \mathbb{E}_{d_1^\pi}[f_{k-1}] - \gamma^2 \cdot \mathbb{E}_{d_2^\pi}[f_{k-2}] - \mathbb{E}[r_1]$$

$$+ \cdots$$

$$+ \gamma^{k-1} \cdot \mathbb{E}_{d_{k-1}^\pi}[f_1] - 0 - \mathbb{E}[r_{k-1}]\gamma^{k-1}$$

$$\rightarrow \mathbb{E}\left[\sum_{t=0}^{k-1} \gamma^t r_t\right]$$

$$= J_{f_k}(\pi_k) - J_k(\pi_{k:k}) \qquad \square$$

(Bellman Flow Equation)

Policy Evaluation $\rightarrow$ holdout validation & testing.

Policy Optimization $\rightarrow$ training

$\pi \rightarrow Q^\pi$ $(J(\pi))$

policy class $\pi \in \Pi$

A naive optimizer: $\arg\max_{\pi \in \Pi} J_{\hat{f}^\pi}(\pi) \triangleq \hat{\pi}$,

$$\mathbb{E}_{s \sim d_0}[\hat{f}^\pi(s, \pi)]$$

(information-theoretically)

$(\hat{f}^\pi \approx Q^\pi)$

Guarantee of BRM $(\hat{f}^\pi)$:

w.p. $\geq 1-\delta$: $J(\pi) - J_{\hat{f}^\pi}(\pi) \lesssim \frac{V_{max}}{1-\gamma} \sqrt{\frac{C_\pi}{n} \log \frac{|\mathcal{F}|}{\delta}}$.

Union Bound over Π

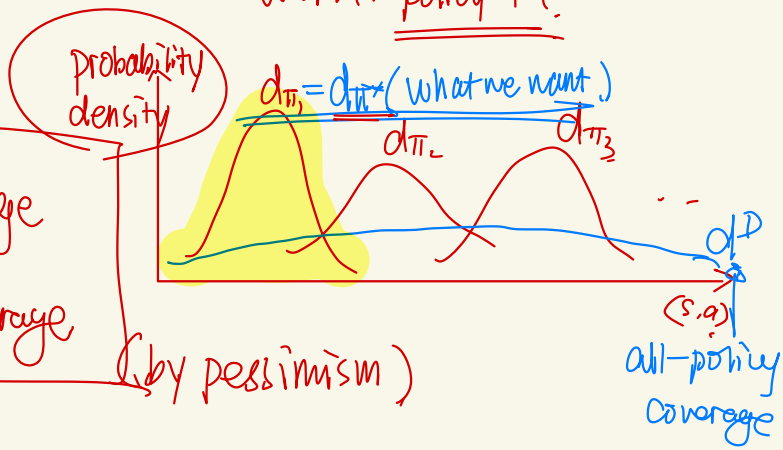
w.p. $\geq 1-\delta$, $\forall$ comparator policy $\pi_{cp} \in \Pi$

$$J(\pi_{cp}) - J(\hat{\pi}) \lesssim \frac{V_{max}}{1-\gamma} \cdot \sqrt{\frac{\max_{\pi\in\Pi} C_\pi}{n} \log \frac{|\mathcal{F}|\cdot|\Pi|}{\delta}}$$

All-Policy Coverage 全策略覆盖:

$\max_{\pi\in\Pi} C_\pi$: data need to be fully exploratory w.r.t. policy π.

$$C_\pi = \left\| \frac{d^\pi}{d^D} \right\|_\infty$$



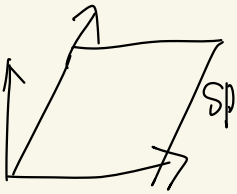
Δ All-policy coverage $\Rightarrow$ Single-policy coverage (by pessimism)

- $\max_{\pi\in\Pi} C_\pi$: ① Good case: low-rank MDP.

$$P(s'|s,a) = \phi(s,a)^T \psi(s')$$

$\Rightarrow d^\pi(s)$ can be written as some linear combination of $\psi_1, \psi_2, \cdots, \psi_d$.

$\Rightarrow d^\pi(s)$: low rank (rank-d)



span

$$\Pi = \{\pi_1, \cdots, \pi_d\} \rightarrow d^{\pi_i}, \; i \in [d]$$

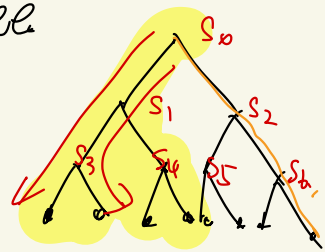
mixture $\Rightarrow d^D$

$$\Rightarrow \max_{\pi \in \Pi} \left\| \frac{d^\pi}{d^D} \right\|_\infty \leq d \cdot |A|$$

② Bad case : exponentially large (in H)

e.g. deterministic tree

policy $\Updownarrow$ path.



$d^D(s) \approx 0$

$d^\pi(s) > 0$

$$\Rightarrow C_\pi = \left\| \frac{d^\pi}{d^D} \right\|_\infty \rightarrow \infty$$

# of paths $|A|^H$.

$$\Rightarrow \max_{\pi \in \Pi} C_\pi \geq |\Pi| = |A|^H$$

Computationally efficient algorithm ?

Approximating Bellman operator !

☆ $\boxed{\text{Approximate Dynamic Programming}}$

π_{cp}

(Fitted-Q)

Bellman equation: $Q^\pi = \mathcal{T}^\pi Q^\pi$.

Bellman optimality equation: $Q^* = \mathcal{T} Q^*$.

Bellman operator ($\mathcal{B}$)

(optimality)

$\forall f \in \mathbb{R}^{S \times A}$: $(\mathcal{T} f)(s,a)$

$$= R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(s,a)}\left[\max_{a' \in A} f(s',a')\right]$$

$$\Longrightarrow \pi^*(s) = \arg\max_{a \in A} Q^*(s,a).$$

optimal return in MDP for all states simultaneously

$\Rightarrow$ FQI (Fitted-Q Iteration)

[Ernst et al. 2005]

$$f_{k+1} \leftarrow \arg\min_{f \in \mathcal{F}} \hat{\mathcal{L}}(f; f_k, *)$$

$\mathcal{F} = \{ f_\theta : \theta \in \Theta \}$

$\Pi = \{ \pi_f : f \in \mathcal{F} \}$

$$\hat{\mathcal{L}}(f; f_k, *) = \frac{1}{|D|} \sum_{(s,a,r,s')} \left(f(s,a) - r - \gamma \cdot \max_{a' \in A} f_k(s',a') \right)^2$$

induced policy class

(semi-SGD)

$$\theta \leftarrow \theta - \alpha \cdot \Big(f_\theta(s,a) - \overbrace{r - \gamma \cdot \max_{a' \in A} f_\theta(s',a')}^{\text{fixed}} \Big) \cdot \nabla_\theta f_\theta(s,a)$$

- f_θ: tabular

$$f(s,a) \leftarrow f(s,a) - \underset{\substack{\uparrow \\ \text{step size.}}}{\alpha} \cdot \underbrace{\left(f(s,a) - r - \gamma \cdot \max_{a' \in A} f(s',a') \right)}_{\text{TD error for } (s,a,r,s')}$$

Q-learning ☆

- f_θ: neural nets $\rightarrow$ (almost) DQN [Mnih et al, 2015].
  - target network θ^- : $\theta^- \leftarrow \theta$.
  - experience replay: replay buffer

    $(s,a,r,s') \sim D$
  - optimizer: SGD, Adam, ...

Approximate Dynamic Programming.

$\mathcal{T}^\pi$ (FQE $\rightarrow$ BRM $\rightarrow$

$\mathcal{T}$ (FQI (iterative): $f_{k+1} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{\mathcal{L}}(f; f_k, *)$

BRM (minimax): $\hat{f} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \max_{g \in \mathcal{F}} \left(\hat{\mathcal{L}}(f; f, *) - \hat{\mathcal{L}}(g; f, *) \right)$

Analysis: (error propagation)

$$J(\pi) - J_f(\pi) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{d^\pi}\left[f - \mathcal{T}^\pi f \right].$$

Lemma $f \in \mathbb{R}^{S \times A}$. $\pi, \pi' : S \to \Delta(A)$

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\Big(\underbrace{\mathbb{E}_{s\sim d^{\pi'}}\big[f(s,\pi') - f(s,\pi) \big]}_{\text{advantage}}$$

$$+ \underbrace{\mathbb{E}_{d^{\pi'}}\big[\mathcal{T}^{\pi} f - f \big]} + \mathbb{E}_{d^{\pi}}\big[f - \mathcal{T}^{\pi} f \big] \Big)$$

Bellman error (under different policies)

$\pi = \pi_{\hat{f}} \Rightarrow \mathcal{T}^{\pi}\hat{f} = \mathcal{T}\hat{f}$.

