CS292F StatRL Lecture 16 Offline RL with function approximation

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

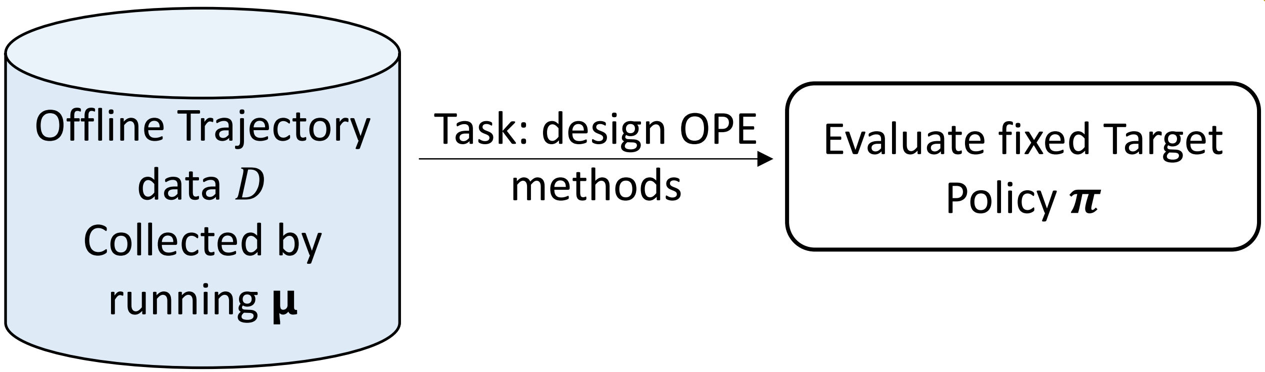
Plan for the remaining sessions

- Two lectures next week
  - Q&A session for me to help you guys with your HWs and projects
  - Will not be recorded

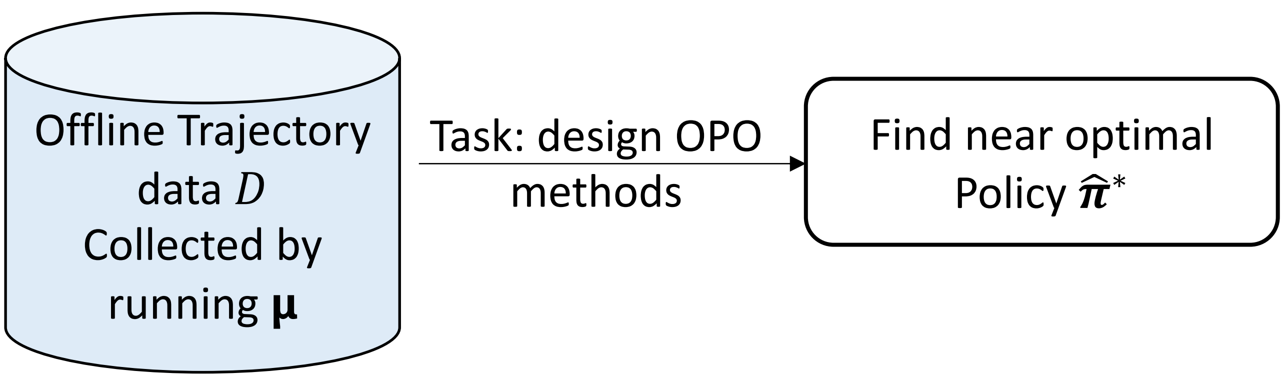
- Jun 2 lecture will be made into a **4-hour-long mini-symposium** of project presentations.

Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)



- Task 2: Offline Policy Learning. (OPL)



Recap: Offline Policy Learning

- Model-based approach
  - ERM to maximize the model-based OPE
  - Possible extensions: Optimism and Pessimism

ERM: $\arg\max_{\pi} \hat{V}^{\pi}$

$\arg\max_{\pi} \hat{V}^{\pi} + b^{\pi}$ $\qquad$ $\arg\max_{\pi} \hat{V}^{\pi} - b^{\pi}$

- Model-free approaches
  - Variance-Reduced Value Iteration
  - Fitted Q-Learning. (We will talk about this today!)

(Yin, Bai, W.) '21

In the tabular setting, the problem is (almost) completely solved.

