Lecture 7: Fitted Q-Iteration

In the previous lectures, we developed a thorough understanding of MDPs and their solutions: we proved the existence and uniqueness of $V^\star$ and $Q^\star$ via the Bellman equations (Lectures 1–2), designed efficient planning algorithms such as Value Iteration and Policy Iteration (Lectures 3–4), and analyzed the sample complexity of the certainty-equivalence algorithm for model-based RL (Lectures 5–6). All of these results, however, are fundamentally *tabular*: they require explicitly representing value functions as tables of size $|\mathcal{S}| \times |\mathcal{A}|$ and transition models as tables of size $|\mathcal{S}|^2 \times |\mathcal{A}|$. In this lecture, we ask: *what happens when the state space is too large—or even infinite—for tabular methods to be feasible?* This question leads us to the study of **function approximation** in RL.

From Tabular Methods to Function Approximation

The Curse of Dimensionality

Recall from Lecture 6 that the sample complexity of the certainty-equivalence algorithm scales as $\widetilde{O}(|\mathcal{S}| \cdot |\mathcal{A}|/(\varepsilon^2(1-\gamma)^2))$ (in terms of total samples). This means we need at least $|\mathcal{S}| \cdot |\mathcal{A}|$ samples just to visit each state-action pair once. In many practical problems, this is completely infeasible:

- **Board games** (e.g., Go): $|\mathcal{S}| \approx 10^{170}$.
- **Video games** (e.g., Atari): states are raw pixel images, giving $|\mathcal{S}| \approx 256^{210\times160\times3}$.
- **Robotics**: states are continuous (joint angles, velocities), so $|\mathcal{S}| = \infty$.
- **Language and dialog**: states are conversation histories, giving a combinatorially large $|\mathcal{S}|$.

The fundamental limitation of tabular methods: both the representation cost (storing $Q(s,a)$ for all (s,a)) and the statistical cost (needing data at every (s,a)) scale with $|\mathcal{S}| \times |\mathcal{A}|$. For large or continuous state-action spaces, we must *generalize* across states and actions, using compact representations that capture the structure of the value function.

Function Approximation for Q-Functions

The key idea is to replace the tabular representation of Q-functions with a **function class** $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$—a set of candidate Q-functions that we believe contains (or approximates) the true optimal Q-function. Each $f \in \mathcal{F}$ maps state-action pairs to real values: $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

Function Approximation: Instead of maintaining a table of $|\mathcal{S}| \times |\mathcal{A}|$ entries, we represent Q-functions using a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ with $|\mathcal{F}|$ potentially much smaller than $|\mathcal{S}| \times |\mathcal{A}|$ (in a statistical sense).

Examples of function classes.

- **Linear function approximation:** Given a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, define $\mathcal{F} = \{(s,a) \mapsto \phi(s,a)^\intercal w : w \in \mathbb{R}^d, \|w\| \le W\}$. The complexity of $\mathcal{F}$ is controlled by d rather than $|\mathcal{S}| \times |\mathcal{A}|$.
- **Neural networks:** $\mathcal{F}$ is the set of functions representable by a neural network of a given architecture. The complexity is controlled by the number of parameters.
- **Decision trees, tile coding, radial basis functions:** various other structured function classes used in practice.

Finite vs. infinite function classes. In this lecture, we will assume $\mathcal{F}$ is **finite** but can be **exponentially large**. This is a simplifying assumption that already captures the essential statistical phenomena. For example, a linear class $\{(s,a) \mapsto \phi(s,a)^\intercal w : w \in \mathbb{R}^d\}$ is infinite, but by discretizing the parameter space to precision δ, we can approximate it with a finite class of size $(1/\delta)^d$. The key quantity $\log|\mathcal{F}|$ then becomes $d\log(1/\delta)$, which is polynomial in d.

Boundedness. We assume that all functions in $\mathcal{F}$ are bounded: $\|f\|_\infty \le V_{\max}$ for all $f \in \mathcal{F}$, where $V_{\max} = R_{\max}/(1-\gamma)$. This is natural since $Q^\star$ satisfies this bound, and we design $\mathcal{F}$ to contain $Q^\star$.

Structural Assumptions

Working with function approximation introduces new challenges: the function class $\mathcal{F}$ may not contain the Bellman backup of every function in $\mathcal{F}$, and the data distribution may not match the distribution under the optimal policy. We formalize these issues with two key structural assumptions.

Definition 1 (Realizability)**.** *We say $\mathcal{F}$ satisfies **realizability** if $Q^\star \in \mathcal{F}$.*

Realizability says that the function class is "rich enough" to contain the optimal Q-function. Without it, there is an inherent approximation error that cannot be reduced by collecting more data.

Definition 2 (Bellman Completeness). *We say $\mathcal{F}$ satisfies **Bellman completeness** if $\forall f \in \mathcal{F}$, $\mathcal{T} f \in \mathcal{F}$, where $\mathcal{T}$ is the Bellman optimality operator $(\mathcal{T} f)(s,a) = \mathbb{E}[r \mid s,a] + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s,a)}[\max_{a'} f(s',a')]$.*

Bellman completeness says that the function class is "closed" under the Bellman operator: applying one step of Bellman backup to any function in $\mathcal{F}$ yields another function in $\mathcal{F}$. For finite $\mathcal{F}$, completeness implies realizability (since $Q^\star$ is the fixed point of $\mathcal{T}$ and the iterates $\mathcal{T}^k\mathbf{0}$ converge to $Q^\star$; by completeness, each iterate is in $\mathcal{F}$, and since $\mathcal{F}$ is finite and closed, $Q^\star \in \mathcal{F}$).

Completeness is stronger than realizability. Realizability only requires $Q^\star \in \mathcal{F}$, while completeness requires that the Bellman backup of *every* $f \in \mathcal{F}$ stays in $\mathcal{F}$. Completeness is naturally satisfied in several important settings:

- **Tabular:** $\mathcal{F} = \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ trivially satisfies completeness.
- **Linear MDPs:** if $\mathcal{P}(s' \mid s,a) = \langle \phi(s,a), \mu(s') \rangle$ and $r(s,a) = \langle \phi(s,a), \theta_r \rangle$ for a feature map ϕ, then the linear class $\mathcal{F} = \{(s,a) \mapsto \phi(s,a)^\top w\}$ satisfies completeness.

However, completeness can fail for generic function classes, and relaxing it is an important research direction (which we will briefly discuss).

The Batch Setting and Data Distribution

We consider a **batch** (or **offline**) setting: we are given a fixed dataset of transition tuples, and our goal is to learn a good policy from this data without further interaction with the environment.