$\pi' = \pi^*$

$$\hat{f}(s,\pi') - \underbrace{\hat{f}(s,\pi_{\hat{f}})}_{\text{Greedy policy}} \le 0$$

$$\Rightarrow J(\pi^*) - J(\pi_{\hat{f}})$$

$$\le \frac{1}{1-\gamma}\Big(\mathbb{E}_{d^{\pi^*}}\big[\mathcal{T}f - f \big] + \mathbb{E}_{d^{\pi_{\hat{f}}}}\big[f - \mathcal{T}f \big] \Big)$$

- We can control: $\mathbb{E}_{d^D}\big[(f - \mathcal{T}f)^2 \big]$
- Coverage: $d^D \to d^{\pi}$ $\begin{cases} d^{\pi^*} & \checkmark \\ d^{\pi_{\hat{f}}} \in \Pi ? \end{cases}$

$\Rightarrow$ d^D cover all $\pi \in \Pi$ (all-policy coverage)

FQI ?

Analysis 1 : control $\|f_{k+1} - \mathcal{T} f_k\|_{2,d^D}$ $\quad \underset{\pi \in \Pi}{\operatorname{argmax}} \, J_{\hat{f}^\pi}(\pi)$

Analysis 2 :

Lemma

sequence of functions : $f_0, \cdots, f_k \in \mathbb{R}^{S \times A}$

sequence of (comparator) policy : $\pi_k, \cdots, \pi_1 = \pi$

non-stationary

$\hat{\pi} = \pi_{f_{k-0}}$

$$\Rightarrow J(\pi) - J(\hat{\pi}) \le \sum_{t=0}^{k-1} \gamma^t \Big(\mathbb{E}_{d_t^\pi}[\mathcal{T} f_{k-t} - f_{k-t+1}] + \mathbb{E}_{d_t^{\hat{\pi}}}[f_{k-t+1} - \mathcal{T} f_{k-t}] \Big)$$

approximation error

$J(\hat{\pi})$

$\gamma^k \cdot V_{max}$

Here: (for FQI) $\pi_k = \cdots = \pi_1 = \pi$.

	Fitted-Q	BRM
Evaluation	finite-horizon (non-stationary)	1 Bellman Error
Optimization		2 Bellman Error

Fitted-Q Iteration $\Leftarrow$ Value Iteration

Fitted Policy Iteration (FPI)

Policy Iteration $\pi_{k+1} \leftarrow \pi_{Q^{\pi_k}}$

policy evaluation policy improvement.

* Pessimistic version:

FPI $\rightarrow$ PSPI

FQI $\rightarrow$ PEVI.

Recall: Policy Optimization

Assumption 1: Bellman Completeness $\forall f \in \mathcal{F}.\ \mathcal{T} f \in \mathcal{F}.$ realizability

Assumption 2: All-policy Coverage $\forall \pi \in \Pi \quad \left\| \frac{d^\pi}{d^D} \right\|_\infty \le C$

$C = \max_{\pi \in \Pi} C_\pi$

single-policy coverage. (d^{π^*})

data are exploratory

Challenge: real-world data often lacks exploration!

- bad behavior $\Downarrow$ Over-estimation

Question: whether we can develop algorithms and guarantees that work for "arbitrary" offline dataset ?

All-policy coverage $\Rightarrow$ Single-policy coverage

☆ **Principle**: PFU (Pessimism in Face of Uncertainty)

in offline RL 悲观原则

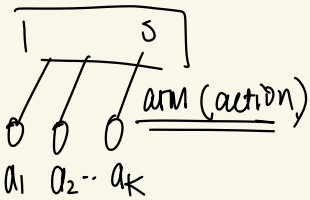
* Online RL: Optimism in Face of Uncertainty (OFU)

- exploration: leaving the current data distribution
- exploitation: staying within the current data distribution

optimism

pessimism

Uncertainty



MAB (Multi-Armed Bandit)

Goal: find the arm with highest mean reward.

Algorithm:

1) estimate the mean reward for each arm

2) pick the arm with highest estimated reward

$r_1 \quad r_2 \quad r_K$

Stochastic / adversarial

$r_i \sim \mu_i$

mean

uncertainty

Uncertainty Quantification

$$\mu(a) \in [R^-(a), R^+(a)] \quad \text{w.h.p}$$

confidence interval

LCB (lower confidence bound): pessimistic algorithm design.

selected arm: $\hat{a} = \arg\max_{a \in A} R^-(a)$

optimal arm: $a^* = \arg\max_{a \in A} \mu(a)$

Guarantee: $\mu(a^*) - \mu(\hat{a})$

$$= \mu(a^*) - R^-(a^*) + R^-(a^*) - \mu(\hat{a})$$

$$\leq \mu(a^*) - R^-(a^*) + \underbrace{R^-(\hat{a}) - \mu(\hat{a})}_{\leq 0}$$

$$\leq \mu(a^*) - R^-(a^*)$$

"single-policy" coverage

confidence radius of a^*. (uncertainty quantification) ☆

If $\hat{a}' = \arg\max_{a \in A} R^+(a)$ (optimistic).

$$\mu(a^*) - \mu(\hat{a}') \leq R^+(\hat{a}) - \mu(\hat{a})$$

unknown (depends on data)

$\Rightarrow$ all-policy coverage

Recall: Pessimism in Face of Uncertainty.

- Online RL: optimism → exploration
- Offline RL: pessimism → exploitation.

Algorithm: LCB (Lower Confidence Bound)

Key: uncertainty quantification

e.g. $\mu(a) \in [R^-(a), R^+(a)]$

MDP

↓

$J(\pi) \in [J^-(\pi), J^+(\pi)]$

$|J(\pi) - J_f(\pi)| \le eb(G_\pi)$

① Recall: (BRM Guarantee)

$$J^{\pm}_{BRM}(\pi) = J_{f^\pi}(\pi) \pm eb(G_\pi)$$

error bound: $eb(G_\pi) = O\left(\frac{V_{max}}{1-\gamma}\sqrt{\frac{G_\pi}{n}\cdot\log\frac{|\mathcal{F}|\cdot|\Pi|}{\delta}}\right)$ unknown

(under Bellman completeness)

LCB Algorithm: $\boxed{\hat{\pi} = \arg\max_{\pi\in\Pi} J^-_{BRM}(\pi)}$

For any comparator policy π_{cp}:

$\forall \pi: J(\pi) \ge J^-_{BRM}(\pi)$

$$J(\pi_{cp}) - J(\hat{\pi}) \le J^-_{BRM}(\pi_{cp}) + eb(G_{\pi_{cp}}) - J^-_{BRM}(\hat{\pi})$$

$$\le eb(G_{\pi_\varphi}) = O\left(\frac{V_{max}}{1-\gamma}\sqrt{\frac{G_{\pi_\varphi}}{n}\log\frac{|\mathcal{F}||\Pi|}{\delta}}\right)$$

Single-policy coverage!

In offline RL,

d^π: inaccessible

$\Rightarrow G_\pi$ unknown

$\Rightarrow$ intractable to optimize LCB (even information-theoretically)

② Some guarantee without knowing G_π:

BRM: $\hat{\mathcal{E}}(f;\pi) = \max_{g\in\mathcal{F}}\left(\hat{\mathcal{L}}(f;f,\pi) - \hat{\mathcal{L}}(g;f,\pi)\right)$.

$$\Rightarrow \quad \hat{\mathcal{E}}(f;\pi) \approx \mathcal{E}(f;\pi) = \|f - \mathcal{T}^\pi f\|^2_{2,d^\mu}.$$

(Goal: estimate Q^π, i.e. $\mathcal{E}(Q^\pi;\pi)=0$)

$\Rightarrow$ w.p. $\ge 1-\delta$:

$$Q^\pi \in \boxed{\mathcal{F}^\pi_{\varepsilon_0}} := \{ f\in\mathcal{F} : \hat{\mathcal{E}}(f;\pi) \le \varepsilon_0 \}.$$

Version-Space:

$$\varepsilon_0 = \frac{V_{max}^2}{n}\log\frac{|\mathcal{F}||\Pi|}{n}$$

set of hypothesis that are consistent with observed training data.

LCB: $\boxed{J_{VS}^{-}(\pi) = \min_{f \in \mathcal{F}_{\varepsilon_0}^{\pi}} J_f(\pi)} \leq J_{Q^\pi}(\pi) = J(\pi)$

$$\Rightarrow \hat{\pi} = \arg\max_{\pi \in \Pi} J_{VS}^{-}(\pi)$$

$\mathcal{E}(f;\pi) \leq 2\varepsilon_0$

For any comparator policy $\pi_{cp} \in \Pi$:

oracle $J(\pi_{cp}) - J(\hat{\pi}) \leq J_{VS}^{-}(\pi_{cp}) + eb'(\pi_{cp}) - J_{VS}^{-}(\hat{\pi})$

$$\leq eb'(\pi_{cp}) = \boxed{\frac{1}{1-\gamma}\sqrt{\frac{2C_{\pi_{cp}}}{n}\log\frac{|\mathcal{F}||\Pi|}{\delta}}}.$$

single-policy coverage!

