# 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>.

---

[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 $\ell_\infty$ 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 $\ell_2$ 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 $\mu$'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 $\nu$ be a distribution over $\mathcal{S} \times \mathcal{A}$ and $\pi$ 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 $\|\nu'/\mu\|_\infty := \max_{s,a} \nu'(s,a)/\mu(s,a) \leq C$. Also note that $\|(d_0 \times \pi)/\mu\|_\infty \leq C$.*

**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 $\ell_2$ 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

---

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

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

---

when we only impose realizability on $\mathcal{F}$, which is why our result may be surprising.

**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 $\ell_\infty$ covering number (Agarwal, 2011), our approach and analysis immediately extend by replacing $\mathcal{F}$ with its $\epsilon$-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 $\epsilon$-optimal policy $\hat{\pi}$, that is, $J(\hat{\pi}) \geq J(\pi^\star) - \epsilon \cdot V_{\max}$, 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 $\epsilon$, and $1/\delta$ where $\delta$ 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 $\ell_2$ 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 $\gamma$-contractions under $\ell_\infty$, 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 $\epsilon_{\text{dct}}$,[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* $\epsilon_{\text{dct}} = \frac{(1-\gamma)^2 \epsilon V_{\max}}{16\sqrt{C}}$ *returns a policy* $\hat{\pi}$ *that satisfies*

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

*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 $\phi$, 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 $\epsilon_{\text{dct}}$ 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 $\phi$ 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 $\phi$.

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 $\phi$ as the coarsest partition that can express both $f$ and $f'$; when $Q^\star \in \{f, f'\}$, $\epsilon_\phi$ will be small. To maintain the modularity of the analysis, however, we will view $\phi$ 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 $f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]$, $\quad \widehat{\mathcal{T}}_\phi^\mu f :=$

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

**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 $\mu$. That is, for any $f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]$,

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

Next, we show that it is possible to define an MDP $M_\phi$, such that $\mathcal{T}_\phi^\mu$ coincides with the Bellman update of $M_\phi$. Readers familiar with state abstractions may find the definition unusual, as the "abstract MDP" associated with $\phi$ is typically defined over the compressed (or abstract) state space instead of the original one (e.g., Ravindran & Barto, 2004). We define $M_\phi$ over $\mathcal{S}$ because (1) our $\phi$ 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* $M_\phi$.

Lemma 3 implies that $\mathcal{T}_\phi^\mu$ is a $\gamma$-contraction under $\ell_\infty$ and has a unique fixed point, namely the optimal $Q$-function of $M_\phi$. 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*

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

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

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

*At the same time,*

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

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 $\phi$ created in Line 7. This forces us to work with the $\mu$-weighted $\ell_2$-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 $M_\phi$ 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 $M_\phi$ no matter what $\phi$ is:

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

### 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 $\epsilon_\phi$ 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 $\epsilon_\phi$.

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

### 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 $\delta$ 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 $\ell_\infty$ 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 $1 - \delta$,*

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

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. $\geq 1 - \frac{\delta}{2}$,*

$\|\widehat{\mathcal{T}}_\phi^\mu f - \mathcal{T}_\phi^\mu f\|_{2,\mu} \leq \tilde{\epsilon}$, *as long as*

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

**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 $\mu$-weighted $\ell_2$ 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 $\phi$ 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. $\geq 1 - \delta/2$, $\forall g \in \mathcal{G}_\phi$, $|\|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu}| \leq \epsilon_1$, as long as*

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

**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 $\nu$ such that $\|\nu/\mu\|_\infty \leq C$, we have $\|Q^\star - f_0\|_{2,\nu} \leq$

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

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 $\|(\cdot)/\mu\|_\infty \leq C$ (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 $\|\nu/\mu\|_\infty \leq C$, and $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ can be related to $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ 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

$$\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu} \leq \|f_0 - \mathcal{T}_\phi^\mu f_0\|_\infty$$
$$\leq \|f_0 - Q^\star\|_\infty + \|\mathcal{T}_\phi^\mu Q^\star - \mathcal{T}_\phi^\mu f_0\|_\infty$$
$$\leq (1+\gamma)\|f_0 - Q^\star\|_\infty. \quad (\gamma\text{-contraction of } \mathcal{T}_\phi^\mu) \quad \square$$

## 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 $\epsilon_\phi$ 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 $\epsilon_\phi$ may be arbitrarily large otherwise.

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

*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 $\delta$ 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:

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

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

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

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

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

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

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

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

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

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

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

To guarantee that $\frac{8\sqrt{C}(\epsilon_{\text{dct}} + \tilde{\epsilon})}{(1-\gamma)^2} \leq \epsilon V_{\max}$, 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. $\square$

# 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 $\ell_\infty$ 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 $\pi$ 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.

---

[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 $\pi$ as long as we change the max operator in $\widehat{\mathcal{T}}_\phi^\mu$ to $\pi$, though Assumption 1 will still be needed.

## 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 $d_t^\pi$ be the distribution of $(s_t, a_t)$ when we start from $s_0 \sim d_0$ and follow policy $\pi$, 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 $\pi$ 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 $\mu$ 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 $\mu$ 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 $\pi$ whose occupancy is covered by $\mu$, but as of now very strong expressivity assumptions are 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

## 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.

## A. Further Discussions of Assumption 1

### A.1. Example of Bounded $C_\mathcal{S}$ in Low-rank Environments

Here we show that in general environments whose transition admits low-rank stochastic factorization, there always exists $\mu$ that satisfies Assumption 1 with a small $C$.

**Example 1.** *Consider a low-rank MDP as defined in Barreto et al. (2014); Jiang et al. (2017, Proposition 9), where the transition matrix $[P(s'|s,a)]_{(s,a),s'} = P_1 \times P_2$, and $P_1 \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times d}$ and $P_2 \in \mathbb{R}^{d \times |\mathcal{S}|}$ are both row-stochastic matrices. Also assume that $d_0$ is a mixture of rows of $P_2$. Then, a data distribution $\mu$, where $\mu(s)$ is the average of $P_2$'s rows and $\mu(a|s)$ is uniform, satisfies Assumption 1 with $C \le d|\mathcal{A}|$.* [12]

*Proof.* $C_\mathcal{A} \le |\mathcal{A}|$ follows from the uniformity of $\mu(a|s)$. For $C_\mathcal{S} \le d$, note that $P(\cdot|s,a)$ and $d_0(\cdot)$ are convex combinations of rows of $P_2$, and $\mu(s)$ is designed to be uniform mixture of these rows, so $P(s'|s,a)/\mu(s')$ is always bounded by the number of rows, which is $d$. $\qquad\square$

**Comparison to Standard Concentrability** The more lenient and popular definitions of concentrability (e.g., Assumption 2) are also found to be satisfiable in low-rank MDPs—in fact, such low-rankness is the only type of general structure known to enable concentrability (Chen & Jiang, 2019, Proposition 10). Comparing our Example 1 with the example given by Chen & Jiang (2019), the most outstanding difference is that in their case, the data distribution $\mu$ can be a mixture of state distributions induced by different policies in the environment; if a hidden factor cannot be reached by any policy, it is possible that any mixture distribution may fail to satisfy Assumption 1 with a reasonably small $C$.

**Comparison to No Inherent Bellman Errors** In the above low-rank MDP scenario, the assumption that $\mathcal{F}$ has no *inherent Bellman error* (Antos et al., 2008; Chen & Jiang, 2019)—which enables polynomial sample complexity for many existing algorithms—can provably hold when the left factorization matrix (the analogy of $P_1$ in Example 1) is known to the learner as state-action features, so it is worth comparing such a setting to ours. In this setting, which is often known as *linear MDPs* (Jin et al., 2020), one can choose $\mathcal{F}$ to be the linear class induced by the left factorization matrix, which is guaranteed to be closed under $\mathcal{T}$, i.e., have no inherent Bellman errors. In contrast, our Assumption 1 holds without relying on knowing the left factorization matrix. The price we pay is that we require a stochastic factorization (Example 1) instead of just low-rankness, and whether Algorithm 1 can hold with just low-rankness (possibly with additional mild assumptions) is an open problem.

### A.2. 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. Among them, the result of (Amortila et al., 2020) is most closely related to ours, and their setup is roughly that (1) $\mathcal{F}$ is a linear function class (induced by feature map $\varphi$) with $Q^\star \in \mathcal{F}$, and (2) the feature covariance matrix, $\mathbb{E}_\mu[\varphi\varphi^\top]$, has lower-bounded eigenvalues under the data distribution. To understand how these results relate to ours, consider the following data assumption:

**Assumption 3** ($\mathcal{F}$-aware Concentrability). We assume that there exists $C < \infty$, such that for any $f, f' \in \mathcal{F}$, $\|f - f'\|_{2, d_t^\pi}^2 \le C\|f - f'\|_{2,\mu}^2$ for any (possibly nonstationary) policy $\pi$ and $t \ge 0$ (see Assumption 2 for the definition of $d_t^\pi$).