Data generation. The dataset $D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ is generated i.i.d. as follows:

$$(s,a) \sim \mu, \quad r \sim R(s,a), \quad s' \sim \mathcal{P}(s,a),$$

where $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ is the **data distribution** (also called the behavior distribution). We do not assume μ is the occupancy of any particular policy—it can be any distribution over state-action pairs.

The distribution mismatch problem. A key challenge in batch RL is that the data distribution μ may differ significantly from the state-action distribution induced by the optimal policy $\pi^\star$. If μ does not cover the states and actions visited by $\pi^\star$, we cannot hope to learn $\pi^\star$ reliably. This motivates the following assumption.

Definition 3 (Concentrability Coefficient). *Let $d_h^\pi(s,a) := \Pr[s_h = s, a_h = a \mid s_0 \sim d_0, \pi]$ be the state-action distribution at step h under policy π (where d_0 is the initial state distribution), for*

*$h = 0, 1, 2, \ldots$ We call a state-action distribution ν **admissible** if it takes the form d_h^π for some step h and some (possibly non-stationary) policy π. The **concentrability coefficient** is defined as*

$$C := \max_{\nu \text{ admissible}} \max_{s,a} \frac{\nu(s,a)}{\mu(s,a)}.$$

The concentrability coefficient C measures how well the data distribution μ covers all state-action distributions that could arise under any policy from the initial distribution d_0. A small C means μ provides good coverage; a large C indicates distribution mismatch.

When is C bounded? The concentrability coefficient is naturally bounded in several scenarios:

- **Uniform exploration:** if μ is uniform over $\mathcal{S} \times \mathcal{A}$, then $C = |\mathcal{S}| \cdot |\mathcal{A}|$.
- **Generative model:** if we have the ability to query any (s, a) pair and can choose μ to cover all relevant distributions.
- **Ergodic MDPs:** in strongly mixing MDPs, any policy's occupancy is close to the stationary distribution, so C can be bounded independently of the horizon.

A fundamental consequence of bounded concentrability is that for any admissible ν and any function g, we have $\|g\|_\nu \le \sqrt{C}\|g\|_\mu$. This "change of measure" inequality will be used repeatedly in our analysis.

Notation

Before proceeding, we collect some notation that will be used throughout.

- For any function $g : \mathcal{X} \to \mathbb{R}$, any distribution $\nu \in \Delta(\mathcal{X})$, and $p \ge 1$, define $\|g\|_{p,\nu} := (\mathbb{E}_{x\sim\nu}[|g(x)|^p])^{1/p}$. We write $\|g\|_\nu$ as shorthand for $\|g\|_{2,\nu}$.
- $V_f(s) := \max_{a\in\mathcal{A}} f(s,a)$ for any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and $\pi_f(s) := \operatorname{argmax}_{a\in\mathcal{A}} f(s,a)$ is the greedy policy w.r.t. f.
- d^π denotes the discounted state-action occupancy of policy π: $d^\pi(s,a) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h d_h^\pi(s,a)$.
- $V_{\max} := R_{\max}/(1-\gamma)$ is the maximum possible value, where $R_{\max}$ is the bound on rewards.

Fitted Q-Iteration

We now introduce the algorithm we will analyze: **Fitted Q-Iteration** (FQI). FQI is one of the most natural and widely-used algorithms for batch RL with function approximation.

The key idea is to iteratively approximate $Q^\star$ by solving a sequence of regression problems.

The Algorithm

Recall that $Q^\star$ is the unique fixed point of the Bellman optimality operator $\mathcal{T}$: $Q^\star = \mathcal{T}Q^\star$. Value iteration computes $Q^\star$ by repeatedly applying $\mathcal{T}$: $f_k = \mathcal{T}f_{k-1}$. In the function approximation setting, we cannot compute $\mathcal{T}f_{k-1}$ exactly (because $\mathcal{T}f_{k-1}$ may not be in $\mathcal{F}$, and because we do not know the MDP model). Instead, we *project* the Bellman backup onto $\mathcal{F}$ using the data.

Define the **empirical loss** for fitting f to the Bellman backup of f':

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

This is a standard squared-error regression loss, where the "targets" are $r + \gamma V_{f'}(s') = r + \gamma \max_{a'} f'(s', a')$.

The **empirical Bellman update** operator is defined as:

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

Algorithm 1 Fitted Q-Iteration (FQI)

```
Require: Dataset $D = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$, function class $\mathcal{F}$, number of iterations $K$
Initialize $f_0 \equiv \mathbf{0}$                                    ▷ assuming $\mathbf{0} \in \mathcal{F}$
for $k = 1, 2, \ldots, K$ do
    $f_k \leftarrow \widehat{\mathcal{T}}_{\mathcal{F}} f_{k-1} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f(s_i, a_i) - r_i - \gamma V_{f_{k-1}}(s'_i)\big)^2$    ▷ least-squares regression
end for
return greedy policy $\widehat{\pi} := \pi_{f_K}$
```

Relationship to value iteration and Q-learning. FQI is the batch function approximation analogue of **value iteration**. Recall that value iteration performs $Q_{k+1} = \mathcal{T}Q_k$ exactly. FQI approximates this by projecting $\mathcal{T}f_{k-1}$ onto $\mathcal{F}$ via regression: $f_k = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_D(f; f_{k-1}) \approx \operatorname{argmin}_{f \in \mathcal{F}} \|f - \mathcal{T}f_{k-1}\|_\mu^2$. **Q-learning** can be viewed as an online, stochastic variant: it uses a single sample to perform an incremental update toward $\mathcal{T}f$, whereas FQI uses the entire dataset in each iteration.

Connection to Supervised Learning

Each iteration of FQI is a standard **least-squares regression** problem: given input-output pairs $\{(s_i, a_i),\ r_i + \gamma \max_{a'} f_{k-1}(s_i', a')\}_{i=1}^n$, find $f \in \mathcal{F}$ that minimizes the squared prediction error. This is exactly the problem solved by supervised learning methods (e.g., linear regression, neural network training).

> **FQI reduces RL to a sequence of supervised learning problems.** At each iteration, we construct regression targets $y_i = r_i + \gamma \max_{a'} f_{k-1}(s_i', a')$ using the current iterate f_{k-1}, then solve a regression problem to obtain f_k. The key difference from standard supervised learning is that the targets y_i change across iterations (they depend on f_{k-1}), creating a "moving target" effect. Under Bellman completeness, the Bellman backup $\mathcal{T} f_{k-1} \in \mathcal{F}$ is the conditional mean of the regression target, and hence the optimal regressor.

