Batch Value-function Approximation with Only Realizability

Tengyang Xie[1] Nan Jiang[1]

Abstract

We make progress in a long-standing problem of batch reinforcement learning (RL): learning $Q^\star$ from an exploratory and polynomial-sized dataset, using a *realizable* and otherwise *arbitrary* function class. In fact, all existing algorithms demand function-approximation assumptions stronger than realizability, and the mounting negative evidence has led to a conjecture that sample-efficient learning is impossible in this setting (Chen & Jiang, 2019). Our algorithm, BVFT, breaks the hardness conjecture (albeit under a stronger notion of exploratory data) via a tournament procedure that reduces the learning problem to pairwise comparison, and solves the latter with the help of a state-action-space partition constructed from the compared functions. We also discuss how BVFT can be applied to model selection among other extensions and open problems.

1. Introduction

What is the minimal function-approximation assumption that enables polynomial sample complexity, when we try to learn $Q^\star$ from an exploratory batch dataset? Existing algorithms and analyses—those that have largely laid the theoretical foundation of modern reinforcement learning—have always demanded assumptions that are substantially stronger than the most basic one: *realizability*, i.e., that $Q^\star$ (approximately) lies in the function class. These strong assumptions have recently compelled Chen & Jiang (2019) to conjecture an information-theoretic barrier, that polynomial learning is impossible in batch RL, even with exploratory data and realizable function approximation.

In this paper, we break this barrier by an algorithm called **B**atch **V**alue-**F**unction **T**ournament (BVFT). Via a tournament procedure, BVFT reduces the learning problem to that of identifying $Q^\star$ from a pair of candidate functions. In this subproblem, we create a piecewise constant function class of statistical complexity $O(1/\epsilon^2)$ that can express both candidate functions up to small discretization errors, and use the projected Bellman operator associated with the class to identify $Q^\star$. We present the algorithm in Section 4 and prove its sample complexity in Sections 5 and 6. A limitation of our approach is the use of a relatively stringent version of *concentrability coefficient* from Munos (2003) to measure the exploratoriness of the dataset (see Assumption 1). Section 7.2 investigates the difficulties in relaxing the assumption, and Appendix D discusses how to mitigate the pathological behavior of the algorithm when the assumption does not hold.

As another limitation, BVFT enumerates over the function class and is computationally inefficient for training. That said, the algorithm is efficient when the function class has a polynomial cardinality, making it applicable to another problem in batch RL: model selection (Farahmand & Szepesvári, 2011).[1] In Section 7.1, we review the literature on this important problem and discuss how BVFT has significantly advanced the state of the art on the theoretical front.

2. Related Work

Stronger Function-Approximation Assumptions in Existing Theory The theory of batch RL has struggled for a long time to provide sample-efficiency guarantees when realizability is the only assumption imposed on the function class. An intuitive reason is that learning $Q^\star$ is roughly equivalent to minimizing the Bellman error, but the latter cannot be estimated from data (Jiang, 2019; Sutton & Barto, 2018, Chapter 11.6), leading to the infamous "double sampling" difficulty (Baird, 1995; Antos et al., 2008). Stronger/additional assumptions have been proposed to circumvent the issue, including low *inherent Bellman errors* (Munos & Szepesvári, 2008; Antos et al., 2008), *averager* classes (Gordon, 1995), and additional function approximation of importance weights (Xie & Jiang, 2020).

State Abstractions State abstractions are the simplest form of function approximation. (They are also special cases of

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, USA. Correspondence to: Nan Jiang <nanjiang@illinois.edu>.

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

[1] We use the phrase "model selection" as in the context of e.g., cross validation, and the word "model" does not refer to MDP dynamics; rather they refer to value functions for our purposes.

the aforementioned averagers.) In fact, certainty equivalence with a state abstraction that can express $Q^\star$, known as $Q^\star$-*irrelevant abstractions*, is known to be consistent, i.e., $Q^\star$ will be correctly learned if each abstract state-action pair receives infinite amount of data (Littman & Szepesvári, 1996; Li et al., 2006, Theorem 4).

While this observation is an important inspiration for our algorithm, making it useful for an arbitrary and unstructured function class is highly nontrivial and is one of the main algorithmic contributions of this paper. Furthermore, our finite-sample analysis significantly deviates from the "tabular"-style proofs in the abstraction literature (Paduraru et al., 2008; Jiang, 2018), where ℓ_∞ concentration bounds are established assuming that each abstract state receives sufficient data (c.f. Footnote 8). In our analysis, the structure of the abstraction is arbitrary, and it is much more convenient to treat them as piecewise constant classes over the original state space and use tools from statistical learning theory to establish concentration results under weighted ℓ_2 norm; see Section 5.2.3 for details.

Tournament Algorithms Our algorithm design also draws inspirations from existing tournament algorithms. Closest related is Scheffé tournament for density estimation (Devroye & Lugosi, 2012), which minimizes the total-variation (TV) distance from the true density among the candidate models, and has been applied to RL by Sun et al. (2019). Interestingly, the main challenge in TV-distance minimization is very similar to ours at a high level, that TV-distance itself of a *single* model cannot be estimated from data when the support of the distribution has a large or infinite cardinality. Similar to Scheffé tournament, our algorithm compares *pairs* of candidate value functions, which is key to overcoming the fundamental unlearnability of Bellman errors.

Tournament algorithms are also found in RL when the goal is to select the best state abstraction from a candidate set (Hallak et al., 2013; Jiang et al., 2015). These works will be discussed in Section 7.1 in the context of model selection.

Lower Bounds Wang et al. (2020); Amortila et al. (2020); Zanette (2020); Chen et al. (2021) have recently proved hardness results under $Q^\star$ realizability in batch RL. These results do not contradict ours because they deploy a weaker data assumption; see Appendix A.2 for discussions. Rather, their negative and our positive results are complementary and together provide a fine-grained characterization of the landscape of batch RL.

3. Preliminaries

3.1. Markov Decision Processes

Consider an infinite-horizon discounted Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$, where $\mathcal{S}$ is the finite state space that can be arbitrarily large, $\mathcal{A}$ is the finite action space, $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $R : \mathcal{S}\times\mathcal{A} \to [0, R_{\max}]$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution.

A (deterministic and stationary) policy $\pi : \mathcal{S} \to \mathcal{A}$ induces a distribution of the infinite trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots,$ as $s_0 \sim d_0, a_0 = \pi(s_0), r_0 = R(s_0, a_0), s_1 \sim P(s_0, a_0), \ldots$. We use $\mathbb{E}[\cdot|\pi]$ to denote taking expectation w.r.t. such a distribution. The expected discounted return of a policy is defined as $J(\pi) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t|\pi]$, and our goal is to optimize $J(\pi)$. Note that the random variable $\sum_{t=0}^{\infty} \gamma^t r_t$ is always bounded in the range $[0, V_{\max}]$ where $V_{\max} = R_{\max}/(1-\gamma)$.