$\boxed{(1-\gamma)^{-1}}$

$N = n \cdot (1-\gamma)^{-1}$

Offline Policy Learning

	H-horizon, Stationary	H-Horizon, Nonstationary	Infinite horizon γ-discounted
Upper bound	$\sqrt{\frac{H^2}{n\, d_m}}$	$\sqrt{\frac{H^3}{n\, d_m}}$	$\sqrt{\frac{(1-\gamma)^{-3}}{N\, d_m}}$
Lower bound	$\sqrt{\frac{H^2}{n\, d_m}}$	$\sqrt{\frac{H^3}{n\, d_m}}$	$\sqrt{\frac{(1-\gamma)^{-3}}{N\, d_m}}$

OPE $\boxed{\frac{1}{\sqrt{n}}\sqrt{\sum_{t=1}^{H}\sum_{s,a}\frac{(d^\pi(s)\pi(s,a))^2}{d^\mu(s,a)} Var[V^\pi_{t+1}(s,a)]}}$

Remaining open research threads:
- More adaptive bounds: More explicit dependence on importance weights.
- Reward free / Task-Agnostic settings
- Function approximation settings

$(s,a,s')_{i=1}^n$

This lecture

- Function approximation in RL in general

- Theoretical analysis ofFitted Q Iteration for offline RL in function approximation setting

Borrowed some ideas / materials from Nan Jiang.
FQI analysis from AJKS Ch 15.

Recap: Large MDPs

- State space can be exponentially large

$\log|S|$ $\cancel{\text{linear}\ |S|}$

- Planning horizon H is large

$\cancel{\exp(H)}$ $\text{poly}(H)$ best α $1-1/\alpha$ is optml

- Typical solutions:
  - use features to denote state (or state-action pairs)
  - use function approximation of various quantities

$s \to \phi(s) \in \mathbb{R}^d$ $f_\theta(\phi(s)) \simeq V^*(s)$

$(s,a) \to \phi(s,a) \in \mathbb{R}^d$ $f_\theta(\phi(s,a)) \simeq Q^*(s,a)$

$f_\theta \in \mathcal{F}_\Theta$ $\theta \in \Theta$

Ideally, if we have access to the MDP exactly then we could solve

$\mathcal{F}: f:(S,A)\to\mathbb{R}$

if $Q^* \in \mathcal{F}$

- We could run value iterations

$$f_t \leftarrow \mathcal{T} f_{t\text{-}1}$$

$$\mathcal{T}f(s,a) := r(s,a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)} \max_{a'\in\mathcal{A}} f(s',a').$$

γ - contraction

$\|f_k - f^*\|_\infty \le \gamma^{k-1} \|f_0 - f^*\|_\infty$

- The issue is that is the updated function still within the function class?
  - Add a projection

$$f_t \leftarrow \prod_F \mathcal{T} f_{t\text{-}1}$$

Fitted Q-Iteration

$$f_t = \arg\min_{f \in \mathcal{F}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \underbrace{\overbrace{\left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s', a') \right)}^{\text{Constant}}}_{\mathbb{E} f_{t-1} \overset{?}{=} \mathbb{E} f_{t-1}} \right)^2$$

if deterministic transition

- Stochastic semigradient updates

$$\theta \leftarrow \theta - \frac{\alpha}{2} \cdot \nabla_\theta \left(f_\theta(s,a) - \left(r + \gamma \overbrace{\max_{a' \in \mathcal{A}} f_\theta(s', a')}^{\text{Treat as constant; don't pass gradient}} \right) \right)^2$$

$$= \theta - \alpha \left(f_\theta(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_\theta(s', a') \right) \right) \nabla_\theta f_\theta(s,a)$$

Same as Q-learning in tabular / linear function approximation case.

Very similar to DQN if we use a neural network function approximation

Questions about convergence?

- If realizable, then we know that Q* is a fixed point.

  $\mathcal{T}Q^* = Q^*$

- But convergence guarantee is not guaranteed in general.
  - Even if it is a linear function approximation.
  - Even if it is realizable.
  - Even it is we know the MDP

2.1 Counter-example for least-square regression [Tsitsiklis and van Roy, 1996]

An MDP with two states x_1, x_2, 1-d features for the two states: $f_{x_1} = 1, f_{x_2} = 2$. Linear Function approximation with $\tilde{V}_\theta(x) = \theta f_x$.