Population Loss

It will be convenient to define the **population loss** (the expected version of $\mathcal{L}_D$):

$$\mathcal{L}_\mu(f; f') := \mathbb{E}_D[\mathcal{L}_D(f; f')] = \mathbb{E}_{(s,a)\sim\mu,\, r\sim R(s,a),\, s'\sim\mathcal{P}(s,a)}\left[(f(s,a) - r - \gamma V_{f'}(s'))^2\right].$$

Note that $(\mathcal{T} f')(s,a) = \mathbb{E}[r + \gamma V_{f'}(s') \mid s, a]$ is the conditional mean of the regression target given (s,a), and hence the minimizer of $\mathcal{L}_\mu(\cdot\,; f')$ over all measurable functions. By Bellman completeness, $\mathcal{T} f' \in \mathcal{F}$, so it is also the minimizer over $\mathcal{F}$.

Analysis of FQI

We now analyze the suboptimality $J(\pi^\star) - J(\widehat{\pi})$ of the policy returned by FQI. This analysis follows the approach of Munos (2003); Antos et al. (2008); Munos and Szepesvári (2008), with simplifications due to our finite-class and completeness assumptions.[1]

Uniform Deviation Bound

Our analysis relies on a uniform convergence assumption that controls how well the empirical loss $\mathcal{L}_D$ approximates the population loss $\mathcal{L}_\mu$.

Assumption 1 (Uniform Deviation). *For the dataset D of size $n = |D|$, we assume*

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

[1] The presentation in this section closely follows notes by Nan Jiang.

This assumption can be justified by standard concentration inequalities and a union bound. Since $\mathcal{F}$ is finite, for each pair (f, f'), $\mathcal{L}_D(f; f')$ is an average of n i.i.d. bounded random variables. By Hoeffding's inequality and a union bound over all $|\mathcal{F}|^2$ pairs, we obtain $\varepsilon = O(V_{\max}^2 \sqrt{\log |\mathcal{F}|/n})$ with high probability. We will later show how to obtain a sharper ("fast rate") bound.

The Performance Difference Lemma

The following identity expresses the value gap between two policies as a weighted sum of per-step suboptimality terms. It is one of the most useful tools in RL theory.

Lemma 1 (Performance Difference Lemma)**.** *For any two policies π' and π, and any state $s \in \mathcal{S}$:*

$$V^{\pi'}(s) - V^{\pi}(s) = \sum_{h=0}^{\infty} \gamma^h \, \mathbb{E}\big[V^{\pi'}(s_h) - Q^{\pi'}(s_h, \pi(s_h)) \mid s_0 = s\big], \tag{1}$$

where $s_{h+1} \sim \mathcal{P}(s_h, \pi(s_h))$, i.e., the trajectory is generated by policy π. Taking expectation over $s_0 \sim d_0$:

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma} \, \mathbb{E}_{s \sim d^{\pi}} \big[V^{\pi'}(s) - Q^{\pi'}(s, \pi(s))\big]. \tag{2}$$

Proof. Define $\delta(s) \coloneqq V^{\pi'}(s) - V^{\pi}(s)$. Decompose:

$$\delta(s) = \big[V^{\pi'}(s) - Q^{\pi'}(s, \pi(s))\big] + \big[Q^{\pi'}(s, \pi(s)) - V^{\pi}(s)\big].$$

For the second term, since $V^{\pi}(s) = Q^{\pi}(s, \pi(s))$:

$$\begin{aligned} Q^{\pi'}(s, \pi(s)) - V^{\pi}(s) &= Q^{\pi'}(s, \pi(s)) - Q^{\pi}(s, \pi(s)) \\ &= \Big(R(s, \pi(s)) + \gamma \, \mathbb{E}_{s' \sim \mathcal{P}(s, \pi(s))}[V^{\pi'}(s')]\Big) \\ &\qquad - \Big(R(s, \pi(s)) + \gamma \, \mathbb{E}_{s' \sim \mathcal{P}(s, \pi(s))}[V^{\pi}(s')]\Big) \\ &= \gamma \, \mathbb{E}_{s' \sim \mathcal{P}(s, \pi(s))}\big[\delta(s')\big]. \end{aligned}$$

Substituting back:

$$\delta(s) = \big[V^{\pi'}(s) - Q^{\pi'}(s, \pi(s))\big] + \gamma \, \mathbb{E}_{s' \sim \mathcal{P}(s, \pi(s))}[\delta(s')].$$

Let $s_0 = s$ and $s_{h+1} \sim \mathcal{P}(s_h, \pi(s_h))$. Unrolling the recursion:

$$\begin{aligned} \delta(s_0) &= \big[V^{\pi'}(s_0) - Q^{\pi'}(s_0, \pi(s_0))\big] + \gamma \, \mathbb{E}\big[\delta(s_1) \mid s_0\big] \\ &= \big[V^{\pi'}(s_0) - Q^{\pi'}(s_0, \pi(s_0))\big] + \gamma \, \mathbb{E}\big[V^{\pi'}(s_1) - Q^{\pi'}(s_1, \pi(s_1)) \mid s_0\big] + \gamma^2 \, \mathbb{E}\big[\delta(s_2) \mid s_0\big] \end{aligned}$$

$$\begin{aligned} &\vdots \\ &= \sum_{h=0}^{H-1} \gamma^h \, \mathbb{E}\big[V^{\pi'}(s_h) - Q^{\pi'}(s_h, \pi(s_h)) \mid s_0\big] + \gamma^H \, \mathbb{E}\big[\delta(s_H) \mid s_0\big]. \end{aligned}$$

Since $|\delta(s)| \le 2V_{\max}$ for all s, the remainder $\gamma^H \mathbb{E}[\delta(s_H) \mid s_0] \to 0$ as $H \to \infty$, yielding (1).

For (2), take $\mathbb{E}_{s_0 \sim d_0}$ of both sides of (1):

$$\begin{aligned} J(\pi') - J(\pi) = \mathbb{E}_{s_0 \sim d_0}[\delta(s_0)] &= \sum_{h=0}^{\infty} \gamma^h \, \mathbb{E}_{s \sim d_h^\pi}\big[V^{\pi'}(s) - Q^{\pi'}(s, \pi(s))\big] \\ &= \frac{1}{1-\gamma} \, \mathbb{E}_{s \sim d^\pi}\big[V^{\pi'}(s) - Q^{\pi'}(s, \pi(s))\big], \end{aligned}$$

where the last equality uses $d^\pi = (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \, d_h^\pi$. □

When $\pi' = \pi^\star$, each summand in (1) is non-negative, since $V^\star(s) = \max_a Q^\star(s,a) \ge Q^\star(s, \pi(s))$ for any π.