In the discounted setting, there is a policy $\pi^\star : \mathcal{S} \to \mathcal{A}$ that simultaneously optimizes the expected return for all starting states. This policy can be obtained as the greedy policy of the $Q^\star$ function, i.e., $\pi^\star(s) = \pi_{Q^\star}(s) := \arg\max_{a\in\mathcal{A}} Q^\star(s, a)$, where we use $\pi_{(\cdot)}$ to denote a policy that greedily chooses actions according to a real-valued function over $\mathcal{S} \times \mathcal{A}$. The optimal Q-value function, $Q^\star$, can be uniquely defined through the Bellman optimality equations: $Q^\star = \mathcal{T} Q^\star$, where $\mathcal{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the optimality operator, defined as $(\mathcal{T} f)(s, a) := R(s, a) + \gamma\mathbb{E}_{s'\sim P(s,a)}[V_f(s', a')]$, where $V_f(s, a) := \max_a f(s, a)$.

3.2. Batch Data

We assume that the learner has access to a batch dataset D consisting of i.i.d. (s, a, r, s') tuples, where $(s, a) \sim \mu, r = R(s, a), s' \sim P(s, a)$. Such an i.i.d. assumption is standard for finite-sample analyses in the ADP literature (Munos & Szepesvári, 2008; Farahmand et al., 2010; Chen & Jiang, 2019), and can often be relaxed at the cost of significant technical burdens and complications (see e.g., Antos et al., 2008). We will also use $\mu(s)$ and $\mu(a|s)$ to denote the marginal of s and the conditional of a given s. To learn a near-optimal policy in batch RL, an exploratory dataset is necessary, and we measure the degree of exploration as follows:

Assumption 1. We assume that $\mu(s, a) > 0\ \forall s, a$. We further assume that
(1) There exists constant $1 \le C_{\mathcal{A}} < \infty$ such that for any $s \in \mathcal{S}, a \in \mathcal{A}$, $\mu(a|s) \ge 1/C_{\mathcal{A}}$.
(2) There exists constant $1 \le C_{\mathcal{S}} < \infty$ such that for any $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$, $P(s'|s, a)/\mu(s') \le C_{\mathcal{S}}$. Also $d_0(s)/\mu(s) \le C_{\mathcal{S}}$.
It will be convenient to define $C = C_{\mathcal{S}} C_{\mathcal{A}}$.

The first statement is very standard, asserting that the data distribution put enough probabilities on all actions. For example, with a small number of actions, a uniformly random policy ensures that $C_{\mathcal{A}} = |\mathcal{A}|$ satisfies this assumption.

The second statement measures the exploratoriness of μ's state marginal by $C_{\mathcal{S}}$, and two comments are in order. First, this is a form of *concentrability* assumption, which not only enforces data to be exploratory, but also implicitly imposes restrictions on the MDP's dynamics (see the reference to P in Assumption 1). While the latter may be undesirable, Chen & Jiang (2019, Theorem 4) shows that such a restriction is *unavoidable* when learning with a general function class. Second, the version of concentrability coefficient we use was introduced by Munos (2003, Eq.(6)), and is more stringent than its more popular variants (e.g., Munos, 2007; Farahmand et al., 2010). That said, (1) hardness results exist under a weaker form of the assumption (see Appendix A.2), and (2) whenever the transition dynamics admit low-rank stochastic factorization, there always exist data distributions that yield small $C_{\mathcal{S}}$ despite that $|\mathcal{S}|$ can be arbitrarily large; see Appendix A.1, where we also discuss how Assumption 1 compares to no inherent Bellman errors in the context of low-rank MDPs. We investigate why it is difficult to work with more relaxed assumptions in Section 7.2, and discuss how to mitigate the negative consequences when the assumption is violated in Appendix D.

A direct consequence of Assumption 1, which we will use later to control error propagation and distribution shift, is the following proposition.

Proposition 1. *Let ν be a distribution over $\mathcal{S} \times \mathcal{A}$ and π be a policy. Let $\nu' = P(\nu) \times \pi$ denote the distribution specified by the generative process $(s', a') \sim \nu' \Leftrightarrow (s, a) \sim \nu, s' \sim P(\cdot|s, a), a' = \pi(s')$. Under Assumption 1, we have $\|\nu'/\mu\|_\infty := \max_{s,a} \nu'(s, a)/\mu(s, a) \le C$. Also note that $\|(d_0 \times \pi)/\mu\|_\infty \le C$.*

Additional Notations For any real-valued function of (s, a, r, s'), we use $\mathbb{E}_\mu[\cdot]$ as a shorthand for taking expectation of the function when $(s, a) \sim \mu, r = R(s, a), s' \sim P(s, a)$. Also for any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define $\|f\|_{2,\mu}^2 := \mathbb{E}_\mu[f^2]$; $\|f\|_{2,\mu}$ is a weighted ℓ_2 norm and satisfies the triangular inequality. We also use $\|f\|_{2,D}^2$ to denote the empirical approximation of $\|f\|_{2,\mu}^2$ based on the dataset D.

3.3. Value-function Approximation

Since the state space $\mathcal{S}$ can be prohibitively large, function approximation is necessary for scaling RL to large and complex problems. In the value-function approximation setting, we are given a function class $\mathcal{F} \subset (\mathcal{S} \times \mathcal{A} \to [0, V_{\max}])$ to model $Q^\star$. Unlike prior works that measure the approximation error of $\mathcal{F}$ using *inherent Bellman errors* (Munos, 2007; Antos et al., 2008)—which amounts to assuming that $\mathcal{F}$ is (approximately) *closed* under $\mathcal{T}$—we will measure the error using Definition 1, where 0 error *only* implies realizability, $Q^\star \in \mathcal{F}$. In fact, given that the assumptions required by all existing algorithms are substantially stronger than realizability, Chen & Jiang (2019, Conjecture 8) conjecture that polynomial sample complexity is unattainable in batch RL when we only impose realizability on $\mathcal{F}$, which is why our result may be surprising.