When $\mathcal{F}$ is linear and $\mathbb{E}_\mu[\varphi\varphi^\top]$ has large eigenvalues (assuming $\|\phi\|_2 \le 1$), Assumption 3 is automatically satisfied with $C$ being the inverse of the smallest eigenvalue of $\mathbb{E}_\mu[\varphi\varphi^\top]$. Therefore, translating the lower bounds of Wang et al. (2020); Amortila et al. (2020) to the setting of generic $\mathcal{F}$, we have:

**Proposition 12** (Corollary of Amortila et al. (2020)). *In the setup of our main text, if we replace Assumption 1 with Assumption 3, no algorithm with finite sample complexity exists.*

*Proof.* This is a direct corollary of Amortila et al. (2020), who show that finite sample complexity is unattainable even with stronger assumptions: linear $\mathcal{F}$ is stronger than generic $\mathcal{F}$, and $\mathbb{E}_\mu[\varphi\varphi^\top]$ having eigenvalues lower-bounded away from $0$ is stronger than Assumption 3. $\qquad\square$

---

[12]We do not particularly attempt to satisfy $\mu(s,a) > 0 \; \forall s, a$ here, because it can always be satisfied by mixing in an infinitesimally small portion of $U(\mathcal{S} \times \mathcal{A})$. In fact, this assumption itself can be removed from the main analysis with some additional care.

Next we consider the relationship between Assumptions 1, 2, and 3 in the following result; the proof is elementary and can be extracted from existing analyses (Munos, 2003; 2007; Chen & Jiang, 2019).

**Proposition 13.** *Assumption 1 $\Rightarrow$ Assumption 2 $\Rightarrow$ Assumption 3.*

With these results, we now can relate the lower bounds Wang et al. (2020); Amortila et al. (2020) to our positive result: indeed they do not contradict each other, as the negative results use the weakest form of concentrability (the $\mathcal{F}$-aware version in Assumption 3), and our positive result uses the strongest form (Assumption 1). Furthermore, to circumvent the hardness in Proposition 12, imposing linear structure on $\mathcal{F}$—which is a very strong structural assumption—does not help, as the hardness results of Wang et al. (2020); Amortila et al. (2020) still apply. On the other hand, making a stronger data assumption as in Assumption 1 would avoid the lower bound.

As a final remark, the hardness conjecture of Chen & Jiang (2019), which uses Assumption 2, remains unsolved. Originally Chen & Jiang (2019) argued that hardness conjecture is highly likely true given the lack of positive results under realizability, but given our work the picture is much less clear now. If we still anticipate a hardness result, our work has substantially narrowed the search space for the lower-bound constructions (if they exist): we will necessarily be able to establish the lower bound with either $|\mathcal{F}| = 2$ or $\mathcal{F}$ being piecewise constant, otherwise our tournament procedure can extend the polynomial upper bounds for these special settings to arbitrary function classes.

# B. Proofs

## B.1. Proof of Lemma 3

Let $\mathcal{T}_{M_\phi}$ denote the Bellman update operator of $M_\phi$. It suffices to show that for any $f$, $\mathcal{T}_{M_\phi} f \in \mathcal{G}_\phi$ and is the argmin in Eq.(2). $\mathcal{T}_{M_\phi} f \in \mathcal{G}_\phi$ follows directly as for any $s, a, \tilde{s}, \tilde{a}$ such that $\phi(s, a) = \phi(\tilde{s}, \tilde{a})$, $R_\phi(s, a) = R_\phi(\tilde{s}, \tilde{a})$ and $P_\phi(\cdot|s, a) = P_\phi(\cdot|\tilde{s}, \tilde{a})$. For the other claim, note that projecting onto a piecewise constant function class means taking a weighted average within each partition, i.e., for any $s, a$,

$$
\begin{aligned}
(\mathcal{T}_\phi^\mu f)(s, a) &= \mathbb{E}_{(\tilde{s}, \tilde{a}, r, s') \sim \mu}[r + \gamma V_f(s')|\phi(\tilde{s}, \tilde{a}) = \phi(s, a)] \\
&= \frac{\sum_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a})=\phi(s,a)} \mu(\tilde{s}, \tilde{a})(R(\tilde{s}, \tilde{a}) + \gamma\mathbb{E}_{s' \sim P(\tilde{s}, \tilde{a})}[V_f(s')])}{\sum_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a})=\phi(s,a)} \mu(\tilde{s}, \tilde{a})} \\
&= R_\phi(s, a) + \gamma\mathbb{E}_{s' \sim P_\phi(s,a)}[V_f(s')] = (\mathcal{T}_{M_\phi} f)(s, a). \qquad \square
\end{aligned}
$$
(6)

## B.2. Proof of Proposition 4

The existence and the uniqueness of the fixed point of $\mathcal{T}_\phi^\mu$ follow from Lemma 3, so it suffices to check $Q^\star = \mathcal{T}_\phi^\mu Q^\star$. For any $(s, a)$, we will calculate $(\mathcal{T}_\phi^\mu Q^\star)(s, a)$ using Eq.(6), which is a convex average of terms in the form of

$$
R(\tilde{s}, \tilde{a}) + \gamma\mathbb{E}_{s' \sim P(\tilde{s}, \tilde{a})}[V_{Q^\star}(s')] = (\mathcal{T}Q^\star)(\tilde{s}, \tilde{a}) = Q^\star(\tilde{s}, \tilde{a}),
$$

where $\phi(\tilde{s}, \tilde{a}) = \phi(s, a)$. Since $\epsilon_\phi = 0$, we have $Q^\star(\tilde{s}, \tilde{a}) = Q^\star(s, a)$, and $(\mathcal{T}_\phi^\mu Q^\star)(s, a)$ is the convex average of terms that are always equal to $Q^\star(s, a)$. Hence, $\mathcal{T}_\phi^\mu Q^\star = Q^\star$. $\qquad \square$

## B.3. Proof of Lemma 6

$M$ and $M_\phi$ share the same initial state distribution $d_0$. Now for any $(s, a)$, $P_\phi(s'|s, a)$ is a convex combination of $P(s'|\tilde{s}, \tilde{a})$ for $\{(\tilde{s}, \tilde{a}) : \phi(s, a) = \phi(\tilde{s}, \tilde{a})\}$, and since $\frac{P(s'|s,a)}{\mu(s')} \leq C_\mathcal{S}$ for every $(s, a)$, replacing the enumerator with any convex combination satisfies the same inequality. The condition about $C_\mathcal{A}$ only concerns the data distribution $\mu$ and does not depend on the MDP. $\qquad \square$

**B.4. Proof of Lemma 7**

Let $g^\star = \arg\min_{g \in \mathcal{G}_\phi} \|g - Q^\star\|_\infty$, and $\|g^\star - Q^\star\|_\infty = \epsilon_\phi$. For any $(s, a)$,

$$\left| Q^\star(s, a) - (\mathcal{T}_\phi^\mu Q^\star)(s, a) \right|$$

$$= \left| Q^\star(s, a) - \frac{\sum_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a}) = \phi(s, a)} \mu(\tilde{s}, \tilde{a})(R(\tilde{s}, \tilde{a}) + \gamma \mathbb{E}_{s' \sim P(\tilde{s}, \tilde{a})}[V^\star(s')])}{\sum_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a}) = \phi(s, a)} \mu(\tilde{s}, \tilde{a})} \right|$$

$$= \left| Q^\star(s, a) - \frac{\sum_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a}) = \phi(s, a)} \mu(\tilde{s}, \tilde{a})Q^\star(\tilde{s}, \tilde{a})}{\sum_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a}) = \phi(s, a)} \mu(\tilde{s}, \tilde{a})} \right| \leq \max_{\tilde{s}, \tilde{a}:\phi(\tilde{s}, \tilde{a}) = \phi(s, a)} |Q^\star(s, a) - Q^\star(\tilde{s}, \tilde{a})|.$$

Now for any $\tilde{s}, \tilde{a}$ s.t. $\phi(\tilde{s}, \tilde{a}) = \phi(s, a)$,

$$|Q^\star(s, a) - Q^\star(\tilde{s}, \tilde{a})|$$
$$= |Q^\star(s, a) - g^\star(s, a) + g^\star(\tilde{s}, \tilde{a}) - Q^\star(\tilde{s}, \tilde{a})| \qquad (g^\star \in \mathcal{G}_\phi \text{ and is piecewise constant})$$
$$\leq |Q^\star(s, a) - g^\star(s, a)| + |g^\star(\tilde{s}, \tilde{a}) - Q^\star(\tilde{s}, \tilde{a})| \leq 2\epsilon_\phi. \qquad \square$$

**B.5. Proof of Lemma 8**

Fixing any $h \in \mathcal{H}$, define random variable $Z(h) := (h(X) - Y)^2 - (h^\star(X) - Y)^2$, which has bounded range $[-V_{\max}^2, V_{\max}^2]$. Noting that $\mathbb{E}[Z(h)] = \mathbb{E}[(h(X) - h^\star(X))^2]$, we bound the variance of $Z(h)$ as

$$\mathbb{V}[Z(h)] \leq \mathbb{E}[Z(h)^2] = \mathbb{E}[((h(X) - Y)^2 - (h^\star(X) - Y)^2)^2]$$
$$= \mathbb{E}[(h(X) - h^\star(X))^2 (h(X) + h^\star(X) - 2Y)^2]$$
$$\leq 4V_{\max}^2 \mathbb{E}[(h(X) - h^\star(X))^2] = 4V_{\max}^2 \mathbb{E}[Z(h)].$$

Next, we derive a uniform derivation bound on $\mathbb{E}[Z(h)] - \hat{\mathbb{E}}[Z(h)]$ for all $h \in \mathcal{H}$. Let $\mathcal{H}' \subset \mathcal{H}$ be the $\epsilon$-cover of $\mathcal{H}$ under $\ell_\infty$. By the definition of covering number, we have that $|\mathcal{H}'| = N$ and for any $h \in \mathcal{H}$, there exists $h' \in \mathcal{H}'$ such that $\|h - h'\|_\infty \leq \epsilon$.