Recall: version-space pessimism.

$$\mathcal{F}_{\varepsilon_0}^{\pi} = \{ f \in \mathcal{F} : \hat{\mathcal{E}}(f;\pi) \leq \varepsilon_0 \}$$

w.p. $\geq 1-\delta$: $Q^\pi \in \mathcal{F}_{\varepsilon_0}^{\pi}$.

$\min_{f \in \mathcal{F}_{\varepsilon_0}^{\pi}} J_f(\pi)$

$$\Rightarrow \text{LCB}: \quad \hat{\pi} = \arg\max_{\pi \in \Pi} J_{VS}^{-}(\pi)$$

Computation efficiency

Jointly efficient

statistical

computational

先和

Oracle efficiency: the computation is efficient if we are given black-box oracles for certain subroutines that

1) we can reasonably approximate in practice. → GD. SGD. Adam. Proximal.

2) the oracle itself can be efficiently implemented (w.o. FA) in special cases such as tabular or linear function class.

e.g. FQE/FQI $\quad f_{k+1} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \hat{\mathcal{L}}(f; f_k, \pi)$

least-square regression oracle

Pessimism: $\quad f^{\pi}_{\min} = \underset{f \in \mathcal{F}^{\pi}_{\varepsilon_0}}{\operatorname{argmin}}\ \boxed{J_f(\pi)}$

$\mathcal{F}^{\pi}_{\varepsilon_0}$ → $\hat{\mathcal{E}}(f;\pi) \le \varepsilon_0$

Constrained Optimization

↳ Lagrangian: $\min_{f \in \mathcal{F}} J_f(\pi) + \lambda \cdot \hat{\mathcal{E}}(f;\pi)$

In linear function approximation, $\mathcal{F} = \{ f \in \mathcal{F} : f = \phi^T \theta \}$

$J_f(\pi)$: linear in ϕ

$\hat{\mathcal{E}}(f;\pi)$: quadratic in ϕ with PSD Hessian.

LSTDQ

⇒ Oracle: quadratic programming

Bellman-consistent Pessimism for Offline Reinforcement Learning.

[Xie et al. 2021]

PSPI (Pessimistic Soft Policy Iteration)

policy evaluation
policy improvement

1. Initialize π_1 as uniformly random over A.

2. For $k = 1, 2, \cdots, K$:

   a) Evaluation: compute $f_k = f_{min}^{\pi_k}$ using the oracle.

   b) Update: $\pi_{k+1}(a|s) \propto \pi_k(a|s) \cdot \exp(\eta \cdot f_k(s,a))$

3. Output uniform mixture of $\pi_1, \cdots, \pi_K$, $\mathrm{Unif}[\pi_{1:K}]$

Last-Iterate Convergence

Remarks:

1) Uniform mixture: trajectory-level mixture.

- randomly sample $i \in [K]$
- using i to generate whole trajectory τ

(s, a, r, s')

$\pi_1, \cdots, \pi_K$

Markov

→ non-Markov. history dependent.

Improper Learning

$h \in H \Rightarrow \hat{h} \notin H$

(e.g. ensemble learning)

common in online learning.

$$J(\mathrm{Unif}[\pi_{1:K}]) = \frac{1}{K}\sum_{k=1}^{K} J(\pi_k)$$

2) Computational efficiency : large state space

$$\pi_{k+1}(a|s) \propto \pi_k(a|s)\cdot \exp(\eta \cdot f_k(s,a))$$

$\rightarrow$ $|S \times A|$ computation ?

$$f_{k+1} = f_{min}^{\pi_{k+1}}$$

evaluate π_{k+1} on data D.

Lazy Evaluation : we only need to update policies for any queried s (in data)

Δ Normalization :

$$\boxed{\pi_{k+1}(a|s) = \frac{\pi_k(a|s)\cdot \exp(\eta \cdot f_k(s,a))}{\sum_{a\in A} \pi_k(a|s)\cdot \exp(\eta \cdot f_k(s,a))}}$$

$\downarrow$

linear in $|A|$

Softmax policy

$\downarrow$

Stochastic policies

Δ ε-Greedy

$(1-\varepsilon) : \frac{\varepsilon}{|A|} : \frac{\varepsilon}{|A|} : \cdots$

☆ Natural Policy Gradient.

(mirror descent in simplex. Hedge. MWU)

3) Continuous action space. ✗ ← policy gradient.

$\Rightarrow$ standalone policy class Π ✗

In PSPI, implicit policy class.

$$\Pi = \left\{ \pi(\cdot|s) \propto \exp\left(\eta \cdot \sum_{i=1}^{k} f_i(s,\cdot)\right) : k \in [K],\ f_{1:k} \in \mathcal{F} \right\}$$

Recall: (PSPI)

$$\begin{cases} f_k = f_{\min}^{\pi_k} = \underset{f \in \mathcal{F}_\varepsilon^{\pi_k}}{\operatorname{argmin}}\ J_f(\pi_k) \\ \pi_{k+1}(a|s) \propto \pi_k(a|s) \cdot \exp(\eta f_k(s,a)) \end{cases}$$

Lemma For any $f \in \mathbb{R}^{S \times A}$, $\pi, \pi' : S \to \Delta(A)$.

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\left(\mathbb{E}_{s \sim d^{\pi'}}\left[f(s,\pi') - f(s,\pi) \right] + \mathbb{E}_{d^{\pi'}}\left[\mathcal{T}^{\pi} f - f \right] + \mathbb{E}_{d^{\pi}}\left[f - \mathcal{T}^{\pi} f \right] \right)$$

$$J(\pi_{cp}) - J(\mathrm{Unif}[\pi_{1:k}]) = \frac{1}{K} \sum_{k=1}^{K} \left(J(\pi_{cp}) - J(\pi_k) \right)$$

π' π

$$J(\pi_{cp}) - J(\pi_k) = \frac{1}{1-\gamma}\Big(\mathbb{E}_{s \sim d^{\pi_{cp}}}\left[f_k(s,\pi_{cp}) - f_k(s,\pi_k) \right] \quad (\mathrm{I})$$

$$+ \mathbb{E}_{d^{\pi_{cp}}}\left[\mathcal{T}^{\pi_k} f_k - f_k \right] \quad (\mathrm{II})$$

$$+ \mathbb{E}_{d^{\pi_k}}\left[f_k - \mathcal{T}^{\pi_k} f_k \right] \Big) \quad (\mathrm{III})$$

(III): $\mathbb{E}_{d^{\pi_k}}\left[f_k - \gamma^{\pi_k} f_k \right] = (1-\gamma)\left(J_{f_k}(\pi_k) - J(\pi_k) \right)$

pessimistic

By pessimism:

$$J_{f_k}(\pi_k) = J_{f_{min}^{\pi_k}}(\pi_k) = \min_{f \in \mathcal{F}_{\varepsilon_0}^{\pi_k}} J_f(\pi_k) \le J_{Q^{\pi_k}}(\pi_k) = J(\pi_k)$$

$\Rightarrow$ (III) ≤ 0

(II): $\mathbb{E}_{d^{\pi_{cp}}}\left[\gamma^{\pi_k} f_k - f_k \right]$ $\overset{\text{Single-policy coverage}}{\Longleftrightarrow}$ $\mathbb{E}_{d^{p}}\left[\left(f_k - \gamma^{\pi_k} f_k \right)^2 \right]$

e.g. $C_{\pi_{cp}}$, $C_{\pi_{cp}}^{sq}$, $C_{\pi_{cp}}^{avg.}$

we can control (via concentration + Bellman completeness)

one-sided error: $f_k = f_{min}^{\pi_k}$

$\widehat{\mathcal{E}}(f_k; \pi_k) \le \varepsilon_0$

A (I): $\mathbb{E}_{s \sim d^{\pi_{cp}, s}}\left[f_k(s, \pi_{cp}) - f_k(s, \pi_k) \right]$

is NOT a one-sided error!

benchmark

($\Rightarrow$ cannot use coverage to translate the error)

Ideally: $\pi_k = \pi_{f_k}$ (i.e. greedy to f_k)

[$f_k = f_{min}^{\pi_k}$: this breaks the causal order!].

