Notes on Fitted Q-iteration

Nan Jiang

October 17, 2022

1 Analysis of FQI

Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ 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 $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, 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. $\mathcal{F}$ is finite but can be exponentially large.
2. Realizability: $Q^\star \in \mathcal{F}$.
3. Bellman completeness: $\forall f \in \mathcal{F}$, $\mathcal{T} f \in \mathcal{F}$. (For finite $\mathcal{F}$, this implies realizability.)
4. The dataset $D = \{(s, a, r, s')\}$ is generated as follows: $(s, a) \sim \mu$, $r \sim R(s, a)$, $s' \sim P(s, a)$. Define the empirical update $\widehat{\mathcal{T}}_{\mathcal{F}} f'$ as

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

   $$\widehat{\mathcal{T}}_{\mathcal{F}} f' := \operatorname*{arg\,min}_{f\in\mathcal{F}} \mathcal{L}_D(f; f'),$$

   where $V_{f'}(s') := \max_{a'} f'(s', a')$. Note that by completeness, $\mathcal{T} f' \in \mathcal{F}$ is the Bayes optimal regressor for the regression problem defined in $\mathcal{L}_D(f; f')$. It will also be useful to define

   $$\mathcal{L}_\mu(f; f') := \mathbb{E}_D[\mathcal{L}_D(f; f')].$$

5. For any function $g : \mathcal{X} \to \mathbb{R}$, any distribution $\nu \in \Delta(\mathcal{X})$, and $p \geq 1$, define $\|g\|_{p,\nu} := (\mathbb{E}_{x\sim\nu}[|g(x)|^p])^{1/p}$, and let $\|g\|_\nu$ be a shorthand for $\|g\|_{2,\nu}$. Such norms are similarly defined for functions over $\mathcal{X}$.
6. Let d_h^π be the distribution of (s_h, a_h) under π, that is, $d_h^\pi(s, a) := \Pr[s_h = s, a_h = a \mid s_1 \sim d_0, \pi]$. d^π is the usual discounted occupancy. The same notations are sometimes abused to refer to the corresponding state marginals, which will be clarified if not clear from the context.

7. We call any state-action distribution **admissible** if it can be generated at some time step from d_0 in the MDP. That is, it takes the form of d_h^π for some h and (possibly non-stationary) policy π. Then, assume that data is exploratory: for any admissible ν,

   $$\forall s \in \mathcal{S},\ \frac{\nu(s,a)}{\mu(s,a)} \le C.$$

   As a consequence, $\|\cdot\|_\nu \le \sqrt{C}\|\cdot\|_\mu$. See slides for example scenarios where C is naturally bounded.

8. Algorithm (simplified for analysis): let $f_0 \equiv \mathbf{0}$ (assuming $\mathbf{0} \in \mathcal{F}$), and for $k \ge 1$, $f_k := \widehat{\mathcal{T}}_{\mathcal{F}} f_{k-1}$.

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

   $$\forall f, f' \in \mathcal{F}, |\mathcal{L}_D(f; f') - \mathcal{L}_\mu(f; f')| \le \epsilon.$$

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

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

Analysis

$$\begin{aligned}
J(\pi^\star) - J(\hat{\pi}) &= \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim d_h^{\hat{\pi}}}[V^\star(s) - Q^\star(s, \hat{\pi})] \\
&\le \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim d_h^{\hat{\pi}}}[Q^\star(s, \pi^\star) - f_k(s, \pi^\star) + f_k(s, \hat{\pi}) - Q^\star(s, \hat{\pi})] \\
&\le \sum_{h=1}^{\infty} \gamma^{h-1} \left(\|Q^\star - f_k\|_{1, d_h^{\hat{\pi}} \times \pi^\star} + \|Q^\star - f_k\|_{1, d_h^{\hat{\pi}} \times \hat{\pi}}\right) \\
&\le \sum_{h=1}^{\infty} \gamma^{h-1} \left(\|Q^\star - f_k\|_{d_h^{\hat{\pi}} \times \pi^\star} + \|Q^\star - f_k\|_{d_h^{\hat{\pi}} \times \hat{\pi}}\right). \qquad (1)
\end{aligned}$$