Applying the one-sided Bernstein and union bounding over all $h' \in \mathcal{H}'$: w.p. $1 - \delta$, $\forall h' \in H'$,

$$\mathbb{E}[Z(h')] - \hat{\mathbb{E}}[Z(h')] \leq \sqrt{\frac{2\mathbb{V}[Z(h')] \log \frac{N}{\delta}}{n}} + \frac{4V_{\max}^2 \log \frac{N}{\delta}}{3n}$$

$$\leq \sqrt{\frac{8V_{\max}^2 \mathbb{E}[Z(h)] \log \frac{N}{\delta}}{n}} + \frac{4V_{\max}^2 \log \frac{N}{\delta}}{3n}.$$

Now for any $h \in \mathcal{H}$, let $h'$ be its closest function in $\mathcal{H}'$, and

$$\mathbb{E}[Z(h)] - \hat{\mathbb{E}}[Z(h)] = \mathbb{E}[Z(h')] - \hat{\mathbb{E}}[Z(h')] + (\mathbb{E}[Z(h)] - \mathbb{E}[Z(h')]) - (\hat{\mathbb{E}}[Z(h)] - \hat{\mathbb{E}}[Z(h')]).$$

We have already bounded the first term, so it suffices to bound the remaining two terms. Consider

$$|Z(h) - Z(h')| = |(h(X) - Y)^2 - (h'(X) - Y)^2|$$
$$= |(h(X) - h'(X))(h(X) + h'(X) - 2Y)| \leq 2V_{\max}\epsilon_0.$$

So we have

$$\mathbb{E}[Z(h)] - \hat{\mathbb{E}}[Z(h)] \leq \sqrt{\frac{8V_{\max}^2 \mathbb{E}[Z(h)] \log \frac{N}{\delta}}{n}} + \frac{4V_{\max}^2 \log \frac{N}{\delta}}{3n} + 4V_{\max}\epsilon_0.$$

Since $\hat{\mathbb{E}}[(h(X) - Y)^2]$ and $\hat{\mathbb{E}}[Z(h)]$ differ by a term that does not depend on $h$, the minimzer of the former, $\hat{h}$, also minimizes the latter. Therefore,

$$\hat{\mathbb{E}}[Z(\hat{h})] \leq \hat{\mathbb{E}}[Z(h^\star)] = 0.$$

This leads to

$$\mathbb{E}[(h(X) - h^\star(X))^2] = \mathbb{E}[Z(\hat{h})] \leq \sqrt{\frac{8V_{\max}^2 \mathbb{E}[Z(\hat{h})] \log \frac{N}{\delta}}{n}} + \frac{4V_{\max}^2 \log \frac{N}{\delta}}{3n} + 4V_{\max}\epsilon_0.$$

Solving for the quadratic formula, we have

$$\mathbb{E}[(h(X) - h^\star(X))^2] = \mathbb{E}[Z(\hat{h})]$$

$$\leq \left( \sqrt{\frac{2V_{\max}^2 \log \frac{N}{\delta}}{n}} + \sqrt{\frac{10V_{\max}^2 \log \frac{N}{\delta}}{3n} + 4V_{\max}\epsilon_0} \right)^2 \leq \frac{22V_{\max}^2 \log \frac{N}{\delta}}{3n} + 8V_{\max}\epsilon_0. \qquad \square$$

## B.6. Proof of Lemma 9

For any $(s, a, r, s')$, let $x = \phi(s, a)$ and $y = r + \gamma V_f(s')$. When we sample $(s, a, r, s')$ according to the data distribution, we use $X$ and $Y$ to denote the random variables whose realizations are $x$ and $y$, respectively. Define $\mathcal{X}$ as the codomain of $\phi$, and $|\mathcal{X}| = |\phi|$. Consider the regression problem $x \mapsto y$ over function class $\mathcal{H} = [0, V_{\max}]^{\mathcal{X}}$. Let $\hat{h}$ and $h^\star$ be the empirical risk minimizer and the Bayes-optimal regressor, respectively, and $h^\star \in \mathcal{H}$ thanks to the full expressivity of $\mathcal{H}$. Also note that $\mathcal{N}_\infty(\mathcal{H}, \epsilon_0) \leq (V_{\max}/\epsilon_0)^{|\mathcal{X}|}$. Invoking Lemma 8 we immediately have that w.p. $\geq 1 - \delta/2$,

$$\mathbb{E}_\mu[(\hat{h}(X) - h^\star(X))^2] \leq \frac{8V_{\max}^2 \log \frac{2(V_{\max}/\epsilon_0)^{|\phi|}}{\delta}}{|D|} + 8V_{\max}\epsilon_0.$$

Next we establish $\mathbb{E}_\mu[(\hat{h}(X) - h^\star(X))^2] = \|\widehat{\mathcal{T}}_\phi^\mu f - \mathcal{T}_\phi^\mu f\|_{2,\mu}^2$. It suffices to show that for any $(s, a, r, s')$, $h^\star(\phi(s, a)) = (\mathcal{T}_\phi^\mu f)(s, a)$ and $\hat{h}(\phi(s, a)) = (\widehat{\mathcal{T}}_\phi^\mu f)(s, a)$. We only show the former and the latter is similar. Let $x = \phi(s, a)$, and

$$\begin{aligned}
(\mathcal{T}_\phi^\mu f)(s, a) &= R_\phi(s, a) + \gamma \mathbb{E}_{s' \sim P_\phi(s,a)}[V_f(s')] \\
&= \frac{\sum_{\tilde{s}, \tilde{a}: \phi(\tilde{s}, \tilde{a}) = \phi(s, a)} \mu(\tilde{s}, \tilde{a})(R(\tilde{s}, \tilde{a}) + \gamma \mathbb{E}_{s' \sim P(\tilde{s}, \tilde{a})}[V_f(s')])}{\sum_{\tilde{s}, \tilde{a}: \phi(\tilde{s}, \tilde{a}) = \phi(s, a)} \mu(\tilde{s}, \tilde{a})} \\
&= \mathbb{E}_{(\tilde{s}, \tilde{a}) \sim \mu}[R(\tilde{s}, \tilde{a}) + \gamma \mathbb{E}_{s' \sim P(\tilde{s}, \tilde{a})}[V_f(s')] \,|\, \phi(\tilde{s}, \tilde{a}) = \phi(s, a)] \\
&= \mathbb{E}_{(\tilde{s}, \tilde{a}) \sim \mu, r = R(\tilde{s}, \tilde{a}), s' \sim P(\tilde{s}, \tilde{a})}[r + \gamma V_f(s') \,|\, \phi(\tilde{s}, \tilde{a}) = x] \\
&= \mathbb{E}_{(s, a) \sim \mu, r = R(s, a), s' \sim P(s, a)}[Y \,|\, X = x]. \qquad \text{(change of variable)} \\
&= h^\star(x).
\end{aligned}$$

Finally we back up the sample size from the generalization error bound. To guarantee $\|\widehat{\mathcal{T}}_\phi^\mu f - \mathcal{T}_\phi^\mu f\|_{2,\mu}^2 \leq \tilde{\epsilon}^2$, we set $\epsilon_0 = \tilde{\epsilon}^2/16V_{\max}$, and it suffices to have

$$\frac{8V_{\max}^2 \log \frac{(16V_{\max}^2/\tilde{\epsilon}^2)^{|\phi|}}{\delta}}{|D|} \leq \tilde{\epsilon}^2/2.$$

Solving for $|D|$ completes the proof. $\qquad \square$

## B.7. Proof of Lemma 10

We prove the concentration inequality in two separate cases depending on the magnitude of $\|f_0 - g\|_{2,\mu}$. In the first case when $\|f_0 - g\|_{2,\mu}$ is small, we may use Bernstein's inequality to obtain fast rate as $2\mathbb{V}_\mu[(f_0(s, a) - g(s, a))^2]$ is controlled by $\|f_0 - g\|_{2,\mu}$. In the second case, large $\|f_0 - g\|_{2,\mu}$ enables us to leverage the inequality of $|\|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu}| \leq \||f_0 - g\|_{2,D}^2 - \|f_0 - g\|_{2,\mu}^2|/\|f_0 - g\|_{2,\mu}$.