The classical form. An equivalent statement, due to Kakade and Langford (2002), rolls out the trajectory under π' instead of π:

$$V^{\pi'}(s) - V^\pi(s) = \sum_{h=0}^{\infty} \gamma^h \, \mathbb{E}_{s_h \sim \pi'}\big[A^\pi(s_h, \pi'(s_h))\big],$$

where $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$ is the *advantage function* of π, and $s_0 = s$ with $s_{h+1} \sim \mathcal{P}(s_h, \pi'(s_h))$. The proof uses an analogous telescoping argument; see Kakade and Langford 2002 or the MDP Preliminaries notes.

Performance Decomposition

Our goal is to bound $J(\pi^\star) - J(\widehat{\pi})$, where $\widehat{\pi} = \pi_{f_K}$ is the greedy policy w.r.t. the final iterate f_K. We use k to denote the number of iterations (and drop the distinction between k and K for notational simplicity).

Step 1: Applying the Performance Difference Lemma. By Lemma 1 with $\pi' = \pi^\star$ and $\pi = \widehat{\pi}$, taking expectation over $s_0 \sim d_0$:

$$J(\pi^\star) - J(\widehat{\pi}) = \sum_{h=0}^{\infty} \gamma^h \, \mathbb{E}_{s \sim d_h^{\widehat{\pi}}}\big[V^\star(s) - Q^\star(s, \widehat{\pi}(s))\big]. \tag{3}$$

Each summand is non-negative since $V^\star(s) = \max_a Q^\star(s,a) \ge Q^\star(s, \widehat{\pi}(s))$.

Step 2: Adding and subtracting f_k. Now we add and subtract f_k terms inside each expectation:

$$\begin{aligned} J(\pi^\star) - J(\widehat{\pi}) &= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s \sim d_h^{\widehat{\pi}}} \big[V^\star(s) - Q^\star(s, \widehat{\pi}(s)) \big] \\ &\le \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s \sim d_h^{\widehat{\pi}}} \big[Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)) + f_k(s, \widehat{\pi}(s)) - Q^\star(s, \widehat{\pi}(s)) \big]. \end{aligned}$$

The inequality holds because $V^\star(s) = Q^\star(s, \pi^\star(s))$ and $f_k(s, \widehat{\pi}(s)) = V_{f_k}(s) \ge f_k(s, a)$ for all a (since $\widehat{\pi}$ is greedy w.r.t. f_k), so $f_k(s, \widehat{\pi}(s)) - f_k(s, \pi^\star(s)) \ge 0$.

Applying the triangle inequality and Cauchy–Schwarz ($\|g\|_{1,\nu} \le \|g\|_\nu$), we obtain:

$$\begin{aligned} J(\pi^\star) - J(\widehat{\pi}) &\le \sum_{h=0}^{\infty} \gamma^h \Big(\|Q^\star - f_k\|_{1, d_h^{\widehat{\pi}} \times \pi^\star} + \|Q^\star - f_k\|_{1, d_h^{\widehat{\pi}} \times \widehat{\pi}} \Big) \\ &\le \sum_{h=0}^{\infty} \gamma^h \Big(\|Q^\star - f_k\|_{d_h^{\widehat{\pi}} \times \pi^\star} + \|Q^\star - f_k\|_{d_h^{\widehat{\pi}} \times \widehat{\pi}} \Big). \end{aligned} \tag{4}$$

Here $d_h^{\widehat{\pi}}$ is treated as a state distribution, and $d_h^{\widehat{\pi}} \times \pi'$ denotes the state-action distribution obtained by sampling $s \sim d_h^{\widehat{\pi}}$ and $a \sim \pi'(\cdot \mid s)$. Both terms in (4) have the form $\|Q^\star - f_k\|_\nu$ for some admissible distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$. So it remains to bound $\|Q^\star - f_k\|_\nu$ for any admissible ν.

A Helper Lemma

The following lemma allows us to relate value-function differences to Q-function differences. It will be used to handle the max operator when we apply the Bellman contraction.

Lemma 2 (V-function to Q-function reduction)**.** *For any $f, f_k \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, define $\pi_{f,f_k}(s) := \operatorname{argmax}_{a \in \mathcal{A}} \max\{f(s,a), f_k(s,a)\}$. Then for all $\tilde{\nu} \in \Delta(\mathcal{S})$,*

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

Proof. We compute:

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

The inequality holds because for $a^\star = \pi_{f,f_k}(s)$:

$$\big|\max_a f(s,a) - \max_a f_k(s,a)\big| \le \big|f(s,a^\star) - f_k(s,a^\star)\big|.$$

To see this, let $M \coloneqq \max_a f(s,a)$, $N \coloneqq \max_a f_k(s,a)$, and $g(a) \coloneqq \max\{f(s,a), f_k(s,a)\}$, so $a^\star \in \operatorname{argmax}_a g(a)$. Assume WLOG $M \ge N$. Then $\max_a g(a) = \max\{M,N\} = M$, so $g(a^\star) = M$. If $M > N$, then $f_k(s,a) \le N < M$ for all a, so the value M at $a^\star$ must be achieved by f:

$$f(s,a^\star) = M, \qquad \text{and thus} \qquad f(s,a^\star) - f_k(s,a^\star) = M - f_k(s,a^\star) \ge M - N.$$

If $M = N$, then $|M-N| = 0$ and the claim is trivial. □

Recursive Error Bound

We now bound $\|f_k - Q^\star\|_\nu$ for any admissible ν. Define $\mathcal{P}(\nu)$ as the next-state distribution induced by ν: $s' \sim \mathcal{P}(\nu)$ means $(s,a) \sim \nu$, $s' \sim \mathcal{P}(s,a)$.

