Fast Rates for the Regret of Offline Reinforcement Learning

Yichun Hu*, Nathan Kallus*, Masatoshi Uehara*
Cornell University

Abstract

We study the regret of reinforcement learning from offline data generated by a fixed behavior policy in an infinite-horizon discounted Markov decision process (MDP). While existing analyses of common approaches, such as fitted Q-iteration (FQI), suggest a $O(1/\sqrt{n})$ convergence for regret, empirical behavior exhibits *much* faster convergence. In this paper, we present a finer regret analysis that exactly characterizes this phenomenon by providing fast rates for the regret convergence. First, we show that given any estimate for the optimal quality function Q^*, the regret of the policy it defines converges at a rate given by the exponentiation of the Q^*-estimate's pointwise convergence rate, thus speeding it up. The level of exponentiation depends on the level of noise in the *decision-making* problem, rather than the estimation problem. We establish such noise levels for linear and tabular MDPs as examples. Second, we provide new analyses of FQI and Bellman residual minimization to establish the correct pointwise convergence guarantees. As specific cases, our results imply $O(1/n)$ regret rates in linear cases and $\exp(-\Omega(n))$ regret rates in tabular cases.

1 Introduction

Offline reinforcement learning (RL) is the problem of learning a reward-maximizing policy in an unknown Markov decision process (MDP) from data generated by running a fixed policy in the same MDP. The problem is particularly relevant in applications where exploration is limited but observational data plentiful. Medicine is one such example: ethical, safety, and operational considerations limit both the application of unproven or random policies and the running of online-updating algorithms, while at the same time rich electronic health records are collected en-masse.

A variety of methods have been proposed for offline RL including fitted Q-iteration (FQI; Ernst et al., 2005; Munos and Szepesvári, 2008), fitted policy iteration (Antos et al., 2008; Lagoudakis and Parr, 2004; Liu et al., 2020), modified Bellman Residual Minimization (Antos et al., 2008; Chen and Jiang, 2019), SBEED (Dai et al., 2018), and MABO (Xie and Jiang, 2020). For all of these, the regret (value suboptimality) bounds obtained are $O(1/\sqrt{n})$, where n is the number of observed transition data (see for example Agarwal et al., 2020a, Chapter 15 for a concise presentation of the analysis of FQI). However, in practice, the regret convergence can actually be *much* faster. For example, we provide a linear-MDP simulation experiment where FQI empirically exhibits an apparent regret convergence rate of $O(1/n)$.

*Alphabetical order.

In this paper, we tightly characterize this phenomenon by theoretically establishing fast rates for the regret convergence of value-based offline RL methods, which directly estimate the optimal quality function, Q^*. These rates leverage the specific noise level of a given problem instance, expressed as the density near zero of the suboptimality of the second-best action (if any), also known as a *margin condition*. RL instances generally satisfy *some* instance-specific nontrivial margin condition. We moreover show that in the linear and tabular cases, we generally have quite strong margin conditions. We show that policies that are greedy with respect to good estimates of Q^* enjoy a regret bounded by the pointwise estimation error raised to a power *larger* than one, thus speeding up convergence for the downstream decision-making task. This analysis can be applied to any value-based offline RL method that have pointwise convergence guarantees for estimating Q^*. As specific examples, we establish that we can achieve such pointwise error bounds for the linear case using FQI and modified BRM (differently from existing analyses of their average error). Together, this means that, under the standard assumptions needed for FQI and modified BRM, *i.e.*, closedneess under Bellman operators (completeness) and sufficient feature coverage, linear FQI and modified BRM generally achieve regret of order $O(1/n)$. Technically, our analysis melds ideas from fast-rate analysis of classification (Audibert and Tsybakov, 2007) with the theoretical analysis of RL (Agarwal et al., 2020a).

1.1 Set Up

We consider a time-homogeneous, finite-action, infinite-horizon, discounted MDP. Namely, we have an arbitrary measurable state space $\mathcal{S}$ (*e.g.*, continuous, discrete, or other), a finite actions space $\mathcal{A}$ (*i.e.*, $|\mathcal{A}| < \infty$), a reward distribution $P_r(\cdot \mid s, a)$ that maps to a probability measure on $\mathbb{R}$, a transition kernel $P_s(\cdot \mid s, a)$ that maps to a probability measure on $\mathcal{S}$, an initial state distribution μ on $\mathcal{S}$, and a discount factor $0 < \gamma < 1$. We let $r(s, a)$ denote the mean of $P_r(\cdot \mid s, a)$.

When we play a policy $\pi(a \mid s)$ in this MDP, the trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \dots$ is given the distribution $s_0 \sim \mu$, $a_0 \sim \pi(\cdot \mid s_0)$, $r_0 \sim P_r(\cdot \mid s_0, a_0)$, $s_1 \sim P_s(\cdot \mid s_0, a_0)$, $a_1 \sim \pi(\cdot \mid s_1)$, $\dots$. Since we consider different policies in the same MDP, we refer to this distribution as $\mathbb{P}^\pi$ and expectations over it as $\mathbb{E}^\pi$. With some abuse of notation, we also identify maps $\pi : \mathcal{S} \to \mathcal{A}$ with the deterministic policy given by Dirac at $\pi(s)$ (*i.e.*, $a_t = \pi(s_t)$). For each policy π, we define the Q-function, V-function, and average state occupancy distribution, respectively, as

$$
\begin{aligned}
Q^\pi(s, a) &= \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right], \\
V^\pi(s) &= \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right], \\
d^\pi(S) &= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t \in S) \quad \text{for measurable } S.
\end{aligned}
$$

The reward of a policy π is

$$V^\pi = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] = \mathbb{E}_{s \sim d^\pi, a \sim \pi(\cdot|s)}[r(s,a)].$$

We also define the optimal value, optimal V-function, and optimal Q-function, respectively, as

$$V^* = \max_\pi V^\pi, \; V^*(s) = \max_\pi V^\pi(s), \; Q^*(s,a) = \max_\pi Q(s,a).$$

We always let π^* be any deterministic policy with $\pi^*(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s,a)$. Notice that $V^* = V^{\pi^*}$, $V^*(s) = V^{\pi^*}(s)$, $Q^*(s,a) = Q^{\pi^*}(s,a)$.

We also define the Bellman optimality operator: for $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

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

Notice that Q^* is the unique fixed point of $\mathcal{T}$ (up to measure zero states).

Two types of MDPs we will sometimes use just as specific examples are tabular and linear MDPs. A *tabular MDP* is one with finite state space (we already assume the action space is finite). A *linear MDP* (Jin et al., 2020) is one where for some known $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ with $\|\phi(s,a)\| \leq 1$ and unknown vector $\theta \in \mathbb{R}^d$ and measures $\nu = (\nu_1, \dots, \nu_d)$, we have

$$r(s,a) = \theta^\intercal \phi(s,a), \quad P_s(S \mid s,a) = \nu(S)^\intercal \phi(s,a) \;\text{ for measurable } S.$$

Notation All unsubscripted norms, $\|\cdot\|$, are Euclidean norms. For a function $f(s,a)$ we define $\|f\|_\infty = \sup_{s \in \mathcal{S}, a \in \mathcal{A}} |f(s,a)|$. Additional norms will be defined in proofs in the appendix as needed. For a square symmetric matrix A we let $\lambda_{\min}(A)$ be its smallest eigenvalue.

1.2 The Offline Reinforcement Learning Problem

The learning problem is as follows. The MDP is unknown and we only observe transitions data from some (possibly unknown) stochastic policy, known as the behavior policy. Namely, for some (possibly unknown) μ_b, we observe n independent and identically distributed (iid) draws $\mathcal{D} = \{(s_i, a_i, r_i, s_i') : i = 1, \dots, n\}$ where each follows

$$(s_i, a_i) \sim \mu_b, \; r_i \sim P_r(\cdot \mid s_i, a_i), \; s_i' \sim P_s(\cdot \mid s_i, a_i).$$

We let $\mathbb{P}_\mathcal{D}$ and $\mathbb{E}_\mathcal{D}$ denote the probability and expectation with respect to the random sampling of the data $\mathcal{D}$. Based on this data, we choose a data-driven policy $\hat{\pi}$. The target is to find one with small average regret,

$$\mathbb{E}_\mathcal{D} \left[V^* - V^{\hat{\pi}} \right].$$

In particular, we will focus on Q-greedy policies that, given some $f(s,a)$, are given by any $\pi_f(s) \in \arg\max_{a \in \mathcal{A}} f(s,a)$. In particular, results will hold for any choice of tie breaking. Note, if $f = Q^*$ then $\pi_f = \pi^*$. Given some hypothesis class $\mathcal{F}$, we also define $\Pi_\mathcal{F} = \{\pi_f : f \in \mathcal{F}\} \subseteq [\mathcal{S} \to \mathcal{A}]$.

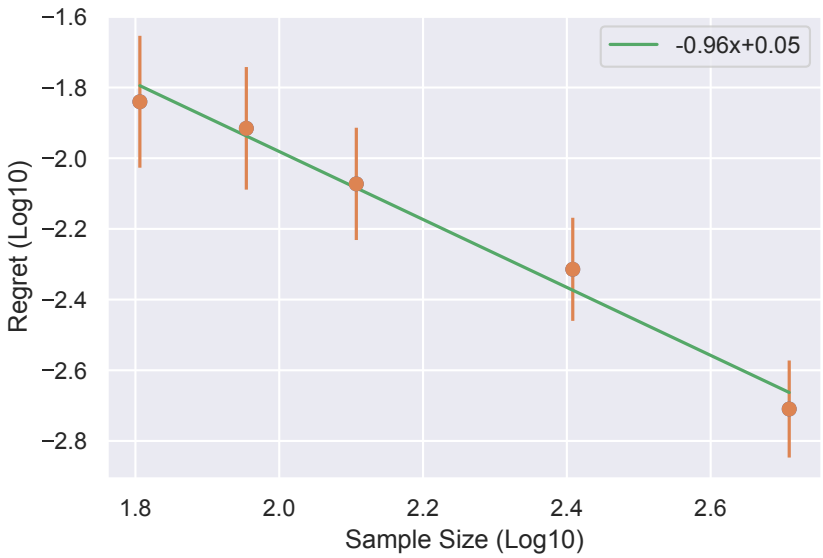

Figure 1: The regret of FQI in a simple example on a log-log scale with a linear trend fit

1.3 Fitted Q-Iteration

We will use FQI as one example of the regret behavior of offline RL. We present modified BRM as an additional example in Section 4.

The FQI algorithm is as follows (Ernst et al., 2005):

1. Start at any $\hat{f}_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (*e.g.*, the zero function).
2. For $k = 1, \ldots, K$:
   (a) Set $y_i = r_i + \gamma \max_{a' \in \mathcal{A}} \hat{f}_{k-1}(s_i', a')$.
   (b) Use any supervised learning algorithm to regress y_i on (s_i, a_i) to obtain $\hat{f}_k$.
3. Return $\hat{f}_K$ and $\hat{\pi} = \pi_{\hat{f}_K}$.