Before discussing these two cases separately, let's first establish some results as preparation. Let $\mathcal{G}_\phi'$ be an $\epsilon_0'$-cover of $\mathcal{G}_\phi$ under $\ell_\infty$, and $N := \mathcal{N}_\infty(\mathcal{G}_\phi, \epsilon_0')$ be the covering number. Thus, for any $g \in \mathcal{G}_\phi$, there exists $g' \in \mathcal{G}_\phi'$ such that $\|g - g'\|_\infty \leq \epsilon_0'$. We also let $n = |D|$.

We apply Bernstein's inequality with a union bound over $\mathcal{G}'_\phi$: w.p. $\geq 1 - \delta/2$, for any $g' \in \mathcal{G}'_\phi$,

$$\left| \|f_0 - g'\|_{2,D}^2 - \|f_0 - g'\|_{2,\mu}^2 \right|$$

$$= \left| \frac{1}{n} \sum_{(s,a,r,s') \in D} (f_0(s,a) - g'(s,a))^2 - \mathbb{E}_\mu \left[ (f_0(s,a) - g'(s,a))^2 \right] \right|$$

$$\leq \sqrt{\frac{2\mathbb{V}_\mu \left[ (f_0(s,a) - g'(s,a))^2 \right] \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n} \qquad \text{(Bernstein's inequality)}$$

$$\leq \sqrt{\frac{2V_{\max}^2 \|f_0 - g'\|_{2,\mu}^2 \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n}, \qquad (7)$$

where the lest inequality follows from the fact of $\mathbb{V}_\mu[(f_0(s,a) - g'(s,a))^2] \leq \mathbb{E}_\mu[(f_0 - g')^4] \leq \|(f_0 - g')^2\|_\infty \mathbb{E}_\mu[(f_0 - g')^2]$.

**Case 1:** $\|f_0 - g\|_{2,\mu}$ **is small**  Now, for any $g \in \mathcal{G}_\phi$, let $g' \in \mathcal{G}'_\phi$ satisfies $\|g - g'\|_\infty \leq \epsilon'_0$.

$$\left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right|$$
$$\leq \left| \|f_0 - g'\|_{2,D} - \|f_0 - g'\|_{2,\mu} \right| + 2\epsilon'_0$$
$$\leq \sqrt{\left| \|f_0 - g'\|_{2,D}^2 - \|f_0 - g'\|_{2,\mu}^2 \right|} + 2\epsilon'_0 \qquad (|a-b|^2 \leq |a^2 - b^2| \text{ for } a,b \geq 0)$$
$$\leq \sqrt{\sqrt{\frac{2V_{\max}^2 \|f_0 - g'\|_{2,\mu}^2 \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n}} + 2\epsilon'_0 \qquad \text{(Eq.(7))}$$
$$\leq \sqrt[4]{\frac{2V_{\max}^2 \|f_0 - g'\|_{2,\mu}^2 \ln \frac{4N}{\delta}}{n}} + \sqrt{\frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n}} + 2\epsilon'_0 \qquad (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \text{ for } a,b \geq 0)$$
$$\leq \sqrt[4]{\frac{2V_{\max}^2 \|f_0 - g\|_{2,\mu}^2 \ln \frac{4N}{\delta}}{n}} + \sqrt[4]{\frac{2V_{\max}^2 \epsilon_0'^2 \ln \frac{4N}{\delta}}{n}} + \sqrt{\frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n}} + 2\epsilon'_0.$$

**Case 2:** $\|f_0 - g\|_{2,\mu}$ **is large**  Similarly, for any $g \in \mathcal{G}_\phi$, let $g' \in \mathcal{G}'_\phi$ satisfies $\|g - g'\|_\infty \leq \epsilon'_0$. Then, as long as $\|f_0 - g\|_{2,\mu} \neq 0$,

$$\left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right| \leq \frac{\left| \|f_0 - g\|_{2,D}^2 - \|f_0 - g\|_{2,\mu}^2 \right|}{\|f_0 - g\|_{2,\mu}}$$

$$\leq \underbrace{\frac{\left| \|f_0 - g'\|_{2,D}^2 - \|f_0 - g'\|_{2,\mu}^2 \right|}{\|f_0 - g\|_{2,\mu}}}_{(I)} + \underbrace{\frac{4\|f_0 - g\|_{2,\mu}\epsilon'_0 + 2\left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right| \epsilon'_0 + 2\epsilon_0'^2}{\|f_0 - g\|_{2,\mu}}}_{(II)}.$$

The last line is obtained by the following argument:

$$\left| \left| \|f_0 - g\|_{2,D}^2 - \|f_0 - g\|_{2,\mu}^2 \right| - \left| \|f_0 - g'\|_{2,D}^2 - \|f_0 - g'\|_{2,\mu}^2 \right| \right|$$
$$\leq \left| \left( \|f_0 - g\|_{2,D}^2 - \|f_0 - g'\|_{2,D}^2 \right) - \left( \|f_0 - g\|_{2,\mu}^2 - \|f_0 - g'\|_{2,\mu}^2 \right) \right|$$
$$\leq \left| \|f_0 - g\|_{2,D}^2 - \|f_0 - g'\|_{2,D}^2 \right| + \left| \|f_0 - g\|_{2,\mu}^2 - \|f_0 - g'\|_{2,\mu}^2 \right|$$
$$= \left| \|f_0 - g\|_{2,D} - \|f_0 - g'\|_{2,D} \right| \cdot \left| \|f_0 - g\|_{2,D} + \|f_0 - g'\|_{2,D} \right|$$
$$\quad + \left| \|f_0 - g\|_{2,\mu} - \|f_0 - g'\|_{2,\mu} \right| \cdot \left| \|f_0 - g\|_{2,\mu} + \|f_0 - g'\|_{2,\mu} \right|$$
$$\leq \epsilon'_0 \cdot (2\|f_0 - g\|_{2,D} + \epsilon'_0) + \epsilon'_0 \cdot (2\|f_0 - g\|_{2,\mu} + \epsilon'_0)$$
$$\leq 4\|f_0 - g\|_{2,\mu}\epsilon'_0 + 2\left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right| \epsilon'_0 + 2\epsilon_0'^2,$$

where all the inequalities follow from the triangle inequality and the fact of $\|g - g'\|_\infty \leq \epsilon'_0$.

We now analyze the two terms above separately.

$$\text{(I)} \le \frac{1}{\|f_0 - g\|_{2,\mu}} \left( \sqrt{\frac{2V_{\max}^2 \|f_0 - g'\|_{2,\mu}^2 \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n} \right) \qquad \text{(by Eq.(7))}$$

$$\le \frac{1}{\|f_0 - g\|_{2,\mu}} \left( (\|f_0 - g\|_{2,\mu} + \epsilon_0') \sqrt{\frac{2V_{\max}^2 \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n} \right)$$

$$\le \sqrt{\frac{2V_{\max}^2 \ln \frac{4N}{\delta}}{n}} + \frac{1}{\|f_0 - g\|_{2,\mu}} \left( \sqrt{\frac{2\epsilon_0'^2 V_{\max}^2 \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n} \right). \qquad (8)$$

$$\text{(II)} = 4\epsilon_0' + \frac{2\epsilon_0'}{\|f_0 - g\|_{2,\mu}} \left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right| + \frac{2\epsilon_0'^2}{\|f_0 - g\|_{2,\mu}}. \qquad (9)$$

**Combine them all** We now unify those two cases above. We first set $\epsilon_0' = \epsilon_1/20$, and $N = \mathcal{N}_\infty(\mathcal{G}_\phi, \epsilon_0') \le (V_{\max}/\epsilon_0')^{|\phi|}$.
When $\|f_0 - g\|_{2,\mu} < 4\epsilon_0'$, we apply the first case and obtain

$$\left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right|$$

$$\le \sqrt[4]{\frac{2V_{\max}^2 \|f_0 - g\|_{2,\mu}^2 \ln \frac{4N}{\delta}}{n}} + \sqrt[4]{\frac{2V_{\max}^2 \epsilon_0'^2 \ln \frac{4N}{\delta}}{n}} + \sqrt{\frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n}} + 2\epsilon_0'$$

$$\le \left( \sqrt[4]{\frac{2}{25}} + \sqrt[4]{\frac{1}{200}} \right) \sqrt[4]{\frac{V_{\max}^2 \epsilon_1^2 \ln \frac{4N}{\delta}}{n}} + \sqrt{\frac{V_{\max}^2 \ln \frac{4N}{\delta}}{3n}} + \frac{\epsilon_1}{10}$$

$$\le \left( \sqrt[4]{\frac{2}{25}} + \sqrt[4]{\frac{1}{200}} \right) \sqrt[4]{\frac{V_{\max}^2 \epsilon_1^2 |\phi| \ln \frac{80V_{\max}}{\epsilon_1 \delta}}{n}} + \sqrt{\frac{V_{\max}^2 |\phi| \ln \frac{80V_{\max}}{\epsilon_1 \delta}}{3n}} + \frac{\epsilon_1}{10}.$$