✗ Obj 1: design π_k that maximize $f_k(s, \cdot)$

✓ Obj 2: design $\pi_{1:k}$ that maximize $\sum_{k=1}^{K} f_k(s, \cdot)$

sequence

At round $k = 1, 2, \cdots, K$:

- learner pick $p_k \in \Delta(N)$
- environment pick $l_k \in \mathbb{R}^N$
- learner suffers loss $\langle l_k, p_k \rangle$

Online Linear Optimization in the Simplex

(Expert Problem)

→ Goal: minimize regret

$$\text{Regret}_K = \sum_{k=1}^{K} \left(\langle l_k, p \rangle - \langle l_k, p_k \rangle \right)$$

benchmark

(best model in hindsight)

$p_k \leftrightarrow \pi_k(\cdot|s) \in \mathbb{R}^A$

$l_k \leftrightarrow f_k(s, \cdot) \in \mathbb{R}^A$.

per-state control

$\Downarrow$

$$\text{Regret}_K = \sum_{k=1}^{K} \left(f_k(s, \pi_{cp}) - f_k(s, \pi_k) \right)$$

$$\leq O\left(\sqrt{K \cdot \log|A|}\right)$$

$$\Rightarrow \frac{1}{K} \sum_{k=1}^{K} \left(f_k(s, \pi_{cp}) - f_k(s, \pi_k) \right)$$

$$\leq O\left(\sqrt{\frac{\log|A|}{K}}\right)$$

A general framework: OMD (online mirror descent)

/ FTRL (Follow-The-Regularized-Leader)

In Expert problem: Hedge / Multiplicative Weight Update (MWU)

$$p_{k+1,i} \propto p_{k,i} \exp(\eta \cdot l_{k,i})$$

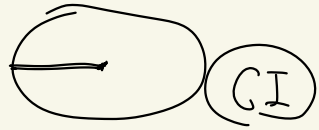
$$\Rightarrow \text{Regret}_K \leq O\left(\sqrt{K \cdot \log N}\right) \quad \text{when } \eta = \sqrt{\frac{\log N}{K}}$$

Is Pessimism Provably Efficient for Offline RL ?

[Jin et al., 2020]

PSPI: oracle $f^{\pi}_{min} = \min_{f \in \mathcal{F}^{\pi}_{\varepsilon_0}} J_f(\pi)$

$\mathcal{D}^{\pi} \in \mathcal{F}^{\pi}_{\varepsilon_0}$ why.



pointwise uncertainty quantification

- Regression: $(s,a) \mapsto r + \gamma \cdot V(s')$ for any $V \in \mathbb{R}^S$.

f^*: Bayes-optimal predictor

$\hat{f} \in \mathcal{F}$: learned predictor

bonus term

$\Rightarrow f^*(s,a) \in [\hat{f}(s,a) - b(s,a), \hat{f}(s,a) + b(s,a)]$

with high probability

optim.: GOLF

pessim.: PEVI

PEVI (PEssimistic Value Iteration)

FQI: $(s,a) \mapsto r + \gamma \cdot \max_{a'} f_{k-1}(s',a')$

for $k = 1, \cdots, K$:

1. fit $(s,a) \mapsto r + \gamma \max_{a'} f^{-}_{k-1}(s',a')$ on $\mathcal{D}$. using the oracle.

   $\rightarrow$ output: $\hat{f}_k \in \mathcal{F}$. b_k

2. pessimistic value function: $f^{-}_k(s',a') = \hat{f}_k(s,a) - b_k(s,a)$

Output: non-stationary policy $\pi_{k:1} = \{\pi_k, \cdots, \pi_1\}$

π_k is greedy w.r.t. f^{-}_k. $(\pi_k = \pi_{f^{-}_k})$

Analysis: **Pessimism**: $f_k^-(s,a) \le Q_k^{\pi_{k-1:1}}$

Proof by induction: $\mathcal{T}^\pi$ is monotone

$\forall f \le f'$ (pointwise) π: $\mathcal{T}^\pi f \le \mathcal{T}^\pi f'$.

If $f_{k-1}^- \le Q_{k-1}^{\pi_{k-2:1}}$ (base case: $f_0^- \equiv Q_0 \equiv 0$)

$$\Rightarrow f_k^- \le f_k^* = \mathcal{T} f_{k-1}^-$$

pessimism

$$= \mathcal{T}^{\pi_{k-1}} f_{k-1}^- \quad (\pi_{k-1} \text{ is greedy wrt } f_{k-1}^-)$$

$$\le \mathcal{T}^{\pi_{k-1}} Q_{k-1}^{\pi_{k-2:1}} \quad (\text{induction hypothesis})$$

$$= Q_k^{\pi_{k-1:1}} \quad (\text{Bellman equation in finite-horizon MDP})$$

$\square$

Bellman error telescoping:

Greedy

$$J_k(\pi_{cp}) - J_k(\pi_{k:1}) \le \sum_{t=0}^{k-1} \gamma^t \cdot \mathbb{E}_{d_t^{\pi_{cp}}}\left[\mathcal{T} f_{k-t-1}^- - f_{k-t}^-\right]$$

$$+ \sum_{t=0}^{k-1} \gamma^t \cdot \mathbb{E}_{d_t^{\pi_{k:1}}}\left[f_{k-t}^- - \mathcal{T} f_{k-t-1}^-\right]$$

≤ 0.

Recall: $f_t^* \in [\hat{f}_t - b, \hat{f}_t + b]$ w.h.p. (pointwise)

$$\mathcal{T} f^-_{k-t-1} = f^*_{k-t} \le \hat{f}_{k-t} + b \le f^-_{k-t} + 2b.$$

$\Rightarrow$ Guarantee of PEVI : $J_K(\pi_\varphi) - J_K(\pi_{k:1})$

$$\le 2 \cdot \sum_{t=0}^{k-1} \gamma^t \cdot \mathbb{E}_{(s,a)\sim d_t^{\pi_\varphi}} \left[b(s,a) \right]$$

bonus term

Note : Bellman Completeness $\left(\forall f \in \mathcal{F} : \mathcal{T} f \in \mathcal{F} \right)$ is sufficient !

$$\hat{f} \in \mathcal{F} \Rightarrow f^- = \hat{f} - b \in \mathcal{F} \ ?$$

Not necessary !

$\Rightarrow$ More strong assumptions : $\forall f \in \mathbb{R}^{S \times A}$: $\mathcal{T} f \in \mathcal{F}$.

Not reasonable !

Linear setting

$\mathcal{F}$ is a linear class induced by feature map ϕ.

$$\mathcal{F}_\phi = \{ \langle \phi, \theta \rangle : \theta \in \mathbb{R}^d \}$$

$$\forall f \in \mathcal{F}_\phi : \exists \theta \in \mathbb{R}^d \text{ s.t. } f(s,a) = \phi(s,a)^T \theta.$$

Linear (Ridge) Regression $\rightarrow \hat{f}_k$.

$(s,a) \mapsto r + \gamma \cdot U(s')$.

$$b(s,a) = \frac{\beta}{\sqrt{n}} \cdot \sqrt{\phi(s,a)^T \left(\Sigma_D^{ridge}\right)^{-1} \phi(s,a)}$$

$$\beta = O\left(dH \cdot \sqrt{\log \frac{dHK}{\delta}}\right)$$

$$\Sigma_D^{ridge} = \frac{1}{n}\left(\sum_{(s,a)\in D} \phi(s,a)\phi(s,a)^T + I\right)$$

Proposition. If $\mathcal{F} = \mathcal{F}_\phi$. and $\mathcal{T}f \in \mathcal{F}.\ \forall f \in \mathbb{R}^{S\times A}$.

then the MDP must be a linear MDP.

(with feature ϕ).

$$J_K(\pi_\varphi) - J_K(\pi_{K:1}) \le \frac{2\beta}{(1-\gamma)\sqrt{n}} \cdot C_{PEVI}.$$

$$C_{PEVI} = (1-\gamma) \cdot \sum_{t=0}^{K-1} \gamma^t \cdot \mathbb{E}_{(s,a)\sim d_t^{\pi_\varphi}}\left[\sqrt{\phi(s,a)\left(\Sigma_D^{ridge}\right)^{-1}\phi(s,a)^T}\right]$$

Coverage C_π^{sq}. C_π^{avg} ✗

$$\left(C_{PEVI}\right)^2 \approx \mathbb{E}_{(s,a)\sim d^{\pi_\varphi}}\left[\underbrace{\phi(s,a)}_{A}\left(\Sigma_D^{ridge}\right)^{-1}\phi(s,a)^T\right]$$

Model-Based Algorithms

$\sigma_{max}\left(\Sigma_\pi^{1/2}\Sigma_D^{-1}\Sigma_\pi^{1/2}\right) = \left(C_\pi^{sq}\right)^2$

BRM. FQI. Pessimistic $\rightarrow$ "model-free" $\left\langle \begin{array}{l} \text{value-based} \\ \text{policy-based} \end{array} \right.$

Model-based: explicitly learning and using the dynamics of MDP.