When the supervised learning algorithm is given by empirical risk minimization of squared loss over a hypothesis class $\mathcal{F}$ (*i.e.*, least squares), that step can be written as

$$\hat{f}_k \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(f(s_i, a_i) - r_i - \gamma \max_{a' \in \mathcal{A}} \hat{f}_{k-1}(s_i', a') \right)^2. \tag{1}$$

1.4 The Empirical Performance of FQI Belies Current Analysis

We next study the performance of FQI in a simple example of a linear MDP. We construct the underlying MDP as follows. We set $\mathcal{S} = [0, 1]^2$, $\mathcal{A} = \{0, 1\}$. We set the initial distribution $\mu(\cdot) = \text{Unif}_{[0,1]^2}(\cdot)$, where $\text{Unif}_{[0,1]^2}(\cdot)$ is the uniform distribution on $[0, 1]^2$. We let

$\phi(s,a) = (s_1(1-a), s_1 a, 1-s_1, s_2(1-a), s_2 a, 1-s_2)/2$, which belongs to the simplex in $\mathbb{R}^6$. We then set $r(s,a) = \theta^\intercal \phi(s,a)$ and

$$P_s(s' \mid s,a) = \sum_{k=1}^{6} \phi_k(s,a)\, \mathrm{Beta}_{10\alpha_{k1},\, 10\beta_{k1}}(s'_1) \times \mathrm{Beta}_{10\alpha_{k2},\, 10\beta_{k2}}(s'_2),$$

where $\mathrm{Beta}_{\alpha,\,\beta}(\cdot)$ is a Beta distribution. We fix some $\theta, \alpha_{1,1}, \beta_{1,1}, \ldots, \alpha_{6,2}, \beta_{6,2}$ by drawing from $\mathrm{Unif}_{[0,1]^{30}}$ (once, after setting the random seed to zero). We let the behavior policy be uniform: $\mu_b(s,a) = \mu(s) \times \mathrm{Ber}_{0.5}(a)$, where $\mathrm{Ber}_{0.5}(\cdot)$ is a Bernoulli distribution with parameter 0.5. We set the discount as $\gamma = 0.9$.

We then apply FQI to data generated by the above MDP and behavior policy using $K = 50$ iterations and a linear function class, $\mathcal{F} = \{\beta^\intercal \phi(s,a) : \beta \in \mathbb{R}^6\}$ (*i.e.*, Eq. (1) becomes ordinary least squares). We vary the sample size $n \in [64, 90, 128, 256, 512]$ and run 70 replications of the experiment for each sample size, in each replication recording the resulting regret $V^* - V^{\hat{\pi}}$. Namely, we calculate the value of a given policy by running the policy on an independent sample of 40000 initial states and truncating at step 50, and we compute π^* by running FQI with $K = 100$ iterations on another independent dataset of size $n = 40000$. We report the results in Fig. 1 on a logarithmic scale along with 75%-confidence intervals for each n and a linear trend fit to the log-log-transformed data. The empirically observed slope with 75%-confidence interval is -0.96 ± 0.08, which is somewhat suggestive of a regret rate of roughly $O(1/n)$. This provides concrete empirical evidence that for some instances, we may be able to get regret convergence that is much faster than the $O(1/\sqrt{n})$ appearing in the existing analyses of FQI and other offline RL algorithms.

2 Fast Rates for Q-Greedy Policies

In this section, we show that any estimate $\hat{f}$ of Q^* with some rate of convergence leads to a Q-greedy policy with regret rate that is the *exponentiation* of this estimation rate, and sometimes even an exponential to this rate. This can possibly speed up the rate considerably. The level of exponentiation depends on the level of noise in the downstream *decision* problem (rather than in the Q-estimation problem), that is, how hard is it to distinguish optimal actions from near-optimal actions (rather than how hard it is to estimate Q^*), also known as a margin in classification and bandit problems.

We define the margin at s as

$$\Delta(s) = \begin{cases} \max_{a \in \mathcal{A}} Q^*(s,a) - \max_{a \notin \mathrm{argmax}_{a \in \mathcal{A}} Q^*(s,a)} Q^*(s,a) & \mathrm{argmax}_{a \in \mathcal{A}} Q^*(s,a) \neq \mathcal{A} \\ 0 & \mathrm{argmax}_{a \in \mathcal{A}} Q^*(s,a) = \mathcal{A} \end{cases}.$$

The margin can be smaller at some s and larger at other s. The larger the margin, the clearer is the choice of the optimal action, the easier it is to learn to make this optimal choice. However, the margin may well be positive and arbitrarily close to 0 in many continuous settings (while a 0 margin leads to trivial decision making). So, motivated by related conditions in classification (Audibert and Tsybakov, 2007; Mammen and Tsybakov, 1999; Tsybakov, 2004) and multi-arm contextual bandits (Hu et al., 2020b; Perchet and Rigollet, 2013), we use the following condition to describe the *density* of $\Delta(s)$ near (but not at) zero.

Condition 1 (Margin)**.** Fix some class of deterministic policies $\Pi \subseteq [\mathcal{S} \to \mathcal{A}]$. There exist constants $\delta_0 > 0, \alpha \in [0, \infty]$ such that for all $\delta > 0$,

$$\sup_{\pi \in \Pi} \mathbb{P}_{s \sim d^\pi}(0 < \Delta(s) \leq \delta) \leq (\delta/\delta_0)^\alpha,$$

where x^∞ is understood as 0 for $x \in [0, 1)$, 1 for $x = 1$, and ∞ for $x > 1$.

We can often just take $\Pi = [\mathcal{S} \to \mathcal{A}]$ to be all deterministic policies for simplicity, but it will be sufficient to take only $\Pi = \Pi_{\mathcal{F}}$ when using a hypothesis class $\mathcal{F}$ for learning Q^*. All instances satisfy Condition 1 with $\alpha = 0$. But, generally, a given instance would satisfy Condition 1 with some $\alpha > 0$. At one extreme, if $\Delta(s)$ is uniformly bounded away from 0 over s then Condition 1 holds with $\alpha = \infty$. We give examples where we can establish a margin below in Section 2.1

Our result applies to Q-greedy policies given a good estimate $\hat{f}$ of Q^*. Our next condition quantifies the quality of the estimate.

Condition 2 (Pointwise error bound)**.** A data-driven $\hat{f}(s, a)$ is given such that for some $C > 0$ and $a_n > 0$ it satisfies that, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\delta \geq a_n$, we have

$$\mathbb{P}_{\mathcal{D}}(|\hat{f}(s, a) - Q^*(s, a)| \geq \delta) \leq C \exp(-\delta^2/a_n^2). \tag{2}$$

Equation (2) is a *pointwise* convergence bound for $\hat{f}$ with rate a_n. In our second main result, in Section 3, we will show that linear FQI satisfies an even stronger *uniform* convergence bound (with the supremum over s, a inside the probability) with $a_n = 1/\sqrt{n}$. This analysis is substantially different from the usual analysis of FQI, which establishes guarantees on the average squared error. In Section 4, we show similar results for Bellman residual minimization.

Using Conditions 1 and 2, we can now establish our key rate-speed-up result.

Theorem 1. *Let a data-driven $\hat{f}$ be given and let Π be given such that $\pi_{\hat{f}} \in \Pi$ almost surely. Suppose Conditions 1 and 2 hold and $\|Q^*\|_\infty \leq Q_{\max}$. Fix any $\delta_1 \geq \delta_0$ with $\delta_1 > 2a_n$. Let $i_{\max}$ be the largest integer such that $2^{i_{\max}+1} a_n < \delta_1$. Then, for $\hat{\pi} = \pi_{\hat{f}}$, we have*

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \leq \frac{2^{\alpha+1}}{(1-\gamma)\delta_0^\alpha} \left(1 + C \sum_{i=1}^{i_{\max}} \exp\left(-2^{2i-2}\right) 2^{(\alpha+1)i+1}\right) a_n^{\alpha+1} + \frac{2Q_{\max}C}{1-\gamma} \exp\left(-\delta_1^2/(4a_n)^2\right).$$

If either $\Pi = [\mathcal{S} \to \mathcal{A}]$ or $\Pi = \Pi_{\mathcal{F}}$ and $\hat{f} \in \mathcal{F}$ then obviously $\pi_{\hat{f}} \in \Pi$ is satisfied. Note that we could also easily only require that $\pi_{\hat{f}} \in \Pi$ with high probability and the failure probability would simply propagate into the regret bound. The key to the result is to leverage the performance difference lemma (Agarwal et al., 2020a, Lemma 1.16) together with a peeling argument to study the behavior at different scales of margin.

We immediately have the following two corollaries to simplify the above expression, one for the case $\alpha < \infty$ and one for the case $\alpha = \infty$.

Corollary 2. *Suppose $\pi_{\hat{f}} \in \Pi$ almost surely, Condition 2 holds, and Condition 1 holds with $\alpha < \infty$. Then, for $\hat{\pi} = \pi_{\hat{f}}$, we have*

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \leq \frac{2^{\alpha+1}}{(1-\gamma)\delta_0^{\alpha}} \left(1 + c(\alpha)C\right) a_n^{\alpha+1},$$

where $c(\alpha) = \sum_{i=1}^{\infty} \exp\left(-2^{2i-2}\right) 2^{(\alpha+1)i+1} \leq \frac{2^{\alpha+1}\Gamma\left(\frac{\alpha+1}{2},1\right)}{\log 2} + 2\left(\frac{2(\alpha+1)}{e}\right)^{(\alpha+1)/2}$, with $\Gamma\left(\frac{\alpha+1}{2},1\right) = \int_1^{\infty} x^{\frac{\alpha-1}{2}} e^{-x} dx < \infty$.

Corollary 2 shows that the *estimation* rate in Condition 2 gets sped up by exponentiation by $1+\alpha$ when applied to the downstream *decision making* problem. Thus, however fast we are able to estimate Q^*, our regret converges *even faster*. Notice that in Corollary 2 we do not actually need to assume a bound on $\|Q^*\|_{\infty}$.

Corollary 3. *Suppose $\pi_{\hat{f}} \in \Pi$ almost surely, Condition 2 holds, Condition 1 holds with $\alpha = \infty$, and $\|Q^*\|_{\infty} \leq Q_{\max}$. Let n be such that $a_n < \delta_0/2$. Then, for $\hat{\pi} = \pi_{\hat{f}}$, we have*

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \leq \frac{2Q_{\max}C}{1-\gamma} \exp\left(-\delta_0^2/(4a_n)^2\right).$$

Corollary 3 shows that in the $\alpha = \infty$ case, our regret vanishes *exponentially* fast. While Corollaries 2 and 3 provide a simple understanding of the behavior in n at any fixed α, if α is finite but very big (*e.g.*, a regime where $\alpha = \omega(1)$ with respect to n) then Theorem 1 with $\delta_1 = \delta_0$ characterizes the correct trade-off between the polynomial and exponential terms.