Solving

$$\left( \sqrt[4]{\frac{2}{25}} + \sqrt[4]{\frac{1}{200}} \right) \sqrt[4]{\frac{V_{\max}^2 \epsilon_1^2 |\phi| \ln \frac{80V_{\max}}{\epsilon_1 \delta}}{n}} + \sqrt{\frac{V_{\max}^2 |\phi| \ln \frac{80V_{\max}}{\epsilon_1 \delta}}{3n}} + \frac{\epsilon_1}{10} \le \epsilon_1$$

implies that it suffices to set

$$|D| = n \ge \frac{16V_{\max}^2 |\phi| \ln \frac{80V_{\max}}{\epsilon_1 \delta}}{\epsilon_1^2},$$

for the case of $\|f_0 - g\|_{2,\mu} < 4\epsilon_0' = \frac{1}{5}\epsilon_1$.

If $\|f_0 - g\|_{2,\mu} \ge 4\epsilon_0'$, we use the second case. The term (I) in Eq.(8) is

$$\text{(I)} \le \frac{5}{4} \sqrt{\frac{2V_{\max}^2 \ln \frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln \frac{4N}{\delta}}{12\epsilon_0' n},$$

and the term (II) in Eq.(9) is

$$\text{(II)} \le \frac{9}{2}\epsilon_0' + \frac{1}{2} \left| \|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu} \right|.$$

Since $|\|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu}| \leq$ (I) + (II), we reorder the terms and obtain

$$\left|\|f_0 - g\|_{2,D} - \|f_0 - g\|_{2,\mu}\right| \leq \frac{5}{2}\sqrt{\frac{2V_{\max}^2 \ln\frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln\frac{4N}{\delta}}{6\epsilon_0 n} + 9\epsilon_0.$$

We still set $\epsilon_0' = \epsilon_1/20$. Thus, solving

$$\frac{5}{2}\sqrt{\frac{2V_{\max}^2 \ln\frac{4N}{\delta}}{n}} + \frac{V_{\max}^2 \ln\frac{4N}{\delta}}{6\epsilon_0' n} + 9\epsilon_0' = \frac{5}{2}\sqrt{\frac{2V_{\max}^2 |\phi| \ln\frac{80V_{\max}}{\epsilon_1\delta}}{n}} + \frac{10V_{\max}^2 |\phi| \ln\frac{80V_{\max}}{\epsilon_1\delta}}{3\epsilon_1 n} + \frac{9\epsilon_1}{20} \leq \epsilon_1$$

provides us with the sufficient sample size $|D|$,

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

Choosing the greater one between the required sample sizes for $\|f_0 - g\|_{2,\mu} < 4\epsilon_0' = \epsilon_1/5$ and $\|f_0 - g\|_{2,\mu} \geq 4\epsilon_0' = \epsilon_1/5$ completes the proof. □

### B.8. Proof of Proposition 5

Since $|D|$ in the proposition statement is chosen to satisfy the sample-size requirements in Lemmas 9 and 10, the statement of each lemma holds with probability at least $1 - \delta/2$, and by union bound they hold simultaneously w.p. $\geq 1 - \delta$.

**Bounding $\|Q^\star - f_0\|$ using $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ (Eq.(3)):**
Define $\pi_{f,f'}$ as the policy $s \mapsto \arg\max_a \max\{f(s,a), f'(s,a)\}$. Consider any $\nu$ such that $\|\nu/\mu\|_\infty \leq C$,

$$\begin{aligned}
\|Q^\star - f_0\|_{2,\nu} &\leq \|Q^\star - \mathcal{T}_\phi^\mu Q^\star\|_{2,\nu} + \|\mathcal{T}_\phi^\mu Q^\star - \mathcal{T}_\phi^\mu f_0\|_{2,\nu} + \|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\nu} \\
&\leq 2\epsilon_\phi + \gamma\|Q^\star - f_0\|_{2,P_\phi(\nu)\times\pi_{f,Q^\star}} + \sqrt{C}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}.
\end{aligned} \tag{10}$$

In Eq.(10), the first term follows from Lemma 7 and that $\|\cdot\|_{2,\nu} \leq \|\cdot\|_\infty$, the second from Chen & Jiang (2019, Lemmas 14 and 15), and the third from Chen & Jiang (2019, Lemma 12).

According to Lemma 6, $P_\phi(\nu) \times \pi_{\hat{f},Q^\star}$ also satisfies $\|(\cdot)/\mu\|_\infty \leq C$, so it can be viewed as one of those $\nu$'s we started with on the LHS, allowing us to expand the inequality indefinitely. Alternatively, we have

$$\begin{aligned}
&\sup_{\nu:\|\nu/\mu\|_\infty\leq C} \|Q^\star - f_0\|_{2,\nu} \\
&\leq \gamma \sup_{\nu:\|\nu/\mu\|_\infty\leq C} \left(\|Q^\star - f_0\|_{2,P_\phi(\nu)\times\pi_{f_0,Q^\star}}\right) + 2\epsilon_\phi + \sqrt{C}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu} \\
&\leq \gamma \sup_{\nu:\|\nu/\mu\|_\infty\leq C} (\|Q^\star - f_0\|_{2,\nu}) + 2\epsilon_\phi + \sqrt{C}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}. && \text{(Lemma 6)}
\end{aligned}$$

So for any $\nu$ such that $\|\nu/\mu\|_\infty \leq C$, $\|Q^\star - \hat{f}\|_{2,\nu} \leq \frac{2\epsilon_\phi + \sqrt{C}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}}{1-\gamma}$.

It then remains to bound $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$:

$$\begin{aligned}
\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu} &\leq \|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,\mu} + \|\widehat{\mathcal{T}}_\phi^\mu f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu} \\
&\leq \|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,\mu} + \tilde{\epsilon} && \text{(Lemma 9)} \\
&\leq \|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D} + \epsilon_1 + \tilde{\epsilon}. && \text{(Lemma 10)}
\end{aligned}$$

**Bounding $\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D}$ using $\|Q^\star - f_0\|$ (Eq.(4)):**

$$
\begin{aligned}
\|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,D} &\le \|f_0 - \widehat{\mathcal{T}}_\phi^\mu f_0\|_{2,\mu} + \epsilon_1 && \text{(Lemma 10)} \\
&\le \|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu} + \tilde{\epsilon} + \epsilon_1 && \text{(Lemma 9)} \\
&\le \|f_0 - \mathcal{T}_\phi^\mu f_0\|_\infty + \tilde{\epsilon} + \epsilon_1 && (\|\cdot\|_{2,\mu} \le \|\cdot\|_\infty) \\
&\le \|f_0 - Q^\star\|_\infty + \|\mathcal{T}_\phi^\mu Q^\star - \mathcal{T}_\phi^\mu f_0\|_\infty + \|Q^\star - \mathcal{T}_\phi^\mu Q^\star\|_\infty + \tilde{\epsilon} + \epsilon_1 \\
&\le (1+\gamma)\|f_0 - Q^\star\|_\infty + 2\epsilon_\phi + \tilde{\epsilon} + \epsilon_1.
\end{aligned}
$$
$$(\gamma\text{-contraction of } \mathcal{T}_\phi^\mu \text{ under } \ell_\infty \text{ (Lemma 3) and Lemma 7)} \quad \square$$

## B.9. Proof of Lemma 11

For the first claim, we may write $\phi(s,a) = (\bar{f}(s,a), \bar{f}'(s,a))$, and the number of equivalent classes induced by $\phi$ is at most the product of the cardinalities of the codomains of $\bar{f}$ and $\bar{f}'$, so the result follows.

For the second claim, recall that $\epsilon_\phi = \epsilon_{\mathcal{G}_\phi} = \min_{g \in \mathcal{G}_\phi} \|g - Q^\star\|_\infty$, so

$$
\begin{aligned}
\epsilon_\phi &\le \|\bar{f}^\star - Q^\star\|_\infty && (f^\star \in \{f, f'\} \text{ and } \bar{f}^\star \in \mathcal{G}_\phi) \\
&\le \|f^\star - Q^\star\|_\infty + \|\bar{f}^\star - f^\star\|_\infty \le \epsilon_{\mathcal{F}} + \epsilon_{\text{dct}}. && \square
\end{aligned}
$$

# C. Obstacles in Relaxing Assumption 1

In this section we discuss the obstacles in relaxing Assumption 1. Before we start, we emphasize that the difficulties have nothing to do with the tournament procedure, **and are entirely about learning $Q^\star$ with a realizable state-action aggregation** $\phi$—a problem so standard, that the difficulties we find may have broader implications beyond the scope of this work; see Appendix C.1 for discussions on the relevance of our findings to existing RL algorithms.

