Notes on Fitted Q-iteration

Nan Jiang
August 30, 2020

Let $M = (S, A, P, R, \gamma, d_0)$ be an MDP, where $d_0$ is the initial distribution over states. Given a dataset $\{(s, a, r, s')\}$ generated from $M$ and a Q-function class $F \subset \mathbb{R}^{S \times A}$, we want to analyze the guarantee of Fitted Q-Iteration. This note is inspired by and scrutinizes the results in Approximate Value/Policy Iteration literature [e.g., 1, 2, 3] under simplification assumptions.

Setup and Assumptions

1. $F$ is finite but can be exponentially large.
2. Realizability: $Q^* \in F$.
3. $F$ is closed under Bellman update: $\forall f \in F, T f \in F$. (For finite $F$, this implies realizability.)
4. The dataset $D = \{(s, a, r, s')\}$ is generated as follows: $(s, a) \sim \mu \times U(A)$ ($U(A)$ is uniform over actions), $r \sim R(s, a)$, $s' \sim P(s, a)$. Define the empirical update $\hat{T}_F f'$ as
   \[ L_D(f; f') := \frac{1}{|D|} \sum_{(s, a, r, s') \in D} (f(s, a) - r - \gamma V_{f'}(s'))^2. \]
   \[ \hat{T}_F f' := \arg \min_{f \in F} L_D(f; f'), \]
   where $V_{f'}(s') := \max_{a'} f'(s', a')$. Note that by completeness, $T f' \in F$ is the Bayes optimal regressor for the regression problem defined in $L_D(f; f')$. It will also be useful to define
   \[ L_{\mu \times U}(f; f') := \mathbb{E}_D[L_D(f; f')]. \]
5. For any function $g : S \to \mathbb{R}$, any distribution $\nu \in \Delta(S)$, and $p \geq 1$, define $\|g\|_{p, \nu} := (\mathbb{E}_{s \sim \nu}[|g(s)|^p])^{1/p}$, and let $\|g\|_\nu$ be a shorthand for $\|g\|_{2, \nu}$. Such norms are similarly defined for functions over $S \times A$.
6. Let $d_h^\pi$ be the distribution of $s_h$ under $\pi$, that is, $d_h^\pi(s) := \Pr[s_h = s | s_1 \sim d_0, \pi]$.
7. $\mu$ is exploratory: for a distribution $\nu \in \Delta(S)$ generated by any (non-stationary) policy at any time step (that is, any distribution $\nu$ of the form $d_h^\pi$ where $\pi$ may be non-stationary),
   \[ \forall s \in S, \frac{\nu(s)}{\mu(s)} \leq C. \]
As a consequence, $\| \cdot \|_\nu \leq \sqrt{C} \| \cdot \|_\mu$. Similarly, when we couple $\mu$ with a uniform distribution over $A$, we have similar results for state-action distributions: $\| \cdot \|_{\nu \times \pi} \leq \sqrt{|A|C} \| \cdot \|_{\mu \times U}$. See slides for example scenarios where $C$ is naturally bounded.
8. Algorithm (simplified for analysis): let \( f_0 \equiv 0 \) (assuming \( 0 \in \mathcal{F} \)), and for \( k \geq 1 \), \( f_k \equiv \widehat{T}_F f_{k-1} \).

9. Uniform deviation bound (can be obtained by concentration inequalities and union bound):

\[
\forall f, f' \in \mathcal{F}, |\mathcal{L}_D(f; f') - \mathcal{L}_{\mu \times U}(f; f')| \leq \epsilon.
\]

(Note: at the end we will show how to obtain fast rates.)

**Goal**  Let \( \hat{\pi} := \pi_{f_k} \). Derive an upper bound on \( J(\pi^*) - J(\hat{\pi}) \).

**Analysis**

\[
J(\pi^*) - J(\hat{\pi}) = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim d_h^\pi} [V^*(s) - Q^*(s, \hat{\pi})]
\]

\[
\leq \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim d_h^\pi} [Q^*(s, \pi^*) - f_k(s, \pi^*) + f_k(s, \hat{\pi}) - Q^*(s, \hat{\pi})]
\]

\[
\leq \sum_{h=1}^{\infty} \gamma^{h-1} \left( \|Q^* - f_k\|_{1,d_h^\pi \times \pi^*} + \|Q^* - f_k\|_{1,d_h^\pi \times \pi} \right)
\]

\[
\leq \sum_{h=1}^{\infty} \gamma^{h-1} \left( \|Q^* - f_k\|_{d_h^\pi \times \pi^*} + \|Q^* - f_k\|_{d_h^\pi \times \pi} \right).
\]

(1)

The last line contains two terms, both in the form of \( \|Q^* - f_k\|_{\nu \times \pi} \). So it remains to bound \( \|Q^* - f_k\|_{\nu \times \pi} \) for any \( \nu \times \pi \in \Delta(S \times A) \) that combines any \( \nu \in \Delta(S) \) that satisfies bullet 4 with any \( \pi : S \to A \).

