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)

Offline Trajectory data D Collected by running $\boldsymbol{\mu}$ $\xrightarrow[\text{methods}]{\text{Task: design OPE}}$ Evaluate fixed Target Policy $\boldsymbol{\pi}$

Via Uniform OPE

- Task 2: Offline Policy Learning. (OPL)

Offline Trajectory data D Collected by running $\boldsymbol{\mu}$ $\xrightarrow[\text{methods}]{\text{Task: design OPO}}$ Find near optimal Policy $\widehat{\boldsymbol{\pi}}^{*}$

Recap: Offline Policy Learning

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

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

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

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	$\sqrt{\frac{H^2}{n\, d_m}}$	$\sqrt{\frac{H^3}{n\, d_m}}$	$\sqrt{\frac{(1-\gamma)^{-3}}{N\, d_m}}$

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

This lecture

- Function approximation in RL in general

- Theoretical analysis of Fitted 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

- Planning horizon H is large

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

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

- We could run value iterations

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

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

- 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

$$f_t = \arg\min_{f \in \mathcal{F}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s',a') \right) \right)^2$$

- 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

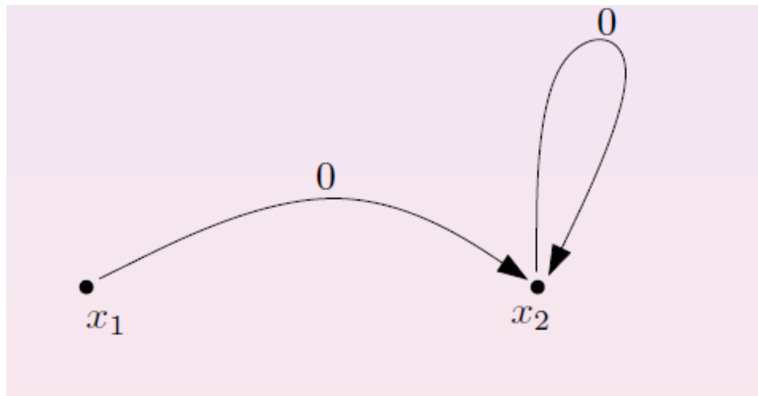
Questions about convergence?

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

- 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 $\tilde{V}_\theta(x) = \theta f_x$.



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

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

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

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

Are they equivalent, as the dataset goes to infinity?

Double sampling issue

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

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

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

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

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

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

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

- 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

- Approximation the occupancy measure of target policy

- Approximation of the importance weights

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

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

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

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

- Goal: find nearly optimal policy

Assumptions

1. Realizability

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

2. Uniform concentrability

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

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

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

- Natural policy of the approximate Q* function

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

FQI “works” under these conditions

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

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

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

Decomposing the error

- By Performance Difference Lemma

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

Further decomposing the error

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

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

Recursive application

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

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

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

Reducing to a uniform convergence problem

- Apply Uniform convergence

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

- 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

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

- 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!