2.1 The Margin Condition in Some Examples

We next discuss some cases where we can explicitly demonstrate a nontrivial margin condition. The heuristic implication is that we should *generically* expect $\alpha = 1$ in continuous-state settings and $\alpha = \infty$ in tabular settings.

The next lemma shows that if Q^* is linear and under a kind of weak concentratability assumption, we have $\alpha = 1$.

Lemma 4. *Suppose $Q^*(s,a) = \beta_a^{\intercal}\psi(s)$ for some $\psi : \mathcal{S} \to \mathbb{R}^d$ with $\|\psi(s)\| \leq 1$ and $\beta \in \mathbb{R}^{\mathcal{A}\times d}$, and that $\psi(s)$ with $s \sim d^{\pi}$ has a density for each $\pi \in \Pi$ and this density is bounded by $\mu_{\max}$. Then, Condition 1 holds with $\alpha = 1$ and $\delta_0 = (6\mu_{\max} \sum_{a\in\mathcal{A}} \max_{a'\in\mathcal{A}:\beta_a\neq\beta_{a'}} \|\beta_a - \beta_{a'}\|^{-1})^{-1}$.*

A linear MDP is sufficient for the condition on Q^* as we can take $\psi(s) = (\phi(s,a)/\sqrt{|\mathcal{A}|})_{a\in\mathcal{A}}$. The assumption of uniformly bounded density is not especially restrictive. In offline RL, we often assume some type of overlap condition; for example, that d^{π} and μ_b have densities (let us overload notation and call these densities d^{π}, μ_b) and that $\sup_{\pi\in\Pi_{\mathcal{F}}} \|d^{\pi}(s)\pi(a \mid s)/\mu_b(s,a)\|_{\infty} \leq C_1$ (*e.g.* Xie and Jiang, 2020). (See also Scherrer, 2014 for a discussion of various overlap conditions.) This implies that the condition in Lemma 4 is satisfied with $\mu_{\max} = C_1\|\mu_b\|_{\infty}$, if $\|\mu_b\|_{\infty} < \infty$.

The result essentially continues to hold even if $s \mapsto (Q^*(s,a))_{a\in\mathcal{A}}$ is nonlinear as long as it has a uniformly nonsingular Jacobian, as it is then locally linear (uniformly). Proving this formally remains future work. At least heuristically, this suggests we should generically

expect to have $\alpha = 1$ in practice in continuous-state-space cases since the violation of nonsingular Jacobian only matters if it occurs on the decision boundary, which, while it might happen would generically not happen.

In the tabular setting, we trivially have $\alpha = \infty$, albeit with a δ_0 that might be small. The following follows trivially by enumeration.

Lemma 5. *Suppose* $|\mathcal{S}| < \infty$. *Set* $\Pi = [\mathcal{S} \to \mathcal{A}]$. *Then, Condition 1 holds with* $\alpha = \infty$ *and* $\delta_0 = \max_{s,a,a':Q^*(s,a)\neq Q^*(s,a')} |Q^*(s,a) - Q^*(s,a')|^{-1}$.

3 Fast Rates for Linear Fitted Q-Iteration

In this section, we show that by applying FQI with a linear function class, we can obtain an estimator $\hat{f}_K$ that has a pointiwse guarantee as in Eq. (2) with $a_n = O(1/\sqrt{n})$. Thus, Theorem 1 ensures that we will obtain regret of order $O(n^{-(1+\alpha)/2})$ when we execute the greedy policy $\pi_{\hat{f}_K}$. When $\alpha = 1$, as in Lemma 4, this means we have regret $O(1/n)$, just as was observed empirically in Section 1.4.

Bounded Linear Class and Assumptions Assume we are given a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ with $\|\phi(s,a)\| \leq 1$. Throughout Section 3, we will consider the hypothesis class $\mathcal{F}$ that is linear in these features with bounded coefficients:

$$\mathcal{F} = \{w^\intercal \phi(s,a) : w \in \mathbb{R}^d,\ \|w\| \leq B\}. \tag{3}$$

Here we introduce two assumptions to ensure the convergence of FQI estimators, which are commonly seen in literature (see Agarwal et al., 2020a, Chapter 15 for a review).

Our first assumption is that our function class $\mathcal{F}$ is closed under the Bellman operator $\mathcal{T}$.

Assumption 1 (Completeness)**.** For any $f \in \mathcal{F}$, $\mathcal{T}f \in \mathcal{F}$.

Since Q^* is the fixed point of $\mathcal{T}$, Assumption 1 directly implies realizability, i.e., $Q^* \in \mathcal{F}$.

As a special example, in the case of linear MDPs, for any $f \in \mathcal{F}$, we have

$$\begin{aligned}
\mathcal{T}f(s,a) &= r(s,a) + \gamma \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') dP(s' \mid s,a) \\
&= \theta^\intercal \phi(s,a) + \gamma \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') d\nu(s')^\intercal \phi(s,a) \\
&= \left(\theta + \gamma \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') d\nu(s')\right)^\intercal \phi(s,a),
\end{aligned}$$

so $\mathcal{T}f$ is a linear function in ϕ as well. Moreover, if $\max\{\|\nu(\mathcal{S})\|, \|\theta\|\} \leq C_{\max}$ and $\gamma C_{\max} < 1$, it is easy to see that $\mathcal{F} = \{w^\intercal \phi(s,a) : w \in \mathbb{R}^d, \|w\| \leq C_{\max}/(1-\gamma C_{\max})\}$ satisfies Assumption 1.

Our second assumption is to ensure sufficient feature coverage in our data set. This assumption is commonly required to guarantee the convergence of ordinary least squares estimators (Wang et al., 2020).

Assumption 2 (Feature Coverage)**.** There exists a constant $\lambda_0 > 0$ such that

$$\lambda_{\min}(\mathbb{E}_{s,a\sim\mu_b}[\phi(s,a)\phi(s,a)^\intercal]) \geq \lambda_0.$$

Uniform Convergence of the Linear FQI Estimator We now apply the FQI algorithm as is stated in Section 1.3 to get $\hat{f}_K$ with $\mathcal{F}$. Here we elaborate on how we regress y_i on (s_i, a_i) to obtain $\hat{f}_k$ in step 2b. Namely, we will set $\hat{f}_k = 0$ whenever the solution to Eq. (1) is not unique.

For any $f \in \mathcal{F}$, by Assumption 1 there exists w_f with $\|w_f\| \le B$ such that $\mathcal{T}f = w_f^\intercal \phi$. Define the empirical design matrix

$$\hat{\Sigma} = \sum_{i=1}^{n} \phi(s_i, a_i)\phi(s_i, a_i)^\intercal.$$

If $\lambda_{\min}(\hat{\Sigma}) > 0$, let $\hat{w}'_f$ be the OLS regressor

$$\begin{aligned}\hat{w}'_f &= \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(w^\intercal \phi(s_i, a_i) - r_i - \gamma \max_{a' \in \mathcal{A}} f\left(s'_i, a'\right) \right)^2 \\ &= \hat{\Sigma}^{-1} \left(\sum_{i=1}^{n} \phi(s_i, a_i) \left(r_i + \gamma \max_{a' \in \mathcal{A}} f(s'_i, a') \right) \right),\end{aligned}$$

and otherwise $\hat{w}'_f = 0$. Finally, let $\hat{w}_f$ be the projection of $\hat{w}'_f$ on to a Euclidean ball $\mathcal{B}(0, B)$. We then set

$$\hat{f}_k = \hat{w}_{\hat{f}_{k-1}}^\intercal \phi.$$

The following Lemma shows that with high probability, our one-step estimator $\hat{w}_f^\intercal \phi$ converges quickly to $\mathcal{T}f$, uniformly over all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and all functions $f \in \mathcal{F}$. In proving this Lemma, we leverage theoretical tools from matrix concentration and empirical process.

Lemma 6. *Assume Assumptions 1 and 2 hold and $|r_t| \le M$ for $t = 1, 2, \ldots$. For any $\delta > 0$, $\hat{w}_f$ estimated from the above procedure satisfies*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left\| \hat{w}_f^\intercal \phi - \mathcal{T}f \right\|_\infty \ge \delta \right) \le 6d \exp\left(-\frac{\lambda_0^2}{5184 d^2 (M+B)^2} n\delta^2 \right).$$

Using Lemma 6, we can then obtain the following convergence guarantee for FQI.

Theorem 7. *Assume Assumptions 1 and 2 hold and $|r_t| \le M$ for $t = 1, 2, \ldots$. Set $a_n = \frac{144d(M+B)}{(1-\gamma)\lambda_0\sqrt{n}}$. Then, for any $K \ge \frac{\log(\lambda_0^2 n/(72d)^2)}{2\log(1/\gamma)}$ and any $\delta \ge a_n$, we have*

$$\mathbb{P}(\|Q^* - \hat{f}_K\|_\infty \ge \delta) \le 6d \exp(-\delta^2/a_n^2).$$

The key to showing Theorem 7 is establishing that (surely)

$$\|Q^* - \hat{f}_K\|_\infty \le \sum_{t=0}^{K-1} \gamma^t \|\hat{f}_{K-t} - \mathcal{T}\hat{f}_{K-t-1}\|_\infty + \frac{\gamma^K M}{1-\gamma}.$$

This means that the estimation error of $\hat{f}_K$ can be controlled by two terms: one is a weighted sum of one-step estimation errors, which can be bounded using Lemma 6, and the other is a diminishing term as we increase the number of iterations. Therefore, as long as our total number of iterations K is large enough, with high probability $\hat{f}_K$ converges to Q^* uniformly.

Fast Rates for Linear Fitted Q-Iteration Since uniform convergence is a stronger condition than pointwise convergence, Theorem 7 shows that $\hat{f}_K$ satisfies Condition 2 with $C = 6d$ and $a_n = \frac{144d(M+B)}{(1-\gamma)\lambda_0\sqrt{n}}$. Then, combined with Corollaries 2 and 3 it immediately implies fast rates for linear fitted Q-interation. We summarize below. (For variable/large α, we should apply Theorem 1 to get the right trade off between the terms.)