In the above equations all the terms in the form of d_h^π should be treated as state distributions, and $d_h^\pi \times \pi'$ refers to a state-action distribution generated as $s \sim d_h^\pi, a \sim \pi'(\cdot|s)$. The last line contains two terms, both in the form of $\|Q^\star - f_k\|_\nu$ with some admissible $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$. So it remains to bound $\|Q^\star - f_k\|_\nu$ for any $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ that satisfies bullet 7.

First a helper lemma:

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

$$\|V_f - V_{f_k}\|_{\tilde{\nu}} \le \|f - f_k\|_{\tilde{\nu} \times \pi_{f,f_k}}.$$

Proof.

$$\begin{aligned}
\|V_f - V_{f_k}\|_{\tilde{\nu}}^2 &= \sum_{s \in \mathcal{S}} \tilde{\nu}(s)(\max_{a \in \mathcal{A}} f(s,a) - \max_{a' \in \mathcal{A}} f_k(s,a'))^2 \\
&\le \sum_{s \in \mathcal{S}} \tilde{\nu}(s)(f(s, \pi_{f,f_k}) - f_k(s, \pi_{f,f_k}))^2 = \|f - f_k\|_{\tilde{\nu} \times \pi_{f,f_k}}^2.
\end{aligned}$$

□

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

$$\begin{aligned}
\|f_k - Q^\star\|_\nu &= \|f_k - \mathcal{T} f_{k-1} + \mathcal{T} f_{k-1} - Q^\star\|_\nu \\
&\le \|f_k - \mathcal{T} f_{k-1}\|_\nu + \|\mathcal{T} f_{k-1} - \mathcal{T} Q^\star\|_\nu \\
&\le \sqrt{C}\, \|f_k - \mathcal{T} f_{k-1}\|_\mu + \gamma \|V_{f_{k-1}} - V^\star\|_{P(\nu)} && \text{(*)} \\
&\le \sqrt{C}\, \|f_k - \mathcal{T} f_{k-1}\|_\mu + \gamma \|f_{k-1} - Q^\star\|_{P(\nu) \times \pi_{f_{k-1}, Q^\star}}. && \text{(Lemma 1)}
\end{aligned}$$

Step (*) holds because:

$$\begin{aligned}
\|\mathcal{T} f_{k-1} - \mathcal{T} Q^\star\|_\nu^2 &= \mathbb{E}_{(s,a)\sim\nu}\left[((\mathcal{T} f_{k-1})(s,a) - (\mathcal{T} Q^\star)(s,a))^2 \right] \\
&= \mathbb{E}_{(s,a)\sim\nu}\left[\left(\gamma \mathbb{E}_{s'\sim P(s,a)}[V_{f_{k-1}}(s') - V^\star(s')]\right)^2 \right] \\
&\le \gamma^2\, \mathbb{E}_{(s,a)\sim\nu, s'\sim P(s,a)}\left[\left(V_{f_{k-1}}(s') - V^\star(s')\right)^2 \right] && \text{(Jensen)} \\
&= \gamma^2\, \mathbb{E}_{s'\sim P(\nu)}\left[\left(V_{f_{k-1}}(s') - V^\star(s')\right)^2 \right] = \gamma^2\, \|V_{f_{k-1}} - V^\star\|_{P(\nu)}^2.
\end{aligned}$$

Note that we can apply the same analysis on $P(\nu) \times \pi_{f_{k-1}, Q^\star}$ since it is also admissible, and expand the inequality k times. It then suffices to upper bound $\|f_k - \mathcal{T} f_{k-1}\|_\mu$.