Step 1: Decompose using the triangle inequality.

$$\begin{aligned}\|f_k - Q^\star\|_\nu &= \|f_k - \mathcal{T}f_{k-1} + \mathcal{T}f_{k-1} - Q^\star\|_\nu \\ &\le \|f_k - \mathcal{T}f_{k-1}\|_\nu + \|\mathcal{T}f_{k-1} - \mathcal{T}Q^\star\|_\nu.\end{aligned}$$

Step 2: Bound the first term via change of measure. Since ν is admissible and $\nu(s,a)/\mu(s,a) \le C$:

$$\|f_k - \mathcal{T}f_{k-1}\|_\nu \le \sqrt{C}\,\|f_k - \mathcal{T}f_{k-1}\|_\mu.$$

Step 3: Bound the second term via Bellman contraction. We claim that

$$\|\mathcal{T}f_{k-1} - \mathcal{T}Q^\star\|_\nu \le \gamma\|V_{f_{k-1}} - V^\star\|_{\mathcal{P}(\nu)}. \tag{5}$$

To see this:

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

Step 4: Apply Lemma 2 to handle the max.

$$\|V_{f_{k-1}} - V^\star\|_{\mathcal{P}(\nu)} \le \|f_{k-1} - Q^\star\|_{\mathcal{P}(\nu)\times\pi_{f_{k-1},Q^\star}}.$$

Combining Steps 1–4:

$$\|f_k - Q^\star\|_\nu \le \sqrt{C}\,\|f_k - \mathcal{T} f_{k-1}\|_\mu + \gamma\|f_{k-1} - Q^\star\|_{\mathcal{P}(\nu)\times\pi_{f_{k-1},Q^\star}}. \tag{6}$$

Since $\mathcal{P}(\nu) \times \pi_{f_{k-1},Q^\star}$ is also an admissible distribution, we can apply the same bound recursively. Expanding k times and using $\|f_0 - Q^\star\|_\nu \le V_{\max}$ (since $f_0 \equiv 0$), we get:

$$\|f_k - Q^\star\|_\nu \le \sqrt{C}\sum_{j=0}^{k-1}\gamma^j\|f_{k-j} - \mathcal{T} f_{k-j-1}\|_\mu + \gamma^k V_{\max}. \tag{7}$$

Bounding the Per-Iteration Error

It remains to bound $\|f_k - \mathcal{T} f_{k-1}\|_\mu$ for each iteration. This is where the Bellman completeness assumption and the uniform deviation bound come into play.

Step 1: Squared loss decomposition. Write the regression target as $Y = r + \gamma V_{f_{k-1}}(s')$. Then $\mathcal{T} f_{k-1}(s,a) = \mathbb{E}[Y \mid s,a]$ is the conditional mean of Y given (s,a). For any f, add and subtract $\mathcal{T} f_{k-1}(s,a)$:

$$\begin{aligned}
\mathcal{L}_\mu(f; f_{k-1}) &= \mathbb{E}_\mu\big[(f(s,a) - Y)^2\big] \\
&= \mathbb{E}_\mu\Big[\big((f(s,a) - \mathcal{T} f_{k-1}(s,a)) + (\mathcal{T} f_{k-1}(s,a) - Y)\big)^2\Big] \\
&= \mathbb{E}_\mu\big[(f(s,a) - \mathcal{T} f_{k-1}(s,a))^2\big] + 2\,\mathbb{E}_\mu\big[(f(s,a) - \mathcal{T} f_{k-1}(s,a))(\mathcal{T} f_{k-1}(s,a) - Y)\big] \\
&\quad + \mathbb{E}_\mu\big[(\mathcal{T} f_{k-1}(s,a) - Y)^2\big].
\end{aligned}$$

The cross term vanishes: conditioned on (s,a), the factor $f(s,a) - \mathcal{T} f_{k-1}(s,a)$ is a constant, and $\mathbb{E}[Y \mid s,a] = \mathcal{T} f_{k-1}(s,a)$ implies $\mathbb{E}[\mathcal{T} f_{k-1}(s,a) - Y \mid s,a] = 0$. So

$$\mathcal{L}_\mu(f; f_{k-1}) = \|f - \mathcal{T} f_{k-1}\|_\mu^2 + \mathcal{L}_\mu(\mathcal{T} f_{k-1}; f_{k-1}),$$

where $\mathcal{L}_\mu(\mathcal{T} f_{k-1}; f_{k-1}) = \mathbb{E}_\mu[(\mathcal{T} f_{k-1}(s,a) - Y)^2]$ is the irreducible noise (independent of f). Rearranging:

$$\|f_k - \mathcal{T} f_{k-1}\|_\mu^2 = \mathcal{L}_\mu(f_k; f_{k-1}) - \mathcal{L}_\mu(\mathcal{T} f_{k-1}; f_{k-1}).$$

Step 2: Uniform deviation and optimality of f_k.

$$\begin{aligned}
\|f_k - \mathcal{T} f_{k-1}\|_\mu^2 &= \mathcal{L}_\mu(f_k; f_{k-1}) - \mathcal{L}_\mu(\mathcal{T} f_{k-1}; f_{k-1}) \\
&\le \mathcal{L}_D(f_k; f_{k-1}) + \varepsilon - (\mathcal{L}_D(\mathcal{T} f_{k-1}; f_{k-1}) - \varepsilon) \\
&\qquad \text{(Assumption 1: } f_k \in \mathcal{F} \text{ by construction, } \mathcal{T} f_{k-1} \in \mathcal{F} \text{ by Bellman completeness)} \\
&= \mathcal{L}_D(f_k; f_{k-1}) - \mathcal{L}_D(\mathcal{T} f_{k-1}; f_{k-1}) + 2\varepsilon \\
&\le 2\varepsilon. \qquad\qquad (f_k = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_D(f; f_{k-1}) \text{ and } \mathcal{T} f_{k-1} \in \mathcal{F})
\end{aligned}$$

Crucially, the right-hand side does not depend on k: we have $\|f_k - \mathcal{T} f_{k-1}\|_\mu \le \sqrt{2\varepsilon}$ for *every* iteration.

Main Result

Substituting $\|f_{k-j} - \mathcal{T} f_{k-j-1}\|_\mu \le \sqrt{2\varepsilon}$ into (7):

Per-iterate error bound: For any admissible $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$,

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

Applying this to (4), we obtain our main theorem:

Theorem 3 (FQI Guarantee)**.** *Under realizability, Bellman completeness, bounded concentrability C, and the uniform deviation bound (Assumption 1), the policy $\widehat{\pi} = \pi_{f_k}$ returned by FQI after k iterations satisfies:*

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

Proof. From (4), each term $\|Q^\star - f_k\|_\nu$ with admissible ν is bounded by the per-iterate error bound. Summing the geometric series $\sum_{h=0}^\infty \gamma^h = 1/(1-\gamma)$ and noting that both terms in (4) have the same bound, we get the factor of $2/(1-\gamma)$. □

Choosing the number of iterations. The bound has two terms: the first grows with k (statistical error accumulation), and the second decays with k (optimization error from initialization). To balance them, we choose k such that $\gamma^k V_{\max} \approx \sqrt{2C\varepsilon}/(1-\gamma)$, which gives $k \approx \frac{1}{1-\gamma}\log\left(\frac{V_{\max}(1-\gamma)}{\sqrt{2C\varepsilon}}\right)$.