Corollary 8 (Fast Rates for Linear Fitted Q-Iteration)**.** *Suppose Condition 1 holds with $\Pi = \Pi_{\mathcal{F}}$, Assumptions 1 and 2 hold, and $|r_t| \le M$ for $t = 1, 2, \ldots$. When $K \ge \frac{\log(\lambda_0^2 n/(72d)^2)}{2\log(1/\gamma)}$, for $\hat{\pi} = \pi_{\hat{f}_K}$, we have:*

1. *if $\alpha < \infty$,*

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \le \frac{288^{\alpha+1}\left(1 + 6dc(\alpha)\right)}{(1-\gamma)\delta_0^{\alpha}} \left(\frac{d(M+B)}{(1-\gamma)\lambda_0}\right)^{\alpha+1} n^{-\frac{\alpha+1}{2}};$$

2. *if $\alpha = \infty$ and $n \ge \left(\frac{288d(M+B)}{(1-\gamma)\lambda_0\delta_0}\right)^2$,*

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \le \frac{12Md}{(1-\gamma)^2} \exp\left(-\left(\frac{(1-\gamma)\lambda_0\delta_0}{576d(M+B)}\right)^2 n\right).$$

3.1 Tabular MDP as a Special Case

The tabular setting ($|\mathcal{S}| < \infty$) is a special case of the linear MDP since we can take

$$\phi(s,a) = (\mathbb{I}\{(s,a) = (s',a')\})_{(s',a') \in \mathcal{S}\times\mathcal{A}}, \quad d = |\mathcal{S}|\,|\mathcal{A}|\,.$$

Moreover, we can satisfy Assumption 2 with $\lambda_0 = \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_b(s,a)$, and given M s.t. $|r_t| \le M$, we can satisfy Assumption 1 with $B = \sqrt{|S||A|}(1-\gamma)^{-1}M$. Thus, Theorem 7 gives us the bound of the model based estimator in a tabular case. This shows that with probability $1-\delta$, $\|\hat{f}_K - Q^*\|_\infty = \frac{144|\mathcal{S}||\mathcal{A}|(M+B)}{(1-\gamma)\lambda_0}\sqrt{\frac{\log(6|\mathcal{S}||\mathcal{A}|/\delta)}{n}}$. A typical way (*e.g.* Singh and Yee, 1994) to translate this into the regret of $\hat{\pi} = \pi_{\hat{f}_K}$ is to note $V^* - V^{\hat{\pi}} \le \frac{1}{1-\gamma}\|\hat{f}_K - Q^*\|_\infty$. Integrating the tail gives

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \le \frac{432\sqrt{\pi}\,|\mathcal{S}|^2\,|\mathcal{A}|^2\,(M+B)}{(1-\gamma)^2\lambda_0\sqrt{n}}.$$

Compared to this, our Lemma 5, Theorem 7, and Corollary 8 together give that, for $n \ge \left(\frac{288|\mathcal{S}||\mathcal{A}|(M+B)}{(1-\gamma)\lambda_0\delta_0}\right)^2$,

$$\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \le \frac{12M\,|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^2} \exp\left(-\left(\frac{(1-\gamma)\lambda_0\delta_0}{576\,|\mathcal{S}|\,|\mathcal{A}|\,(M+B)}\right)^2 n\right),$$

where $\delta_0 = \max_{s,a,a':Q^*(s,a)\neq Q^*(s,a')} |Q^*(s,a) - Q^*(s,a')|^{-1}$. The above regret bound vanishes *exponentially*, much faster than the usual $O(1/\sqrt{n})$ result.

4 Fast Rates for Modified Bellman Residual Minimization

Modified Bellman Residual Minimization (BRM) is a common offline Q-function estimation method (Antos et al., 2008), which approximates the Bellman error by introducing another maximization problem, thus avoiding the need to iterate. The original BRM was for offline policy evaluation (estimate Q^π for a given π); recently Chen and Jiang (2019) adapted it to offline policy learning (estimate Q^*) in a method called MSBO. They establish the convergence of the Bellman residual errors of MSBO when the hypothesis classes are finite ($|\mathcal{F}| < \infty$). In this section, we show the convergence of MSBO in terms of uniform error to Q^* for a linear function class. Using our results, we conclude that MSBO enjoys fast rates as well, similarly to FQI.

Given, a class $\mathcal{F}_w \subseteq [\mathcal{S} \times \mathcal{A} \to \mathbb{R}]$ and $\zeta > 0$, MSBO is defined as follows:

$$\hat{f} \in \operatorname*{argmin}_{q \in \mathcal{F}} \max_{w \in \mathcal{F}_w} \sum_{i=1}^{n} \left(\left(r_i - q(s_i, a_i) + \max_{a' \in \mathcal{A}} q(s_i', a') \right) w(s_i, a_i) - \zeta w^2(s_i, a_i) \right).$$

Note MSBO was originally proposed using $\zeta = 0.5$. We consider general $\zeta > 0$.

Here, given a feature map $\phi : \mathcal{S} \times \mathcal{A} \in \mathbb{R}^d$ with $\|\phi(s,a)\| \le 1$ we consider

$$\mathcal{F} = \{\theta^\top \phi(s,a) : \|\theta^\top \phi(s,a)\|_\infty \le M', \theta \in \mathbb{R}^d\},$$
$$\mathcal{F}_w = \{\theta^\top \phi(s,a) : \|\theta^\top \phi(s,a)\|_\infty \le M', \theta \in \mathbb{R}^d\}.$$

Theorem 9. *Suppose $|r_t| \le M$ for $t = 1, 2, \cdots$ and $M' = (1-\gamma)^{-1}M$. Moreover, assume $Q^* \in \mathcal{F}$ (realizability), $(\mathcal{T} - I)\mathcal{F} \subset \mathcal{F}_w$ (modified completeness), and (modified feature coverage)*

$$\lambda_{\min}(\mathbb{E}_{(s,a)\sim\mu_b}[(\mathcal{T} - I)\phi(s,a)\{(\mathcal{T} - I)\phi(s,a)\}^\top]) \ge \lambda_0'.$$

Then, there exists a universal constant $c > 0$, such that, letting $a_n = c(\sqrt{d} + M'\sqrt{M'^2/\zeta + M' + \zeta + 1}/\lambda_0')\sqrt{\log n / n}$, we have for all $\delta \ge a_n$,

$$\mathbb{P}(\|\hat{f} - Q^*\|_\infty \ge \delta) \le \exp\left(-\delta^2 / a_n^2\right),$$

The assumptions are similar to FQI in the sense that we need realizability, completeness, and feature coverage, but the latter two are slightly different. As in Section 3, by combining our results (Theorem 1, Corollaries 2 and 3, and Lemmas 4 and 5) with Theorem 9, we obtain a fast regret rate for MSBO. Specifically, if $\alpha < \infty$, we obtain a regret rate of $(\log n/n)^{-(\alpha+1)/2}$, which is faster than the rate of the $O(1/\sqrt{n})$ rate in the analysis of Chen and Jiang (2019).

5 Related Literature

Tabular offline RL Agarwal et al. (2020b); Gheshlaghi Azar et al. (2013) analyze the tabular case where we have a generative model (Agarwal et al., 2020b; Gheshlaghi Azar et al., 2013), that is an oracle for drawing from the MDP's reward and transition distributions, but assuming a generative model is much stronger than our offline setting, where we

just see data passively. In the tabular offline setting, the minimax optimal regret rate is obtained by Yin et al. (2020) and is $O(1/\sqrt{n})$. Our result in the tabular case is $\exp(-\Omega(n))$, but it depends on instance parameters such as δ_0, *i.e.*, it is *not* minimax if δ_0 is allowed to vary, and in particular approach 0.

Value-based offline RL Value-based offline RL is an approach to offline RL where we estimate Q^* and then use the corresponding greedy (*i.e.*, argmax) policy (some also consider a smoothed softmax version). This is the approach we studied here. The most common way to estimate Q^* is FQI (Ernst et al., 2005), which we analyze in Section 3. Chen and Jiang (2019); Fan et al. (2020); Munos and Szepesvári (2008) have analyzed finite-sample error bounds for FQI. Since they obtain bounds on the *average* error, their analysis is not directly applicable to our setting. We therefore established uniform convergence in order to derive fast regret rates for the resulting greedy policy.

Another common method to estimate Q^* is modified BRM and its variants, which we analyze in Section 4. The finite-sample error bound of the Q^*-function has been analyzed in Chen and Jiang (2019); Xie and Jiang (2020) when the hypothesis class is finite. In a general function approximation setting (such as linear), a slow rate of $O(n^{-1/4})$ is obtained by Antos et al. (2008) for policy evaluation (estimate Q^π for a given π). Since existing analyses are for average errors and/or not tight, it is not directly applicable. We therefore established uniform convergence.

Beyond FQI and BRM/MSBO, there are many offline estimators for Q^* such as SBEED (Dai et al., 2018), which uses a smoothed version of the max operator in modified RBM, and MABO (Xie and Jiang, 2020), which is derived by a conditional moment equation formulation of Q^*. In all of the aforementioned value-based offline RL work including these two, the final regret is $O(1/\sqrt{n})$. Our analysis shows that we can obtain the faster rate depending on the margin condition.

Policy-based offline RL Policy-based offline RL is an approach where we directly optimize a policy among some restricted policy class. One common approach is fitted policy iteration (Lagoudakis and Parr, 2004; Lazaric et al., 2010). Finite-sample regret bound have been analyzed by Antos et al. (2008); Liu et al. (2020), which show that the final regret is $O(1/\sqrt{n})$. Another common way is offline policy gradient (Nachum et al., 2019). Kallus and Uehara (2020a) showed the asymptotic regret of a offline policy gradient method based on efficient estimation is $O(1/\sqrt{n})$. Note our analysis here does not apply to the policy-based approach. We leave it as future work.

Offline policy evaluation Offline policy evaluation (OPE) is the task evaluating the policy value of a single policy (Precup et al., 2000), *i.e.*, estimating V^π for a given π. The typical error rate is $O(1/\sqrt{n})$ (Duan et al., 2020; Kallus and Uehara, 2019; Liu et al., 2018; Thomas and Brunskill, 2016). Kallus and Uehara (2019, 2020b) focuses on how to reduce the constant in the leading $1/\sqrt{n}$ term. Note that our work does *not* imply a fast rate for OPE since regret is different from estimation error. In particular, our fast rate leverages the impact of the downstream decision-making problem, after estimation.

Margin Condition and Fast Rate In classification, Audibert and Tsybakov (2007); Tsybakov (2004) showed that both empirical risk minimization and plug-in methods can achieve $o(1/\sqrt{n})$ fast rates under a margin condition that quantifies the concentration of mass of $P(Y = 1 \mid X)$ near 1/2. An analogous condition has been used in contextual bandits to quantify the separation between arms and get low regret (Goldenshluger and Zeevi, 2013; Hu et al., 2020b; Perchet and Rigollet, 2013). In particular, such margin conditions for $\alpha < \infty$ can be much weaker than assuming strict separation of arms ($\alpha = \infty$), which is often unrealistic. Hu et al. (2020a) used a margin condition that characterizes the distribution of near-degeneracy in contextual linear optimization problems and obtained fast regret rates in that problem for both plug-in policies and empirical risk minimization.

