Risk Bounds and Rademacher Complexity in Batch Reinforcement Learning

Yaqi Duan [1] **Chi Jin** [2] **Zhiyuan Li** [3]

Abstract

This paper considers batch Reinforcement Learning (RL) with general value function approximation. Our study investigates the minimal assumptions to reliably estimate/minimize Bellman error, and characterizes the generalization performance by (local) Rademacher complexities of general function classes, which makes initial steps in bridging the gap between statistical learning theory and batch RL. Concretely, we view the Bellman error as a surrogate loss for the optimality gap, and prove the followings: (1) In double sampling regime, the excess risk of Empirical Risk Minimizer (ERM) is bounded by the Rademacher complexity of the function class. (2) In the single sampling regime, sample-efficient risk minimization is not possible without further assumptions, regardless of algorithms. However, with completeness assumptions, the excess risk of FQI and a minimax style algorithm can be again bounded by the Rademacher complexity of the corresponding function classes. (3) Fast statistical rates can be achieved by using tools of local Rademacher complexity. Our analysis covers a wide range of function classes, including finite classes, linear spaces, kernel spaces, sparse linear features, etc.

1. Introduction

Statistical learning theory, since its introduction in the late 1960's, has become one of the most important frameworks in machine learning, to study problems of inference or function estimation from a given collection of data (Hastie et al., 2009; Vapnik, 2013; James et al., 2013). The development of statistical learning has led to a series of new popular algorithms including support vector machines (Cortes & Vapnik, 1995; Suykens & Vandewalle, 1999), boosting (Freund et al., 1996; Schapire, 1999), as well as many successful applications in fields such as computer vision (Szeliski, 2010; Forsyth & Ponce, 2012), speech recognition (Juang & Rabiner, 1991; Jelinek, 1997), and bioinformatics (Baldi et al., 2001).

Notably, in the area of supervised learning, a considerable amount of effort has been spent on obtaining sharp risk bounds. These are valuable, for instance, in the problem of model selection—choosing a model of suitable complexity. Typically, these risk bounds characterize the excess risk—the suboptimality of the learned function compared to the best function within a given function class, via proper complexity measures of that function class. After a long line of extensive research (Vapnik, 2013; Vapnik & Chervonenkis, 2015; Bartlett et al., 2005; 2006), risk bounds are proved under very weak assumptions which do not require realizability—the prespecified function class contains the ground-truth. The complexity measures for general function classes have also been developed, including but not limited to metric entropy (Dudley, 1974), VC dimension (Vapnik & Chervonenkis, 2015) and Rademacher complexity (Bartlett & Mendelson, 2002). (See e.g. (Wainwright, 2019) for a textbook review.)

Concurrently, batch reinforcement learning (Lange et al., 2012; Levine et al., 2020)—a branch of Reinforcement Learning (RL) that learns from offline data, has been independently developed. This paper considers the value function approximation setting, where the learning agent aims to approximate the optimal value function from a restricted function class that encodes the prior knowledge. Batch RL with value function approximation provides an important foundation for the empirical success of modern RL, and leads to the design of many popular algorithms such as DQN (Mnih et al., 2015) and Fitted Q-Iteration with neural networks (Riedmiller, 2005; Fan et al., 2020).

Despite being a special case of supervised learning, batch RL also brings several unique challenges due to the additional requirement of learning the rich temporal structures within the data. Addressing these unique challenges has been the main focus of the field so far (Levine et al., 2020). Consequently, the field of statistical learning and batch RL

*Equal contribution [1]Department of Operations Research and Financial Engineering, Princeton University [2]Department of Electrical and Computer Engineering, Princeton University [3]Department of Computer Science, Princeton University. Correspondence to: Yaqi Duan <yaqid@princeton.edu>, Chi Jin <chij@princeton.edu>, Zhiyuan Li <zhiyuanli@cs.princeton.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

have been developed relatively in parallel. In contrast to the mild assumptions required and the generic function class allowed in classical statistical learning theory, a majority of batch RL results (Munos & Szepesvári, 2008; Antos et al., 2008; Lazaric et al., 2012; Chen & Jiang, 2019) remain under rather strong assumptions which rarely hold in practice, and are applicable only to a restricted set of function classes. This raises a natural question: can we bring the rich knowledge in statistical learning theory to advance our understanding in batch RL?

This paper makes initial steps in bridging the gap between statistical learning theory and batch RL. We investigate the minimal assumptions required to reliably estimate or minimize the Bellman error, and characterize the generalization performance of batch RL algorithms by (local) Rademacher complexities of general function classes. Concretely, we establish conditions when the Bellman error can be viewed as a surrogate loss for the optimality gap in values. We then bound the excess risk measured in Bellman errors. We prove the followings:

- In the double sampling regime, the excess risk of a simple Empirical Risk Minimizer (ERM) is bounded by the Rademacher complexity of the function class, under almost no assumptions.
- In the single sampling regime, without further assumptions, no algorithm can achieve small excess risk in the worse case unless the number of samples scales up polynomially with respect to the number of states.
- In the single sampling regime, under additional completeness assumptions, the excess risks of Fitted Q-Iteration (FQI) algorithm and a minimax style algorithm can be again bounded by the Rademacher complexity of the corresponding function classes.
- Fast statistical rates can be achieved by using tools of local Rademacher complexity.

Finally, we specialize our generic theory to concrete examples, and show that our analysis covers a wide range of function classes, including finite classes, linear spaces, kernel spaces, sparse linear features, etc.

1.1. Related Work

We restrict our discussions in this section to the RL results under function approximation.

Batch RL There exists a stream of literature regarding finite sample guarantees for batch RL with value function approximation. Among the works, fitted value iteration (Munos & Szepesvári, 2008) and policy iteration (Antos et al., 2008; Farahmand et al., 2008; Lazaric et al., 2012; Farahmand et al., 2016; Le et al., 2019) are canonical and popular approaches. When using a linear function space, the sample complexity for batch RL is shown to depend on the dimension (Lazaric et al., 2012). When it comes to general function classes, several complexity measures of function class such as metric entropy and VC dimensions have been used to bound the performance of fitted value iteration and policy iteration (Munos & Szepesvári, 2008; Antos et al., 2008; Farahmand et al., 2016).