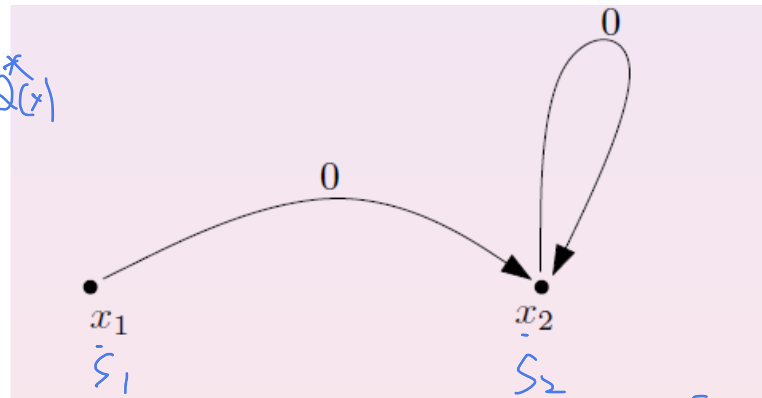
$\theta = 0$

$\phi(s_1) = 1 \in \mathbb{R}^1$ $\quad \phi(s_2) = 2 \in \mathbb{R}^1$

$\theta f_x = 0 = V^*_{(x)} = Q^*(x)$

$V^* = 0$

$-B \leq \theta \leq B$



initialize at θ

$(x_1, a, x_2, 0)$

$(x_2, a, x_2, 0)$

credit: course notes from Shipra Agrawal

$$
\begin{aligned}
\theta_k &:= \arg\min_\theta \frac{1}{2}(\theta - \text{target}_1)^2 + (2\theta - \text{target}_2)^2 \\
&= \arg\min_\theta \frac{1}{2}(\theta - \gamma\theta^{k-1} f_{x_2})^2 + (2\theta - \gamma\theta^{k-1} f_{x_2})^2 \\
&= \arg\min_\theta \frac{1}{2}(\theta - \gamma 2\theta^{k-1})^2 + (2\theta - \gamma 2\theta^{k-1})^2
\end{aligned}
$$

$$(\theta - \gamma 2\theta^{k-1}) + 2(2\theta - \gamma 2\theta^{k-1}) = 0 \Rightarrow 5\theta = 6\gamma\theta^{k-1}$$

$$\theta_k = \frac{6}{5}\gamma\theta_{k-1}$$

This diverges if $\gamma \geq 5/6$.

Direct Bellman Residual Minimization

$$\text{minimize } \| f - \mathcal{T} f \|^2 \text{ over } f \in F$$

$$\sum_{s,a} \mu_{s,a} \left(f(s,a) - \left(r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} f(s',a') \right) \right)^2$$

- We do not have access to the transition kernel
- So what we can minimize is the following

$$\min_{f \in F} \sum_{(s,a,s',r)} \mathbb{E}_{\substack{(s,a)\sim\mu \\ r\sim R(s,a) \\ s'\sim P(s,a)}} \left[\left(f(s,a) - (r + \gamma \max_{a'} f(s', a')) \right)^2 \right]$$

GD

Are they equivalent, as the dataset goes to infinity?

Double sampling issue

$$\mathbb{E}\left[f(s,a) - (\mathcal{T}f)(s,a) \right] \left[(\mathcal{T}f)(s,a) - \left(r + \gamma \max_{a'} f(s',a') \right) \right] = 0$$

$$- (\mathcal{T}f)(s,a) + (\mathcal{T}f)(s,a)$$

$$\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2\right]$$

$$= \mathbb{E}_{(s,a)\sim\mu}\left[(f(s,a) - (\mathcal{T}f)(s,a))^2\right] + \mathbb{E}_{(s,a)\sim\mu}\left[\left((\mathcal{T}f)(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2\right]$$

- Workaround #1: If you can get two samples from the same s,a then you could do stochastic approximation for both.

$$\mathbb{E}\left(f(s,a) - \left(r_A + \gamma \max_{a' \in \mathcal{A}} f(s'_A, a')\right)\right) \mathbb{E}\left(f(s,a) - \left(r_B + \gamma \max_{a' \in \mathcal{A}} f(s'_B, a')\right)\right)$$

Workaround #2: Solve a saddle point problem instead (Antos et al. 08)

- Idea: let us estimate the second term and subtract it away.