**Negative Result: Contrived $\mu$** We present our first counterexample in Figure 1, where Assumption 1 is violated but Assumption 2 is satisfied due to missing data in state-action pairs unreachable from the initial state $s_0$. A state-action aggregation $\phi$, which is guaranteed to express $Q^\star$, results in a projected Bellman operator $\mathcal{T}_\phi^\mu$ that has multiple fixed points other than $Q^\star$ even when sample size goes to infinity, and many such fixed points produce suboptimal policies. Therefore, $\|f - \mathcal{T}_\phi^\mu f\|_{2,\mu}$, which is the surrogate loss that plays a central role in our algorithm, cannot control the performance of $\pi_f$ in this setting; see Appendix C.1 for details.

**Negative Result: Admissible $\mu$** An issue with Figure 1 is that its $\mu$ cannot be admitted in the MDP (i.e., generated by some behavior policy from $d_0$), and assuming admissible $\mu$ (which is reasonable) excludes this counterexample. While this may look promising, here we show that our analysis still faces substantial obstacles if we replace Assumption 1 with Assumption 2 plus admissible $\mu$. Below we explain in more details.

As Section 5.2.1 has alluded to, the error propagates according to the dynamics of $M_\phi$ instead of $M$ in our analysis, so the purpose of Assumption 1 is really to guarantee the following type of assumption:[13]

**Assumption 4.** There exists $C < \infty$, such that for any $\phi$, we have $\|\nu/\mu\|_\infty \le C$ for any $\nu$ that is admissible in $M_\phi$.

Unfortunately, via a carefully constructed example, we show in Figure 2 that Assumption 4 cannot be implied by Assumption 2 plus admissible $\mu$. In particular, even when Assumption 2 is satisfied with a constant $C$ and $\mu$ is admissible, we can use the aggregation to "leak" probabilities from easy-to-reach states in $\mu$ to hard-to-reach states gradually over time steps, causing an exponential blow-up of $\|\nu/\mu\|_\infty$. See Appendix C.2 for details.

**Positive Result: On-policy Data** Despite the discouraging counterexamples, we show a slightly positive result, implying that Assumption 4 *could* be much weaker than Assumption 1, leaving the possibility of something weaker than Assumption 1 but more natural and interpretable than Assumption 4. In particular, we show a scenario where Assumptions 2 and 4 can be

---

[13]The actual assumption we need is slightly more complicated, that we need $\|\nu/\mu\|_\infty \le C$ for any $\nu$ admissible in $M_\phi$ but using any other admissible $\nu'$ from $M$ as the initial distribution. This complication, however, does not affect our counterexample. Neither does it affect the next positive result.
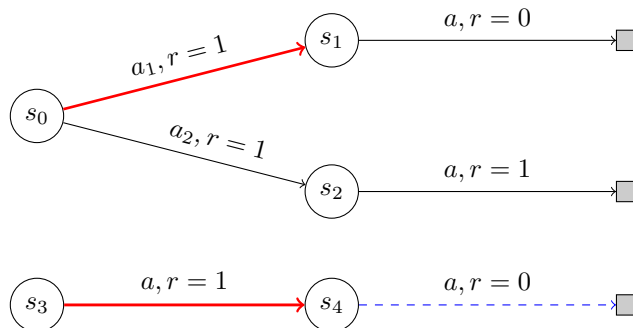
Figure 1: MDP construction for Section 7.2; see Appendix C.1 for details. $s_0$ is the deterministic initial state. The gray squares represent known absorbing states. Actions $a_1, a_2$ have the same effects (and abbreviated as "$a$") in all states except in $s_0$. The **red thick** arrows for $(s_0, a_1)$ and $(s_3, a)$ indicate that these state-action pairs are aggregated together (they share the same $Q^\star$ value). The blue dashed arrow for $(s_4, a)$ indicates that data is missing for this state-action pair, which violates Assumption 1; Assumption 2 still holds since no policy can visit $s_3$ or $s_4$ from $s_0$. The missing data leads to high uncertainty in $Q^\star(s_3, a)$, and such uncertainty propagates to $Q^\star(s_0, a_1)$ and causes the failure of learning.

satisfied with a small $C$, yet Assumption 1 may be violated badly, implying the looseness and unnecessity of Assumption 1 for our analysis.

Consider the uncontrolled case where there is only one action, and we may treat $P$ as an $|\mathcal{S}| \times |\mathcal{S}|$ transition matrix. We consider the "on-policy" case, where $d_0 = \mu$ is an invariant distribution w.r.t. $P$, i.e., $\mu^\top P = \mu^\top$. Since an invariant distribution always exists, this does not impose any restriction on $P$, so we can make the $C$ in Assumption 1 very large: for example, when $P$ is identity, $C$ must be as large as $|\mathcal{S}|$ to satisfy Assumption 1. On the other hand, Assumption 2 is trivially satisfied with $C = 1$, as $\mu$ is the only admissible distribution in $M$. Perhaps surprisingly, this is also true for Assumption 4: regardless of $\phi$, we have $\mu^\top P_\phi = \mu^\top P = \mu$, because $P_\phi$ is defined by averaging the dynamics of $P$ with weights proportional to $\mu$, and this averaging step can be ignored when the incoming distribution is $\mu$ itself.

### C.1. Details of Figure 1

We construct an MDP, a data distribution $\mu$, and a realizable function class $\mathcal{F}$ with $|\mathcal{F}| = 2$, such that (1) Assumption 1 is violated; (2) Assumption 2 is satisfied; (3) Algorithm 1 may output a suboptimal policy even with infinite data.

See Figure 1 for an illustration of the MDP, where the transition dynamics and the rewards are deterministic. Let $\gamma = 0.9$, and $s_0$ be the deterministic initial state.

Let $\mu$ be uniform over all state-action pairs other than $(s_4, a)$,[14] which violates Assumption 1.[15] However, since no policy can visit $s_3$ or $s_4$ from the starting state $s_0$, lacking data in $s_4$ does not affect the validity of Assumption 2. It remains to specify $\mathcal{F}$ and show that our algorithm fails.

Our $\mathcal{F}$ consists of two functions, $Q^\star$ and $Q$. We specify them by writing down their values on $(s_0, a_1), (s_0, a_2), (s_1, a), (s_2, a), (s_3, a), (s_4, a)$ as a vector: $Q^\star = (1, 1.9, 0, 1, 1, 0)$ and $Q = (7, 1.9, 0, 1, 7, 10)$. The red and blue colors correspond to the color schemes in Figure 1 to facilitate understanding.

When we run Algorithm 1, the only nontrivial aggregation $\phi$ performs is grouping together $(s_0, a_1)$ and $(s_3, a)$; $(s_0, a_2)$ and every other state-action pair are kept in their own equivalence classes, respectively. ($a_1, a_2$ under the same state are by default aggregated except in $s_0$.)

Now that the construction is complete, we can verify that our loss $\|f - \mathcal{T}_\phi^\mu f\|$ is zero for both $f = Q^\star$ and $f = Q$. Therefore,

---

[14]The probability assigned to $(s_3, a)$ is twice as much as that to $(s_0, a_1)$ by $\mu$, because the former is the abbreviation of two state-action pairs. This detail is of minor importance, and $\mu$ can be changed to many other distributions, as long as the later values of $Q$ are set in a way consistent with $\mu$.

[15]The fact that $\mu(s, a) > 0 \ \forall s, a$ is violated is of minor concern here: we can add exponentially small probabilities to $(s_4, a)$ in $\mu$, and with a polynomially large $D$, the non-uniqueness of the fixed point of $\widehat{\mathcal{T}}_\phi^\mu$ still persists. What is really important is that no finite $C_\mathcal{S}$ satisfies $P(s'|s, a)/\mu(s') \leq C_\mathcal{S}$.
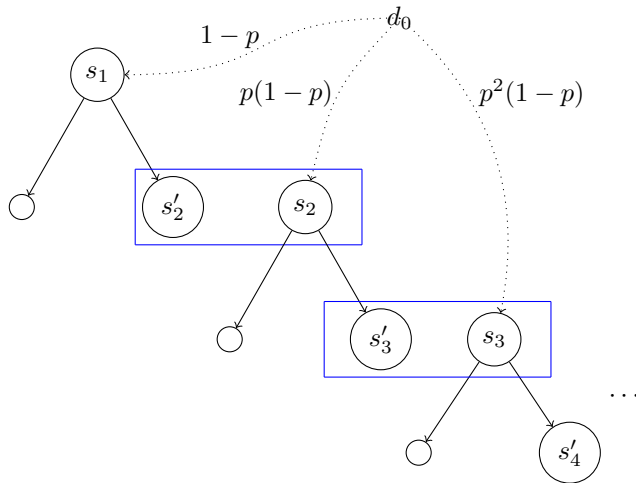
Figure 2: The example in Figure 1 requires that $\mu$ visits unreachable states. While assuming that $\mu$ can be naturally generated in the MDP using a behavior policy ("admissible") would exclude the pathology in Figure 1, here we show that in a more carefully constructed example, dynamics of $M_\phi$ allow us to visit a state exponentially more likely than in $\mu$, while Assumption 2 is satisfied with a constant $C$. Roughly speaking, $\mu(s_t) \propto p^t$ but it is possible to produce a distribution in $M_\phi$ that visits $s_t$ w.p.$\propto (2p)^t$ for $p \ll 1$; see Appendix C.2.

the algorithm may choose to output the greedy policy of either function, but the greedy policy of $Q$ is suboptimal since it chooses $a_1$ in $s_0$.