$\text{MDP}(S, A, P, R, \gamma, d_0)$

e.g. Certainty Equivalence

Core: estimate P (assume known R)

$\boxed{\text{Given a class of candidate dynamics } \mathcal{P}.}$

$\forall \tilde{P} \in \mathcal{P}.$

MLE: $\log\text{-loss}_D(\tilde{P}) = \frac{1}{|D|} \cdot \sum_{(s,a,s')\in D} -\log \tilde{P}(s'|s,a)$

$\hat{P}$: minimizer of log-loss.

$$\Rightarrow \mathbb{E}_{(s,a)\sim d^n}\left[\| \hat{P}(\cdot|s,a) - P(\cdot|s,a) \|_1^2\right] \lesssim \frac{\log |\mathcal{P}|/\delta}{n}.$$

estimation error

$$\Rightarrow Q^{\pi}_{\hat{p}} \quad (\hat{f})$$

Bellman error: $\left| Q^{\pi}_{\hat{p}}(s,a) - \mathcal{T}^{\pi} Q^{\pi}_{\hat{p}}(s,a) \right|$

$$= \left| (\mathcal{T}^{\pi}_{\hat{p}} Q^{\pi}_{\hat{p}})(s,a) - (\mathcal{T}^{\pi} Q^{\pi}_{\hat{p}})(s,a) \right|$$

$$= \Big| R(s,a) + \gamma \cdot \langle \hat{p}(s,a), V^{\pi}_{\hat{p}} \rangle - R(s,a) - \gamma \cdot \langle p(s,a), V^{\pi}_{\hat{p}} \rangle \Big|$$

$$= \gamma \cdot \left| \langle \hat{p}(s,a) - p(s,a), V^{\pi}_{\hat{p}} \rangle \right|$$

Hölder

$$\leq \gamma \cdot \| \hat{p}(s,a) - p(s,a) \|_1 \cdot V_{max}$$

$$\mathcal{E}(Q^{\pi}_{\hat{p}}; \pi) = \mathbb{E}_{d^D}\left[(Q^{\pi}_{\hat{p}} - \mathcal{T}^{\pi} Q^{\pi}_{\hat{p}})^2 \right] \lesssim V_{max} \cdot \frac{\log \frac{|P|}{\delta}}{n}$$

Pessimism via version space:

$$\left| \{ f \in F : \hat{\mathcal{E}}(f;\pi) \leq \varepsilon_0 \} \right|$$

BRM

$$\hat{\mathcal{L}}(f; f, \pi) - \min_{g \in F} \hat{\mathcal{L}}(g; f, \pi))$$

Model version space:

$$\mathcal{P}_D = \{ \tilde{p} \in \mathcal{P} : \text{log-loss}_D(\tilde{p}) \leq \text{log-loss}_D(\hat{p}) + \varepsilon_0 \}$$

$\Rightarrow$ Model-based pessimistic estimator:

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \min_{\tilde{p} \in P_D} J_{\tilde{p}}(\pi).$$

Robust MDP (RMDP) [Wiesemann et al. 2012]

distributional robustness

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \min_{M \in \mathcal{M}} J_M(\pi)$$

worst-case return

uncertainty set $\mathcal{M}$

$\Updownarrow$

$$\min_{M \in \mathcal{M}} J_M(\pi_{cp})$$

Fact: true model dynamics are unknown

data-driven estimation has uncertainty

Note: in PSPI, $f^{\pi}_{min} = \min_{f \in \mathcal{F}_{\varepsilon_0}} J_f(\pi)$

in RMDP: assume oracle $M(\pi) = \arg\min_{M \in \mathcal{M}} J_M(\pi)$

π^*_{rb}: optimal robust policy

$\hat{\pi}$: PSPI output. ($\mathrm{Unif}[\pi_{1:K}]$)

$$\Rightarrow J_{M(\pi^*_{rb})}(\pi^*_{rb}) - J_{M(\hat{\pi})}(\hat{\pi})$$

$M(\pi^*_{rb}) = \arg\min_{M \in \mathcal{M}} J_M(\pi^*_{rb})$

$$= J_{M(\pi_{rb}^*)}(\pi_{rb}^*) - \frac{1}{K}\sum_{k=1}^{K} J_{M(\pi_k)}(\pi_k)$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\left(J_{M(\pi_k)}(\pi_{rb}^*) - J_{M(\pi_k)}(\pi_k)\right)$$

Performance Difference

$$= \frac{1}{K(1-\gamma)}\sum_{k=1}^{K} \mathbb{E}_{d^{\pi_{rb}^*}_{M(\pi_k)}}\left[Q^{\pi_k}_{M(\pi_k)}(s,\pi_{rb}^*) - Q^{\pi_k}_{M(\pi_k)}(s,\pi_k)\right]$$

$$\boxed{\mathbb{E}_{d^{\pi_{cp}}}\left[f(s,\pi_{cp}) - f(s,\pi_k)\right]}$$

[If all $M \in \mathcal{M}$ share the same transition dynamics.

$d^{\pi_k}_{M(\pi_k)}$ will be same for all k. $\Rightarrow$ PSPI analysis]

Uncertainty Set $M \in \mathcal{M}$?

$$M^\circ = (S, A, P^\circ, R, \gamma, d_0)$$

true dynamics

$\Downarrow$

$P \in \mathcal{P}$ (candidate dynamics)

(s,a)-rectangularity condition

$$\mathcal{P} = \bigotimes_{(s,a)\in S\times A} \mathcal{P}_{s,a}$$

distance measure.

$$\mathcal{P}_{s,a} = \{ P_{s,a} \in \Delta(S) : D(P_{s,a}, P^\circ_{s,a}) \leq \rho_{s,a} \}$$

$\mathcal{P}(\mathcal{M})$: the set of all models around $P^\circ(\mathcal{M}^\circ)$ within a particular distance.

estimated from data.

e.g. TV : $D_{TV}(P, P') = \frac{1}{2}\|P - P'\|_1$

Wasserstein : $D_W(P, P') = \inf_{\nu \in M(P, P')} \int \ell(x, y)\, d\nu(dx, dy)$

KL : $D_{KL}(P, P') = \sum_{s'} P(s') \cdot \log \frac{P(s')}{P'(s')}$

Chi-squared : $D_C(P, P') = \sum_{s'} \frac{(P(s') - P'(s'))^2}{P'(s')}$

Robust value function : $V_{\mathcal{M}}^{\pi} = \inf_{M \in \mathcal{M}} V_M^{\pi}$

$V_{\mathcal{M}}^{*} = \sup_{\pi} V_{\mathcal{M}}^{\pi}$

Robust optimal policy : $\pi_{\mathcal{M}}^{*}$

Robust Bellman operator : $\mathcal{T}_{\mathcal{M}}$

$$(\mathcal{T}_{\mathcal{M}} Q)(s, a) = R(s, a) + \gamma \cdot \inf_{M_{s,a} \in \mathcal{M}_{s,a}} \mathbb{E}_{s'}\left[\max_{a'} Q(s', a')\right]$$

$s' \sim P_{s,a}$

e.g. Robust Q-Iterations :

$$f_{k+1} \leftarrow \mathcal{T}_{\mathcal{M}} f_k \qquad (: \; f_k \to Q_{\mathcal{M}}^{*})$$

Recall: 2 core assumptions in offline RL

BRM

Fitted-Q

Function Approx: $\boxed{\text{Bellman completeness}}$: $\forall f \in \mathcal{F}.\ \mathcal{T}^{\pi} f \in \mathcal{F}$.

Data: Coverage condition: $\forall \pi \in \Pi.\ \left\| \frac{d^{\pi}}{d^{D}} \right\|_{\infty} \le C$.

policy optimization: all-policy coverage $\longrightarrow$ single-policy coverage.

$Q^{\pi} \in \mathcal{F}$: realizability

Learning under realizability

Value Function Selection $\longleftarrow$ Model-Selection

Input: $\boxed{\mathcal{F} = \{f_1, f_2, \cdots, f_m\}}$. $f_i \in \mathbb{R}^{S \times A}$.

Output: f_i such that $f_i \approx Q^{\pi}$

Highly non-trivial!

$m = 2$.

one of f_1, f_2, is exactly Q^{π} $\times$

BRM: $\hat{f}^{\pi} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\max_{g \in \mathcal{G}} \hat{\mathcal{L}}(f; f, \pi) - \hat{\mathcal{L}}(g; f, \pi)}_{\hat{\mathcal{E}}(f; \pi)}$

$\hat{\mathcal{L}}(\mathcal{T}^{\pi} f; f, \pi)$

$\mathcal{G}$ is to realize $\{\mathcal{T}^\pi f_i\}_{i=1}^m$.

Double Sampling

$d^D(s,a) > 0$.

FQE / FQI :

$$f_k \leftarrow \underset{f \in \mathcal{F}}{\arg\min}\ \hat{\mathcal{L}}(f; f_{k-1}, \pi)$$

hyperparameter tuning

oracle (least-square regression)

e.g. $\mathcal{F} = \{\mu(W^T\theta) : \theta \in \mathbb{R}^{S \times A_2}\}$

$f_k \in \mathcal{F}$. $f_k - \eta \cdot g \notin \mathcal{F}$. $\rightarrow$ projection

---

Realizability $\mathcal{F} = \{f_1, \cdots, f_m\}$: one of $f_i \approx Q^\pi$.