Exponential Lower Bounds with only Realizability and Concentratability Both our results for FQI and MSBO and the previous analyses mentioned above assume some form of a completeness condition. It has been shown that when the hypothesis class is linear, if we only assume realizability and concentratability (*e.g.*, feature coverage) and we do *not* assume completeness, then the sample complexity must have an exponential dependence on the dimension of linear class or horizon (*i.e.*, $1/(1-\gamma)$) (Wang et al., 2020; Weisz et al., 2020). Completeness, as we have used here, is one sufficient condition to avoid this and make offline RL feasible.

6 Concluding Remarks

In this paper we established the first sub-$1/\sqrt{n}$ regret guarantees for offline reinforcement learning. In particular, we showed that, given an estimate of Q^*, the resulting Q-greedy policy has a regret rate given by exponentiating the estimation rate, where the exponent depends on a margin condition. We also showed that quite strong margin conditions generally hold in linear and tabular MDPs, and argued a nontrivial margin should usually hold for a given instance in practice. Our rate-speed-up result relied on pointwise convergence guarantees for Q^* estimates. Since no such exist, we derived new uniform convergence guarantees for FQI and a BRM variant called MSBO (and this implied pointwise convergence). The rates our theory predict are almost exactly what is observed in practice in a simulation example.

Acknowledgments

Th authors thank Yoav Kallus for providing the geometric argument to improve the dependence of the δ_0 constant on $|\mathcal{A}|$ in the proof of Lemma 4, which is based on a proof by Noam Elkies of the monotonicity of perimeter for convex sets under containment (`https://mathoverflow.net/questions/71502/circumference-of-convex-shapes/71505#71505`). This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. Technical report, 2020a.

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 67–83, 09–12 Jul 2020b.

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.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1042–1051, 2019.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1125–1134, 2018.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2701–2709, 2020.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 486–489, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325, 2013. ISSN 0885-6125.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *arXiv preprint arXiv:2011.03030*, 2020a.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010. PMLR, 2020b.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019.

Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy policy gradients. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5089–5100, 2020a.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020b.

Michail Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4(6):1107–1149, 2004.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Remi Munos. Finite-sample analysis of lstd. pages 615–622, 2010.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: infinite-horizon off-policy estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5361–5371, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

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

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Annals of statistics*, 41(2):693–721, 2013.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

D. Precup, R. Sutton, and S Singh. Eligibility traces for off-policy policy evaluation. *In Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1314–1322, 2014.

Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. 16(3):227–233, 1994.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *In Proceedings of the 33rd International Conference on Machine Learning*, pages 2139–2148, 2016.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline rl with linear function approximation?. *arXiv preprint arXiv:2010.11895*, 2020.

Gellert Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*, 2020.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. *UAI2020*, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*, 2020.

Technical Appendix for Fast Rates for the Regret of Offline Reinforcement Learning

A Proofs

A.1 Proofs for Section 2

Proof of Theorem 1. For any policy π, define $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$, and $A^*(s,a) = A^{\pi^*}(s,a) \le 0$. By the performance difference lemma (Agarwal et al. (2020a, Lemma 1.16)),

$$\begin{aligned}(1-\gamma)(V^* - V^{\hat\pi}) =& \mathbb{E}_{s\sim d^{\hat\pi}}[-A^*(s,\hat\pi(s)] \\ =& \mathbb{E}_{s\sim d^{\hat\pi}}[Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s))]. \end{aligned} \tag{4}$$

Define the events

$$\begin{aligned} B_0 &= \{0 < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)) \le 2a_n\}, \\ B_i &= \{2^i a_n < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)) \le 2^{i+1} a_n\} \quad i \ge 1, \\ B' &= \{2^{i_{\max}+1} a_n < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s))\}. \end{aligned}$$

Peeling on $Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s))$ we get

$$\begin{aligned} &\mathbb{E}_{s\sim d^{\hat\pi}}[Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s))] \\ =& \mathbb{E}_{s\sim d^{\hat\pi}}[\sum_{i=0}^{i_{\max}} (Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)))\mathbb{I}\{B_i\} + (Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)))\mathbb{I}\{B'\}] \\ \le& 2a_n \sum_{i=0}^{i_{\max}} 2^i \mathbb{P}_{s\sim d^{\hat\pi}}(B_i) + Q_{\max}\mathbb{P}_{s\sim d^{\hat\pi}}(B'). \end{aligned} \tag{5}$$

In what follows, we control $\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat\pi}}[\mathbb{I}\{B_i\}]$ for $i \in [0, i_{\max}]$ and $\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat\pi}}[\mathbb{I}\{B'\}]$, which, combined with Eqs. (4) and (5), would give an upper bound on $\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat\pi}]$.

First of all, by Condition 1,

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat\pi}}[\mathbb{I}\{B_0\}] \le& \sup_\pi \mathbb{P}_{s\sim d^\pi}(0 < \Delta(s) \le 2a_n) \\ \le& 2^\alpha a_n^\alpha / \delta_0^\alpha. \end{aligned}$$

Second, for $i \ge 1$,

$$\begin{aligned} \mathbb{I}\{B_i\} =& \mathbb{I}\{\hat f(s,\hat\pi(s)) \ge \hat f(s,\pi^*(s)), 2^i a_n < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)) \le 2^{i+1} a_n\} \\ \le& \mathbb{I}\{Q^*(s,\pi^*(s)) - \hat f(s,\pi^*(s)) + \hat f(s,\hat\pi(s)) - Q^*(s,\hat\pi(s)) - 2^i a_n > 0, \\ & 0 < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)) \le 2^{i+1} a_n\} \\ \le& \mathbb{I}\{Q^*(s,\pi^*(s)) - \hat f(s,\pi^*(s)) > 2^{i-1} a_n, 0 < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)) \le 2^{i+1} a_n\} \\ & + \mathbb{I}\{\hat f(s,\hat\pi(s)) - Q^*(s,\hat\pi(s)) > 2^{i-1} a_n, 0 < Q^*(s,\pi^*(s)) - Q^*(s,\hat\pi(s)) \le 2^{i+1} a_n\} \\ \le& \mathbb{I}\{Q^*(s,\pi^*(s)) - \hat f(s,\pi^*(s)) > 2^{i-1} a_n, 0 < \Delta(s) \le 2^{i+1} a_n\} \\ & + \mathbb{I}\{\hat f(s,\hat\pi(s)) - Q^*(s,\hat\pi(s)) > 2^{i-1} a_n, 0 < \Delta(s) \le 2^{i+1} a_n\}. \end{aligned}$$

where the second inequality comes from a union bound, and the last one comes from the definition of Δ. Therefore, we have for $i \geq 1$,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat{\pi}}}[\mathbb{I}\{B_i\}] \\
\leq&\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat{\pi}}}[\mathbb{I}\{Q^*(s,\pi^*(s)) - \hat{f}(s,\pi^*(s)) > 2^{i-1}a_n, 0 < \Delta(s) \leq 2^{i+1}a_n\}] \\
&+ \mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat{\pi}}}[\mathbb{I}\{\hat{f}(s,\hat{\pi}(s)) - Q^*(s,\hat{\pi}(s)) > 2^{i-1}a_n, 0 < \Delta(s) \leq 2^{i+1}a_n\}] \\
\leq&\sup_{\pi}\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\pi}}[\mathbb{I}\{Q^*(s,\pi^*(s)) - \hat{f}(s,\pi^*(s)) > 2^{i-1}a_n, 0 < \Delta(s) \leq 2^{i+1}a_n\}] \\
&+ \sup_{\pi}\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\pi}}[\mathbb{I}\{\hat{f}(s,\pi(s)) - Q^*(s,\pi(s)) > 2^{i-1}a_n, 0 < \Delta(s) \leq 2^{i+1}a_n\}] \\
\leq&\sup_{\pi}\mathbb{E}_{s\sim d^{\pi}}[\mathbb{I}\{0 < \Delta(s) \leq 2^{i+1}a_n\}\mathbb{P}_{\mathcal{D}}(Q^*(s,\pi^*(s)) - \hat{f}(s,\pi^*(s)) > 2^{i-1}a_n)] \\
&+ \sup_{\pi}\mathbb{E}_{s\sim d^{\pi}}[\mathbb{I}\{0 < \Delta(s) \leq 2^{i+1}a_n\}\mathbb{P}_{\mathcal{D}}(\hat{f}(s,\pi(s)) - Q^*(s,\pi(s)) > 2^{i-1}a_n)].
\end{aligned}
$$

Then from Eq. (2) and Condition 1 we have for $i \in [1, i_{\max}]$,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat{\pi}}}[\mathbb{I}\{B_i\}] \leq& 2C\exp(-2^{2i-2})\sup_{\pi\in\Pi}\mathbb{P}_{s\sim d^{\pi}}(0 < \Delta(s) \leq 2^{i+1}a_n) \\
\leq& 2C\exp(-2^{2i-2})(2^{i+1}a_n/\delta_0)^{\alpha}.
\end{aligned}
$$

Finally, since

$$
\begin{aligned}
\mathbb{I}\{B'\} =&\mathbb{I}\{\hat{f}(s,\hat{\pi}(s)) \geq \hat{f}(s,\pi^*(s)), 2^{i_{\max}+1}a_n < Q^*(s,\pi^*(s)) - Q^*(s,\hat{\pi}(s))\} \\
\leq&\mathbb{I}\{Q^*(s,\pi^*(s)) - \hat{f}(s,\pi^*(s)) + \hat{f}(s,\hat{\pi}(s)) - Q^*(s,\hat{\pi}(s)) - 2^{i_{\max}+1}a_n > 0\} \\
\leq&\mathbb{I}\{Q^*(s,\pi^*(s)) - \hat{f}(s,\pi^*(s)) > 2^{i_{\max}}a_n\} + \mathbb{I}\{\hat{f}(s,\hat{\pi}(s)) - Q^*(s,\hat{\pi}(s)) > 2^{i_{\max}}a_n\},
\end{aligned}
$$

by Eq. (2) we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\mathbb{E}_{s\sim d^{\hat{\pi}}}[\mathbb{I}\{B'\}] \leq& \sup_{\pi\in\Pi}\mathbb{E}_{s\sim d^{\pi}}[\mathbb{P}_{\mathcal{D}}(Q^*(s,\pi^*(s)) - \hat{f}(s,\pi^*(s)) > 2^{i_{\max}}a_n)] \\
&+ \sup_{\pi\in\Pi}\mathbb{E}_{s\sim d^{\pi}}[\mathbb{P}_{\mathcal{D}}(\hat{f}(s,\pi(s)) - Q^*(s,\pi(s)) > 2^{i_{\max}}a_n)] \\
\leq& 2C\exp\left(-2^{2i_{\max}}\right) \\
\leq& 2C\exp(-\delta_1^2/(4a_n)^2),
\end{aligned}
$$

where the last inequality comes from the definition of $i_{\max}$.