Throughout the existing theoretical studies of batch RL, people commonly use concentrability, realizability and completeness assumptions to prove polynomial sample complexity. Chen & Jiang (2019) justify the necessity of low concentrability and hold a debate on realizability and completeness. Xie & Jiang (2020a) develop an algorithm that only relies on the realizability of optimal Q-function and circumvents completeness condition. However, they use a stronger concentrability assumption and the error bound has a slower convergence rate. While the analyses in Chen & Jiang (2019) and Xie & Jiang (2020a) are restricted to discrete function classes with a finite number of elements, Wang et al. (2020a) investigate value function approximation with linear spaces. It is shown that data coverage and realizability conditions are not sufficient for polynomial sample complexity in the linear case.

Off-policy evaluation Off-policy evaluation (OPE) refers to the estimation of value function given offline data (Precup, 2000; Precup et al., 2001; Xie et al., 2019; Uehara et al., 2020; Kallus & Uehara, 2020; Yin et al., 2020; Uehara et al., 2021), which can be viewed as a subroutine of batch RL. Combining OPE with policy improvement leads to policy-iteration-based or actor-critic algorithms (Dann et al., 2014). OPE is considered as a simpler problem than batch RL and its analyses cannot directly translate to guarantees in batch RL.

Online RL RL in online mode is in general a more difficult problem than batch RL. The role of value function approximation in online RL remains largely unclear. It requires better tools to measure the capacity of function class in an online manner. In the past few years, there are some investigations in this direction, including using Bellman rank (Jiang et al., 2017) and Eluder dimension (Wang et al., 2020b) to characterize the hardness of RL problem.

1.2. Notation

For any integer $K > 0$, let $[K]$ be the collection of $1, 2, \ldots, K$. We use $\mathbb{1}[\cdot]$ to denote the indicator function. For any function $q(\cdot)$ and any measure ρ over the domain of q, we define norm $\|\cdot\|_\rho$ where $\|q\|_\rho^2 := \mathbb{E}_{x\sim\rho} q^2(x)$. Let ρ_1 be a measure over $\mathcal{X}_1$ and $\rho_2(\cdot \mid x_1)$ be a conditional distribution over $\mathcal{X}_2$. Define $\rho_1 \times \rho_2$ as a joint distribution over

$\mathcal{X}_1 \times \mathcal{X}_2$, given by $(\rho_1 \times \rho_2)(x_1, x_2) := \rho_1(x_1)\rho_2(x_2 \mid x_1)$. For any finite set $\mathcal{X}$, let $\mathrm{Unif}(\mathcal{X})$ define a uniform distribution over $\mathcal{X}$.

2. Preliminaries

We consider the setting of episodic Markov decision process $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the set of states which possibly has infinitely many elements; $\mathcal{A}$ is a finite set of actions with $|\mathcal{A}| = A$; H is the number of steps in each episode; $\mathbb{P}_h(\cdot \mid s, a)$ gives the distribution over the next state if action a is taken from state s at step $h \in [H]$; and $r_h \colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function at step h. [1]

In each episode of an MDP, we start with **a fixed initial state** s_1. Then, at each step $h \in [H]$, the agent observes state $s_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, receives reward $r_h(s_h, a_h)$, and then transitions to the next state s_{h+1}, which is drawn from the distribution $\mathbb{P}_h(\cdot \mid s_h, a_h)$. Without loss of generality, we assume there is a terminating state s_{end} which the environment will *always* transit to at step $H + 1$, and the episode terminates when s_{end} is reached.

A (non-stationary, stochastic) policy π is a collection of H functions $\{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h \in [H]}$, where $\Delta_{\mathcal{A}}$ is the probability simplex over action set $\mathcal{A}$. We denote $\pi_h(\cdot \mid s)$ as the action distribution for policy π at state s and time h. Let $V_h^\pi \colon \mathcal{S} \to \mathbb{R}$ denote the value function at step h under policy π, which gives the expected sum of remaining rewards received under policy π, starting from $s_h = s$, until the end of the episode. That is,

$$V_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\middle|\, s_h = s \right].$$

Accordingly, the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at step h is defined as,

$$Q_h^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\middle|\, s_h = s, a_h = a \right].$$

Since the action spaces, and the horizon, are all finite, there always exists (see, e.g., (Puterman, 2014)) an optimal policy $\pi^\star$ which gives the optimal value $V_h^\star(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

For notational convenience, we take shorthands $\mathbb{P}_h, \mathbb{P}_h^\pi, \mathbb{P}_h^\star$ as follows, where (s, a) is the state-action pair for the current step, while (s', a') is the state-action pair for the next step,

$$\begin{aligned}
[\mathbb{P}_h V](s, a) &:= \mathbb{E}\big[V(s') \,\big|\, s, a\big], \\
[\mathbb{P}_h^\pi Q](s, a) &:= \mathbb{E}_\pi\big[Q(s', a') \big| s, a\big], \\
[\mathbb{P}_h^\star Q](s, a) &:= \mathbb{E}\big[\max_{a'} Q(s', a') \,\big|\, s, a\big].
\end{aligned}$$

We further define Bellman operators $\mathcal{T}_h^\pi, \mathcal{T}_h^\star : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ for $h \in [H]$ as

$$\begin{aligned}
(\mathcal{T}_h^\pi Q)(s, a) &:= (r_h + \mathbb{P}_h^\pi Q)(s, a), \\
(\mathcal{T}_h^\star Q)(s, a) &:= (r_h + \mathbb{P}_h^\star Q)(s, a).
\end{aligned}$$

Then the Bellman equation and the Bellman optimality equation can be written as:

$$Q_h^\pi(s, a) = (\mathcal{T}_h^\pi Q_{h+1}^\pi)(s, a), \; Q_h^\star(s, a) = (\mathcal{T}_h^\star Q_{h+1}^\star)(s, a).$$

The objective of RL is to find a near-optimal policy, where the sub-optimality is measured by $V_1^\star(s_1) - V_1^\pi(s_1)$. Accordingly, we have the following definition of ϵ-optimal policy.