$$\begin{aligned}
\|f_k - \mathcal{T} f_{k-1}\|_\mu^2 &= \mathcal{L}_\mu(f_k; f_{k-1}) - \mathcal{L}_\mu(\mathcal{T} f_{k-1}; f_{k-1}) && (\mathcal{L} \text{ squared loss + } \mathcal{T} f_{k-1} \text{ Bayes optimal}) \\
&\le \mathcal{L}_D(f_k; f_{k-1}) - \mathcal{L}_D(\mathcal{T} f_{k-1}; f_{k-1}) + 2\epsilon && (\mathcal{T} f_{k-1} \in \mathcal{F}) \\
&\le 2\epsilon. && (f_k \text{ minimizes } \mathcal{L}_D(\cdot\,; f_{k-1}))
\end{aligned}$$

Note that the RHS does not depend on k, so we conclude that for any admissible ν,

$$\|f_k - Q^\star\|_\nu \le \frac{1-\gamma^k}{1-\gamma}\sqrt{2C\epsilon} + \gamma^k V_{\max}.$$

Apply this to Equation (1) and we get

$$J(\pi^\star) - J(\pi_{f_k}) \le \frac{2}{1-\gamma}\left(\frac{1-\gamma^k}{1-\gamma}\sqrt{2C\epsilon} + \gamma^k V_{\max} \right).$$

Extension: fast rate The previous bound should have $O(n^{-1/4})$ dependence on sample size $n := |D|$, because ϵ in bullet 9 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

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

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

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

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

$$\|\widehat{\mathcal{T}}_{\mathcal{F}} f' - \mathcal{T} f'\|_\mu^2 \equiv \mathbb{E}[Y(\widehat{\mathcal{T}}_{\mathcal{F}} f'; f')] = O(1/n).$$

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

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

$$\mathbb{E}[Y(f; f')] = \mathcal{L}_\mu(f; f') - \mathcal{L}_\mu(\mathcal{T} f'; f') = \|f - \mathcal{T} f'\|_\mu^2,$$

thanks to realizability and squared loss. Next we bound variance of Y:

$$\begin{aligned}
\mathbb{V}[Y(f; f')] &\le \mathbb{E}[Y(f; f')^2] \\
&= \mathbb{E}\left[\left(\left(f(s,a) - r - \gamma V_{f'}(s')\right)^2 - \left((\mathcal{T} f')(s,a) - r - \gamma V_{f'}(s')\right)^2\right)^2\right] \\
&= \mathbb{E}\left[\left(f(s,a) - (\mathcal{T} f')(s,a)\right)^2 \left(f(s,a) + (\mathcal{T} f')(s,a) - 2r - 2\gamma V_{f'}(s')\right)^2\right] \\
&\le 4V_{\max}^2\, \mathbb{E}\left[\left(f(s,a) - (\mathcal{T} f')(s,a)\right)^2\right] \\
&= 4V_{\max}^2\, \|f - \mathcal{T} f'\|_\mu^2 = 4V_{\max}^2\, \mathbb{E}[Y(f; f')],
\end{aligned}$$

where $V_{\max} = R_{\max}/(1-\gamma)$ 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 $1-\delta$, $\forall f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}[Y(f; f')] - \frac{1}{n}\sum_{i=1}^n Y_i(f; f') &\le \sqrt{\frac{2\mathbb{V}[Y(f; f')] \log \frac{N}{\delta}}{n}} + \frac{4V_{\max}^2 \log \frac{N}{\delta}}{3n} \qquad (Y_i \in [-V_{\max}^2, V_{\max}^2]) \\
&\le \sqrt{\frac{8V_{\max}^2 \mathbb{E}[Y(f; f')] \log \frac{N}{\delta}}{n}} + \frac{4V_{\max}^2 \log \frac{N}{\delta}}{3n}.
\end{aligned}$$

Since $\widehat{\mathcal{T}}_{\mathcal{F}} f'$ 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 $\mathcal{L}_D(\mathcal{T} f'; f')$. Hence,

$$\frac{1}{n}\sum_{i=1}^n Y_i(\widehat{\mathcal{T}}_{\mathcal{F}} f'; f') \le \frac{1}{n}\sum_{i=1}^n Y_i(\mathcal{T} f'; f') = 0.$$