$\mathbb{E}[r + \gamma \max_{a' \in \mathcal{A}} f(s', a')] \quad \forall f$

$\mathcal{L}f$

$$\arg\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}}\left(\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a)-\left(r+\gamma\max_{a'\in\mathcal{A}} f(s',a')\right)\right)^2 - \left(g(s,a)-\left(r+\gamma\max_{a'\in\mathcal{A}} f(s',a')\right)\right)^2\right]\right)$$

- If function class G is sufficiently expressive, then it can make the second term 0 for all f.

Quick checkpoint

- Idea: Approximate Q* function
  - Minimize the best approximation error by
  - Value iterations?
  - Direct Bellman Residual Minimization?

- Still an active area of research in both theory and practice

- Standard techniques seem to work (especially when you have a simulator and can restart..)

Function approximations of other quantities

- Approximating the occupancy measure of logging policy

$$\mathbb{E}_{(s,a,s',r)\sim\mu} \frac{d_0^{\pi}(s)}{d^{\mu}(s)} \frac{\pi(a|s)}{\mu(a|s)} \cdot r(s,a) = V^{\pi}$$

$$\max_{f_1\in F,\ f_2} \frac{f_1(s)\,\pi(a|s)}{f_1(s)\,\mu(a|s)}\, r(s,a)$$

s.t. Bellman equation of V, f_1

"Breaking the Curse of Horizon" 2018

"DualDice"

- Approximation the occupancy measure of target policy

- Approximation of the importance weights

$$\max_{W\in F} \mathbb{E}\, W(s,a)\, r(s,a)$$

s.t. linear constraints on W

Remaining part of the lecture

- A theoretical analysis of Fitted Q-Iterations

- Under a number of additional conditions to make it tractable

Fitted Q Iterations: Problem Setup

- Infinite Horizon Discounted MDP

d_0, μ, S_0

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, P, r, \rho\}$$

$s_i' \sim P(\cdot|s_i, a_i)$

$r_i \sim r(s_i, a_i)$

$$\sup_{s,a} r(s,a) \in [0,1]$$

$a_i \sim \mu(a|s_i)$

i.i.d: by $s_i \sim d^\mu(s)$

$$\mathcal{D} := \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$$

$d^\mu(s) = (1-\gamma) \cdot V^\mu(s)$

- Goal: find nearly optimal policy $\hat\pi$

$V^* - V^{\hat\pi} < \varepsilon$

Assumptions

$|\mathcal{F}| < +\infty$

1. Realizability

We assume $\mathcal{F}$ is rich enough such that $Q^\star \in \mathcal{F}$.

2. Uniform concentrability

$\boxed{d^\mu(s,a) > d_m \text{ for all } s,a}$

$$\forall \pi, h, x, a : \frac{d^\pi(s,a)}{\mu(s,a)} \leq C.$$

s

3. Bellman completeness

We assume that for any $f \in \mathcal{F}$, $\mathcal{T} f \in \mathcal{F}$.

Recap: the FQI algorithm

- Initialize at any function f_0 in the function class $\in F$

$$\text{FQI:} \qquad f_t \in \operatorname{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} \left(f(s_i', a_i) - r_i - \gamma \max_{a'\in\mathcal{A}} f_{t-1}(s_i, a_i) \right)^2 .$$

- Natural policy of the approximate Q* function

$$\pi_k(s) := \operatorname{argmax}_a f_k(s,a), \forall s.$$

$$V^{*} - V^{\pi_k} \xrightarrow{(k\to\infty)} 0 \quad O\left(\frac{1}{\sqrt{n}}\right)$$

FQI "works" under these conditions

$\|V = n \cdot H\|$ $\sqrt{\frac{H^3}{N d_m}}$

$\log|F| = |S| H \log A$

$|F| = A^{SH}$

Theorem 15.4 (AJKS): Assume Assumptions 1--3, with probability at least 1- δ

Uniform Convergence

$$V^\star - V^{\pi_k} \le \mathcal{O}\left(\frac{1}{(1-\gamma)^3}\sqrt{\frac{C\log(|\mathcal{F}|/\delta)}{n}}\right) + \frac{2\gamma^k}{(1-\gamma)^2}.$$

+ Approx. error

$f^* \in F$

$\frac{\|f^* - Q^*\|_\infty}{1-\gamma}$

?