View 1 : learning view $\mathcal{F}$: large, structured Q^π, Q^*

View 2 : selection view $\mathcal{F}$: small, unstructured

Learning under realizability ?

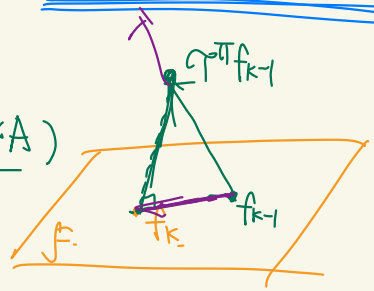
Divergence of FQE :

$$f_k \leftarrow \underset{f \in \mathcal{F}}{\arg\min}\ \hat{\mathcal{L}}(f; f_{k-1}, \pi)$$

$$f_k \leftarrow \mathrm{Proj}_{\mathcal{F}}\, \mathcal{T}^{\pi} f_{k-1}$$

(infinite data
$\Rightarrow$ d^D on support on $S \times A$)

$$\| f - \mathcal{T}^{\pi} f_{k-1} \|_{2, d^D}$$



Without projection:

$$f_k \leftarrow \mathcal{T}^{\pi} f_{k-1} \quad \text{(Value Iteration)}$$

$$\| f_k - Q^{\pi} \|_{\infty} = \| \mathcal{T}^{\pi} f_{k-1} - \mathcal{T}^{\pi} Q^{\pi} \|_{\infty}$$
$$\leq \gamma \cdot \| f_{k-1} - Q^{\pi} \|_{\infty}$$

Completeness:
$\forall f \in \mathcal{F}$: $\mathcal{T}^{\pi} f \in \mathcal{F}$

A natural idea: Projection is non-expansive!

非扩张性

$\forall f, f' \in \mathbb{R}^{S \times A}$:

$$\| \mathrm{Proj}_{\mathcal{F}} f - \mathrm{Proj}_{\mathcal{F}} f' \| \leq \| f - f' \|$$

$$f_k \leftarrow \mathrm{Proj}_{\mathcal{F}}\, \mathcal{T}^{\pi} f_{k-1}$$

non-expansive. contraction

e.g. $\mathcal{F}$ linear class

$$\| \cdot \| = \| \cdot \|_{2, d^D}$$

$\rightarrow$ $\mathrm{Proj}_{\mathcal{F}}$ is non-expansive.

Norm-mismatch!

$\mathcal{T}^\pi$ is contraction wrt. $\|\cdot\|_\infty$.

$\mathrm{Proj}_{\mathcal{F}}$ is non-expansive wrt. $\|\cdot\|_{2,d^D}$

for some distribution $\nu \in \Delta(S \times A)$:

$$\|\mathcal{T}^\pi f - \mathcal{T}^\pi f'\|_{2,\nu}^2$$

$$= \gamma^2 \cdot \mathbb{E}_{(s,a)\sim\nu}\left[\left(\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s',\pi) - f'(s',\pi)\right]\right)^2\right]$$

$$\leq \gamma^2 \cdot \mathbb{E}_{\substack{(s,a)\sim\nu \\ s'\sim P(\cdot|s,a) \\ a'\sim\pi(\cdot|s)}}\left[\left(f(s,a) - f'(s,a)\right)^2\right]$$

$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$

$$:= \gamma^2 \cdot \|f - f'\|_{2,\boxed{P(\nu)\times\pi} = \nu}^2$$

: ν is an invariant distribution of $P(\nu)\times\pi$.

Contraction:

$(s,a)\sim\nu \to d^D$

$s'\sim P(\cdot|s,a)$

$a'\sim\pi(\cdot|s')$

$\Rightarrow (s',a')\sim\nu \to d^D$

On-policy 同策略

off-policy evaluation: $\begin{cases} \text{target policy } \pi \to d^{\pi} \\ \text{behavior policy } \pi_b \to d^{D} \end{cases}$

FQE: $f_k \leftarrow \mathrm{Proj}_F \, T^{\pi} f_{k-1}$

Last time: Proj_F. T^{π} under $\|\cdot\|_{2,d^D}$

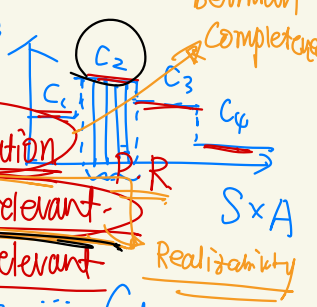
This time: Proj_F. T^{π} under $\|\cdot\|_{\infty}$

F: piecewise-constant function

$$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

feature ϕ: $S \times A \to \mathbb{R}^d$.

$\phi(s,a)$ is an one-hot vector



state-aggregation

bisimulation

Q^*-irrelevant

π^*-irrelevant

$C_1, C_2, \dots, C_d$

$\boxed{\text{State-abstraction}}$ 状态抽象

$$\|\mathrm{Proj}_F f - \mathrm{Proj}_F f'\|_{\infty} \le \|f - f'\|_{\infty}$$

Proof: $(\mathrm{Proj}_F f)(x) = \sum_{y \in C(x)} w(y) f(y)$ weighted average.

$$\forall x \in X : \left| (\mathrm{Proj}_F f)(x) - (\mathrm{Proj}_F f')(x) \right|$$

$$= \left| \sum_{y \in C(x)} w(y) \cdot (f(y) - f'(y)) \right|$$

$[0,1]$,

$$\leq \sum_{y \in C(x)} w(y) \cdot \left| f(y) - f'(y) \right|$$

$$\leq \max_{y \in C(x)} \left| f(y) - f'(y) \right| \leq \| f - f' \|_\infty \qquad \square,$$

Abstract MDP : $M_\phi = (S_\phi, A, P_\phi, R_\phi, \gamma)$

d

aggregated data distribution

$d_\phi^D(s,a) \propto d^D(s,a)$

"distribution-aware analysis"

$$P_\phi(s' \mid s, a) = \sum_{\substack{\tilde{s}, \tilde{a} : \\ \phi(\tilde{s}, \tilde{a}) = \phi(s,a)}} d_\phi^D(\tilde{s}, \tilde{a}) \cdot P(s' \mid \tilde{s}, \tilde{a})$$

original state space.

- $Q^\pi = Q^\pi_{M_\phi}$: M_ϕ can "realize" Q^π.
- $\mathrm{Proj}_F \mathcal{T}^\pi f = \mathcal{T}^\pi_{M_\phi} f$: projected Bellman update = Bellman update in the abstract space.

piecewise-constant

$\Rightarrow$ Error propagation analysis in M_ϕ

Fitted-Q in M $\Leftrightarrow$ VI in M_ϕ

Remark: Concentrability coefficient: $\left\| \frac{d^{\pi}_{M_\phi}}{d^D} \right\|_\infty \le C_\pi^\phi$

$\left\| \frac{d^{\pi}}{d^D} \right\|_\infty \le C_\pi$

Takeaway: learning under realizability is possible under state abstraction.

Now: selection view

$\mathcal{F} = \{f_1, \cdots, f_m\}$

piecewise-constant function class

trade-off

expressivity

structure

strong structure $\rightarrow$ unstructure

Value function selection $m=2$ $Q^\pi \in \{f_1, f_2\}$

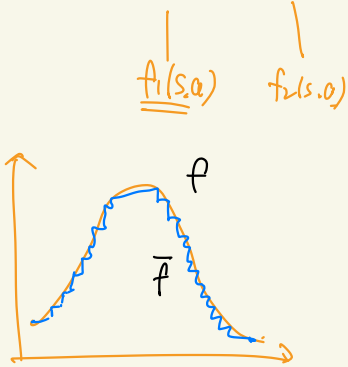
for each (s,a), group identifier "cell"

- $[0, V_{max}]$: discretization by ε

- $f_1(s,a)$, $f_2(s,a)$

$\downarrow$ ① discretization

$\bar{f}_1(s,a)$, $\bar{f}_2(s,a)$



if (s,a), (s',a'): $\left.\begin{array}{l}\bar{f}_1(s,a) = \bar{f}_1(s',a') \\ \bar{f}_2(s,a) = \bar{f}_2(s',a')\end{array}\right\} \Rightarrow$ ② aggregate (s,a) & (s',a')

$\Rightarrow$ construct a piecewise-constant function class $\bar{\mathcal{F}}$

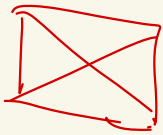
s.t. $\bar{f}_1, \bar{f}_2 \in \bar{\mathcal{F}}$; $Q^\pi \in_\varepsilon \bar{\mathcal{F}}$

approximation error : $O(\varepsilon)$

$$|\bar{\mathcal{F}}| = O\left(\left(\frac{V_{max}}{\varepsilon}\right)^2\right)$$ independent to $|S|$

BVFT (Batch Value-Function Tournament)

$|\mathcal{F}| = m$: tournament of pairwise comparisons



C.f. Batch Value-function approximation with only realizability. [Xie and Jiang, 2021]

Guarantee : ① $\log m \longleftrightarrow$ (learning) $\log |\mathcal{F}|$

② Assumption: $\forall \phi$: $\left\| \frac{d^{\pi}_{M_\phi}}{d^{D}} \right\|_\infty \le C_\pi^{\phi}$.