**Broader Implications** The non-uniqueness of the fixed point of $\mathcal{T}_\phi^\mu$ not only affects our algorithm, but also affects many popular algorithms when they are applied with the piecewise constant class induced by $\phi$, as *any* fixed point of $\mathcal{T}_\phi^\mu$ is a valid output of these algorithms. Such algorithms include iterative algorithms such as FQI and (both $\ell_1$ and $\ell_2$) Bellman residual minimization-style algorithms (Xie & Jiang, 2020). Moreover, related algorithms that do not directly approximate $Q^\star$ but instead perform policy iteration (e.g., LSPI; Lagoudakis & Parr, 2003) are also subject to this counterexample, as the $Q^\pi$ learned at the policy evaluation step is equal to $Q^\star$ regardless of the policy in this MDP.[16]

This phenomenon is particularly interesting when we notice that, everything will work fine if we remove the data on $(s_3, a)$: the algorithm will still have high uncertainty in the values of $s_3$ and $s_4$, but such uncertainty does not incorrectly propagate to $s_0$ and hence does not affect our ability to choose the optimal action there. Therefore, the pathological behavior is due to having data with *more* coverage than necessary, which may be surprising. To our best knowledge, this pathology—which affects a wide range of batch RL algorithms—is documented for the first time. It will be interesting to see if we can obtain a deeper understanding of this issue and possibly circumvent it.

### C.2. Counterexample Against Admissible $\mu$

One weakness of the counterexample in Figure 1 is that $\mu$ cannot be generated by a behavior policy starting from $d_0$, as such *admissible distributions* never visit $s_3$. Therefore, we can exclude the pathology by assuming that $\mu$ is (a mixture) of admissible distributions, on top of Assumption 2.

**Assumption 5** ($\mu$ is admissible). In addition to Assumption 2, assume that $\mu$ is a mixture of $d_t^\pi$ for a set of $(\pi, t)$ pairs, where $\pi$ may be nonstationary and/or stochastic.

This seemingly mild additional assumption leads to some powerful corollaries. For example, any distribution induced from $\mu$ (say, with policy $\pi'$ in $t'$ steps) as the initial distribution (instead of $d_0$) will still be covered by $\mu$, since the distribution is essentially a mixture of $d_{\pi \circ \pi'}^{t+t'}$. Furthermore, creating a situation like Figure 1 becomes very difficult (see below for detailed reasons); in fact, we believe that it is impossible to induce hardness with a constant-depth construction as in Figure 1.

Despite the power of Assumption 5, we show below that our analysis still faces substantial obstacles even under Assumption 5.

---
[16]This is because the only action choice is in the initial state $s_0$.

In particular, our proof around Eq.(10) requires that distribution induced in $M_\phi$ should also be well covered by $\mu$ (i.e., Assumption 4), and this is guaranteed by Assumption 1 via Lemma 6. If we replace Assumption 1 with Assumption 5, however, below we show in a counterexample that $M_\phi$ can induce a distribution that visits a state exponentially more likely compared to its likelihood in $\mu$, thus breaking Assumption 4.

**Counterexample** See Figure 2. The true MDP $M$ is essentially a 2-armed contextual bandit, where the initial state (or context) distribution is $d_0(s_t) = (1-p)p^{t-1}$ for $t$ between 1 and a sufficiently large integer $N$ (and the rest probability goes to $s_{N+1}$; we will not need it). $0 < p < 1$ is a parameter to be set later. For each $s_t$, we will call the two actions $L$ (for "left") and $R$ ("right"), respectively. All states other than $\{s_t\}$ are absorbing states. During data collection, the learner randomly starts in some $s_t$ according to $d_0$, and take actions uniformly at random. This guarantees that $C = 2$ for Assumption 2. Note that we do not specify the rewards as our goal is only to show that a distribution with large density ratio against $\mu$ can be induced in $M_\phi$, and not to directly show the failure of the algorithm.

**Aggregation and Dynamics of $M_\phi$** The dynamics of $M_\phi$ (Definition 4) can be equivalently described as the following: given $(s, a)$, the next-state $s'$ is generated in two steps,

1. Re-draw $(\tilde{s}, \tilde{a})$ from $\{(\tilde{s}, \tilde{a}) : \phi(\tilde{s}, \tilde{a}) = \phi(s, a)\}$ with probability proportional to $\mu$.

2. Sample $s' \sim P(\tilde{s}, \tilde{a})$.

In Figure 2, we aggregate $s'_t$ and $s_t$ together for every $t$, and the "re-drawing" step happens after the agent lands in $s'_t$. That is, before the next time step, there is some probability that it will teleport to $s_t$. Strictly speaking we are aggregating states instead of state-action pairs, but the effects can be reproduced by e.g., adding a dummy state with only 1 action above each $s_t$, and aggregating this state-action pair with $(s_{t-1}, R)$. We opt for a slightly different mechanism for simplicity of the construction.

**Calculation** We will show that with the effect of aggregation, we can induce a distribution whose density ratio against $\mu$ is exponentially large, using the policy that always takes action "$R$". Let $P(s_t)$ be the probability of visiting $s_t$ at time step $t-1$ by this policy, and let $\mu(s_t)$ be the mass of $s_t$ in $\mu$. Our goal is to calculate $P(s_t)/\mu(s_t)$ and show that it is exponential in $t$. First, $\mu(s_t) = p^{t-1}(1-p)/2$. The division by 2 is because $\mu$ contains data from two different time steps ($\{s_t\}$ at time step 0, and $\{s_{t'}\}$ and their sibling states at time step 1). Note that the data is collected in the true MDP $M$ and the aggregation plays no role here.

Now we calculate $P(s_t)$. The only probability path of visiting $s_t$ at time step $t-1$ is $s_1 \to s'_2 \to s_2 \to s'_3 \to s_3 \to \ldots \to s'_t \to s_t$. Along this path, $P(s_1) = 1-p$, and $s_{t-1} \to s'_t$ is deterministic, so we focus on the probability of $s'_t \to s_t$.

Recall from the above that whenever at $s'_t$, we will redraw a state from $\{s'_t, s_t\}$ according to their probabilities in $\mu$. Therefore,

$$P(s'_t \to s_t) = \frac{\mu(s_t)}{\mu(s'_t) + \mu(s_t)} = \frac{p^{t-1}(1-p)/2}{p^{t-2}(1-p)/4 + p^{t-1}(1-p)/2} = \frac{2p}{1 + 2p},$$

where $\mu(s'_t)$ is calculated based on the fact that the data collection policy is uniformly random. This gives

$$P(s_t) = (1-p)\left(\frac{2p}{1+2p}\right)^{t-1}, \quad \frac{P(s_t)}{\mu(s_t)} = \frac{1}{2}\left(\frac{2}{1+2p}\right)^{t-1}.$$

Therefore, when $p < 1/2$, $P(s_t)/\mu(s_t)$ will be exponential in $t$.

## D. What If Assumption 1 is Violated?

When Assumption 1 is violated, the guarantees in Theorem 2 does not hold in general. However, this does not mean that the algorithm is useless or there is nothing we can do in this case. Below we discuss a few actionable items and briefly sketch the theoretical analyses that justifies them.