Definition 2.1 (ϵ-optimal policy)**.** We say a policy π is **ϵ-optimal** if $V_1^\star(s_1) - V_1^\pi(s_1) \le \epsilon$.

2.1. (Local) Rademacher complexity

In this paper, we leverage Rademacher complexity to characterize the complexity of a function class. For a generic real-valued function space $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and n fixed data points $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, the empirical Rademacher complexity is defined as

$$\widehat{\mathcal{R}}_X(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \,\middle|\, X \right],$$

where $\sigma_i \sim \mathrm{Uniform}(\{-1, 1\})$ are *i.i.d.* Rademacher random variables and the expectation is taken with respect to the uncertainties in $\{\sigma_i\}_{i=1}^n$. Let ρ be the underlying distribution of x_i. We further define a population Rademacher complexity $\mathcal{R}_n^\rho(\mathcal{F}) := \mathbb{E}_\rho[\widehat{\mathcal{R}}_X(\mathcal{F})]$ with expectation taken over data samples X. Intuitively, $\mathcal{R}_n^\rho(\mathcal{F})$ measures the complexity of $\mathcal{F}$ by the extent to which functions in the class $\mathcal{F}$ correlate with random noise σ_i.

This paper further uses the tools of local Rademacher complexity to obtain results with fast statistical rate. For a generic real-valued function space $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, and data distribution ρ. Let T be a functional $T : \mathcal{F} \to \mathbb{R}^+$, we study the local Radmacher complexity in the form of

$$\mathcal{R}_n^\rho(\{f \in \mathcal{F} \mid T(f) \le r\}).$$

A crucial quantity that appears in the generalization error bound using local Rademacher complexity is the critical radius (Bartlett et al., 2005). We define as follows.

[1] While we study deterministic reward functions for notational simplicity, our results generalize to randomized reward functions. Note that we are assuming that rewards are in $[0, 1]$ for normalization.

Definition 2.2 (Sub-root function)**.** A function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is sub-root if it is nondecreasing, and $r \to \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Definition 2.3 (Critical radius of local Radmacher complexity)**.** The critical radius of the local Radmacher complexity $\mathcal{R}_n^\rho(\{f \in \mathcal{F} \mid T(f) \le r\})$ is the infimum of the set $\mathfrak{B}$, where set $\mathfrak{B}$ is defined as follows: for any $r^\star \in \mathfrak{B}$, there exists a sub-root function ψ such that $r^\star$ is the fixed point of ψ, and for any $r \ge r^\star$ we have

$$\psi(r) \ge \mathcal{R}_n^\rho(\{f \in \mathcal{F} \mid T(f) \le r\}). \tag{1}$$

We typically obtain an upper bound of this critical radius by constructing one specific sub-root function ψ satisfying (1).

3. Batch RL with Value Function Approximation

This paper focuses on the offline setting where the data in form of tuples $\mathcal{D} = \{(s, a, r, s', h)\}$ are collected beforehand, and are given to the agent. In each tuple, (s, a) are the state and action at the h^{th} step, r is the resulting reward, and s' is the next state sampled from $\mathbb{P}_h(\cdot|s, a)$. For each $h \in [H]$, we have access to n data, that are *i.i.d* sampled with marginal distribution μ_h over (s, a) at the h^{th} step. We denote $\mu = \mu_1 \times \mu_2 \times \ldots \times \mu_H$. For each $h \in [H]$, we further denote the marginal distribution of s' in tuple (s, a, s', h) as ν_h , and let $\nu = \nu_1 \times \nu_2 \times \ldots \times \nu_H$. Throughout this paper, we will consistently use μ and ν to only denote the probability measures defined above.

We assume data distribution μ is well-behaved and satisfies the following assumption.

Assumption 1 (Concentrability)**.** Given a policy π, let P_h^π denote the marginal distribution at time step h, starting from s_1 and following π. There exists a parameter C such that

$$\sup_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{\mathrm{d}P_h^\pi}{\mathrm{d}\mu_h}(s, a) \le C \qquad \text{for any policy } \pi.$$

Assumption 1 requires that for any state-action pair (s, a), if there exists a policy π that reaches (s, a) with some descent amount of probability, then the chance that sample (s, a) appears in the dataset would not be low. Intuitively, Assumption 1 ensures that the dataset $\mathcal{D}$ is representative for all the "reachable" state-action pairs. The assumption is frequently used in the literature of batch RL, e.g. equation (7) in Munos (2003), Definition 5.1 in Munos (2007), Proposition 1 in Farahmand et al. (2010), Assumption 1 in Chen & Jiang (2019), etc. We remark that Assumption 1 here is the only assumption of this paper regarding the properties of the batch data.

We consider the setting of value function approximation, where at each step h we use a function f_h in class $\mathcal{F}_h$ to approximate the optimal Q-value function. For notational simplicity, we denote $f := (f_1, \cdots, f_H) \in \mathcal{F}$ with $\mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$. Since no reward is collected in the $(H+1)^{\text{th}}$ steps, we will always use the convention that $f_{H+1} = 0$ and $\mathcal{F}_{H+1} = \{0\}$. We assume $f_h \in [-H, H]$ for any $f_h \in \mathcal{F}_h$. Each $f \in \mathcal{F}$ induces a greedy policy $\pi_f = \{\pi_{f_h}\}_{h=1}^H$ where

$$\pi_{f_h}(a \mid s) = \mathbb{1}\Big[a = \arg\max_{a'} f_h(s, a')\Big].$$

In valued-based batch RL, we take the offline dataset $\mathcal{D}$ as input and output an estimated optimal Q-value function f and the associated policy π_f. We are interested in the performance of π_f, which is measured by suboptimality in values, i.e., $V_1^\star(s_1) - V_1^{\pi_f}(s_1)$. However, this gap is highly nonsmooth in f, which is similar to the case of supervised learning where the $0-1$ losses for classification tasks are also highly nonsmooth and intractable. To mitigate this issue, a popular approach is to use a surrogate loss—the Bellman error.