The slow rate problem. With the Hoeffding-based uniform deviation bound $\varepsilon = \widetilde{O}(V_{\max}^2/\sqrt{n})$, the FQI guarantee becomes

$$J(\pi^\star) - J(\widehat{\pi}) \le \widetilde{O}\left(\frac{\sqrt{C} \cdot V_{\max}}{(1-\gamma)^2 \cdot n^{1/4}}\right).$$

The $n^{-1/4}$ rate arises because the bound depends on $\sqrt{\varepsilon}$ and ε itself scales as $n^{-1/2}$. This is a *slow rate*—can we do better? In the next section, we show how to achieve the faster $n^{-1/2}$ rate by exploiting the variance structure of the problem.

Fast Rate via Bernstein's Inequality

The slow rate in the previous section arose because we used a crude uniform deviation bound that treats $\mathcal{L}_D(f; f') - \mathcal{L}_\mu(f; f')$ uniformly for all f. However, the per-iteration error $\|f_k - \mathcal{T} f_{k-1}\|_\mu^2$ has a special self-bounding structure: the variance of the relevant random variable is proportional to its expectation. By exploiting this via Bernstein's inequality, we can achieve a *fast rate* of $O(n^{-1/2})$ in the final bound.

The Y-Variable Trick

The key idea is to work with a different random variable that captures the excess loss. Define:

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

For each $(s_i, a_i, r_i, s_i') \in D$, plugging in gives i.i.d. random variables $Y_1(f; f'), Y_2(f; f'), \ldots, Y_n(f; f')$.

Empirical average. It is easy to verify that

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

since Y is the difference of squared losses and the term $\mathcal{L}_D(\mathcal{T} f'; f')$ is f-independent.

Population mean. By realizability and the squared-loss decomposition:

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

So the population mean of Y is exactly the quantity we want to bound: $\|\widehat{\mathcal{T}}_{\mathcal{F}} f' - \mathcal{T} f'\|_\mu^2$ (when $f = \widehat{\mathcal{T}}_{\mathcal{F}} f'$).

Variance Bound

The crucial property of Y is that its variance is controlled by its mean.

Lemma 4 (Variance-mean relationship). *For any $f, f' \in \mathcal{F}$,*

$$\mathrm{Var}[Y(f;f')] \leq 4V_{\max}^2 \cdot \mathbb{E}[Y(f;f')].$$

Proof. We bound the variance by the second moment:

$$\begin{aligned}\mathrm{Var}[Y(f;f')] &\leq \mathbb{E}[Y(f;f')^2] \\ &= \mathbb{E}\Big[\Big(\big(f(s,a) - r - \gamma V_{f'}(s')\big)^2 - \big((\mathcal{T}f')(s,a) - r - \gamma V_{f'}(s')\big)^2\Big)^2\Big].\end{aligned}$$

Using the factorization $A^2 - B^2 = (A-B)(A+B)$ with $A = f(s,a) - r - \gamma V_{f'}(s')$ and $B = (\mathcal{T}f')(s,a) - r - \gamma V_{f'}(s')$:

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

□

Applying Bernstein's Inequality

We now apply the one-sided Bernstein inequality together with a union bound over all $f \in \mathcal{F}$. Let $N = |\mathcal{F}|$.

Theorem 5 (Bernstein's Inequality, one-sided). *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $|X_i| \leq b$. Then with probability at least $1-\delta$:*

$$\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^n X_i \leq \sqrt{\frac{2\mathrm{Var}[X]\log(1/\delta)}{n}} + \frac{b\log(1/\delta)}{3n}.$$

For any fixed f', applying Bernstein's inequality to $Y_1(f;f'), \ldots, Y_n(f;f')$ for each $f \in \mathcal{F}$ (note $|Y_i| \leq 4V_{\max}^2$) and taking a union bound, we get that with probability at least $1-\delta$, for all $f \in \mathcal{F}$:

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

$$\leq \sqrt{\frac{8V_{\max}^2\,\mathbb{E}[Y(f;f')]\log\frac{N}{\delta}}{n}} + \frac{4V_{\max}^2\log\frac{N}{\delta}}{3n},$$

where the second line uses the variance bound from Lemma 4.

Using optimality of $\widehat{\mathcal{T}}_{\mathcal{F}} f'$. Recall that $\widehat{\mathcal{T}}_{\mathcal{F}} f' = \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}_D(f;f')$. Since $\widehat{\mathcal{T}}_{\mathcal{F}} f'$ minimizes $\mathcal{L}_D(\cdot\,;f')$ over $\mathcal{F}$, it also minimizes $\frac{1}{n}\sum_{i=1}^n Y_i(\cdot\,;f')$ (because the two objectives differ by the f-independent constant $\mathcal{L}_D(\mathcal{T}f';f')$). Therefore:

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

where the equality holds because $Y(\mathcal{T}f';f') = \mathcal{L}_D(\mathcal{T}f';f') - \mathcal{L}_D(\mathcal{T}f';f') = 0$ for each sample.

Substituting into the Bernstein bound with $f = \widehat{\mathcal{T}}_{\mathcal{F}} f'$:

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

Solving the self-bounding inequality. Let $x = \mathbb{E}[Y(\widehat{\mathcal{T}}_{\mathcal{F}} f';f')]$ and $\beta = V_{\max}^2\log(N/\delta)/n$. The inequality becomes $x \leq \sqrt{8\beta x} + \frac{4}{3}\beta$. By AM-GM, $\sqrt{8\beta x} \leq \frac{x}{2} + 4\beta$, so

$$x \leq \frac{x}{2} + 4\beta + \frac{4}{3}\beta = \frac{x}{2} + \frac{16}{3}\beta.$$

Rearranging gives $x \leq \frac{32}{3}\beta = \frac{32}{3}\cdot\frac{V_{\max}^2\log(N/\delta)}{n}$. Taking the union bound over all $f' \in \mathcal{F}$ (so $N = |\mathcal{F}|$ and replacing δ by $\delta/|\mathcal{F}|$):

Fast rate for empirical Bellman update: With probability at least $1-\delta$,

$$\|\widehat{\mathcal{T}}_{\mathcal{F}} f' - \mathcal{T}f'\|_\mu^2 = \mathbb{E}[Y(\widehat{\mathcal{T}}_{\mathcal{F}} f';f')] \leq \frac{32}{3}\cdot\frac{V_{\max}^2\log\frac{|\mathcal{F}|}{\delta}}{n} = O\left(\frac{V_{\max}^2\log|\mathcal{F}|}{n}\right).$$