First a helper lemma:

**Lemma 1.** Define \( \pi_{f, f_k}(s) := \max_{a \in A} \max \{ f(s, a), f_k(s, a) \} \). Then we have \( \forall \nu \in \Delta(S), \)

\[
\|V_f - V_{f_k}\|_{\nu} \leq \|f - f_k\|_{\nu \times \pi_{f, f_k}}.
\]

**Proof.**

\[
\|V_f - V_{f_k}\|_{\nu}^2 = \sum_{s \in S} \nu(s) (\max_{a \in A} f(s, a) - \max_{a' \in A} f_k(s, a'))^2
\]

\[
\leq \sum_{s \in S} \nu(s) (f(s, \pi_{f, f_k}) - f_k(s, \pi_{f, f_k}))^2 = \|f - f_k\|_{\nu \times \pi_{f, f_k}}^2.
\]

Now we can bound \( \|Q^* - f_k\|_{\nu \times \pi} \) using Lemma 1. Define \( P(\nu \times \pi) \) as a distribution over \( S \) generated as \( s' \sim P(\nu \times \pi) \iff (s, a) \sim \nu \times \pi, s' \sim P(s, a) \), and

\[
\|f_k - Q^*\|_{\nu \times \pi} = \|f_k - \mathcal{T}_f k - 1 + \mathcal{T}_f k - 1 - Q^*\|_{\nu \times \pi}
\]

\[
\leq \|f_k - \mathcal{T}_f k - 1\|_{\nu \times \pi} + \|\mathcal{T}_f k - 1 - \mathcal{T} Q^*\|_{\nu \times \pi}
\]

\[
\leq \sqrt{|A|^C} \|f_k - \mathcal{T}_f k - 1\|_{\mu \times U} + \gamma \|V_{f_k - 1} - V^*\|_{P(\nu \times \pi)} \tag{(*)}
\]

\[
\leq \sqrt{|A|^C} \|f_k - \mathcal{T}_f k - 1\|_{\mu \times U} + \gamma \|f_k - 1 - Q^*\|_{P(\nu \times \pi) \times \pi_{f, f_k - 1}, Q} \tag{Lemma 1}
\]

2
Step (*) holds because:

\[ \|T f_{k-1} - TQ^*\|^2_{\nu \times \pi} = E_{(s,a) \sim \nu \times \pi} \left[ \left( (T f_{k-1})(s,a) - (TQ^*)(s,a) \right)^2 \right] \]

\[ = E_{(s,a) \sim \nu \times \pi} \left[ (\gamma E_{s' \sim P(s,a)}[V_{f_{k-1}}(s') - V^*(s')])^2 \right] \leq \gamma^2 E_{(s,a) \sim \nu \times \pi, s' \sim P(s,a)} \left[ (V_{f_{k-1}}(s') - V^*(s'))^2 \right] \]

\[ = \gamma^2 E_{s' \sim P(\nu \times \pi)} \left[ (V_{f_{k-1}}(s') - V^*(s'))^2 \right] = \gamma^2 \|V_{f_{k-1}} - V^*\|_{P(\nu \times \pi)}^2. \]

Note that we can apply the same analysis on \( P(\nu \times \pi) \times \pi_{f_{k-1},Q} \) and expand the inequality \( k \) times. It then suffices to upper bound \( \|f_k - T f_{k-1}\|_{\mu \times U} \).

\[ \|f_k - T f_{k-1}\|^2_{\mu \times U} = \mathcal{L}_{\mu \times U}(f_k; f_{k-1}) - \mathcal{L}_{\mu \times U}(T f_{k-1}; f_{k-1}) \] (\( \mathcal{L} \) squared loss + \( T f_{k-1} \) Bayes optimal)

\[ \leq \mathcal{L}_D(f_k; f_{k-1}) - \mathcal{L}_D(T f_{k-1}; f_{k-1}) + 2\epsilon \]

(f_k minimizes \( \mathcal{L}_D(\cdot; f_{k-1}) \))

Note that the RHS does not depend on \( k \), so we conclude that

\[ \|f_k - Q^*\|_{\nu \times \pi} \leq \frac{1 - \gamma^k}{1 - \gamma} \sqrt{2|A|C\epsilon} + \gamma^k R_{\max}. \]

Apply this to Equation [1] and we get

\[ J(\pi^*) - J(\pi_{f_k}) \leq \frac{2}{1 - \gamma} \left( \frac{1 - \gamma^k}{1 - \gamma} \sqrt{2|A|C\epsilon} + \gamma^k R_{\max} \right). \]

**Extension: fast rate** The previous bound should have \( O(n^{-1/4}) \) dependence on sample size \( n := |D| \), because \( \epsilon \) in bullet 6 should be \( O(n^{-1/2}) \) using Hoeffding’s, and the final bound depends on \( \sqrt{\epsilon} \). Here we exploit realizability to achieve fast rate so that the final bound is \( O(n^{-1/2}) \).