Definition 3.1 (Bellman error)**.** Under data distribution μ, we define the *Bellman error* of function $f = (f_1, \cdots, f_H)$ as

$$\mathcal{E}(f) := \frac{1}{H}\sum_{h=1}^H \|f_h - \mathcal{T}_h^\star f_{h+1}\|_{\mu_h}^2. \tag{2}$$

Bellman error $\mathcal{E}(f)$ appears in many classical RL algorithms including Bellman risk minimization (BRM) (Antos et al., 2008), least-square temporal difference (LSTD) learning (Bradtke & Barto, 1996; Lazaric et al., 2012), etc.

The following lemma shows that under Assumption 1, one can control the suboptimality in values by the Bellman error.

Lemma 3.2 (Bellman error to value suboptimality)**.** *Under Assumption 1, for any $f \in \mathcal{F}$, we have that ,*

$$V_1^\star(s_1) - V_1^{\pi_f}(s_1) \le 2H\sqrt{C \cdot \mathcal{E}(f)}, \tag{3}$$

where C is the concentrability coefficient in Assumption 1.

Therefore, the Bellman error $\mathcal{E}(f)$ is indeed a surrogate loss for the suboptimality of π_f under mild conditions. In the next two sections, we will focus on designing efficient algorithms that minimize the Bellman error.

4. Results for Double Sampling Regime

As a starting point for Bellmen error minimization, we consider an empirical version of $\mathcal{E}(f)$ computed from samples. A natural choice of this empirical proxy is as follows

$$\hat{L}_{\text{B}}(f) := \frac{1}{nH}\sum_{(s,a,r,s',h)\in\mathcal{D}} \big(f_h(s, a) - r - V_{f_{h+1}}(s')\big)^2, \tag{4}$$

where $V_{f_{h+1}}(s) := \max_{a\in\mathcal{A}} f_{h+1}(s,a)$. Unfortunately, the estimator $\hat{L}_{\mathrm{B}}$ is biased due to the error-in variable situation (Bradtke & Barto, 1996). In particular, we have the following decomposition.

$$\mathcal{E}(f) = \mathbb{E}_\mu \hat{L}_{\mathrm{B}}(f) - \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}_{\mu_h} \mathrm{Var}_{s'\sim\mathbb{P}_h(\cdot|s,a)}(V_{f_{h+1}}(s')). \tag{5}$$

That is, the Bellman error and the expectation of $\hat{L}_{\mathrm{B}}$ differ by a variance term. This variance term is due to the stochastic transitions in the system, which is non-negligible even when f approximates the optimal value function $Q^\star$. A direct fix of this problem is to estimate the variance by double samples, where two independent samples of s_{h+1} are drawn when being in state s_h (Baird, 1995).

Formally, in this section, we consider the setting where for any (s,a,r,s',h) in dataset $\mathcal{D}$, there exists a paired tuple $(s,a,r,\tilde{s}',h)$ which share the same state-action pair (s,a) at step h, while $s',\tilde{s}'$ being two independent samples of the next state. Such data can be collected for instance if a simulator is avaliable, or the system allows an agent to revert back to the previous step. For simplicity, we denote this dataset as $\widetilde{\mathcal{D}} = \{(s,a,r,s',\tilde{s}',h)\}$ without placing additional constraints.

We construct the following empirical risk, which further estimates the variance term in (5) via double samples,

$$\hat{L}_{\mathrm{DS}}(f) := \frac{1}{nH}\sum_{(s,a,r,s',\tilde{s}',h)\in\widetilde{\mathcal{D}}} \Big[\big(f_h(s,a) - r - V_{f_{h+1}}(s')\big)^2 - \frac{1}{2}\big(V_{f_{h+1}}(s') - V_{f_{h+1}}(\tilde{s}')\big)^2 \Big].$$

We can show that, for any fixed $f\in\mathcal{F}$, $\mathbb{E}\hat{L}_{\mathrm{DS}}(f) = \mathcal{E}(f)$, i.e., $\hat{L}_{\mathrm{DS}}$ is an unbiased estimator of the Bellman error. Our algorithm for this setting is simply the Empirical Risk Minimizer (ERM), and we prove the following guarantee.

Theorem 4.1. *There exists an absolute constant $c>0$, with probability at least $1-\delta$, the ERM estimator $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{L}_{DS}(f)$ satisfies the following:*

$$\mathcal{E}(\hat{f}) \le \min_{f\in\mathcal{F}} \mathcal{E}(f) + cH^2\sqrt{\frac{\log(1/\delta)}{n}} + c\sum_{h=1}^{H}\big(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + \mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}})\big).$$

Here, we use shorthand $V_{\mathcal{F}_{h+1}} := \{V_{f_{h+1}} \mid f_{h+1}\in\mathcal{F}_{h+1}\}$ for any $h\in[H]$. Theorem 4.1 asserts that, in the double sampling regime, simple ERM has its excess risk $\mathcal{E}(\hat{f}) - \min_{f\in\mathcal{F}}\mathcal{E}(f)$ upper bounded by the Rademacher complexity of function class $\{\mathcal{F}_h\}_{h=1}^H$, $\{V_{\mathcal{F}_{h+1}}\}_{h=1}^H$ and a small concentration term that scales as $\tilde{\mathcal{O}}(1/\sqrt{n})$.

Most importantly, we remark that Theorem 4.1 holds without any assumption on the input data distribution or the properties of the MDP. Function class $\mathcal{F}$ can also be completely misspecificed in the sense the optimal value function $Q^\star$ may be very far from $\mathcal{F}$. This allows Theorem 4.1 to be widely applicable to a large number of applications.

However, a major limitation of Theorem 4.1 is its reliance on double samples. Double samples are not available in most dynamical systems that have no simulators or can not be reverted back to the previous step. In next section, we analyze algorithms in the standard single sampling regime.

