# Notes on Fitted Q-iteration

Nan Jiang

October 16, 2018

Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho_0)$ be an MDP, where $\rho_0$ is the initial distribution over states.[1] 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. $\mathcal{F}$ is closed under Bellman update: $\forall f \in \mathcal{F}$, $\mathcal{T} f \in \mathcal{F}$. (For finite $\mathcal{F}$, this implies realizability.)
4. The dataset $D = \{(s, a, r, s')\}$ is generated as follows: $(s, a) \sim \mu \times U$ ($U$ is uniform over actions), $r \sim R(s, a)$, $s' \sim P(s, a)$. Define the empirical update $\widehat{\mathcal{T}}_{\mathcal{F}} f'$ as

$$\mathcal{L}_D(f; f') := \frac{1}{|D|} \sum_{(s,a,r,s')\in D} \left(f(s, a) - r - \gamma V_{f'}(s')\right)^2 .$$
$$\widehat{\mathcal{T}}_{\mathcal{F}} f' := \arg\min_{f\in\mathcal{F}} \mathcal{L}_D(f; f'),$$

   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

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

5. For any function $g : \mathcal{S} \to \mathbb{R}$, any distribution $\nu \in \Delta(\mathcal{S})$, 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 $\mathcal{S}\times\mathcal{A}$.
6. Let $\eta_h^\pi$ be the distribution of $s_h$ under $\pi$, that is, $\eta_h^\pi(s) := \Pr[s_h = s \mid s_1 \sim \rho_0, \pi]$.
7. $\mu$ is exploratory: for a distribution $\nu \in \Delta(\mathcal{S})$ generated by any (non-stationary) policy at any time step (that is, any distribution $\nu$ of the form $\eta_h^\pi$ where $\pi$ may be non-stationary),

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

[1] In previous notes we have used $\mu$ for the initial distribution; for this note we will reserve $\mu$ for the data distribution.

As a consequence, $\|\cdot\|_\nu \leq \sqrt{C}\|\cdot\|_\mu$. Similarly, when we couple $\mu$ with a uniform distribution over $\mathcal{A}$, we have similar results for state-action distributions: $\|\cdot\|_{\nu\times\pi} \leq \sqrt{|\mathcal{A}|C}\,\|\cdot\|_{\mu\times U}$. See slides for example scenarios where $C$ is naturally bounded.

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

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

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

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

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

**Analysis**

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

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

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 $\forall \nu \in \Delta(\mathcal{S})$,*

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

*Proof.*

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

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

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

Step (*) holds because:

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

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

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

Note that the RHS does not depend on $k$, so we conclude that

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

Apply this to Equation (1) and we get

$$v^\star - v^{\pi_{f_k}} \le \frac{2}{1-\gamma}\left(\frac{1-\gamma^k}{1-\gamma}\sqrt{2|\mathcal{A}|C\epsilon} + \gamma^k\frac{R_{\max}}{1-\gamma}\right).$$

**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

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

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

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

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 $\forall f\in\mathcal{F}$,

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

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

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

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 $1-\delta$, $\forall f \in \mathcal{F}$,

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

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,

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

# References


[1] Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 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] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 2008.

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