Putting all pieces together we get

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[V^* - V^{\hat{\pi}}] \leq& \frac{2^{\alpha+1}}{(1-\gamma)\delta_0^{\alpha}}\left(1 + \sum_{i=1}^{i_{\max}} C\exp\left(-2^{2i-2}\right)2^{(\alpha+1)i+1}\right)a_n^{\alpha+1} \\
&+ \frac{2Q_{\max}C}{1-\gamma}\exp\left(-\delta_1^2/(4a_n)^2\right).
\end{aligned}
$$

□

Proof of Corollary 2. The statement comes from setting $\delta_1 = \infty$ in Theorem 1. We now provide an upper bound on $c(\alpha) = \sum_{i=1}^{\infty} \exp\left(-2^{2i-2}\right) 2^{(\alpha+1)i+1}$.

Define $f(x) = \exp(-2^{2x-2})2^{(\alpha+1)x}$. The maximizer of $f(x)$ is $x_0 = \frac{\log(\alpha+1)+\log 2}{2\log 2}$, and $f(x_0) = \left(\frac{2(\alpha+1)}{e}\right)^{(\alpha+1)/2}$. Therefore,

$$\begin{aligned}
\sum_{i=1}^{\infty} \exp\left(-2^{2i-2}\right) 2^{(\alpha+1)i+1} =& 2\sum_{i=1}^{\lfloor x_0 \rfloor - 1} f(i) + 2f(\lfloor x_0 \rfloor) + 2\sum_{i=\lfloor x_0 \rfloor + 1}^{\infty} f(i) \\
\leq& 2\int_1^{\lfloor x_0 \rfloor} f(x)dx + 2f(x_0) + 2\int_{\lfloor x_0 \rfloor}^{\infty} f(x)dx \\
=& 2\int_1^{\infty} f(x)dx + 2f(x_0) \\
=& \frac{2^{\alpha+1}\Gamma\left(\frac{\alpha+1}{2}, 1\right)}{\log 2} + 2\left(\frac{2(\alpha+1)}{e}\right)^{(\alpha+1)/2}.
\end{aligned}$$

□

Proof of Lemma 4. First, for clarity, we provide a very short proof of a weaker result where $\delta_0 = (6\mu_{\max} \sum_{a\in\mathcal{A}} \sum_{a'\in\mathcal{A}:\beta_a\neq\beta_{a'}} \|\beta_a - \beta_{a'}\|^{-1})^{-1}$. For simplicity suppose all $\{\beta_a : a \in \mathcal{A}\}$ are distinct; otherwise we can simply eliminate duplicates. Letting $V_d(R)$ be the volume of the R-radius d-ball, we have for any $\pi \in \Pi$,

$$\begin{aligned}
\mathbb{P}_{s\sim d^\pi}\left(0 < \Delta(s) \leq \delta\right) &= \mathbb{P}_{s\sim d^\pi}\left(\exists a \neq a' : 0 < Q(s,a) - Q(s,a') \leq \delta\right) \\
&\leq \sum_{a\neq a'} \mathbb{P}_{s\sim d^\pi}\left(0 < (\beta_a - \beta_{a'})^\intercal \psi(s) \leq \delta\right) \\
&\leq \mu_{\max} \sum_{a\neq a'} \int_0^{\delta/\|\beta_a - \beta_{a'}\|} V_{d-1}((1-u^2)_+^{1/2})du \\
&\leq \mu_{\max} \sum_{a\neq a'} 6\delta / \left\|\beta_a - \beta_{a'}\right\|,
\end{aligned}$$

since the volume of a unit ball in any dimension is always less than $\frac{8\pi^2}{15} \leq 6$.

Now we present an argument to tighten the above so that the inner sum in δ_0 becomes a max. Again suppose all $\{\beta_a : a \in \mathcal{A}\}$ are distinct; else eliminate duplicates. Letting Vol

denote the Lebesgue measure and $\mathcal{B}_d = \{\|v\| \le 1\}$ the unit ball, we have

$$\begin{aligned}
&\mathbb{P}_{s\sim d^\pi}\left(0 < \Delta(s) \le \delta\right) \le \sum_{a'\in\mathcal{A}} \mathbb{P}_{s\sim d^\pi}\left(0 < \Delta(s) \le \delta,\, a' \in \operatorname*{argmax}_a Q^*(s,a)\right) \\
&\le \sum_{a'\in\mathcal{A}} \mathbb{P}_{s\sim d^\pi}\left(\forall a \ne a' : (\beta_{a'} - \beta_a)^\intercal \psi(s) \ge 0,\ \exists a \ne a' : (\beta_{a'} - \beta_a)^\intercal \psi(s) \le \delta\right) \\
&\le \mu_{\max} \sum_{a'\in\mathcal{A}} \mathrm{Vol}\left(\bigcup_{a\ne a'}\left(\mathcal{B}_d \cap \bigcap_{a''\in\mathcal{A}} \{(\beta_{a'} - \beta_{a''})^\intercal v \ge 0\} \cap \{(\beta_{a'} - \beta_a)^\intercal v \le \delta\}\right)\right) \\
&\le \mu_{\max} \sum_{a'\in\mathcal{A}} \mathrm{Vol}\left(\bigcup_{a\ne a'}\left(\mathcal{B}_d \cap \bigcap_{a''\ne a} \left\{\bar\beta_{a',a''}^\intercal v \ge 0\right\} \cap \left\{\bar\beta_{a',a}^\intercal v \le \delta / \min_{a\ne a'} \|\beta_{a'} - \beta_a\|\right\}\right)\right),
\end{aligned}$$

where $\bar\beta_{a',a} = \frac{\beta_{a'} - \beta_a}{\|\beta_{a'} - \beta_a\|}$. The first inequality is by union bound, the second by definition of Δ and including 0 in the event, the third by the uniform upper bound on d^π, and the fourth by inclusion as we are only increasing the half space in the last term for each a', a. We will next show that the inner volume term is upper bounded by $6\delta / \min_{a\ne a'} \|\beta_{a'} - \beta_a\|$, yielding the result.

We now study the inner volume term. To abstract things, consider $\beta_1, \dots, \beta_m$ with $\|\beta_i\| = 1$ and the polyhedral cones $K^{(k)} = \{v : \beta_i^\intercal v \ge 0\, \forall i = 1, \dots, k\}$ for every $k = 1, \dots, m$. We are then concerned with

$$V = \mathrm{Vol}\left(\bigcup_{i=1}^m \left(\mathcal{B}_d \cap K^{(m)} \cap \left\{\beta_i^\intercal v \le \delta'\right\}\right)\right).$$

Placing a prism of height δ' (in the direction of β_i) on top of $\mathcal{B}_d \cap K^{(m)} \cap \{\beta_i^\intercal v = 0\}$ for each i, we see that the sum of the prims' volumes upper bounds V: we are only overcounting volume outside the sphere and any overlaps between the prisms placed at different faces. Let $\partial\mathcal{B}_d = \{\|v\| = 1\}$ be the unit sphere shell and let $\rho = d - 1$ be the proportionality between the volume inside the $(d-1)$-dimensional unit sphere and its area. Notice that the sum of the prisms' volume is equal to δ' times ρ^{-1} times the perimeter of the spherical polyhedron that $K^{(m)}$ defines on the unit sphere, that is, $\left|\partial K^{(m)} \cap \partial\mathcal{B}_d\right|$. We claim that $\left|\partial K^{(m)} \cap \partial\mathcal{B}_d\right| \le \left|\partial K^{(m-1)} \cap \partial\mathcal{B}_d\right|$. If β_m does not intersect the interior of $K^{(m)}$ then this is trivial. Suppose it does intersect. Let $H = \{\beta_m^\intercal v \ge 0\}$ and $H' = \{\beta_m^\intercal v \le 0\}$. Then $\left|K^{(m-1)} \cap \partial H \cap \partial\mathcal{B}_d\right| \le \left|\partial K^{(m-1)} \cap H' \cap \partial\mathcal{B}_d\right|$ because if we project $\partial K^{(m-1)} \cap H' \cap \partial\mathcal{B}_d$ onto $\partial H \cap \partial\mathcal{B}_d$ then we obtain $K^{(m-1)} \cap \partial H \cap \partial\mathcal{B}_d$ (that is, one side of a spherical polyhedron cannot be larger than the sum of the other sides). Therefore, by adding β_m to $K^{(m-1)}$ to obtain $K^{(m)}$, we have lost more perimeter than we have gained, as was claimed. By repeating this, we obtain $\left|\partial K^{(m)} \cap \partial\mathcal{B}_d\right| \le \left|\partial K^{(1)} \cap \partial\mathcal{B}_d\right|$. Next, notice that $\left|\partial K^{(1)} \cap \partial\mathcal{B}_d\right| / \rho$ is just the volume of the $(d-1)$-dimensional unit ball, which is $\frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)} \le 6$. Hence, $V \le 6\delta'$. □

A.2 Proofs for Section 3

Proof of Lemma 6.

Preliminary. Let $y_i^f = r_i + \gamma \max_{a' \in \mathcal{A}} f(s_i', a')$. We can write

$$y_i^f = w_f^\intercal \phi(s_i, a_i) + \epsilon_i^f,$$

where $\epsilon_i^f = r_i + \gamma \max_{a' \in \mathcal{A}} f(s_i', a') - \mathcal{T} f(s_i, a_i)$. Note that $\mathbb{E}_{r_i, s_i' \sim P_s(\cdot | s_i, a_i)}[\epsilon_i^f \mid s_i, a_i] = 0$.

When the event $\lambda_{\min}(\hat{\Sigma}) > n\lambda_0/2$ holds,

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \left\| \hat{w}_f' - w_f \right\| &= \sup_{f \in \mathcal{F}} \left\| \hat{\Sigma}^{-1} \left(\sum_{i=1}^n \phi(s_i, a_i) \epsilon_i^f \right) \right\| \\
&\leq \sup_{f \in \mathcal{F}} \left\| \hat{\Sigma}^{-1} \right\|_2 \left\| \sum_{i=1}^n \phi(s_i, a_i) \epsilon_i^f \right\| \\
&\leq \frac{2}{n\lambda_0} \sup_{f \in \mathcal{F}} \left\| \sum_{i=1}^n \phi(s_i, a_i) \epsilon_i^f \right\|.
\end{aligned}$$

Therefore, from a union bound we get

$$\begin{aligned}
\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left\| \hat{w}_f' - w_f \right\| \geq \delta \right) \leq & \mathbb{P}\left(\lambda_{\min}(\hat{\Sigma}) \leq n\lambda_0/2 \right) + \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left\| \sum_{i=1}^n \phi(s_i, a_i) \epsilon_i^f \right\| \geq n\lambda_0 \delta / 2 \right) \\
\leq & \underbrace{\mathbb{P}\left(\lambda_{\min}(\hat{\Sigma}) \leq n\lambda_0/2 \right)}_{\textbf{(a)}} + \sum_{j=1}^d \underbrace{\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \phi_j(s_i, a_i) \epsilon_i^f \right| \geq \frac{n\lambda_0 \delta}{2\sqrt{d}} \right)}_{\textbf{(b)}},
\end{aligned}$$

where ϕ_j is the j-th component of ϕ. We now aim to bound the two terms on the right hand side.