Algorithm 1 **B**atch **V**alue-**F**unction **T**ournament (BVFT)

```
Input: Dataset $D$, function class $\mathcal{F}$, discretization parameter $\epsilon_{\text{dct}} \in (0, V_{\max})$.
for $f \in \mathcal{F}$ do
    $\bar{f} \leftarrow$ discretize the output of $f$ with resolution $\epsilon_{\text{dct}}$. (see Footnote 4).
end for
for $f \in \mathcal{F}$ do
    for $f' \in \mathcal{F}$ do
        Define $\phi$ s.t. $\phi(s, a) = \phi(s', a')$ iff $\bar{f}(s, a) = \bar{f}(s', a')$ and $\bar{f}'(s, a) = \bar{f}'(s', a')$.
        $\mathcal{E}(f; f') \leftarrow \|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,D}$ (see Eq.(1) for def of $\widehat{\mathcal{T}}_\phi^\mu$; the dependence on $f'$ is only through $\phi$).
    end for
end for
$\hat{f} \leftarrow \arg\min_{f \in \mathcal{F}} \max_{f' \in \mathcal{F}} \mathcal{E}(f; f')$.
Output: $\hat{\pi} = \pi_{\hat{f}}$.
```

Definition 1. Let $\epsilon_{\mathcal{F}} := \inf_{f \in \mathcal{F}} \|f - Q^\star\|_\infty$. [2] Let $f^\star$ denote the f that attains the infimum.

We assume $\mathcal{F}$ is finite but exponentially large (as in Chen & Jiang (2019)), i.e., we can only afford poly $\log |\mathcal{F}|$ in the sample complexity. For continuous function classes that admit a finite ℓ_∞ covering number (Agarwal, 2011), our approach and analysis immediately extend by replacing $\mathcal{F}$ with its ϵ-net at the cost of slightly increasing $\epsilon_{\mathcal{F}}$.

3.4. Polynomial Learning

Our goal is to devise a statistically efficient algorithm with the following kind of guarantee: with high probability we can learn an ϵ-optimal policy $\hat{\pi}$, that is, $J(\hat{\pi}) \ge J(\pi^\star) - \epsilon \cdot V_{\max}$, when $\mathcal{F}$ is realizable and the dataset D is only polynomially large. The polynomial may depend on the effective horizon $1/(1 - \gamma)$, the statistical complexity of the function class $\log |\mathcal{F}|$, the concentrability coefficient C, (the inverse of) the suboptimality gap ϵ, and $1/\delta$ where δ is the failure probability. Our results can also accommodate the more general setting when $\mathcal{F}$ is not exactly realizable, in which case the suboptimality of $\hat{\pi}$ is allowed to contain an additional term proportional to the approximation error $\epsilon_{\mathcal{F}}$ up to a polynomial multiplicative factor.

[2] It is possible to define $\epsilon_{\mathcal{F}}$ under weighted ℓ_2 norm, though making this change in our current proof yields a worse (albeit polynomial) sample complexity; see Appendix E for more details.

4. Algorithm and the Guarantee

In this section, we introduce and provide intuitions for our algorithm, and state its sample complexity guarantee which will be proved in the subsequent sections.

The design of our algorithm is based on an important observation inspired by the state-abstraction literature (see Section 2): when the function class $\mathcal{F}$ is piecewise constant and realizable, batch learning with exploratory data is consistent using e.g., Fitted Q-Iteration. This is because piecewise constant classes are very stable, and their associated projected Bellman operators are always γ-contractions under ℓ_∞, implying that $Q^\star$ is the only fixed point of such operators when statistical errors are ignored;[3] we will actually establish these properties in Section 5.1.

While realizability is the only expressivity condition assumed, being piecewise constant is a major structural assumption, and is too restrictive to accommodate practical function-approximation schemes such as linear predictors or neural networks, let alone the completely unstructured set of functions one would encounter in model selection (Section 7.1). How can we make use of this observation?

An immediate idea is *improper learning*, i.e., augmenting $\mathcal{F}$—which is not piecewise constant in general and may have an arbitrary structure—to its smallest superset that *is* piecewise constant, which automatically inherits realizability from $\mathcal{F}$. To do so, we may first discretize the output of each function $f \in \mathcal{F}$ up to a small discretization error ϵ_{dct},[4] and partition $\mathcal{S} \times \mathcal{A}$ by grouping state-action pairs together *only when the output $f \in \mathcal{F}$ (after discretization) is constant across them.* The problem is that, the resulting function class is way too large compared to $\mathcal{F}$; its statistical complexity—measured by the number of groups—can be as large as $(V_{\max}/\epsilon_{\text{dct}})^{|\mathcal{F}|}$, *doubly exponential* in poly $\log|\mathcal{F}|$ which is what we can afford!

To turn this idea into a polynomial algorithm, we note that the statistical complexity of the superset is affordable when $|\mathcal{F}|$ is constant, say, $|\mathcal{F}| = 2$. This provides us with a procedure that identifies $Q^\star$ out of two candidate functions. To handle an exponentially large $\mathcal{F}$, we simply perform pairwise comparisons between all pairs of $f, f' \in \mathcal{F}$, and output the function that has survived all pairwise comparisons involving it. Careful readers may wonder what happens when $Q^\star \notin \{f, f'\}$, as realizability is obviously violated. As we will show in Section 6, the outcomes of these "bad" comparisons simply do not matter: $Q^\star$ is never involved in such comparisons, and any other function f will always be checked against $f' = Q^\star$, which is enough to expose the deficiency of a bad f.

The above reasoning ignores approximation and estimation errors, which we handle in the actual algorithm and its analysis; see Algorithm 1. Below we state its sample complexity guarantee, which is the main theorem of this paper.

Theorem 2. *Under Assumption 1, with probability at least $1-\delta$, BVFT (Algorithm 1) with $\epsilon_{dct} = \frac{(1-\gamma)^2\epsilon V_{\max}}{16\sqrt{C}}$ returns a policy $\hat{\pi}$ that satisfies*

$$J(\pi^\star) - J(\hat{\pi}) \le \frac{(4+8\sqrt{C})\epsilon_{\mathcal{F}}}{(1-\gamma)^2} + \epsilon \cdot V_{\max},$$

with a sample complexity of[5]

$$|D| = \tilde{O}\left(\frac{C^2 \ln\frac{|\mathcal{F}|}{\delta}}{\epsilon^4(1-\gamma)^8}\right).$$

The most outstanding characteristic of the sample complexity is the $1/\epsilon^4$ rate. In fact, the poor dependencies on C and $1/(1-\gamma)$ are both due to $1/\epsilon^4$: when we rewrite the guarantee in terms of suboptimality gap as a function of $n = |D|$, we see an $O(\sqrt{C}n^{-1/4}/(1-\gamma)^2)$ estimation-error term, featuring the standard $\sqrt{C}$ penalty due to distribution shift and quadratic-in-horizon error propagation.