Define

\[ Y(f; f') := (f(s,a) - r - \gamma V_{f'}(s'))^2 - (\mathcal{T} f')(s,a) - r - \gamma V_f(s')). \]

Plug each \((s,a,r,s') \in D\) into \( Y(f; f') \) and we get i.i.d. variables \( Y_1(f; f'), Y_2(f; f'), \ldots, Y_n(f; f') \) where \( n = |D| \). It is easy to see that

\[ \frac{1}{n} \sum_{i=1}^n Y_i(f; f') = \mathcal{L}_D(f; f') - \mathcal{L}_D(T f'; f'), \]

so we only shift our objective \( \mathcal{L}_D \) by a \( f \)-independent constant. Our goal is to show that

\[ \|\hat{T}_f f' - T f'\|^2_{\mu \times U} = \mathbb{E}[Y(\hat{T}_f f'; f')] = O(1/n). \]

Note that this result can be directly plugged into the previous analysis by letting \( f' = f_{k-1} \) (hence \( \hat{T}_f f' = f_k \)), and we immediately obtain a final bound of \( O(n^{-1/2}) \).

To prove the result, first notice that \( \forall f \in \mathcal{F}, \)

\[ \mathbb{E}[Y(f; f')] = \mathcal{L}_{\mu \times U}(f; f') - \mathcal{L}_{\mu \times U}(T f'; f') = \|f - T f'\|^2_{\mu \times U}, \]
thanks to realizability and squared loss. Next we bound variance of $Y$:

\[
\mathbb{V}[Y(f; f')] \leq \mathbb{E}[Y(f; f')^2] \\
= \mathbb{E}
\left[
\left((f(s, a) - r - \gamma V_{f'}(s'))^2 - ((T f')(s, a) - r - \gamma V_{f'}(s'))^2
\right)^2
\right] \\
= \mathbb{E}
\left[
(f(s, a) - (T f')(s, a))^2(f(s, a) + (T f')(s, a) - 2r - 2\gamma V_{f'}(s'))^2
\right] \\
\leq 4V_{\text{max}}^2 \mathbb{E}
\left[
(f(s, a) - (T f')(s, a))^2
\right] \\
= 4V_{\text{max}}^2 \|f - T f'\|_{\mu \times U}^2 = 4V_{\text{max}}^2 \mathbb{E}[Y(f; f')],
\]

where $V_{\text{max}} = R_{\text{max}}/(1 - \gamma)$ is a constant.

Next we apply (one-sided) Bernstein’s inequality (see [4]) and union bound over all $f \in \mathcal{F}$. Let $N = |\mathcal{F}|$. For any fixed $f'$, with probability at least $1 - \delta$, $\forall f \in \mathcal{F}$,

\[
\mathbb{E}[Y(f; f')] - \frac{1}{n} \sum_{i=1}^{n} Y_i(f; f') \leq \sqrt{\frac{2V[Y(f; f')] \log N}{n}} + \frac{4V_{\text{max}}^2 \log N}{3n} \\
= \sqrt{\frac{8V_{\text{max}}^2 \mathbb{E}[Y(f; f')] \log N}{n}} + \frac{4V_{\text{max}}^2 \log N}{3n}.
\]

Since $\tilde{T}_{\mathcal{F}} f'$ minimizes $\mathcal{L}_D(\cdot; f')$, it also minimizes $\frac{1}{n} \sum_{i=1}^{n} Y_i(\cdot; f')$ because the two objectives only differ by a constant $\mathcal{L}_D(T f'; f')$. Hence,

\[
\frac{1}{n} \sum_{i=1}^{n} Y_i(\tilde{T}_{\mathcal{F}} f'; f') \leq \frac{1}{n} \sum_{i=1}^{n} Y_i(T f'; f') = 0.
\]

Then,

\[
\mathbb{E}[Y(\tilde{T}_{\mathcal{F}} f'; f')] \leq \sqrt{\frac{8V_{\text{max}}^2 \mathbb{E}[Y(\tilde{T}_{\mathcal{F}} f'; f')] \log N}{n}} + \frac{4V_{\text{max}}^2 \log N}{3n}.
\]

Solving for the quadratic formula,

\[
\mathbb{E}[Y(\tilde{T}_{\mathcal{F}} f'; f')] \leq \left(\sqrt{2 + \sqrt{\frac{10}{3}}}\right)^2 \frac{V_{\text{max}}^2 \log N}{n}.
\]

References