[Foster et al., 2021] $C_\pi = \left\| \frac{d^{\pi}}{d^{D}} \right\|_\infty$

Realizability $\rightarrow$ almost impossible to accurately estimate $J(\pi)$

Learning under Realizability

→ with state abstraction

BVFT

→ with LSTDQ

LSTD(λ)

(Least-Square Temporal Difference Q)

linear class $\mathcal{F}$ induced by some feature vector ϕ.

+ FQE (fixed point) = LSTDQ

$f_\theta(s,a) = \phi(s,a)^T \theta$

Assume $\boxed{Q^\pi(s,a) = \phi(s,a)^T \theta^* \in \mathcal{F}}$

Solve for θ^*

Bellman equation, $Q^\pi = \mathcal{T}^\pi Q^\pi$ ← Realizability

$$\phi(s,a)^T\theta^* = r + \gamma \cdot \phi(s',\pi)^T\theta^*$$

$$\Rightarrow \left(\phi(s,a) - \gamma\cdot\phi(s',\pi)\right)^T\theta^* = r(s,a)$$

$$\phi(s,a)\left(\phi(s,a) - \gamma\cdot\phi(s',\pi)\right)^T\theta^* = \phi(s,a)\cdot r$$

$(s,a,r,s') \sim D$

Least-Square : $\theta = A^{-1}B$

LSTDQ

$$A = \mathbb{E}_D\left[\phi(s,a)\cdot\left(\phi(s,a) - \gamma\cdot\phi(s',\pi)\right)^T\right]$$

$$= \boxed{\Sigma_D} - \gamma\cdot\mathbb{E}_D\left[\phi(s,a)\cdot\phi(s',\pi)\right]$$

$$B = \mathbb{E}_D\left[\phi(s,a)\cdot r\right]$$

empirical $\hat{A}$, $\hat{B}$. $\hat{\theta} = \hat{A}^{-1}\hat{B}$

θ is valid when A is invertible ($\sigma_{min}(A) > 0$)

BRM
MIS
PSPI.

linear → LSTDQ

$Q^\pi \in \mathcal{F}$: linear

$\mathcal{F}$: expressivity

Why learning under realizability is possible with LSTDQ ?

$\hat{A}, \hat{B} \xrightarrow{\text{concentration}} \|A-\hat{A}\|_2, \|B-\hat{B}\|_2$

(operator norm)

① $Q^\pi(s,a) = \phi(s,a)^T\theta^*$.

② $\sigma_{min}(A) > 0$.

BRM guarantee:

$$\frac{V_{max}}{1-\gamma}\sqrt{\frac{C_\pi}{n}\log\frac{|F|}{n}}$$

$C_\pi = \left\|\frac{d^\pi}{d^D}\right\|_\infty$

$C_\pi = \Sigma_D^{-\frac{1}{2}}\Sigma_\pi\Sigma_D^{-\frac{1}{2}}$

$$\Rightarrow \|\hat{\theta}-\theta^*\|_2 = \|\hat{\theta} - A^{-1}B\|_2$$

$$= \|A^{-1}A\hat{\theta} - A^{-1}B\|_2$$

$$\le \frac{1}{\sigma_{min}(A)}\|A\hat{\theta} - B\|_2$$

$$= \frac{1}{\sigma_{min}(A)}\cdot\|A\hat{\theta} - \hat{A}\hat{\theta} + \hat{B} - B\|_2$$

$$\le \frac{1}{\sigma_{min}(A)}\left(\|\hat{B}-B\|_2 + \|\hat{A}-A\|_2\|\hat{\theta}\|_2\right)$$

$F \Leftrightarrow \Theta$: bounded

Assumptions

① expressivity of function class $\checkmark$ ($Q^\pi \in F$)

② structure of function class $\checkmark$ (linear)

③ coverage of data

Here: $\sigma_{min}(A)$ is the "coverage" notion.

- on-policy: $\nu = P(\pi) \times \nu$ $\rightarrow \sigma_{min}(A) > 0$,

  $\rightarrow \phi(s,a) - \phi(s',\pi)$ marginal distribution

  $\Sigma_D = \mathbb{E}_D[\phi(s,a)\phi(s,a)^T]$ PD matrix

- off-policy: $\sigma_{min}(A)$

  $$A = \mathbb{E}_D[\phi(s,a)\cdot(\phi(s,a) - \gamma\cdot\phi(s',\pi))^T]$$

  ① A can be singular.

  ② $\sigma_{min}(A)$ is not scalar-invariant.

Marginalized Importance Sampling (MIS)

Recall: (Importance Sampling)

$$\mathbb{E}_{x\sim p}[f(x)] = \mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\cdot f(x)\right]$$

importance weight.

Q^π / Q^*

$\rightarrow \frac{d^\pi}{d^p}$ $\rightarrow p$, $\rightarrow q$

G_π

MDP

$$\mathbb{E}_{\pi}[r_t] = \mathbb{E}_{\pi_b}\left[\frac{p^{\pi}(s_1, a_1, \cdots, s_t, a_t)}{p^{\pi_b}(s_1, a_1, \cdots, s_t, a_t)} \cdot r_t\right]$$

$$\hookrightarrow \prod_{t'=1}^{t} \frac{\pi(a_{t'} \mid s_{t'})}{\pi_b(a_{t'} \mid s_{t'})}$$

Markovian

s_t, a_t $\quad s_1, a_1, \cdots, s_{t-1}, a_{t-1}$

$\downarrow \quad \downarrow$

$(x, y) \sim p. \quad (x, y) \sim q.$

$$p(x, y) = p(x) \cdot p(y \mid x)$$

$$\mathbb{E}_{(x,y)\sim p}[f(x)] = \mathbb{E}_{(x,y)\sim q}\left[\frac{p(x,y)}{q(x,y)} f(x)\right].$$

Marginalize y

$$\mathbb{E}_{(x,y)\sim q}\left[\frac{p_x(x)}{q_x(x)} f(x)\right]$$

occupancy

$$\mathbb{E}_{\pi}[r_t] = \mathbb{E}_{\pi_b}\left[\frac{d^{\pi}(s_t, a_t)}{d^{\pi_b}(s_t, a_t)} \cdot r_t\right]$$

$w(s,a)$

"Marginalized" Importance Weight

- Uncertainty Set: $w \in \mathcal{W} = \{ w \in \mathbb{R}^{S \times A} : \mathbb{E}_{d^D}[w] = 1, \; 0 \le w(s,a) \le B \}$

$$\mathbb{E}_{\pi}[r_t] = \mathbb{E}_{d^D}\left[\frac{d^{\pi}(s_t, a_t)}{d^{D}(s_t, a_t)} \cdot r_t\right]$$

$$\triangleq \mathbb{E}_{d^D}[w^{\pi}(s_t, a_t) \cdot r_t].$$

Assumption: (realizability) $w^{\pi} = \frac{d^{\pi}}{d^{D}} \in \mathcal{W}$

Recall: MIS

$\hookrightarrow$ realizability $\begin{cases} Q^\pi \in \mathcal{F} \\ w^\pi = \frac{d^\pi}{d^D} \in \mathcal{W} \end{cases}$

$$\left| J_f(\pi) - J(\pi) \right| = \frac{1}{1-\gamma} \cdot \left| \mathbb{E}_{d^\pi}\left[f - \mathcal{T}^\pi f \right] \right|$$

≥ 0 or ≤ 0.

$$= \frac{1}{1-\gamma} \cdot \left| \mathbb{E}_{d^D}\left[\frac{d^\pi}{d^D} \cdot (f - \mathcal{T}^\pi f) \right] \right|$$

$w^\pi \in \mathcal{W}$

$$= \frac{1}{1-\gamma} \cdot \left| \mathbb{E}_{d^D}\left[w^\pi \cdot (f - \mathcal{T}^\pi f) \right] \right|$$

$$\leq \max_{w \in \mathcal{W}} \frac{1}{1-\gamma} \cdot \left| \mathbb{E}_{d^D}\left[w \cdot (f - \mathcal{T}^\pi f) \right] \right|$$

worst-case

$$\triangleq \max_{w \in \mathcal{W}} L_q(w, f)$$

BRM:

$$\leq \| f - \mathcal{T}^\pi f \|_{1, d^\pi}$$

$$\leq \| f - \mathcal{T}^\pi f \|_{2, d^\pi}$$

Coverage

(non-negativity)

$$\leq \sqrt{C_\pi} \cdot \| f - \mathcal{T}^\pi f \|_{2, d^D}$$

can be controlled by concentration

$\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$

(double sampling problem)

$$L_q(w, f) = \frac{1}{1-\gamma} \cdot \left| \mathbb{E}_{(s,a) \sim d^D}\left[w(s,a) \cdot \left(f(s,a) - r - \gamma \cdot f(s', \pi) \right) \right] \right|$$