The $1/\epsilon^4$ rate comes from two sources: $1/\epsilon^2$ of it is due to the worst-case statistical complexity of the piecewise constant classes created during pairwise comparisons. The other $1/\epsilon^2$ is the standard statistical rate. While standard, proving $O(1/\epsilon^2)$ concentration bounds in our analysis turns out to be technically challenging and requires some clever tricks. We refer mathematically inclined readers to Section 5.2.3 for how we overcome those challenges.

We prove Theorem 2 in the next two sections. Section 5 establishes the essential properties of the pairwise comparison step in Line 8, where we view the problem at a somewhat abstract level to attain proof modularity. Section 6 uses the results in Section 5 to prove the final guarantee.

5. Value-function Validation using a Piecewise Constant Function Class

In this section we analyze a subproblem that is crucial to our algorithm: given a piecewise constant class $\mathcal{G}_\phi \subset (\mathcal{S}\times\mathcal{A} \to [0, V_{\max}])$ (induced by ϕ, a partition of $\mathcal{S} \times \mathcal{A}$)[6] with small

[3] $Q^\star$ is always *a* fixed point of the projected Bellman update operators associated with any realizable function class, but there is no uniqueness guarantee in general.

[4] When $V_{\max}/\epsilon_{\text{dct}}$ is an odd integer, discretization onto a regular grid $\{\epsilon_{\text{dct}}, 3\epsilon_{\text{dct}}, \ldots, V_{\max} - \epsilon_{\text{dct}}\}$ guarantees at most ϵ_{dct} approximation error, and the cardinality of the set is $V_{\max}/2\epsilon_{\text{dct}}$. For arbitrary $\epsilon_{\text{dct}} \in (0, V_{\max})$, a similar discretization yields a cardinality of $\lceil V_{\max}/2\epsilon_{\text{dct}}\rceil$, and we upper-bound it by $V_{\max}/\epsilon_{\text{dct}}$ throughout the analysis for convenience.

[5] $\tilde{O}(\cdot)$ suppresses poly-logarithmic dependencies.

[6] We treat ϕ as mapping $\mathcal{S} \times \mathcal{A}$ to an arbitrary finite codomain, and $g(s,a) = g(s',a')\ \forall g \in \mathcal{G}_\phi$ iff $\phi(s,a) = \phi(s',a')$.

realizability error $\epsilon_\phi := \epsilon_{\mathcal{G}_\phi}$, we show that we can compute a statistic for any given function $f_0 : \mathcal{S}\times\mathcal{A} \to [0, V_{\max}]$, and the statistic will be a good surrogate for $\|f_0 - Q^\star\|$ as long as Assumption 1 holds and the sample size is polynomially large. We use $|\phi|$ to denote the number of equivalence classes induced by ϕ.

As Section 4 and Algorithm 1 have already alluded to, later we will invoke this result when comparing two candidate value functions f and f' (with $f_0 = f$), and define ϕ as the coarsest partition that can express both f and f'; when $Q^\star \in \{f, f'\}$, ϵ_ϕ will be small. To maintain the modularity of the analysis, however, we will view ϕ as an arbitrary partition of $\mathcal{S}\times\mathcal{A}$ in this section.

The statistic we compute is $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ (c.f. Line 8 of Algorithm 1), where $\widehat{\mathcal{T}}_\phi^\mu$ is defined as follows:

Definition 2. Define $\widehat{\mathcal{T}}_\phi^\mu$ as the sample-based projected Bellman update operator associated with $\mathcal{G}_\phi$: for any $f : \mathcal{S}\times\mathcal{A} \to [0, V_{\max}]$, $\widehat{\mathcal{T}}_\phi^\mu f :=$

$$\arg\min_{g\in\mathcal{G}_\phi} \frac{1}{|D|}\sum_{(s,a,r,s')\in D} [(g(s,a) - r - \gamma V_f(s'))^2]. \quad (1)$$

5.1. Warm up: $|D| \to \infty$ and $\epsilon_\phi = 0$

To develop intuitions, we first consider the special case of $|D| \to \infty$ and $\epsilon_\phi = 0$. In this scenario, we can show that $Q^\star$ is the unique fixed point of $\widehat{\mathcal{T}}_\phi^\mu$, which justifies using $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|$ as a surrogate for $\|f_0 - Q^\star\|$. The concepts and lemmas introduced here will also be useful for the later analysis of the general case.

We start by defining $\mathcal{T}_\phi^\mu$ as $\widehat{\mathcal{T}}_\phi^\mu$ when $|D| \to \infty$.

Definition 3. Define $\mathcal{T}_\phi^\mu$ as the projected Bellman update where the projection is onto $\mathcal{G}_\phi$, weighted by μ. That is, for any $f : \mathcal{S}\times\mathcal{A} \to [0, V_{\max}]$,

$$\mathcal{T}_\phi^\mu f := \arg\min_{g\in\mathcal{G}_\phi} \mathbb{E}_\mu[(g(s,a) - r - \gamma V_f(s'))^2]. \quad (2)$$

Next, we show that it is possible to define an MDP M_ϕ, such that $\mathcal{T}_\phi^\mu$ coincides with the Bellman update of M_ϕ. Readers familiar with state abstractions may find the definition unusual, as the "abstract MDP" associated with ϕ is typically defined over the compressed (or abstract) state space instead of the original one (e.g., Ravindran & Barto, 2004). We define M_ϕ over $\mathcal{S}$ because (1) our ϕ is an arbitrary partition of $\mathcal{S}\times\mathcal{A}$, which does not necessarily induce a consistent notion of abstract states, and (2) even when it does, the MDPs defined over $\mathcal{S}$ are dual representations to and share many important properties with the classical notion of abstract MDPs (Jiang, 2018).

Definition 4. Define $M_\phi = (\mathcal{S}, \mathcal{A}, P_\phi, R_\phi, \gamma, d_0)$, where

$$R_\phi(s,a) = \frac{\sum_{\tilde{s},\tilde{a}:\phi(\tilde{s},\tilde{a})=\phi(s,a)} \mu(\tilde{s},\tilde{a})R(\tilde{s},\tilde{a})}{\sum_{\tilde{s},\tilde{a}:\phi(\tilde{s},\tilde{a})=\phi(s,a)} \mu(\tilde{s},\tilde{a})}.$$

$$P_\phi(s'|s,a) = \frac{\sum_{\tilde{s},\tilde{a}:\phi(\tilde{s},\tilde{a})=\phi(s,a)} \mu(\tilde{s},\tilde{a})P(s'|\tilde{s},\tilde{a})}{\sum_{\tilde{s},\tilde{a}:\phi(\tilde{s},\tilde{a})=\phi(s,a)} \mu(\tilde{s},\tilde{a})}.$$

Lemma 3. *$\mathcal{T}_\phi^\mu$ is the Bellman update operator of M_ϕ.*

Lemma 3 implies that $\mathcal{T}_\phi^\mu$ is a γ-contraction under ℓ_∞ and has a unique fixed point, namely the optimal Q-function of M_ϕ. It then suffices to show that $Q^\star$ is such a fixed point.