$(1-\gamma)^{-1} \asymp H$

Statistical

$O\left(\frac{1}{\sqrt{n}}\right)$

Optimization error

$\log|F| \asymp d$

$\sqrt{|S| \cdot H^3}$ Suboptimal when applying it to the Tabular Setting

Sketch of the proof

- Use contraction of properties of Bellman updates
- Use uniform convergence
  - the approximate Bellman update using the offline dataset is similar to the actual exact Bellman update
  - simultaneously for all functions within the function class

- By Performance Difference Lemma

$$\pi_k = \arg\max_a f_k(\cdot, a)$$

$$(1-\gamma)\left(V^\star - V^{\pi_k}\right) = \mathbb{E}_{s\sim d^{\pi_k}}\left[-A^\star(s, \pi_k(s))\right]$$

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

$$\leq \mathbb{E}_{s\sim d^{\pi_k}}\Big[Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)) + \underbrace{\cancel{f_k(s, \pi^\star(s)) - f_k(s, \pi_k(s))}}_{\leq 0} + f_k(s, \pi_k(s)) - Q^\star(s, \pi_k(s)) \Big]$$

$$\|\cdot\|_{1,\nu} \longrightarrow = \mathbb{E}_{s\sim d^{\pi_k}_{\cdot\pi_k}}\left[|Q^\star - f_k|(s, \pi^\star(s)) \right] + \mathbb{E}_{s\sim d^{\pi_k}_{\cdot\pi_k}}\left[|Q^\star - f_k|(s, \pi_k(s)) \right]$$

$$\leq \sqrt{\mathbb{E}_{s\sim d^{\pi_k}_{\cdot\pi_k}}\left[(Q^\star - f_k)^2(s, \pi^\star(s)) \right]} + \sqrt{\mathbb{E}_{s\sim d^{\pi_k}_{\cdot\pi_k}}\left[(Q^\star - f_k)^2(s, \pi_k(s)) \right]}$$

$$\| Q^\star - f_k \|_{2, d^{\pi_k}_{\cdot\pi^\star}} \qquad\qquad \| Q^\star - f_k \|_{2, d^{\pi_k}_{\cdot\pi_k}}$$

$$(\mathbb{E}X)^2 \le \mathbb{E}X^2$$

$$\left(\max_a Q^\star - \max_a f_{k-1}\right)^2 \le \max_a (Q^\star - f_{k-1})^2$$

$$\frac{\nu}{\mu} \le C$$

$$\|Q^\star - f_k\|_{2,\nu} \le \|Q^\star - \mathcal{T} f_{k-1}\|_{2,\nu} + \|f_k - \mathcal{T} f_{k-1}\|_{2,\nu}$$

$$\le \gamma \sqrt{\mathbb{E}_{s,a\sim\nu}\left[\left(\mathbb{E}_{s'\sim P(\cdot|s,a)} \max_a Q^\star(s',a) - \max_a f_{k-1}(s',a)\right)^2\right]} + \|f_k - \mathcal{T} f_{k-1}\|_{2,\nu}$$

$$\le \gamma \sqrt{\mathbb{E}_{s,a\sim\nu, s'\sim P(\cdot|s,a)} \max_a \left(Q^\star(s',a) - f_{k-1}(s',a)\right)^2} + \sqrt{C}\|f_k - \mathcal{T} f_{k-1}\|_{2,\mu},$$

$$\|Q^\star - f_{k-1}\|_{2,\nu'}$$

$$\nu'(s',a') = \sum_{s,a} \nu(s,a) P(s'|s,a) \mathbf{1}\{a' = \mathrm{argmax}_a \left(Q^\star(s',a) - f_{k-1}(s',a)\right)^2\},$$

Roll in with ν, then take action a'