Bounding (a). Note that

$$\begin{aligned}
&\lambda_{\max}\left(\phi(s_i, a_i) \phi(s_i, a_i)^\intercal \right) = \max_{\|u\|=1} u^\intercal \phi(s_i, a_i) \phi(s_i, a_i)^\intercal u \leq 1, \\
&\lambda_{\min}\left(\sum_{i=1}^n \mathbb{E} \phi(s_i, a_i) \phi(s_i, a_i)^\intercal \right) \geq \sum_{i=1}^n \lambda_{\min}(\mathbb{E} \phi(s_i, a_i) \phi(s_i, a_i)^\intercal) = n\lambda_0.
\end{aligned}$$

By matrix Chernoff inequality (Tropp (2015, Theorem 5.1.1)),

$$\mathbb{P}\left(\lambda_{\min}(\hat{\Sigma}) \leq n\lambda_0/2 \right) \leq d \exp(-n\lambda_0/8).$$

Bounding (b). Let $h^f(s, a, r, s') = \phi_j(s, a)(r + \gamma \max_{a' \in \mathcal{A}} f(s', a') - \mathcal{T} f(s, a))$. We have $\mathbb{E}[h^f(s, a, r, s')] = 0$. Define $h_i^f = h(s_i, a_i, r_i, s_i')$, $\mathbf{h}^f = \left(h_1^f, \ldots, h_n^f \right)$, and $\mathbf{H} = \{\mathbf{h}^f : f \in \mathcal{F}\}$. Note that $\sup_{f \in \mathcal{F}} \left| h_i^f \right| \leq M + B$ for each i.

By Pollard (1990, Theorem 2.2), for any convex increasing Φ,

$$\begin{aligned}\mathbb{E}\Phi\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\phi_j(s_i,a_i)\epsilon_i^f\right|\right) =&\mathbb{E}\Phi\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}h_i^f\right|\right)\\ \leq&\mathbb{E}\mathbb{E}_\sigma\Phi\left(2\sup_{\mathbf{h}\in\mathbf{H}}|\langle\sigma,\mathbf{h}\rangle|\right).\end{aligned}$$

Let $\Psi(x)=\frac{1}{5}\exp(x^2)$. By Pollard (1990, Theorem 3.5),

$$\mathbb{E}\mathbb{E}_\sigma\Psi\left(\frac{1}{J}\sup_{\mathbf{h}\in\mathbf{H}}|\langle\sigma,\mathbf{h}\rangle|\right)\leq 1,\quad \text{where } J=9\int_0^{\sqrt{n}(M+B)}\sqrt{\log D(\delta,\mathbf{H})}d\delta.$$

Since

$$\left\|\mathbf{h}^f-\mathbf{h}^g\right\|\leq 2\gamma\sqrt{n}\left\|f-g\right\|_\infty,$$

we have

$$\begin{aligned}\log D(\delta,\mathbf{H})\leq&\log D\left(\frac{\delta}{2\gamma\sqrt{n}},\mathcal{F},\|\cdot\|_\infty\right)\\ \leq&d\log\left(1+8\gamma B\sqrt{n}/\delta\right).\end{aligned}$$

where the last inequality comes from Wainwright (2019, Lemma 5.5 and Example 5.8). This implies

$$\begin{aligned}J\leq&9\sqrt{nd}(M+B)\int_0^1\sqrt{\log\left(1+8/\delta'\right)}d\delta'\\ \leq&18\sqrt{nd}(M+B).\end{aligned}$$

Combining all pieces we get for all $\delta>0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\phi_j(s_i,a_i)\epsilon_i^f\right|>\delta\right)\leq 5\exp\left(-\frac{\delta^2}{1296nd(M+B)^2}\right).$$

Bounding the Error $\sup_{f\in\mathcal{F}}\left\|\hat{w}_f^\intercal\phi-\mathcal{T}f\right\|_\infty$**.** Recall that $\hat{w}_f$ is the projection of $\hat{w}'_f$ onto $\mathcal{B}(0,B)$, so we naturally have $\mathbb{P}\left(\sup_{f\in\mathcal{F}}\|\hat{w}_f-w_f\|\geq\delta\right)=0$ for $\delta>2B$. On the other hand, when $\delta\leq 2B$,

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left\|\hat{w}_f^\intercal\phi-\mathcal{T}f\right\|_\infty\geq\delta\right)\leq&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\|\hat{w}_f-w_f\|\geq\delta\right)\\ \leq&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left\|\hat{w}'_f-w_f\right\|\geq\delta\right)\\ \leq&5d\exp\left(-\frac{n\lambda_0^2\delta^2}{5184d^2(M+B)^2}\right)+d\exp\left(-n\lambda_0/8\right)\\ \leq&6d\exp\left(-\frac{\lambda_0^2}{5184d^2(M+B)^2}n\delta^2\right).\end{aligned}$$

where we used the fact that $\lambda_0\leq 1$. Our conclusion then follows. □

Proof of Theorem 7. Note that for any $k \in [K]$,

$$Q^*(s,a) = \mathcal{T}Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q^*(s',a'),$$
$$\mathcal{T}\hat{f}_{k-1}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} \hat{f}_{k-1}(s',a'),$$

so we have

$$\begin{aligned}
\|Q^* - \mathcal{T}\hat{f}_{k-1}\|_\infty \leq & \sup_{s,a} \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \mid \max_{a' \in \mathcal{A}} Q^*(s',a') - \max_{a' \in \mathcal{A}} \hat{f}_{k-1}(s',a')| \\
\leq & \sup_{s,a} \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} |Q^*(s',a') - \hat{f}_{k-1}(s',a')| \\
\leq & \gamma \|Q^* - \hat{f}_{k-1}\|_\infty.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|Q^* - \hat{f}_k\|_\infty \leq & \|Q^* - \mathcal{T}\hat{f}_{k-1}\|_\infty + \|\hat{f}_k - \mathcal{T}\hat{f}_{k-1}\|_\infty \\
\leq & \gamma \|Q^* - \hat{f}_{k-1}\|_\infty + \|\hat{f}_k - \mathcal{T}\hat{f}_{k-1}\|_\infty.
\end{aligned}$$

We can recursively repeat the same process for $\|Q^* - \hat{f}_{k-1}\|_\infty$ till $k = 0$, and get

$$\begin{aligned}
\|Q^* - \hat{f}_K\|_\infty \leq & \sum_{t=0}^{K-1} \gamma^t \|\hat{f}_{K-t} - \mathcal{T}\hat{f}_{K-t-1}\|_\infty + \gamma^K \|Q^* - \hat{f}_{K-1}\|_\infty \\
\leq & \sum_{t=0}^{K-1} \gamma^t \|\hat{f}_{K-t} - \mathcal{T}\hat{f}_{K-t-1}\|_\infty + \frac{\gamma^K M}{1-\gamma}.
\end{aligned}$$

Note that $K \geq \frac{\log(\lambda_0^2 n/(5184 d^2))}{2\log(1/\gamma)}$ implies that $\frac{\gamma^K M}{(1-\gamma)} \leq \frac{72d(M+B)}{(1-\gamma)\lambda_0\sqrt{n}}$. Moreover, by Lemma 6, we know that for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=0}^{k-1} \gamma^t \|f_{k-t} - \mathcal{T}f_{k-t-1}\|_\infty \geq \frac{\delta}{2}\right) \leq & \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left\|\hat{w}_f^\intercal \phi - \mathcal{T}f\right\|_\infty \geq \frac{\delta(1-\gamma)}{2}\right) \\
\leq & 6d \exp(-\frac{(1-\gamma)^2 \lambda_0^2}{20736 d^2 (M+B)^2} n\delta^2).
\end{aligned}$$

Therefore, for any $\delta \geq \frac{144d(M+B)}{(1-\gamma)\lambda_0\sqrt{n}}$,

$$\mathbb{P}(\|Q^* - f_k\|_\infty \geq \delta) \leq 6d \exp(-\frac{(1-\gamma)^2 \lambda_0^2}{20736 d^2 (M+B)^2} n\delta^2).$$

□

A.3 Proof for Section 4

Proof of Theorem 9.

Preliminary We introduce $\mathbb{E}_n[f(\cdot)] = 1/n \sum_i f(s_i, a_i, r_i, s_i')$. In addition, we use $\mathbb{E}[\cdot]$ to present $\mathbb{E}_{\mu_b}[\cdot]$. We define

$$\begin{aligned}
\Phi(q; w) &= \mathbb{E}[\{r - q(s,a) + \gamma \max_{a'} q(s', a')\} w(s,a)], \\
\Phi_n(q; w) &= \mathbb{E}_n[\{r - q(s,a) + \gamma \max_{a'} q(s', a')\} w(s,a)], \\
\Phi_n^\zeta(q; w) &= \Phi_n(q; w) - \zeta \mathbb{E}_n[w^2], \\
\Phi^\zeta(q; w) &= \Phi(q; w) - \zeta \mathbb{E}[w^2].
\end{aligned}$$

Let η_n be the upper bound of the critical radius of $\mathcal{F}_w$ and

$$\mathcal{G}_q = \{(s,a) \mapsto w(s,a)\{-q(s,a) + \gamma \max_{a'} q(s', a') + Q^*(s,a) + \gamma \max_{a'} Q^*(s', a')\} : w \in \mathcal{F}_w, q \in \mathcal{F}\}.$$

From a standard analysis, $\eta_n = c\sqrt{d \log n / n}$. Then, from Wainwright (2019, Theorem 14.1), with $1 - c_0 \exp(-c_1 n \eta^2 / M'^2)$, for any $\eta \geq \eta_n$, we have

$$\forall w(s,a) \in \mathcal{F}_w,\ |\mathbb{E}_n[w^2] - \mathbb{E}[w^2]| \leq 0.5 \mathbb{E}[w^2] + \eta^2 \tag{6}$$

noting η_n upper bounds the critical radius of $\mathcal{F}_w$.