Proposition 4. *When $\epsilon_\phi = 0$, $Q^\star$ is the unique fixed point of $\mathcal{T}_\phi^\mu$.*

5.2. The General Case

In the general case, we want to show that $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ and $\|f_0 - Q^\star\|$ control each other. The central result of this section is the following proposition:

Proposition 5. *Fixing any $\epsilon_1, \tilde{\epsilon}$. Suppose*

$$|D| \ge \frac{32V_{\max}^2|\phi|\ln\frac{8V_{\max}}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2} + \frac{50V_{\max}^2|\phi|\ln\frac{80V_{\max}}{\epsilon_1\delta}}{\epsilon_1^2}.$$

Then, with probability at least $1-\delta$, for any $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$ such that $\|\nu/\mu\|_\infty \le C$,

$$\|f_0 - Q^\star\|_{2,\nu} \le \frac{2\epsilon_\phi + \sqrt{C}(\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D} + \epsilon_1 + \tilde{\epsilon})}{1-\gamma}. \quad (3)$$

At the same time,

$$\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D} \le (1+\gamma)\|f_0 - Q^\star\|_\infty + 2\epsilon_\phi + \tilde{\epsilon} + \epsilon_1. \quad (4)$$

Proving the proposition requires quite some preparations. We group the helper lemmas according to their nature in Sections 5.2.1 to 5.2.3, and prove Proposition 5 in Section 5.2.4.

5.2.1. Error Propagation

The first two lemmas allow us to characterize error propagation in later proofs. That is, it will help answer the question: if we find $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ to be small (but nonzero), why does it imply that $\|f_0 - Q^\star\|$ is small?

While results of similar nature exist in the state-abstraction literature, they often bound $\|f_0 - Q^\star\|$ with $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_\infty$, where error propagation is easy to handle (Jiang, 2018). However, $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_\infty$ can only be reliably estimated if each group of state-action pairs receives a sufficient portion of the data, which is not guaranteed in our setting due to the arbitrary nature of ϕ created in Line 7. This forces us to work with the μ-weighted ℓ_2-norm and carefully characterize how error propagation shifts the distributions.

In fact, it is precisely this analysis that demands the strong definition of concentrability coefficient $C_{\mathcal{S}}$ in Assumption 1: as we will later show in the proof of Proposition 5 (Section 5.2.4), the error propagates according to the dynamics of M_ϕ instead of that of M (c.f. the $P_\phi(\nu)$ term in Eq.(10)). Therefore, popular definitions of concentrability coefficient (e.g., Munos, 2007; Antos et al., 2008; Farahmand et al., 2010; Xie & Jiang, 2020)—which all consider state distributions induced in M—do not fit our analysis. Fortunately, the $C_{\mathcal{S}}$ defined in Assumption 1 has a very nice property, that it automatically carries over to M_ϕ no matter what ϕ is:

Lemma 6. *Any $C < \infty$ that satisfies Assumption 1 for the true MDP M also satisfies the same assumption in M_ϕ. As a further consequence, Proposition 1 is also satisfied when P is replaced by P_ϕ.*

5.2.2. Error of $Q^\star$ under $\mathcal{T}_\phi^\mu$

The next lemma parallels Proposition 4 in Section 5.1, where we showed that $\|Q^\star - \mathcal{T}_\phi^\mu Q^\star\| = 0$ when $\epsilon_\phi = 0$. When ϵ_ϕ is non-zero, we need a more robust version of this result showing that $\|Q^\star - \mathcal{T}_\phi^\mu Q^\star\|$ is controlled by ϵ_ϕ.

Lemma 7. $\|Q^\star - \mathcal{T}_\phi^\mu Q^\star\|_\infty \le 2\epsilon_\phi$.

5.2.3. Concentration Bounds

We need two concentration events: that $\widehat{\mathcal{T}}_\phi^\mu f_0$ is close to $\mathcal{T}_\phi^\mu f_0$, and that $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ is close to $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,\mu}$. We will split the failure probability δ evenly between these events.

Concentration of $\widehat{\mathcal{T}}_\phi^\mu f_0$ We begin with the former, which requires a standard result for realizable least-square regression. The proof is deferred to Appendix B.5.

Lemma 8 (Concentration Bound for Least-Square Regression)**.** *Consider a real-valued regression problem with feature space $\mathcal{X}$ and label space $\mathcal{Y} \subset [0, V_{\max}]$. Let $(x_i, y_i) \sim P_{X,Y}$ be n i.i.d. data points. Let $\mathcal{H} \subset (\mathcal{X} \to \mathcal{Y})$ be a hypothesis class with ℓ_∞ covering number $N = \mathcal{N}_\infty(\mathcal{H}, \epsilon_0)$ and that realizes the Bayes-optimal regressor, i.e., $h^\star = (x \mapsto \mathbb{E}[Y|X = x]) \in \mathcal{H}$. Let $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\mathbb{E}}[(h(X) - Y)^2]$ be the empirical risk minimizer (ERM), where $\hat{\mathbb{E}}$ is the empirical expectation based on $\{(x_i, y_i)\}_{i=1}^n$. Then, with probability at least $1 - \delta$,*

$$\mathbb{E}[(h^\star(X) - \hat{h}(X))^2] \le \frac{8V_{\max}^2 \log\frac{N}{\delta}}{n} + 8V_{\max}\epsilon_0.$$

We then use Lemma 8 to prove that $\|\widehat{\mathcal{T}}_\phi^\mu f - \mathcal{T}_\phi^\mu f\|_{2,\mu}$ is small.

Lemma 9. *Fixing $f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]$. W.p. $\ge 1 - \frac{\delta}{2}$, $\|\widehat{\mathcal{T}}_\phi^\mu f - \mathcal{T}_\phi^\mu f\|_{2,\mu} \le \tilde{\epsilon}$, as long as*

$$|D| \ge \frac{16V_{\max}^2(2|\phi|\log(4V_{\max}/\tilde{\epsilon}) + \log(2/\delta))}{\tilde{\epsilon}^2}.$$

Technical Challenge & Proof Idea Recall that $\widehat{\mathcal{T}}_\phi^\mu f$ is the ERM (in $\mathcal{G}_\phi$) of the least-square regression problem $(s, a) \mapsto r + \gamma V_f(s')$, so $\|\widehat{\mathcal{T}}_\phi^\mu f - \mathcal{T}_\phi^\mu f\|_{2,\mu}$ essentially measures the μ-weighted ℓ_2 distance between the ERM and the population risk minimizer. Proving this is straightforward when the regression problem is realizable, as Lemma 8 would be directly applicable.[7] In our case, however, the regression problem is in general non-realizable (except for $f = Q^\star$) and can incur arbitrarily large approximation errors, as $\mathcal{G}_\phi$ does not necessarily contain the Bayes-optimal regressor $\mathcal{T}f$.