5. Results for Single Sampling Regime

In this section, we focus on batch RL in the standard single sampling regime, where each tuple (s,a,r,s',h) in dataset $\mathcal{D}$ has a single next step s' following (s,a). We first present a sample complexity lower bound for minimizing the Bellman error, showing that in order to acheive an excess risk that does not scale polynomially with respect to the number of states, it is inevitable to have additional structural assumptions on function class $\mathcal{F}$ and the MDP. Then we analyze fitted Q-iteration (FQI) and a minimax estimator respectively, under different completeness assumptions. In addition to Rademacher complexity upper bounds similar to Theorem 4.1, we also utilize localization techniques and prove bounds with faster statistical rate in these two schemes.

5.1. Lower bound

Recall that when double samples are available, the excess risk of ERM estimator is controlled by Rademacher complexities of function classes (Theorem 4.1). In the single sampling regime, one natural question to ask is whether there exists an algorithm with a similar guarantee (i.e. the excess risk is upper bounded by certain complexity measure of the function class). Unfortunately, without further assumptions, the answer is negative.

Theorem 5.1. *Let $\mathfrak{A}$ be an arbitrary algorithm that takes any dataset $\mathcal{D}$ and function class $\mathcal{F}$ as input and outputs an estimator $\hat{f}\in\mathcal{F}$. For any $S\in\mathbb{N}^+$ and sample size $n\ge 0$, there exists an S-state, single-action MDP paired with a function class $\mathcal{F}$ with $|\mathcal{F}| = 2$ such that the $\hat{f}$ output by algorithm $\mathfrak{A}$ satisfies*

$$\mathbb{E}\mathcal{E}(\hat{f}) \ge \min_{f\in\mathcal{F}}\mathcal{E}(f) + \Omega\left(\min\left\{1, \frac{S^{1/2}}{n}\right\}\right). \tag{6}$$

Here, the expectation is taken over the randomness in $\mathcal{D}$.

Theorem 5.1 reveals a fundamental difference between the single sampling regime and the double sampling regime. The lower bound in inequality (6) depends polynomially on S—the cardinality of state space, which is considered to be intractably large in the setting of function approximation.

Algorithm 1 FQI

initialize $\hat{f}_{H+1} \leftarrow 0$.
for $h = H, H-1, \ldots, 1$ **do**
  $\hat{f}_h \leftarrow \arg\min_{f_h \in \mathcal{F}_h} \hat{\ell}_h(f_h, \hat{f}_{h+1}) := \frac{1}{n}\sum_{(s,a,r,s',h)\in\mathcal{D}_h} \left(f_h(s,a) - r - V_{\hat{f}_{h+1}}(s')\right)^2$.
return $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_H)$.

In batch RL with single sampling, despite the use of function class $\mathcal{F}$, the hardness of Bellman error minimization is still determined by the size of state space. This also suggests that minimizing Bellman error in the single sampling regime, is intrinsically different from the classic supervised learning due to the additional temporal correlation structure presented within the data.

We remark that unlike most lower bounds of similar type in prior works (Sutton & Barto, 2018; Sun et al., 2019), which only apply to certain restrictive classes of algorithms, Theorem 5.1 is completely information-theoretic, and applies to any algorithm. Recently, (Jiang, 2019) proved the impossibility of estimating the bellman error in single sampling regime for all algorithms. Our results is stronger since the impossibility of minimization implies the impossibility of estimation, but not vice versa.

To circumvent the hardness result in Theorem 5.1, additional structural assumptions are necessary. In the following, we provide statistical gurantees for two batch RL algorithms, where different completeness assumptions on $\mathcal{F}$ are used.

5.2. Fitted Q-iteration (FQI)

We consider the classical FQI algorithm. We assume that function class $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$ is (approximately) closed under the optimal Bellman operators $\mathcal{T}_1^\star, \ldots, \mathcal{T}_H^\star$, which is commonly adopted by prior analyses of FQI (Munos & Szepesvári, 2008; Chen & Jiang, 2019).

Assumption 2. There exists $\epsilon > 0$ such that, for all $h \in [H]$, $\sup_{f_{h+1}\in\mathcal{F}_{h+1}} \inf_{f_h\in\mathcal{F}_h} \|f_h - \mathcal{T}_h^\star f_{h+1}\|_{\mu_h}^2 \le \epsilon$.

The FQI algorithm is closely related to approximate dynamic programming (Bertsekas & Tsitsiklis, 1995). It starts by setting $\hat{f}_{H+1} := 0$ and then recursively computes Q-value functions at $h = H, H-1, \ldots, 1$. Each iteration in FQI is a least squares regression problem based on data collected at that time step. For $h \in [H]$, we denote $\mathcal{D}_h$ as set of data at the h^{th} step. The details of FQI are specified in Algorithm 1.

In the following Theorem 5.2, we upper bound the excess risk of the output of FQI in terms of Rademacher complexity.

Theorem 5.2 (FQI, Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumption 2, with probability at least $1-\delta$, the output of FQI $\hat{f}$ satisfies*

$$\mathcal{E}(\hat{f}) \le \epsilon + c\sum_{h=1}^{H} \mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + cH^2\sqrt{\frac{\log(H/\delta)}{n}}. \quad (7)$$

We remark that Assumption 2 immediately implies that $\min_{f\in\mathcal{F}} \mathcal{E}(f) \le \epsilon$. Therefore, although the minimal Bellman error $\min_{f\in\mathcal{F}} \mathcal{E}(f)$ does not explicitly appear on the right hand side, inequality (7) is still a variant of excess risk bound.

For typical parametric function classes, the Rademacher complexity scales as $n^{-1/2}$ (see Section 6). Therefore, Theorem 5.2 guarantees that the excess risk decrease as $n^{-1/2}$, up to a constant error ϵ due to the approximate completeness (in Assumption 2). However, since Bellman error is the average of squared L^2-norms (Definition 3.1), one may expect a faster statistical rate in this setting, similar to the case of linear regression. For this reason, we take advantage of the localization techniques and develop sharper error bounds in Theorem 5.3.