First Step By definition of $\hat{f}$ and $Q^* \in \mathcal{F}_q$, we have

$$\sup_{w \in \mathcal{F}_w} \Phi_n^\zeta(\hat{f}; w) \leq \sup_{w \in \mathcal{F}_w} \Phi_n^\zeta(Q^*; w) \tag{7}$$

From Wainwright (2019, Theorem 14.20), with probability $1 - c_0 \exp(-c_1 n \eta^2 / M'^2)$, for any $\eta \geq \eta_n$, we have

$$\forall w \in \mathcal{F}_w : |\Phi_n(Q^*; w) - \Phi(Q^*; w)| \leq c C_1 \{\eta \mathbb{E}[w^2]^{1/2} + \eta^2\}. \tag{8}$$

Here, we use $l(a_1, a_2) := a_1 a_2, a_1 = w(s,a), a_2 = r - q(s,a) + \gamma \max_{a'} q(s', a')$ is $2(1+\gamma)B$-Lipschitz with respect to a_1 by defining $C_1 = 2(1+\gamma)B$, that is,

$$|l(a_1, a_2) - l(a_1', a_2)| \leq C_1 |a_1 - a_1'|.$$

Thus,

$$\begin{aligned}
\sup_{w \in \mathcal{F}_w} \Phi_n^\zeta(Q^*; w) &= \sup_{w \in \mathcal{F}_w} \left\{\Phi_n(Q^*; w) - \zeta \mathbb{E}_n[w^2]\right\} && \text{Definition} \\
&\leq \sup_{w \in \mathcal{F}_w} \left\{\Phi(Q^*; w) + c C_1 \eta \mathbb{E}[w^2]^{1/2} + c C_1 \eta^2 - \zeta \mathbb{E}_n[w^2]\right\} && \text{From Eq. (8)} \\
&\leq \sup_{w \in \mathcal{F}_w} \left\{\Phi(Q^*; w) + c C_1 \eta \mathbb{E}[w^2]^{1/2} + c C_1 \eta^2 - 0.5\zeta \mathbb{E}[w^2] + \zeta \eta^2\right\} && \text{From Eq. (6)} \\
&\leq \sup_{w \in \mathcal{F}_w} \left\{\Phi(Q^*; w) + (4c^2 C_1^2 / \zeta + c C_1 + \zeta) \eta^2\right\}. && (9)
\end{aligned}$$

In the last line, we use a general inequality, $a, b > 0$:

$$\sup_{w \in \mathcal{F}_w} (a \mathbb{E}[w^2]^{1/2} - b \mathbb{E}[w^2]) \leq a^2 / 4b.$$

Moreover,

$$\begin{aligned}
\sup_{w\in\mathcal{F}_w}\{\Phi_n^\zeta(\hat{f};w)\} &= \sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)+\Phi_n(Q^*;w)-\zeta\mathbb{E}_n[w^2]\}\\
&\geq \sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\}+\inf_{w\in\mathcal{F}_w}\{\Phi_n(Q^*;w)+\zeta\mathbb{E}_n[w^2]\}\\
&= \sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\}+\inf_{-w\in\mathcal{F}_w}\{\Phi_n(Q^*;-w)+\zeta\mathbb{E}_n[w^2]\}\\
&= \sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\}+\inf_{-w\in\mathcal{F}_w}\{-\Phi_n(Q^*;w)+\zeta\mathbb{E}_n[w^2]\}\\
&= \sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\}-\sup_{-w\in\mathcal{F}_w}\{\Phi_n(Q^*;w)-\zeta\mathbb{E}_n[w^2]\}\\
&= \sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\}-\sup_{w\in\mathcal{F}_w}\Phi_n^\zeta(Q^*;w).
\end{aligned}$$

Here, we use $\mathcal{F}_w$ is symmetric. Therefore,

$$\begin{aligned}
\sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\} &\leq \sup_{w\in\mathcal{F}_w}\left\{\Phi_n^\zeta(Q^*;w)\right\}+\sup_{w\in\mathcal{F}_w}\{\Phi_n^\zeta(\hat{f};w)\}\\
&\leq 2\sup_{w\in\mathcal{F}_w}\Phi_n^\zeta(Q^*;w)\\
&\leq \sup_{w\in\mathcal{F}_w}\left\{\Phi(Q^*;w)+(c^2 4C_1^2/\zeta+vC_1+\zeta)\eta^2\right\}\\
&= (c^2 4C_1^2/\zeta+vC_1+\zeta)\eta^2.
\end{aligned}$$

Second Step Define

$$w_q=(\mathcal{T}-I)q.$$

Suppose $\{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2}\geq\eta$, and let $\kappa=\eta/\{2\{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2}\}\in[0,0.5]$. Then, noting $\mathcal{F}_w$ is star-convex,

$$\sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\zeta\mathbb{E}_n[w^2]\}\geq\kappa\{\Phi_n(\hat{f},w_{\hat{f}})-\Phi_n(Q^*,w_{\hat{f}})\}-2\kappa^2\zeta\mathbb{E}_n[w_{\hat{f}}^2].$$

since $\kappa w_{\hat{f}}\in\mathcal{F}_w$. Then,

$$\begin{aligned}
\kappa^2\mathbb{E}_n[w_{\hat{f}}^2] &\leq \kappa^2\{1.5\mathbb{E}[w_{\hat{f}}^2]+0.5\eta^2\} && (\text{ Eq. (6)})\\
&\leq 3\eta^2. && (\text{Definition of }\kappa)
\end{aligned}$$

Therefore,

$$\sup_{w\in\mathcal{F}_w}\{\Phi_n(\hat{f};w)-\Phi_n(Q^*;w)-2\mathbb{E}_n[w^2]\}\geq\kappa\{\Phi_n(\hat{f},w_{\hat{f}})-\Phi_n(Q^*,w_{\hat{f}})\}-2\zeta\eta^2.$$

Using

$$\Phi_n(q,w_q)-\Phi_n(Q^*,w_q)=\mathbb{E}_n[\{-q(s,a)+Q^*(s,a)+\gamma\max_{a'}q(s',a')-\gamma\max_{a'}Q^*(s',a')\}w_q(s,a)],$$

from Wainwright (2019, Theorem 14.20), with probability $1 - c_0 \exp(-c_1 n\eta^2/M'^2)$, for any $\eta \geq \eta_n$, for any $q \in \mathcal{F}_q$,

$$
\begin{aligned}
&|\Phi_n(q, w_q) - \Phi_n(Q^*, w_q) - \{\Phi(q, w_q) - \Phi(Q^*, w_q)\}| \\
&= |\mathbb{E}_n[\{-q(s,a) + Q^*(s,a) + \gamma \max_{a'} q(s', a') - \gamma \max_{a'} Q^*(s', a')\} w_q(s,a)] \\
&- \mathbb{E}[\{-q(s,a) + Q^*(s,a) + \gamma \max_{a'} q(s', a') - \gamma \max_{a'} Q^*(s', a')\} w_q(s,a))]| \\
&\leq (\eta \mathbb{E}[\{-q(s,a) + Q^*(s,a) + \gamma \max_{a'} q(s', a') - \gamma \max_{a'} Q^*(s', a')\}^2 w_q^2(s,a)]^{1/2} + \eta^2) \\
&\leq (\eta M' \{\mathbb{E}[w_q]\}^{1/2} + \eta^2).
\end{aligned}
$$

Here, we invoke Wainwright (2019, Theorem 14.20) by treating $l(a_1, a_2) = a_1$, $a_1 = \{q(s,a) - Q^*(s,a) - \gamma \max_{a'} q(s', a') + \gamma \max_{a'} Q^*(s', a')\} w_q(s,a)\}$.

Thus,

$$
\begin{aligned}
\kappa\{\Phi_n(\hat{f}, w_{\hat{f}}) - \Phi_n(Q^*, w_{\hat{f}})\} &\geq \kappa\{\Phi(\hat{f}, w_{\hat{f}}) - \Phi(Q^*, w_{\hat{f}})\} - \kappa(M' \eta \{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2} + \eta^2). \\
&\qquad (\kappa \leq 0.5) \\
&\geq \kappa\{\Phi(\hat{f}, w_{\hat{f}}) - \Phi(Q^*, w_{\hat{f}})\} - \kappa(M' \eta \{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2}) - 0.5\eta^2 \\
&\overset{\text{(a)}}{=} \kappa \mathbb{E}[w_{\hat{f}}(s,a)(\mathcal{T} - I)\hat{f}(s,a)] - \kappa(M' \eta \{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2}) - 0.5\eta^2 \\
&= \frac{\eta}{2\mathbb{E}[w_{\hat{f}}^2]^{1/2}} \{\mathbb{E}[w_{\hat{f}}^2]^{1/2} - M' \eta \{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2}\} - 0.5\eta^2 \\
&\geq 0.5\eta \mathbb{E}[\{(\mathcal{T} - I)\hat{f}\}^2]^{1/2} - (0.5 + M')\eta^2.
\end{aligned}
$$

For (a), we use

$$
\begin{aligned}
\Phi(\hat{f}, w_{\hat{f}}) - \Phi(Q^*, w_{\hat{f}}) &= \mathbb{E}[w_{\hat{f}}(s,a)\{-\hat{f}(s,a) + Q^*(s,a) + \gamma \hat{\max}_{a'} f(s', a') - \gamma \max_{a'} Q^*(s', a')\}] \\
&= \mathbb{E}[w_{\hat{f}}(s,a)(\mathcal{T} - I)\hat{f}(s,a)].
\end{aligned}
$$

Third Step Thus, for $\eta > \eta_n$ with probability $1 - c_0 \exp(-c_1 n\eta^2/M'^2)$, $\{\mathbb{E}[w_{\hat{f}}^2]\}^{1/2} \leq \eta$ or

$$
\eta \mathbb{E}[\{(\mathcal{T} - I)\hat{f}\}^2]^{1/2} - (M' + \zeta)\eta^2 \leq c_2 \times (M'^2/\zeta + M' + \zeta)\eta^2.
$$

Therefore, for $\eta \geq \eta_n$, with $1 - c_0 \exp(-c_1 n\eta^2/M'^2)$, we have

$$
\|\hat{w} - w_0\| \lambda_0' \leq \mathbb{E}[\{(\mathcal{T} - I)\hat{f}\}^2]^{1/2} \leq c_2(M'^2/\zeta + M' + \zeta + 1)\eta.
$$

This implies for any $\delta \geq \eta_n$, we have

$$
\mathbb{P}(\|\hat{f} - Q^*\|_\infty \geq \delta) \leq \exp\left(-c_1 \frac{n\lambda_0^2 \delta^2}{M'^2\{(M'^2/\zeta + M' + \zeta + 1)\}}\right).
$$

In the end, for $\delta \geq a_n, a_n = c\{\sqrt{d} + M'\sqrt{M'^2/\zeta + M' + \zeta + 1}/\lambda_0\}\sqrt{\log n/n}$,

$$
\mathbb{P}(\|\hat{f} - Q^*\|_\infty \geq \delta) \leq \exp\left(-\delta^2/a_n\right).
$$

□