The key proof idea is to leverage a special property of piecewise constant classes[8] to reduce the analysis to the realizable case: regressing $(s, a) \mapsto r + \gamma V_g(s')$ over $\mathcal{G}_\phi$ is equivalent to regressing $x \mapsto r + \gamma V_g(s')$ (with $x = \phi(s, a)$) over a "tabular" function class, where the $s, a|x$ portion of the data generation process is treated as part of the inherent label noise. After switching to this alternative view, the tabular class over the codomain of ϕ is fully expressive and always realizable, which makes Lemma 8 applicable. See Appendix B.6 for the full proof of Lemma 9.

Concentration of $\|f_0 - g\|_{2,D}$ The second concentration result we need is an upper bound on $|\|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu}|$ for all $g \in \mathcal{G}_\phi$ simultaneously. We need to union bound over $g \in \mathcal{G}_\phi$ because our statistic is $\|f_0 - g\|_{2,D}$ with $g = \widehat{\mathcal{T}}_\phi^\mu f_0$, which is a data-dependent function.

Lemma 10. *W.p. $\ge 1 - \delta/2$, $\forall g \in \mathcal{G}_\phi$, $|\|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu}| \le \epsilon_1$, as long as*

$$|D| \ge \frac{50V_{\max}^2|\phi| \ln\frac{80V_{\max}}{\epsilon_1\delta}}{\epsilon_1^2}.$$

Technical Challenge & Proof Idea It is straightforward to bound $|\|f_0 - g\|_{2,D}^2 - \|f_0 - g\|_{2,\mu}^2|$ (note the squares), but a naïve conversion to a bound on the desired quantity (difference without squares) would result in $O(n^{-1/4})$ rate. To obtain $O(n^{-1/2})$ rate, we consider two situations separately, depending on whether $\|f_0 - g\|_{2,\mu}$ is below or above certain threshold: when it is below the threshold, we can use

[7] See e.g., Lemma 16 of Chen & Jiang (2019), where the (approximate) realizability of any such regression problem is guaranteed by the assumption of low inherent Bellman error.

[8] An alternative (and much messier) approach is to prove scalar-valued concentration bounds for $\widehat{\mathcal{T}}_\phi^\mu f$ in each group of state-action pairs. Those groups with few data points will have high uncertainty, but they also contribute little to $\|\cdot\|_{2,\mu}$. Compared to this approach, our proof is much simpler.

Bernstein's to exploit the low variance of $(f_0 - g)^2$; when it is above the threshold, we obtain the bound by factoring the difference of squares. Combining these two cases with an $O(\epsilon_1)$ threshold yields a clean $O(n^{-1/2})$ result; see proof details in Appendix B.7.

5.2.4. Proof of Proposition 5

We are now ready to prove Proposition 5. Due to space limit we only provide a proof sketch in the main text.

Proof Sketch. To prove Eq.(3), define $\pi_{f,f'}$ as the policy $s \mapsto \arg\max_a \max\{f(s,a), f'(s,a)\}$. Consider any ν such that $\|\nu/\mu\|_\infty \le C$, we have $\|Q^\star - f_0\|_{2,\nu} \le$

$$\|Q^\star - \mathcal{T}_\phi^\mu Q^\star\|_{2,\nu} + \|\mathcal{T}_\phi^\mu Q^\star - \mathcal{T}_\phi^\mu f_0\|_{2,\nu} + \|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\nu}.$$

The first term can be bounded via Lemma 7. The second term is bounded by $\gamma\|Q^\star - f_0\|_{2,P_\phi(\nu)\times\pi_{\hat{f},Q^\star}}$, where $P_\phi(\nu) \times \pi_{\hat{f},Q^\star}$ is a distribution that also satisfies $\|(\cdot)/\mu\|_\infty \le C$ (Proposition 1) and hence can be handled by recursion. The third can be bounded by $\sqrt{C}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ due to $\|\nu/\mu\|_\infty \le C$, and $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ can be related to $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ by the concentration bounds established in Section 5.2.3, which are satisfied due to the choice of $|D|$ in the proposition statement.

To prove Eq.(4), we can similarly relate $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ to $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ via the concentration bounds, and

$$\begin{aligned}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu} &\le \|f_0 - \mathcal{T}_\phi^\mu f_0\|_\infty \\ &\le \|f_0 - Q^\star\|_\infty + \|\mathcal{T}_\phi^\mu Q^\star - \mathcal{T}_\phi^\mu f_0\|_\infty \\ &\le (1+\gamma)\|f_0 - Q^\star\|_\infty. \quad (\gamma\text{-contraction of } \mathcal{T}_\phi^\mu)\end{aligned}$$ □

6. Proof of Theorem 2

With the careful analysis of the pairwise-comparison step given in Section 5, we are now ready to analyze Algorithm 1. Roughly speaking, we will make the following arguments:

- For the output $\hat{f}$, if $\max_{f'} \mathcal{E}(\hat{f}; f')$ is small, then $\hat{f} \approx Q^\star$. (Eq.(3) of Proposition 5)
- That $\max_{f'} \mathcal{E}(\hat{f}; f')$ will be small, because $\max_{f'} \mathcal{E}(f^\star; f')$ is small, where $f^\star \in \mathcal{F}$ is the best approximation of $Q^\star$ in Definition 1. (Eq.(4) of Proposition 5)

Before we delve into the proof of Theorem 2, we need yet another lemma, which connects ϵ_ϕ in Section 5 to the approximation error of $\mathcal{F}$. As Section 4 has suggested, this is feasible because we are only concerned with the comparisons involving $f^\star$, and ϵ_ϕ may be arbitrarily large otherwise.