This result can be directly plugged into the analysis of Section 3. Setting $f' = f_{k-1}$ (so $\widehat{\mathcal{T}}_{\mathcal{F}} f' = f_k$), we get $\|f_k - \mathcal{T}f_{k-1}\|_\mu^2 = O(V_{\max}^2\log|\mathcal{F}|/n)$ instead of $O(\varepsilon) = O(V_{\max}^2\sqrt{\log|\mathcal{F}|/n})$. The final suboptimality bound becomes:

Theorem 6 (FQI with Fast Rate). *Under the same assumptions as Theorem 3, but using the*

Bernstein-based bound, we have with probability at least $1 - \delta$:

$$J(\pi^\star) - J(\widehat{\pi}) \leq \widetilde{O}\left(\frac{\sqrt{C} \cdot V_{\max}}{(1-\gamma)^2 \cdot \sqrt{n}}\right),$$

which achieves the faster $O(n^{-1/2})$ rate.[2]

Slow rate vs. fast rate: The improvement from $O(n^{-1/4})$ to $O(n^{-1/2})$ is significant. The slow rate arises from a naive uniform deviation bound that ignores the variance structure. The fast rate exploits the key insight that **the variance of the excess loss is proportional to its mean** (a "self-bounding" property), which enables Bernstein's inequality to yield a tighter bound. This is conceptually similar to the improvement from Hoeffding to Bernstein in the tabular analysis of Lecture 6.

Relaxing the Concentrability Coefficient

The concentrability coefficient C defined as $\max_\nu \|\nu/\mu\|_\infty$ (where ν ranges over admissible distributions) can be quite large, as it measures the worst-case density ratio over all state-action pairs. However, inspecting our analysis, we see that C always appears in the form:

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

for some $f, f' \in \mathcal{F}$. We can therefore *redefine* C as a tighter, function-class-dependent quantity:

Relaxed concentrability: Define

$$C_{\mathcal{F}} := \max_{\nu \text{ admissible}} \max_{f,f' \in \mathcal{F}} \frac{\|f - \mathcal{T}f'\|_\nu^2}{\|f - \mathcal{T}f'\|_\mu^2}.$$

All previous results hold with C replaced by $C_{\mathcal{F}}$.

When $\mathcal{F}$ has structural properties, $C_{\mathcal{F}}$ can be significantly tighter than the raw density-ratio-based C. For example, when $\mathcal{F}$ is linear and Bellman completeness holds, $f - \mathcal{T}f'$ is also a linear function of the features, and $C_{\mathcal{F}}$ measures coverage in the linear feature space rather than over the raw state-action pairs.

[2] Formally, applying a union bound over all k iterations. Since we run $K = O(1/(1-\gamma) \cdot \log n)$ iterations, this adds at most a $\log n$ factor.

Alternative Analysis via Performance Difference Lemma

We now sketch an alternative proof of the FQI guarantee that addresses two aesthetic shortcomings of the analysis in the previous sections:[3]

(1) **"Ugly" error-propagation policies.** The recursive error bound in (6) propagates errors along the policy $\pi_{f_{k-1},Q^\star}$, which in each state takes the action that "witnesses" the inequality $|\max_a f(s,a) - \max_a f'(s,a)| \le \max_a |f(s,a) - f'(s,a)|$. This is an artificial policy that does not correspond to any natural decision rule. In contrast, the ADP literature (Munos and Szepesvári, 2008) typically propagates errors along "simple" policies such as π_f for $f \in \mathcal{F}$.

(2) **Old-style recursive expansion.** The previous analysis proceeds by recursively expanding the error bound k times. Modern approaches in RL theory use cleaner tools based on the performance difference lemma.

The Performance Difference Lemma for Bellman Error

The following lemma provides a direct connection between the suboptimality of a policy π_f and the Bellman error of f.

Lemma 7 (Performance Difference via Bellman Error)**.** *For any policy π and any function $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

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

Proof left as an exercise.

Intuition. The first term $\mathbb{E}_{d^\pi}[\mathcal{T}f - f]$ measures the Bellman error of f under the comparator policy's distribution. The second term $\mathbb{E}_{d^{\pi_f}}[f - \mathcal{T}f]$ measures the Bellman error under the learned policy's distribution. If f is self-consistent (i.e., $\mathcal{T}f = f$, meaning $f = Q^{\pi_f}$), both terms vanish and π_f is at least as good as π. This lemma is well-aligned with modern RL theory because it only involves "simple" policies (π and π_f), making the coverage requirements more interpretable.

Non-Stationary FQI

A major difficulty in applying Lemma 7 to FQI is that it requires control over the *self-consistency* error $\|f - \mathcal{T}f\|$, i.e., the learned function should be consistent with its own Bellman backup. However, in FQI, we only control $\|f_k - \mathcal{T}f_{k-1}\|$: the iterate f_k is consistent with the *previous* iterate f_{k-1}, not with itself.

[3] This section follows Xie and Jiang (2020) and Chen and Jiang (2019); see also Scherrer and Lesner (2012).

To overcome this, we consider a different output policy: the **non-stationary policy**

$$\pi_{f_{k:0}} \coloneqq \pi_{f_k} \circ \pi_{f_{k-1}} \circ \cdots \circ \pi_{f_0},$$

which applies π_{f_k} at step 0, $\pi_{f_{k-1}}$ at step 1, and so on, with π_{f_0} at step k and arbitrary actions thereafter. This policy is greedy with respect to the **non-stationary value function** $f_{k:0} \coloneqq f_k \circ f_{k-1} \circ \cdots \circ f_0$, which assigns f_{k-h} as the Q-function at step h.

The key observation is that $f_{k:0}$ *is* self-consistent in the non-stationary sense: at each step h, the function f_{k-h} satisfies $f_{k-h} \approx \mathcal{T} f_{k-h-1}$, and f_{k-h-1} is exactly the next-step function. This allows us to apply a non-stationary variant of Lemma 7:

Lemma 8 (Non-stationary Performance Difference)**.** *Given an arbitrary sequence of functions $f_0, \ldots, f_k \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and any (non-stationary) comparator policy π, let $\widehat{\pi} \coloneqq \pi_{f_{k:0}}$ (followed by arbitrary actions after $k+1$ steps). Then:*

$$J(\pi) - J(\widehat{\pi}) \leq \sum_{h=0}^{k-1} \gamma^h \Big(\mathbb{E}_{d_h^\pi}[\mathcal{T} f_{k-h-1} - f_{k-h}] + \mathbb{E}_{d_h^{\widehat{\pi}}}[f_{k-h} - \mathcal{T} f_{k-h-1}] \Big) + \gamma^k V_{\max}.$$

Proof left as a homework exercise.