Then,

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

Solving for the quadratic formula,

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

Relaxing the definition of C The assumption $\|\nu/\mu\|_\infty \le C$ for all admissible ν can be relaxed. In particular, note that we always use this assumption in the form of

$$\|f - \mathcal{T} f'\|_\nu \le \sqrt{C}\|f - \mathcal{T} f'\|_\mu$$

for some $f, f' \in \mathcal{F}$. We can therefore literally redefine C as an upper bound of

$$\max_{f, f' \in \mathcal{F}} \frac{\|f - \mathcal{T} f'\|_\nu^2}{\|f - \mathcal{T} f'\|_\mu^2}$$

for all admissible ν. Despite the straightforward relaxation, when $\mathcal{F}$ has some nice structural properties, this new definition can be significantly tighter than the old definition based on raw density ratios. For example, when $\mathcal{F}$ is induced from a bisimulation state abstraction (which satisfies completeness), the new definition measures density ratio between the distributions over *abstract* state-action pairs, which can be much smaller than that between the raw state-action pairs. More generally, $\mathcal{F}$ is linear and Bellman completeness is satisfied, $f - \mathcal{T} f'$ is also a linear function, and the new definition measures coverage in the linear feature space. See further discussion on this in Akshay's note.

2 Alternative Analysis

Below we sketch an alternative proof to the FQI guarantee. There are two motivations:

Error propagation along "simple" policies The error propagation in the above analysis of FQI was along a somewhat "ugly" set of policies in the form of $\pi_{f_k, Q^\star}$, which in each state takes the action that "witnesses" the inequality $|\max_a f(s,a) - \max_a f'(s,a)| \le \max_a |f(s,a) - f'(s,a)|$ for $f = f_k$ and $f' = Q^\star$. However, the error propagation in the ADP literature (e.g., [3]) only involved "simple" policies, such as π_f for some $f \in \mathcal{F}$ (and the concatenation of such policies at different time steps to form a non-stationary policy).

"Modern" error-propagation analysis Error propagation in RL theory were often done by recursive expansion in the "old" literature, and the above analysis also follows this style. However, we have also seen alternative proofs based on cleaner and more elegant tools. For example, in the class we showed how easy it is to analyze the error propagation of the "minimax algorithm" $\arg\min_{f\in\mathcal{F}} \max_{g\in\mathcal{F}} \mathcal{L}_D(f;f) - \mathcal{L}_D(g;f)$ [5, 6] using the following lemma (which you are asked to prove in the homework): $\forall \pi, f$

$$J(\pi) - J(\pi_f) \le \frac{1}{1-\gamma}(\mathbb{E}_{d^\pi}[\mathcal{T}f - f] + \mathbb{E}_{d^{\pi_f}}[f - \mathcal{T}f]). \tag{2}$$

Using this lemma is also well aligned with the first motivation, as it often produces simple policies on the RHS (which the data distribution needs to cover).

2.1 Performance guarantee for non-stationary FQI

A major difficulty in applying Eq.(2) to FQI is that it requires control over the Bellman error $\|f - \mathcal{T}f\|$, i.e., the learned function should be "self-consistent". However, in FQI, we only have control over $\|f_t - \mathcal{T}f_{t-1}\|$, i.e., the output function f_k is not necessarily consistent with itself, but consistent with its previous iterate f_{k-1}, which is further consistent with *its* previous iterate f_{k-2}, and so on.

To overcome this difficulty, we first consider a different output policy: $\pi_{f_{k:1}} := \pi_{f_k} \circ \pi_{f_{k-1}} \circ \cdots \circ \pi_{f_0}$. This is a non-stationary policy, and after π_{f_0} we take arbitrary actions.[1] We call FQI with such a policy (instead of π_{f_k}) *non-stationary FQI*. Similar to the situation in value iteration (see note1), such a non-stationary policy actually has better guarantees than the usual FQI policy and saves a factor of horizon[2], and is also easier to analyze.