Lemma 11. *The ϕ induced from Line 7 satisfies $|\phi| \le (V_{\max}/\epsilon_{dct})^2$. When $f^\star \in \{f, f'\}$, we further have $\epsilon_\phi \le \epsilon_{\mathcal{F}} + \epsilon_{dct}$.*

Proof of Theorem 2. Among the (f, f') pairs enumerated in Lines 5 and 6, we will only be concerned with the cases when either $f = f^\star$ or $f' = f^\star$, and there are $2|\mathcal{F}|$ such pairs. We require that w.p. $1-\delta$, Proposition 5 holds for all these $2|\mathcal{F}|$ pairs. To guarantee so, we set the sample size $|D|$ to the expression in the statement of Proposition 5, with $|\phi|$ replaced by its upper bound in Lemma 11 and δ replaced by $\delta/2|\mathcal{F}|$ (union bound). We also let $\epsilon_1 = \tilde{\epsilon}$ to simplify the expressions, and will set the concrete value of $\tilde{\epsilon}$ later. The following is a sample size that satisfies all the above:

$$|D| \ge \frac{82 V_{\max}^4 \ln \frac{160 V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2 \epsilon_{\text{dct}}^2}. \tag{5}$$

Let ϕ be the partition induced by $\hat{f}$ and $f^\star$. According to Eq.(3), for any ν s.t. $\|\nu/\mu\|_\infty \le C$,

$$\begin{aligned}\|\hat{f} - Q^\star\|_{2,\nu} &\le \frac{2\epsilon_\phi + \sqrt{C}(\|\hat{f} - \widehat{\mathcal{T}}_\phi^\mu \hat{f}\|_{2,D} + 2\tilde{\epsilon})}{1-\gamma} \\ &= \frac{2\epsilon_\phi + \sqrt{C}(\mathcal{E}(\hat{f}; f^\star) + 2\tilde{\epsilon})}{1-\gamma} \\ &\le \frac{2\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + \sqrt{C}(\max_{f'\in\mathcal{F}} \mathcal{E}(\hat{f}; f') + 2\tilde{\epsilon})}{1-\gamma}.\end{aligned}$$

It then remains to bound $\max_{f'} \mathcal{E}(\hat{f}; f')$. Note that

$$\max_{f'\in\mathcal{F}} \mathcal{E}(\hat{f}; f') = \min_{f\in\mathcal{F}} \max_{f'\in\mathcal{F}} \mathcal{E}(f; f') \le \max_{f'\in\mathcal{F}} \mathcal{E}(f^\star; f').$$

For any f', let ϕ' be the partition of $\mathcal{S} \times \mathcal{A}$ induced by $f^\star$ and f'. Then

$$\begin{aligned}\mathcal{E}(f^\star; f') &= \|f^\star - \widehat{\mathcal{T}}_{\phi'}^\mu f^\star\|_{2,D} \\ &\le (1+\gamma)\|f^\star - Q^\star\|_\infty + 2\epsilon_{\phi'} + 2\tilde{\epsilon} \quad (\text{Eq.}(4)) \\ &\le 4\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + 2\tilde{\epsilon}.\end{aligned}$$

Combining the above results, we have for any ν s.t. $\|\nu/\mu\|_\infty \le C$,

$$\begin{aligned}\|\hat{f} - Q^\star\|_{2,\nu} &\le \frac{2\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + \sqrt{C}(4\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + 2\tilde{\epsilon} + 2\tilde{\epsilon})}{1-\gamma} \\ &\le \frac{(2+4\sqrt{C})\epsilon_{\mathcal{F}} + 4\sqrt{C}(\epsilon_{\text{dct}} + \tilde{\epsilon})}{1-\gamma}.\end{aligned}$$

Finally, since any state-action distribution induced by any (potentially non-stationary) policy always satisfies Proposition 5, by Chen & Jiang (2019, Lemma 13) we have

$$\begin{aligned}J(\pi^\star) - J(\hat{\pi}) &\le \frac{2}{1-\gamma} \sup_{\nu:\|\nu/\mu\|_\infty \le C} \|\hat{f} - Q^\star\|_\nu \\ &\le \frac{(4+8\sqrt{C})\epsilon_{\mathcal{F}} + 8\sqrt{C}(\epsilon_{\text{dct}} + \tilde{\epsilon})}{(1-\gamma)^2}.\end{aligned}$$

To guarantee that $\frac{8\sqrt{C}(\epsilon_{\text{dct}}+\tilde{\epsilon})}{(1-\gamma)^2} \le \epsilon V_{\max}$, we set $\epsilon_{\text{dct}} = \tilde{\epsilon} = \frac{(1-\gamma)^2 \epsilon V_{\max}}{16\sqrt{C}}$. Plugging this back into Eq.(5) yields the sample complexity in the theorem statement. □

7. Discussions and Conclusions

7.1. Application to Model Selection

When learning $Q^\star$ from a batch dataset in practice, one would like to try different algorithms, different function approximators, and even different hyperparameters for a fixed algorithm and see which combination gives the best result, as is always the case in machine-learning practices. In supervised learning, this can be done by a simple cross-validation procedure on the holdout dataset. In batch RL, however, how to perform such a *model-selection* step in a provably manner has been a widely open problem.[9]

There exists a limited amount of theoretical work on this topic, which often consider a restrictive setting when the base algorithms are model-based learners using nested state abstractions (Hallak et al., 2013; van Seijen et al., 2014; Jiang et al., 2015).[10] The only finite-sample guarantee we are aware of, given by Jiang et al. (2015), provides an oracle inequality with respect to an upper bound of $\|f - Q^\star\|$ based on how much the base state abstractions violate *bisimulation* (or model-irrelevance) criterion (Whitt, 1978; Even-Dar & Mansour, 2003; Li et al., 2006) and ℓ_∞ concentration bounds, and the guarantee does not scale to the case where the number of base algorithms is super constant.

In comparison, BVFT provides a more direct approach with a much stronger guarantee: let $Q_1, \ldots, Q_m$ be the output of different base algorithms. We can simply run BVFT on the holdout dataset with $\mathcal{F} = \{Q_i\}_{i=1}^m$. The only function-approximation assumption we need is that one of Q_i's is a good approximation of $Q^\star$, which is hardly an assumption as there is little we can do if all the base algorithms produce bad results. Compared to prior works, our approach is much more agnostic w.r.t. the details of the base algorithms, our loss and guarantees are directly related to $\|f - Q^\star\|$ as opposed to relying on (possibly loose) upper bounds based on bisimulation, and our statistical guarantee scales to an exponentially large $\mathcal{F}$ as opposed to a constant-sized one.

Another common approach to model selection is to estimate $J(\pi)$ for each candidate π via off-policy evaluation (OPE).[11] OPE-based model selection has very different characteristics compared to BVFT, and they may be used together to complement each other; see a more detailed comparison and discussion in Appendix F.