Theorem 5.3 (FQI, local Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumption 2, with probability at least $1-\delta$, the output of FQI $\hat{f}$ satisfies*

$$\mathcal{E}(\hat{f}) \le \epsilon + c\sqrt{\epsilon\cdot\Delta} + c\Delta\,, \quad (8)$$
$$\Delta := H\sum_{h=1}^{H} r_h^\star + H^2\frac{\log(H/\delta)}{n}\,.$$

Here $r_h^\star$ is the critical radius of local Rademacher complexity $\mathcal{R}_n^{\mu_h}(\{f_h \in \mathcal{F}_h \mid \|f_h - f_h^\dagger\|_{\mu_h}^2 \le r\})$ with $f_h^\dagger := \arg\min_{f_h\in\mathcal{F}_h} \|f_h - \mathcal{T}_h^\star \hat{f}_{h+1}\|_{\mu_h}$.

On the RHS of inequality (8), the first term ϵ measures model misspecification. The other two terms $c(\sqrt{\epsilon\cdot\Delta} + \Delta)$ can be viewed as statistical errors since $\Delta \to 0$ as sample size $n \to \infty$. For typical parametric function classes, the critical radius of the local Rademacher complexity scales as n^{-1} (see Section 6), which decreases much faster than standard Rademacher complexity. That is, Theorem 5.3 indeed guarantees faster statistical rate comparing to Theorem 5.2.

Finally, we remark that $f_h^\dagger$ in Theorem 5.3 depends on $\hat{f}_{h+1}$ and therefore is random. We will show later in Section 6 for many examples, the critical radius can be upper bounded independent of the choice of $f_h^\dagger$, in which case the randomness in $f_h^\dagger$ does not affect the final results.

5.3. Minimax Algorithm

The (approximate) completeness of $\mathcal{F}$ in Assumption 2 can be stringent sometimes. For instance, if there is a new function f_h attached to $\mathcal{F}_h$, for the sake of completeness, we need to enlarge $\mathcal{F}_{h-1}$ by adding several approximations of

$\mathcal{T}^\star_{h-1} f_h$. The same goes for $\mathcal{F}_{h-2}, \ldots, \mathcal{F}_1$. After amplifying the function classes one by one for each step, we may obtain an exceedingly large $\mathcal{F}$.

To avoid the issue above posted by the completeness assumptions on $\mathcal{F}$, we introduce a new function class $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_H$, where $\mathcal{G}_h$ consists of functions mapping from $\mathcal{S} \times \mathcal{A}$ to $[-H, H]$. We assume that for each $f_{h+1} \in \mathcal{F}_{h+1}$, one can always find a good approximation of $\mathcal{T}^\star_h f_{h+1}$ in this helper function class $\mathcal{G}_h$.

Assumption 3. There exists $\epsilon > 0$ such that, for all $h \in [H]$, $\sup_{f_{h+1} \in \mathcal{F}_{h+1}} \inf_{g_h \in \mathcal{G}_h} \|g_h - \mathcal{T}^\star_h f_{h+1}\|^2_{\mu_h} \le \epsilon$.

According to (5), we can approximate Bellman error $\mathcal{E}(f)$ by subtracting the variance term from $\hat{L}_{\mathrm{B}}(f)$. If g_h is close to $\mathcal{T}^\star_h f_{h+1}$, then $\big(g_h(s,a) - r - V_{f_{h+1}}(s')\big)^2$ averaged over data provides a good estimator of the variance term. Following this intuition, we define a new loss

$$\hat{L}_{\mathrm{MM}}(f, g) := \frac{1}{nH} \sum_{(s,a,r,s',h) \in \mathcal{D}} \Big[\big(f_h(s,a) - r - V_{f_{h+1}}(s')\big)^2 - \big(g_h(s,a) - r - V_{f_{h+1}}(s')\big)^2 \Big].$$

The minimax algorithm (Antos et al., 2008; Chen & Jiang, 2019) then computes

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \hat{L}_{\mathrm{MM}}(f, g).$$

Now we are ready to state our theoretical guarantees for the minimax algorithms.

Theorem 5.4 (Minimax algorithm, Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumption 3, with probability at least $1 - \delta$, the minimax estimator $\hat{f}$ satisfies:*

$$\mathcal{E}(\hat{f}) \le \min_{f \in \mathcal{F}} \mathcal{E}(f) + \epsilon + cH^2 \sqrt{\frac{\log(1/\delta)}{n}} + c \sum_{h=1}^{H} \big(\mathcal{R}_n^{\mu_h}(\mathcal{F}_h) + \mathcal{R}_n^{\mu_h}(\mathcal{G}_h) + \mathcal{R}_n^{\nu_h}(V_{\mathcal{F}_{h+1}}) \big).$$

As is shown in Theorem 5.4, the excess risk is simultaneously controlled by the Rademacher complexities of $\{\mathcal{F}_h\}_{h=1}^H$, $\{\mathcal{G}_h\}_{h=1}^H$ and $\{V_{\mathcal{F}_{h+1}}\}_{h=1}^H$.

Similar to the results for FQI, we can also develop risk bounds with faster statistical rate using the localization techniques. For technical reasons that will be soon discussed, we introduce the following assumption, which can be viewed a variant of the concentrability coefficient in Assumption 1 under different initial distributions.

Assumption 4. For any policy π and $h \in [H]$, let $P^\pi_{h,t}$ (or $\widetilde{P}^\pi_{h,t}$) denote the marginal distribution at $t > h$, starting from μ_h at time step h (or from $\nu_h \times \mathrm{Unif}(\mathcal{A})$ at $h+1$) and following π. There exists a parameter $\widetilde{C}$ such that

$$\sup_{\substack{(s,a) \in \mathcal{S} \times \mathcal{A} \\ h \in [H], t > h}} \left(\frac{\mathrm{d}P^\pi_{h,t}}{\mathrm{d}\mu_t} \vee \frac{\mathrm{d}\widetilde{P}^\pi_{h,t}}{\mathrm{d}\mu_t} \right)(s,a) \le \widetilde{C} \quad \text{for any policy } \pi.$$

For notational convenience, we define

$$f^\dagger = (f_1^\dagger, \ldots, f_H^\dagger) := \arg\min_{f \in \mathcal{F}} \mathcal{E}(f) \qquad \text{and}$$
$$g_h^\dagger := \arg\min_{g_h \in \mathcal{G}_h} \|g_h - \mathcal{T}^\star_h f^\dagger_{h+1}\|_{\mu_h}.$$

Now we are ready to state the excess risk bound of the minimax algorithm in terms of local Rademacher complexity as follows.