In particular, we now can use Eq.(2), because $\pi_{f_{k:1}}$ is greedy w.r.t. a self-consistent function, namely $f_{k:1} := f_k \circ f_{k-1} \circ \cdots \circ f_0$! The unusual aspect here is that we typically consider all value functions as only functions of states and actions, i.e., they are stationary. Here, the function $f_k \circ f_{k-1} \circ \cdots \circ f_0$ itself is a non-stationary object. The correct way to handle this is to

[1] The result might be improved (though it is unclear if the improvement is significant) if we produce a periodic policy that simply repeats $\pi_{f_{k:1}}$ forever [7].

[2] Another way to save this factor of horizon is to run the minimax algorithm [6].

1. Define a new finite-horizon MDP which is the k-step truncated version of the original MDP (c.f. the alternative proof of VI in note1).[3] The truncation will cause an $O(\gamma^k V_{\max})$ error between value functions in the original MDP and in the finite-horizon MDP, which exactly corresponds to the $\gamma^k V_{\max}$ term in the FQI bound.

2. Prove the variant of Eq.(2) for the finite-horizon MDP, and apply it to $f_{k:1}$.

Reading off the distributions on the RHS of Eq.(2), we should need data μ to cover the state distributions induced by two types of policies: $(\pi^\star)^t$, $\forall t \le k$, and $\pi^{f_{k:k'}}$, $\forall 0 \le k' \le k$. Caution: when we analyze the minimax algorithm using Eq.(2), we only need the data μ to cover the discounted occupancy as a whole, instead of covering the per-step distributions that contribute to the occupancy separately. Here we do not enjoy such a property, because our algorithm controls $\|f_t - \mathcal{T} f_{t-1}\|_{2,\mu}$ separately for each t separately, so change of measure must happen in each step instead of over the entire occupancy as a whole.

2.2 Performance guarantee for FQI

The previous section sketches an analysis of non-stationary FQI. To relate FQI to the analysis of its non-stationary variant, we use performance difference lemma

$$\begin{aligned} J(\pi^\star) - J(\pi_{f_k}) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{f_k}}}[V^\star(s) - Q^\star(s, \pi_{f_k})] \\ &\le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{f_k}}}[V^\star(s) - V^{\pi_{f_{k:1}}}(s)]. \end{aligned}$$

Here the inequality is due to $Q^\star(s, \pi_{f_k}) \ge V^{\pi_{f_{k:1}}}(s)$, because $Q^\star(s, \pi_{f_k})$ is the expected return of starting in s, takes π_{f_k} immediately (which coincides with $V^{\pi_{f_{k:1}}}(s)$ so far because π_{f_k} is $\pi_{f_{k:1}}$'s first-step policy), and acts optimally thereafter.

Now, the RHS looks like the performance guarantee of $\pi_{f_{k:1}}$, which we can directly apply the analysis in the previous section! We can also see that compared to the guarantee of non-stationary FQI, here we paying an extra $1/(1-\gamma)$ factor. The only unusual aspect is that here $d^{\pi_{f_k}}$ is treated as the initial distribution for the non-stationary FQI analysis, so finally, the distributions that need to be covered are those induced by the policies mentioned at the end of Section 2.1, but from $d_{\pi_{f_k}}$ as the initial distribution.

References

[1] Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

[2] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008.

[3] The truncated MDP should still have a discount factor. Also, one may need to define f_0 as the "terminal value" of this finite-horizon MDP. For simplicity we can consider the case where $f_0 \equiv 0$.

[3] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

[4] Sham Kakade. *Hoeffding, Chernoff, Bennet, and Bernstein Bounds*, 2011. `http://stat.wharton.upenn.edu/~skakade/courses/stat928/lectures/lecture06.pdf`.

[5] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1042–1051, 2019.

[6] Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

[7] Bruno Scherrer and Boris Lesner. On the use of non-stationary policies for stationary infinite-horizon markov decision processes. In *Advances in Neural Information Processing Systems*, pages 1826–1834, 2012.