Recursive application

$$\|Q^\star - f_k\|_{2,\nu} \le \gamma \|Q^\star - f_{k-1}\|_{2,\nu'} + \sqrt{C}\|f_k - \mathcal{T} f_{k-1}\|_{2,\mu}.$$

for over $t = k-1, \dots, 0$

$$\|Q^\star - f_k\|_{2,\nu} \le \sqrt{C} \sum_{t=0}^{k-1} \gamma^t \|f_{k-t} - \mathcal{T} f_{k-t-1}\|_{2,\mu} + \gamma^k \|Q^\star - f_0\|_{2,\widetilde{\nu}},$$

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

$V_{max} \le \frac{1}{1-\gamma}$

$$\gamma \|Q^\star - f_0\|_{2,\widetilde{\nu}} \le \gamma^k V_{\max}.$$

Reducing to a uniform convergence problem

- Apply Uniform convergence

$\hat{\mathcal{T}} f_{k-t-1}$

$$\sqrt{C}\sum_{t=0}^{k-1}\gamma^t \|f_{k-t} - \mathcal{T} f_{k-t-1}\|_{2,\mu} \le \mathcal{O}\left(\sqrt{C}\sum_{t=0}^{k-1}\gamma^k \sqrt{\frac{V_{\max}^2 \ln(|\mathcal{F}|/\delta)}{n}}\right) = \mathcal{O}\left(\frac{V_{\max}\sqrt{C\ln(|\mathcal{F}|/\delta)}}{(1-\gamma)\sqrt{n}}\right)$$

True Bellman Operator M

Bellman Operator by solving FQI

- Finally

$$\|Q^\star - f_k\|_{2,\nu} = \mathcal{O}\left(\frac{V_{\max}\sqrt{C\ln(|\mathcal{F}|/\delta)}}{(1-\gamma)\sqrt{n}}\right) + \gamma^k V_{\max},$$

Uniform convergence

Lemma 15.5 (Least Square Generalization Error)**.** *Given* $f \in \mathcal{F}$*, denote* $\hat{f}_f := \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(s_i, a_i) - r_i - \gamma \max_{a'} f(s'_i, a'))^2$*. With probability at least* $1 - \delta$*, for all* $f \in \mathcal{F}$*, we have:*

$$\mathbb{E}_{s,a\sim\mu}\left(\hat{f}_f(s,a) - \mathcal{T}f(s,a)\right)^2 = \mathcal{O}\left(\frac{V_{\max}^2 \ln\left(\frac{|\mathcal{F}|}{\delta}\right)}{n}\right).$$

- Implies that

$$\mathbb{E}_{s,a\sim\mu}\left(f_t(s,a) - \mathcal{T}f_{t-1}(s,a)\right)^2 = \mathcal{O}\left(\frac{V_{\max}^2 \ln\left(\frac{|\mathcal{F}|}{\delta}\right)}{n}\right).$$

Proof sketch

- Bernstein + Union bound

$$z_i^f := \left(f(s_i,a_i) - r_i - \gamma \max_{a'\in\mathcal{A}} f'(s_i',a')\right)^2 - \left(\mathcal{T}f'(s_i,a_i) - r_i - \gamma \max_{a'\in\mathcal{A}} f'(s_i',a')\right)^2.$$

- Boundedness

$$|z_i^f| \le V_{\max}^2,$$

- Small variance

$$V_{\max}^2 \mathbb{E}_{s,a\sim\mu}\left[(f(s,a) - \mathcal{T}f'(s,a))^2\right]$$

A self-bounding trick

$$\mathbb{E}_{s,a\sim\mu}\left(\hat{f}(s,a) - \mathcal{T}f'(s,a)\right)^2 \le \sqrt{\frac{8V_{\max}^2 \mathbb{E}_{s,a\sim\mu}\left[(\hat{f}(s,a) - \mathcal{T}f'(s,a))^2\right] \ln(|\mathcal{F}|/\delta)}{n}} + \frac{4V_{\max}^2 \ln(|\mathcal{F}|/\delta)}{3n}.$$

- Solve a quadratic equation

$$\mathbb{E}_{s,a\sim\mu}\left(\hat{f}(s,a) - \mathcal{T}f'(s,a)\right)^2 \le \left(\sqrt{2} + \sqrt{10/3}\right)^2 \frac{V_{\max}^2 \ln(|\mathcal{F}|/\delta)}{n}.$$

Thank you very much!