Applying the bound. Setting $\pi = \pi^\star$ and using $|\mathbb{E}_\nu[g]| \leq \|g\|_\nu$ (Cauchy–Schwarz) followed by change of measure $\|g\|_\nu \leq \sqrt{C}\|g\|_\mu$, each term $\mathbb{E}_{d_h^\pi}[\mathcal{T} f_{k-h-1} - f_{k-h}]$ and $\mathbb{E}_{d_h^{\widehat{\pi}}}[f_{k-h} - \mathcal{T} f_{k-h-1}]$ is bounded by $\sqrt{C}\|f_{k-h} - \mathcal{T} f_{k-h-1}\|_\mu$. Summing over h:

$$J(\pi^\star) - J(\widehat{\pi}) \leq \frac{1-\gamma^k}{1-\gamma} \cdot 2\sqrt{C} \max_{1 \leq j \leq k} \|f_j - \mathcal{T} f_{j-1}\|_\mu + \gamma^k V_{\max}.$$

Plugging in $\|f_j - \mathcal{T} f_{j-1}\|_\mu \leq \sqrt{2\varepsilon}$ (slow rate) or $\|f_j - \mathcal{T} f_{j-1}\|_\mu = O(V_{\max}\sqrt{\log|\mathcal{F}|/n})$ (fast rate):

$$J(\pi^\star) - J(\widehat{\pi}) \leq \begin{cases} \dfrac{1-\gamma^k}{1-\gamma} \cdot 2\sqrt{2C\varepsilon} + \gamma^k V_{\max} & \text{(slow rate)}, \\[2ex] \dfrac{1-\gamma^k}{1-\gamma} \cdot O\left(\sqrt{C} \cdot V_{\max}\sqrt{\dfrac{\log|\mathcal{F}|}{n}}\right) + \gamma^k V_{\max} & \text{(fast rate)}. \end{cases}$$

Non-stationary FQI saves a $1/(1-\gamma)$ factor. Comparing with the guarantee for stationary FQI (Theorem 3), the non-stationary FQI bound is better by a factor of $1/(1-\gamma)$. The improvement comes from the fact that the non-stationary policy $\pi_{f_{k:0}}$ automatically terminates after k steps, avoiding the infinite-horizon accumulation of errors. This mirrors the observation (from Lecture 3) that non-stationary value iteration converges faster than stationary value iteration.

Coverage requirement. The distributions that need to be covered by μ are those induced by two types of policies from d_0: $(\pi^\star)^t$ for $t \le k$, and $\pi_{f_{k:k'}}$ for $0 \le k' \le k$. Importantly, when analyzing the minimax algorithm (Xie and Jiang, 2020; Chen and Jiang, 2019) using Lemma 7, we only need μ to cover the *discounted occupancy as a whole*, rather than the per-step distributions separately. Here, we do not enjoy this luxury because our algorithm only controls $\|f_t - \mathcal{T} f_{t-1}\|_\mu$ for each t separately.

From Non-Stationary to Stationary FQI

The non-stationary policy $\pi_{f_{k:0}}$ requires remembering all past iterates $f_0, \ldots, f_k$ and using a different policy at each step. In practice, we often prefer the simpler stationary policy π_{f_k}. We can relate the performance of stationary FQI to non-stationary FQI using Lemma 1.

By the Performance Difference Lemma (equation (2)) with $\pi' = \pi^\star$ and $\pi = \pi_{f_k}$:

$$\begin{aligned} J(\pi^\star) - J(\pi_{f_k}) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{f_k}}}\left[V^\star(s) - Q^\star(s, \pi_{f_k}(s))\right] \\ &\le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{f_k}}}\left[V^\star(s) - V^{\pi_{f_{k:0}}}(s)\right]. \end{aligned}$$

The inequality uses the fact that $Q^\star(s, \pi_{f_k}(s)) \ge V^{\pi_{f_{k:0}}}(s)$, because $Q^\star(s, \pi_{f_k}(s))$ is the expected return of starting at s, taking action $\pi_{f_k}(s)$ (which coincides with the first action of $\pi_{f_{k:0}}$), and then acting optimally thereafter—which is at least as good as following $\pi_{f_{k:0}}$.

The right-hand side is precisely the suboptimality of $\pi_{f_{k:0}}$ (with $d^{\pi_{f_k}}$ playing the role of the initial distribution), which we can bound using the non-stationary FQI analysis. Compared to the non-stationary guarantee, the stationary FQI pays an extra $1/(1-\gamma)$ factor.

Summary

We analyzed Fitted Q-Iteration (FQI) for batch RL with function approximation, under the assumptions of realizability, Bellman completeness, and bounded concentrability.

Summary of Results:

Result	Bound	Key Idea
Slow rate (Hoeffding)	$\widetilde{O}\left(\frac{\sqrt{C} \cdot V_{\max}}{(1-\gamma)^2 \cdot n^{1/4}}\right)$	Uniform deviation + recursion
Fast rate (Bernstein)	$\widetilde{O}\left(\frac{\sqrt{C} \cdot V_{\max}}{(1-\gamma)^2 \cdot n^{1/2}}\right)$	Self-bounding variance + Bernstein
Non-stationary FQI	$\widetilde{O}\left(\frac{\sqrt{C} \cdot V_{\max}}{(1-\gamma) \cdot n^{1/2}}\right)$	Performance difference lemma

Key Takeaways:

- **Each FQI iteration is a regression problem:** under Bellman completeness, the Bayes optimal regressor (the conditional mean $\mathcal{T} f_{k-1}$) is in $\mathcal{F}$, so the per-iteration error is controlled by $\log|\mathcal{F}|$ rather than $|\mathcal{S}| \times |\mathcal{A}|$. However, unlike standard supervised learning, the target changes each iteration and the data distribution μ may differ from the deployment distribution, which is captured by the concentrability coefficient C.
- **Fast rates are achievable:** by exploiting the self-bounding variance structure of the excess loss (Lemma 4), we apply Bernstein's inequality to improve from $O(n^{-1/4})$ to $O(n^{-1/2})$.
- **Concentrability C is the key complexity measure:** it quantifies the distribution mismatch between data and deployment, and can be relaxed to a function-class-dependent quantity $C_{\mathcal{F}}$ that is potentially much smaller.
- **Non-stationary FQI has better guarantees:** it saves a factor of $1/(1-\gamma)$ compared to stationary FQI, at the cost of requiring a non-stationary policy.

References

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*, 71(1):89–129, 2008.

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.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pages 267–274, 2002.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, pages 560–567, 2003.

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

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 1835–1843, 2012.

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.