$r = R(s,a)$

$s' \sim P(\cdot | s, a)$

$$\hat{L}_q(w, f) = \frac{1}{1-\gamma} \cdot \frac{1}{N} \sum_{i=1}^{N} \left| w(s_i, a_i) \cdot f(s_i, a_i) - r_i - \gamma \cdot f(s_i', \pi) \right|$$

$N \to \infty$

Minimax Q Learning (MQL): $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \max_{w \in \mathcal{W}} \hat{L}_q(w, f)$

Why? Realizability: $Q^\pi \in \mathcal{F}$. $w^\pi = \frac{d^\pi}{d^D} \in \mathcal{W}$.

$$L_q(w, Q^\pi) = \frac{1}{1-\gamma} \left| \mathbb{E}\left[w \cdot (Q^\pi - \mathcal{T}^\pi Q^\pi) \right] \right| = 0. \quad \forall w \in \mathcal{W}$$

c.f. 1) Robust MDP $\quad \arg\max_{\pi \in \Pi} \min_{M \in \mathcal{M}} J_M(\pi)$

2) offline game $\begin{cases} f \text{ player} \\ w \text{ player} \end{cases}$

Thm $w^\pi \in \mathcal{W}$. $Q^\pi \in \mathcal{F}$. w.p. $\geq 1-\delta$:

$$\left| J_{\hat{f}}(\pi) - J(\pi) \right| \lesssim \frac{V_{max} \|w\|_\infty}{1-\gamma} \sqrt{\frac{1}{n} \cdot \log \frac{|\mathcal{F}| \cdot |\mathcal{W}|}{\delta}},$$

Coverage C_π

$$\|w\|_\infty = \max_{w \in \mathcal{W}} \max_{s,a} |w(s,a)|$$

$$C_\pi = \left\| \frac{d^\pi}{d^D} \right\|_\infty = \|w^\pi\|_\infty \leq \|w\|_\infty$$

$\sqrt{C_\pi}$

$C_\pi \; (C_\pi \approx \|w\|_\infty)$

$\mathbb{E}_{d^D}[w] = 1$

when $\mathcal{W}$ is importance weight class.

variance reduction $\mathbb{E}_{d^D}[w^2] \leq \|w\|_\infty^2$

↳ Bernstein's inequality

Recall: Minimax Q-function Learning (MQL)

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \max_{w \in \mathcal{W}} \hat{L}_q(w, f) \qquad (Q^\pi \in \mathcal{F},\ w^\pi \in \mathcal{W})$$

① $L_q(w, f) = \frac{1}{1-\gamma} \left| \mathbb{E}_{d^D}[\, w \cdot (f - \mathcal{T}^\pi f) \,] \right|$

→ convex function wrt. w.

⇒ only need $w^\pi \in \mathrm{conv}(\mathcal{W})$

② computational efficiency : minimax optimizer.

MIS. BRM lacks practice.

③ uncertainty quantification. → pessimism.

---

discriminator-generator.

MIS: (originally) $\mathcal{F}$ as the discriminator.

$$J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi}[\, R(s,a) \,]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d^D}[\, w^\pi(s,a) \cdot R(s,a) \,] := J_{w^\pi}(\pi)$$

$J(\pi) = \mathbb{E}_{s\sim d_0}[\, Q^\pi(s,\pi) \,]$

$J_f(\pi) = \mathbb{E}_{s\sim d_0}[\, f(s,\pi) \,]$ $\Longleftrightarrow$ $J_w(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d^D}[\, w \cdot r \,]$

telescoping lemma: $|J(\pi) - J_f(\pi)| = \frac{1}{1-\gamma} \left| \mathbb{E}_{d^\pi}[\, f - \mathcal{T}^\pi f \,] \right|$

"Bellman error"

$Q^\pi = \mathcal{T}^\pi Q^\pi$

Lemma $\forall w \in \mathbb{R}^{S \times A}$: $J(\pi) - J_w(\pi)$

"Bellman flow error"

$$= \mathbb{E}_{s \sim d_0}[Q^\pi(s,\pi)] + \frac{1}{1-\gamma} \cdot \mathbb{E}_{d^p}[w \cdot (\gamma \cdot Q^\pi(s',\pi) - Q^\pi(s,a))]$$

Proof (sketch):

$$J(\pi) - J_w(\pi) = \mathbb{E}_{s \sim d_0}[Q^\pi(s,\pi)] - \frac{1}{1-\gamma} \cdot \mathbb{E}_{d^p}[w(s,a) R(s,a)]$$

By Bellman equation: $R(s,a) = Q^\pi(s,a) - \gamma \cdot Q^\pi(s',\pi)$ □

Bellman flow equation (for distribution):

$$d^\pi(s,a) = (1-\gamma) \cdot d_0(s,\pi) + \gamma \cdot \sum_{(s',a')} d^\pi(s',a') \cdot P(s|s',a') \cdot \pi(a|s)$$

When $w = w^\pi$: RHS $= 0$. (satisfy the Bellman flow equation)

$$L_w(w,f) = \left| \mathbb{E}_{s \sim d_0}[f(s,\pi)] + \frac{1}{1-\gamma} \cdot \mathbb{E}_{d^p}[w(\gamma \cdot f(s',\pi) - f(s,a))] \right|$$

↳ (Dual Algorithm) Minimax Weight Learning (MWL)

MWL: $\hat{f} = \arg\min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} \hat{L}_w(w,f)$

MQL: $\hat{f} = \arg\min_{f \in \mathcal{F}} \max_{w \in \mathcal{W}} \hat{L}_q(w,f)$

duality

c.f. Minimax Weight and Q-function learning for Off-policy evaluation [Uehara, 2020]

Policy Optimization

Q^π, W^π

$\downarrow$

Q^*, W^*

oracle

↳ Linear Programming

↳ saddle point : $\left(V^*(s) = Q^*(s, \pi^*),\ d^{\pi^*} \right)$

$\Downarrow$

optimal policy π^*

$\frac{d^{\pi^*}}{d^D} \in W$, $V^* \in V$

→ MIS

Summary of Offline RL (this series)

[Levine et al. 2020] Offline RL: Tutorial, Review .... (empirical)

OPE

(Off-policy Evaluation)

Model-Selection

1. Importance Sampling → WIS, DR, MIS, MQL, MWL

↳ Policy Gradient

↓

Curse of Horizon

Policy Optimization

2. Value-Function Estimation < Fitted-Q, BRM

Fitted-Q → BVFT, LSTDQ

Assumptions

- function approximation : Bellman Completeness → Realizability
- data : coverage < All-policy coverage, Single-policy coverage

pessimism (LCB) → PSPI, PEVI

"coverability" $\inf_{d^b} \max_{\pi \in \Pi} C_\pi$

Data: "coverage"

① with Online RL

online RL : exploration ↔ optimism

offline RL : exploitation ↔ pessimism

Technique: uncertainty quantification

[The role of coverage in online RL, ICLR'23]

hybrid RL : online + offline

imitation learning (IL)

② with Deep Learning Theory

Deep RL / RLHF = Neural Net + RL

Nash Equilibrium

Complexity Measure: $\log |\mathcal{F}|$ → covering number.

FQE / FQI

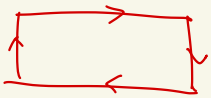
$\mathcal{I}(f'; f, \pi)$

$(s,a) \mapsto r + \gamma \cdot f(s', \pi)$

③ with Multi-Agent RL (MARL)

learner

"adversarial"



RL

Game Theory: equilibrium, PPAD complexity, replay

Two-Player Zero-Sum Markov Games

[When does offline TPZSMG solvable?]

cover (μ^*, ν^*) × → unilateral coverage

$\forall \mu$: (μ, ν^*)

$\forall \nu$: (μ^*, ν)

④ with Partial Observability

"non-Markov"

MDP → POMDP. (history-dependent)

PSR (Predictive State Repretation)

IS. ✓

belief coverage

outcome coverage

latent variable → "confounded" data

Equilibrium Gap:

$$\mathrm{Gap}(\mu, \nu) = \max_{\mu^\dagger} J(\mu^\dagger, \nu) - \min_{\nu^\dagger} J(\mu, \nu^\dagger)$$

$$\le \max_{\mu^\dagger} J^+_{VS}(\mu^\dagger, \nu) \le \varepsilon$$

$$- \min_{\nu^\dagger} J^-_{VS}(\mu, \nu^\dagger)$$