Theorem 5.5 (Minimax algorithm, local Rademacher complexity)**.** *There exists an absolute constant $c > 0$, under Assumptions 3 and 4, with probability at least $1 - \delta$, the minimax estimator $\hat{f}$ satisfies:*

$$\mathcal{E}(\hat{f}) \le \min_{f \in \mathcal{F}} \mathcal{E}(f) + \epsilon + c\sqrt{\big(\min_{f \in \mathcal{F}} \mathcal{E}(f) + \epsilon\big)\Delta} + c\Delta\,, \quad (9)$$

$$\Delta := H^3 \sum_{h=1}^{H} \left[\widetilde{C}\big(r^\star_{f,h} + r^\star_{g,h} + \tilde{r}^\star_{f,h}\big) + \sqrt{\widetilde{C} r^\star_{g,h} \epsilon} \right] + H^2 \frac{\log(H/\delta)}{n}\,.$$

where $\widetilde{C}$ is the concentrability coefficient in Assumption 4, and $r^\star_{f,h}, r^\star_{g,h}, \tilde{r}^\star_{f,h}$ are the critical radius of the following local Rademacher complexities respectively:

$$\mathcal{R}_n^{\mu_h}\big(\{f_h \in \mathcal{F}_h \mid \|f_h - f_h^\dagger\|^2_{\mu_h} \le r\}\big)\,,$$
$$\mathcal{R}_n^{\mu_h}\big(\{g_h \in \mathcal{G}_h \mid \|g_h - g_h^\dagger\|^2_{\mu_h} \le r\}\big)\,,$$
$$\mathcal{R}_n^{\nu_h}\big(\{V_{f_{h+1}} \mid f_{h+1} \in \mathcal{F}_{h+1}, \|f_{h+1} - f^\dagger_{h+1}\|^2_{\nu_h \times \mathit{Unif}(\mathcal{A})} \le r\}\big)\,.$$

Similar to Theorem 5.3, our upper bound in (9) can also be viewed as a combination of model misspecification error ($\min_{f \in \mathcal{F}} \mathcal{E}(f) + \epsilon$) and statistical error ($c\sqrt{\big(\min_{f \in \mathcal{F}} \mathcal{E}(f) + \epsilon\big)\Delta} + c\Delta$). As $n \to \infty$, the model misspecification error is nonvanishing and the statistical error tends to zero. Again for typical parametric function classes, the critical radius of the local Rademacher complexity scales as n^{-1} (see Section 6), and Theorem 5.5 claims the excess risk of the minimax algorithm also decreases as n^{-1} except a constant model misspecification error ϵ.

Intuitively, Assumption 4 is required in Theorem 5.5 to allow that $\mathcal{E}(f)$ close to $\mathcal{E}(f^\dagger)$ implies f_h in the neighborhood of $f_h^\dagger$ for each step $h \in [H]$. We conjecture such additional assumption is unavoidable if we would like to upper bound

the excess risk using the local Rademacher complexity of $\mathcal{F}_h$, $\mathcal{G}_h$ and $V_{\mathcal{F}_{h+1}}$ for the minimax algorithm.

In Appendix C, we present an alternative version of Theorem 5.3, which does not require Assumption 4 but bound the excess risk using the local Rademacher complexity of a composite function class depending on the loss, $\mathcal{F}$, and $\mathcal{G}$. The alternative version recovers the sharp result in (Chen & Jiang, 2019) when the function classes $\mathcal{F}$ and $\mathcal{G}$ both have finite elements.

Finally, our upper bounds for the minimax algorithm contain Rademacher complexities of the function class $V_{\mathcal{F}}$. We can conveniently control them using the Radermacher complexities of function class $\mathcal{F}$ as follows.

Proposition 5.6. *Let $\mathcal{F}$ be a set of functions over $\mathcal{S}\times\mathcal{A}$ and ρ be a measure over $\mathcal{S}$. We have the following inequality,*

$$\mathcal{R}_n^{\rho}(V_{\mathcal{F}}) \le \sqrt{2}\, A\, \mathcal{R}_n^{\rho\times \mathit{Unif}(\mathcal{A})}(\mathcal{F}),$$

where A is the cardinality of the set $\mathcal{A}$.

6. Examples

Below we give four examples of function classes, each with an upper bound on Rademacher complexity, as well as the critical radius of the local Rademacher complexity. Throughout this section we use notation $r_n^{\star,\rho}(\mathcal{F}, f_o)$ to denote the critical radius of local Rademacher complexity $\mathcal{R}_n^{\rho}(\{f\in\mathcal{F} \mid \|f-f_o\|_\rho^2 \le r\})$.

Function class with finite element. First, we consider the function class $\mathcal{F}$ with $|\mathcal{F}|<\infty$. Under the normalization that $f\in[0,H]$ for any $f\in\mathcal{F}$, we have the following.

Proposition 6.1. *For function class $\mathcal{F}$ defined above, for any data distribution ρ and any anchor function $f_o\in\mathcal{F}$:*

$$\mathcal{R}_n^{\rho}(\mathcal{F}) \le 2H\max\left\{\sqrt{\frac{\log|\mathcal{F}|}{n}}, \frac{\log|\mathcal{F}|}{n}\right\},$$
$$r_n^{\star,\rho}(\mathcal{F}, f_o) \le \frac{2H\log|\mathcal{F}|}{n}.$$

Linear functions. Let $\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ be the feature map to a d-dimensional Euclidean space, and consider the function class $\mathcal{F}\subset\{w^\top\phi \mid w\in\mathbb{R}^d, \|w\|\le H\}$. Under the normalization that $\|\phi(s,a)\|\le 1$ for any (s,a), we have