7.2. On the Assumption of Exploratory Data

As noted in Section 3.2, our Assumption 1 adopts a relatively stringent definition of concentrability coefficient. A more standard definition is the following, as appeared in the hardness conjecture of Chen & Jiang (2019):

Assumption 2. Let d_t^π be the distribution of (s_t, a_t) when we start from $s_0 \sim d_0$ and follow policy π, which we will call an *admissible distribution*. We assume that there exists $C < \infty$ such that $\|d_t^\pi / \mu\|_\infty \leq C$ for any (possibly nonstationary) policy π and $t \geq 0$.

In Appendix C we construct 3 scenarios to illustrate the difficulties (and sometimes possibilities) in extending our algorithm and its guarantees to a weaker data assumption such as Assumption 2; due to space limit we only include a high-level summary of the results below. In the first construction, we show that BVFT fails under Assumption 2 in a very simple MDP if we are allowed to provide a contrived μ distribution to the learner where data is unnaturally missing in certain states (Figure 1). Motivated by the unnaturalness of the construction, we attempt to circumvent the hardness by imposing an additional mild assumption on top of Assumption 2, that μ must itself be "admissible" . While it becomes much more difficult to construct a counterexample against the algorithm, it is still possible to design a scenario where our analysis breaks down seriously (Figure 2). We conclude with a positive result showing that the actual assumption we need is somewhere in between Assumptions 1 and 2, for that our algorithm and analysis work for a simple and natural "on-policy" case which obviously violates Assumption 1; formulating a tighter version of the assumption in a natural and interpretable manner remains future work.

7.3. Conclusions

We conclude the paper with a few open problems:

- Is it possible to circumvent the failure modes discussed in Section 7.2 with novel algorithmic ideas, so that a variant of BVFT only requires a weaker assumption on data? On a related note, the original hardness conjecture of Chen & Jiang (2019) remains unsolved: our positive result assumes a stronger data assumption, and the negative results of Wang et al. (2020); Amortila et al. (2020) assume weaker ones.
- When the data is seriously under-exploratory, to the extent that it is impossible to compete with $\pi^\star$ (Fujimoto et al., 2019; Liu et al., 2019; 2020), what is the minimal function-approximation assumption that enables polynomial learning? In particular, requiring that $\mathcal{F}$ realizes $Q^\star$ no longer makes sense as we do not even attempt to compete with $\pi^\star$. Recent works often suggest that we compete with π whose occupancy is covered by μ, but as of now very strong expressivity assumptions are needed

[9] See Mandel et al. (2014) and Paine et al. (2020) for empirical advances on this problem.

[10] An exception is the work of Farahmand & Szepesvári (2011), which requires the additional assumption that a regression procedure can approximate $\mathcal{T}f$ and uses it to compute $\|f - \mathcal{T}f\|$.

[11] As a side note, BVFT can be adapted to OPE when $Q^\pi \in \mathcal{F}$ for target policy π as long as we change the max operator in $\widehat{\mathcal{T}}_\phi^\mu$ to π, though Assumption 1 will still be needed.

to achieve such an ambitious goal (e.g., Jiang & Huang, 2020, Proposition 9). It will be interesting to explore more humble objectives and see if the algorithmic and analytical ideas in this work extend to the more realistic setting of learning with non-exploratory data.

Acknowledgments

The authors thank Akshay Krishnamurthy, Yu Bai, Yu-Xiang Wang for discussions related to Lemma 10, and Alekh Agarwal for numerous comments and discussions after the initial draft of the paper was released. Nan Jiang acknowledges support from the DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Agarwal, S. *E0 370 Statistical Learning Theory: Covering Numbers, Pseudo-Dimension, and Fat-Shattering Dimension*. Indian Institute of Science, 2011. https://www.shivani-agarwal.net/Teaching/E0370/Aug-2011/Lectures/5.pdf.

Amortila, P., Jiang, N., and Xie, T. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

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.

Barreto, A. d. M. S., Pineau, J., and Precup, D. Policy iteration based on stochastic factorization. *Journal of Artificial Intelligence Research*, 50:763–803, 2014.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1042–1051, 2019.

Chen, L., Scherrer, B., and Bartlett, P. L. Infinite-horizon offline reinforcement learning with linear function approximation: Curse of dimensionality and algorithm. *arXiv preprint arXiv:2103.09847*, 2021.

Devroye, L. and Lugosi, G. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Even-Dar, E. and Mansour, Y. Approximate equivalence of Markov decision processes. In *Learning Theory and Kernel Machines*, pp. 581–594. Springer, 2003.

Farahmand, A.-m. and Szepesvári, C. Model selection in reinforcement learning. *Machine learning*, 85(3):299–332, 2011.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error Propagation for Approximate Policy and Value Iteration. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2010.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. In *Proceedings of the 37th International Conference on Machine Learning (ICML-20)*, 2020.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

Gordon, G. J. Stable function approximation in dynamic programming. In *Proceedings of the twelfth international conference on machine learning*, pp. 261–268, 1995.

Hallak, A., Di-Castro, D., and Mannor, S. Model selection in markovian processes. In *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data mining*, pp. 374–382, 2013.

Jiang, N. *CS 598: Notes on State Abstractions*. University of Illinois at Urbana-Champaign, 2018. http://nanjiang.cs.illinois.edu/files/cs598/note4.pdf.

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

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiang, N. and Li, L. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 652–661, 2016.

Jiang, N., Kulesza, A., and Singh, S. Abstraction Selection in Model-based Reinforcement Learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 179–188, 2015.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4: 1107–1149, 2003.

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pp. 531–539, 2006.

Littman, M. L. and Szepesvári, C. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pp. 310–318, 1996.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1077–1084, 2014.

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.

Paduraru, C., Kaplow, R., Precup, D., and Pineau, J. Model-based reinforcement learning with state aggregation. In *8th European Workshop on Reinforcement Learning*, 2008.

Paine, T. L., Paduraru, C., Michi, A., Gulcehre, C., Zolna, K., Novikov, A., Wang, Z., and de Freitas, N. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766, 2000.

Ravindran, B. and Barto, A. Approximate homomorphisms: A framework for nonexact minimization in Markov decision processes. In *Proceedings of the 5th International Conference on Knowledge-Based Computer Systems*, 2004.

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*, 2019.

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

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.

van Seijen, H., Whiteson, S., and Kester, L. Efficient abstraction selection in reinforcement learning. *Computational Intelligence*, 30(4):657–699, 2014.

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*, 2020.

Whitt, W. Approximations of dynamic programs, I. *Mathematics of Operations Research*, 3(3):231–243, 1978.

Xie, T. and Jiang, N. $Q^\star$ Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 550–559, 2020.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *arXiv preprint arXiv:2012.08005*, 2020.