Proposition 6.2. *For linear function class $\mathcal{F}$ defined above, for any data distribution ρ and any anchor function $f_o\in\mathcal{F}$:*

$$\mathcal{R}_n^{\rho}(\mathcal{F}) \le H\sqrt{\frac{2d}{n}},$$
$$r_n^{\star,\rho}(\mathcal{F}, f_o) \le \frac{2d}{n}.$$

Functions in RKHS. Consider a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ associated with a positive kernel $k:(\mathcal{S}\times\mathcal{A})\times(\mathcal{S}\times\mathcal{A})\to\mathbb{R}$. Suppose that $k\big((s,a),(s,a)\big)\le 1$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. Consider the function class $\mathcal{F}\subseteq\{f\in\mathcal{H} \mid \|f\|_{\mathcal{K}}\le H\}$, here $\|\cdot\|_{\mathcal{K}}$ denotes the RKHS norm. Define an integral operator $\mathscr{T}: L^2(\rho)\to L^2(\rho)$ as

$$\mathscr{T}f := \mathbb{E}_{(s,a)\sim\rho}\big[k\big(\cdot,(s,a)\big)f(s,a)\big].$$

Suppose that $\mathbb{E}_{(s,a)\sim\rho}\big[k\big((s,a),(s,a)\big)\big] < +\infty$. Let $\{\lambda_i(\mathscr{T})\}_{i=1}^{\infty}$ be the eigenvalues of $\mathscr{T}$, arranging in a non-increasing order. Then

Proposition 6.3. *For kernel function class $\mathcal{F}$ defined above, for any data distribution ρ and any anchor function $f_o\in\mathcal{F}$:*

$$\mathcal{R}_n^{\rho}(\mathcal{F}) \le H\sqrt{\frac{2}{n}\sum_{i=1}^{\infty}1\wedge\big(4\lambda_i(\mathscr{T})\big)},$$
$$r_n^{\star,\rho}(\mathcal{F}, f_o) \le 2\min_{j\in\mathbb{N}}\left\{\frac{j}{n} + H\sqrt{\frac{2}{n}\sum_{i=j+1}^{\infty}\lambda_i(\mathscr{T})}\right\}.$$

Sparse linear functions. Let $\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ be the feature map to a d-dimensional Euclidean space, and consider the function class $\mathcal{F}\subset\{w^\top\phi \mid w\in\mathbb{R}^d, \|w\|_0\le s\}$. Assume that when $(s,a)\sim\rho$, $\phi(s,a)$ satisfies a Gaussian distribution with covariance Σ. Assume $\|f\|_\rho\le H$ for any $f\in\mathcal{F}$. Furthermore, denote $\kappa_s(\Sigma)$ to be the upper bound such that $\kappa_s(\Sigma)\ge\lambda_{\max}(M)/\lambda_{\min}(M)$ for any matrix M that is a $s\times s$ principal submatrix of Σ. Then

Proposition 6.4. *There exists an absolute constant $c>0$, for sparse linear function class $\mathcal{F}$ defined above, assume the data distribution ρ satisfies the conditions specified above, then for any anchor function $f_o\in\mathcal{F}$:*

$$\mathcal{R}_n^{\rho}(\mathcal{F}) \le cH\sqrt{\kappa_s(\Sigma)}\sqrt{\frac{s\log d}{n}},$$
$$r_n^{\star,\rho}(\mathcal{F}, f_o) \le c^2\kappa_s(\Sigma)\cdot\frac{s\log d}{n}.$$

End-to-end results. Finally, to obtain an end-to-end result that upper bounds the suboptimality in values for specific function classes listed above, we can simply combine (a) the result that upper bound the value suboptimality using the Bellman error (Lemma 3.2); (b) the results that upper bound the Bellman error in terms of (local) Rademacher complexity (Theorems 4.1, 5.2-5.5); (c) the upper bounds of (local) Rademacher complexity for specific function classes (Propositions 6.1-6.4).

7. Conclusion

This paper studies batch RL with general value function approximation from the lens of statistical learning theory.

We identify the intrinsic difference between batch reinforcement learning and classical supervised learning (Theorem 5.1) due to the additional temporal correlation structure presented in the RL data. Under mild conditions, this paper also provides upper bounds on the generalization performance of several popular batch RL algorithms in terms of the (local) Rademacher complexities of general function classes. We hope our results shed light on the future research in further bridging the gap between statistical learning theory and RL.

Acknowledgement

ZL acknowledges support from NSF, ONR, Simons Foundation, Schmidt Foundation, Microsoft Research, Mozilla Research, Amazon Research, DARPA and SRC.

References

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995.

Baldi, P., Brunak, S., and Bach, F. *Bioinformatics: the machine learning approach*. MIT press, 2001.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Bertsekas, D. P. and Tsitsiklis, J. N. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, volume 1, pp. 560–564. IEEE, 1995.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dann, C., Neumann, G., Peters, J., et al. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

Dudley, R. M. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Farahmand, A. M., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration. In *nips*, pp. 441–448, 2008.

Farahmand, A. M., Munos, R., and Szepesvári, C. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Forsyth, D. A. and Ponce, J. *Computer vision: a modern approach*. Pearson,, 2012.

Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pp. 148–156. Citeseer, 1996.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, volume 112. Springer, 2013.

Jelinek, F. *Statistical methods for speech recognition*. MIT press, 1997.

Jiang, N. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1704–1713. JMLR. org, 2017.

Juang, B. H. and Rabiner, L. R. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

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

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Mendelson, S. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pp. 29–43. Springer, 2002.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Munos, R. Error bounds for approximate policy iteration. In *ICML*, volume 3, pp. 560–567, 2003.

Munos, R. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

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

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Precup, D., Sutton, R. S., and Dasgupta, S. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Riedmiller, M. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pp. 317–328. Springer, 2005.

Schapire, R. E. A brief introduction to boosting. In *Ijcai*, volume 99, pp. 1401–1406. Citeseer, 1999.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pp. 2898–2933. PMLR, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Suykens, J. A. and Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters*, 9 (3):293–300, 1999.

Szeliski, R. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

van Handel, R. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020a.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020b.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020a.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. *arXiv preprint arXiv:2003.03924*, 2020b.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

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