## D.1. Diagnosing the Uniqueness of $\widehat{\mathcal{T}}_\phi^\mu$'s Fixed Point

As Section 7.2 has shown, one possible consequence of not having Assumption 1 is that the fixed point of $\widehat{\mathcal{T}}_\phi^\mu$ may be non-unique. This suggests a diagnostic procedure that checks if such pathology occurs: let $\widehat{\mathcal{G}_\phi} := \{g \in \mathcal{G}_\phi : \|g - \widehat{\mathcal{T}}_\phi^\mu g\| \le \epsilon'\}$ be the set of approximate fixed points of $\widehat{\mathcal{T}}_\phi^\mu$ where $\epsilon'$ is some small threshold. Then, we may compute $\max_{g,g' \in \widehat{\mathcal{G}_\phi}} \|g - g'\|_{2,D}$ as a statistic for the diagnosis. If Assumption 1 is satisfied, such a maximum distance should be small as all functions in $\widehat{\mathcal{G}_\phi}$ should be close to the fixed point of $\mathcal{T}_\phi^\mu$ under $\|\cdot\|_{2,\mu}$,[17] and it is important to note that such a claim does not depend on $f^\star \in \{f, f'\}$ (see e.g., Lemma 3). On the other hand, if the maximum distance within $\widehat{\mathcal{G}_\phi}$ is observed to be small when we actually run the algorithm, we can rest assured that $\|\hat{f} - Q^\star\|_{2,\mu}$ is small (assuming $\max_{f' \in \mathcal{F}} \mathcal{E}(\hat{f}; f')$ is small), and we only need Assumption 2 to further guarantee the near-optimality of $\pi_{\hat{f}}$, regardless of whether Assumption 1 holds or not.

As an example, consider the counterexample in Figure 1 and Appendix C.1, where both $Q^\star$ and $Q$ are fixed points of $\mathcal{T}_\phi^\mu$ and $\|Q^\star - Q\|_{2,\mu}$ is large. As Section 7.2 suggests, the pathology goes away if we remove the data from $(s_3, a)$. Note that $\mathcal{T}_\phi^\mu$ still has many fixed points (the value of $(s_4, a)$ can be anything), but their distance from each other under $\|\cdot\|_{2,\mu}$ is always 0 because $\mu$ is only supported on $s_0, s_1, s_2$, and all these fixed points induce an optimal policy from $s_0$.

## D.2. Tweaking $\phi$

Our main analysis treats the discretization step (Line 7) very casually. However, the structure of $\phi$ plays an important role in error propagation, and small changes in discretization can produce significantly different $\phi$'s, so one may want to search for a favorable $\phi$ among all possibilities. What should be the guideline for such a search?

We propose searching for $\phi$ that maximizes the least number of data points that fall into a group of state-action pairs.[18] If each group of state-action pairs receives enough data, $\mu_{\min}^\phi := \sum_{\tilde{s},\tilde{a}:\phi(\tilde{s},\tilde{a})=\phi(s,a)} \mu(\tilde{s}, \tilde{a})$ will be lower bounded away from 0 for any $(s, a)$, which guarantees well-controlled error propagation. To see this, recall that Eq.(10) is the key step that characterizes distribution shift error propagation, and we need Assumption 1 and Lemma 6 to upper bound $\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\nu}$ with $\sqrt{C}\|f_0 - \mathcal{T}_\phi^\mu f_0\|_{2,\mu}$ for any $\nu$ produced by the dynamics of $M_\phi$. If $\mu_{\min}^\phi$ is large, however, we do not need to reason about the properties of $\nu$. Instead, the fact that both $f_0$ and $\mathcal{T}_\phi^\mu f_0$ are (approximately) in $\mathcal{G}_\phi$ (recall that $\phi$ is produced by $f$ and $f'$ in Line 7 with $f_0 = f$) and $\mu_{\min}^\phi$ is lower bounded already allows us to state the desired bound with $C$ replaced by $1/\mu_{\min}^\phi$, up to small additive errors. Therefore, if we can always guarantee a lower-bounded $\mu_{\min}^\phi$ (via the sample size in each group as a surrogate) whenever in Line 7, the algorithm enjoys its sample complexity guarantee *without any concentrability assumptions*, which is very appealing.

Of course, there is no guarantee that we can find such a well-behaved $\phi$ for a single pair of $f, f'$, let alone the $|\mathcal{F}|^2$ pairs that all need to be handled. Therefore, this suggestion is more suitable for the model-selection scenario where $|\mathcal{F}|$ is small (Section 7.1). To produce a rich set of possible $\phi$'s, one can consider various designs of the discretization grid (see Footnote 4): changing the offset of the grid, using a non-regular grid, using soft aggregations instead of hard ones, or even setting $\epsilon_{\text{dct}}$ to be slightly greater than intended. Whether it is easy to find a well-behaved $\phi$ is more of an empirical question, and we leave further investigation to future work.

# E. Defining $\epsilon_\mathcal{F}$ in Weighted $\ell_2$ Norm

It is possible to define $\epsilon_\mathcal{F}$ as $\inf_{f \in \mathcal{F}} \|f - Q^\star\|_{2,\mu}$, and still prove a polynomial sample complexity result. However, making this change leads to a suboptimal dependence on $C$ if we still follow the same proof structure. Below we briefly explain the challenges.

In Section 5 we characterize $\phi$ with only two parameters, $|\phi|$ and $\epsilon_\phi$, and $\epsilon_\phi$ is later bounded in Lemma 11 by $\epsilon_\mathcal{F} + \epsilon_{\text{dct}}$. When we define $\epsilon_\mathcal{F}$ as $\inf_{f \in \mathcal{F}} \|f - Q^\star\|_{2,\mu}$, the $\epsilon_\mathcal{F}$ part of $\epsilon_\phi$ is subject to the $\sqrt{C}$ penalty when there is distribution shift,

---

[17] In an early stage of the project, our algorithm actually checks whether $f \in \widehat{\mathcal{G}_\phi}$ instead of the current Line 8. This alternative algorithm also enjoys polynomial sample complexity.

[18] One may recall that this is also a preferred situation in the analyses of state abstractions (Paduraru et al., 2008; Jiang et al., 2015; Jiang, 2018), as $\ell_\infty$ quantities are only statistically stable when each abstract state receives enough data. However, this is not the primary reason for our recommendation: our statistic $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,D}$ is stable and generalizes well regardless of the structure of $\phi$, thanks to the use of weighted $\ell_2$ norm instead of $\ell_\infty$ norm; see Section 5.2.3.

but the $\epsilon_{\text{dct}}$ part of $\epsilon_\phi$ is an $\ell_\infty$ discretization error and intuitively should not be affected by distribution shift. However, since we bundle $\epsilon_{\mathcal{F}}$ and $\epsilon_{\text{dct}}$ together in $\epsilon_\phi$, $\epsilon_{\text{dct}}$ will suffer an additional $\sqrt{C}$ penalty compared to our current analysis, leading to additional dependence on $C$ in the sample complexity.

One possible solution is to define $\epsilon_{\mathcal{F}}$ as $\inf_{f \in \mathcal{F}} \sup_{\nu:\|\nu/\mu\|_\infty \leq C} \|f - Q^\star\|_{2,\nu}$, i.e., to capture distribution shift inside the definition of $\epsilon_{\mathcal{F}}$, so that $\epsilon_{\mathcal{F}}$ is still effectively (a soft version of) $\ell_\infty$ error. The same change needs to be made in the definition of $\epsilon_\phi$. The difficulty with this approach is that such a "soft $\ell_\infty$" $\epsilon_{\mathcal{F}}$ does not play well with Lemma 7: to avoid $\epsilon_\phi$ being amplified by $\sqrt{C}$, we need to also replace the LHS of Lemma 7 with $\sup_\nu \|Q^\star - \mathcal{T}_\phi^\mu Q^\star\|_{2,\nu}$. However, it is unclear whether this quantity can be bounded by $\epsilon_\phi$ without paying $\sqrt{C}$, as $\mathcal{T}_\phi^\mu Q^\star = \arg\min_{g \in \mathcal{G}_\phi} \|g - \mathcal{T}Q^\star\|_{2,\mu} = \arg\min_{g \in \mathcal{G}_\phi} \|g - Q^\star\|_{2,\mu}$ is the best approximation of $Q^\star$ in $\mathcal{G}_\phi$ under weighting $\mu$, but ideally we want the best approximation under worst-case $\nu$ to avoid $\sqrt{C}$ penalty. We leave the resolution of this technical issue to future work.

## F. Comparison to OPE in Model Selection

Section 7.1 discussed the application of BVFT to model selection. Another common approach to model selection is to estimate $J(\pi)$ for each candidate $\pi$ via off-policy evaluation (OPE). Unfortunately, unbiased OPE with importance sampling (IS) (Precup et al., 2000) incurs exponential variance in horizon when the behavior policy is significantly different from the ones being evaluated (Jiang & Li, 2016), and recent marginalized IS (MIS) methods that overcome such a "curse of horizon" require some nontrivial function-approximation assumptions. For example, even if one has a function class $\mathcal{Q}$ that realizes $Q^\pi$ for *every* $\pi$ being evaluated, the state-or-the-art approaches only provide an interval that contains $J(\pi)$ without tightness guarantees (Jiang & Huang, 2020; Feng et al., 2020), and tightness requires further assumptions on realizing marginalized importance weights (e.g., Liu et al., 2018; Uehara et al., 2020). On a related note, if a rich $\mathcal{Q}$ is used to better satisfy these additional assumptions, one has to reserve a large amount of data as holdout dataset due to the statistical complexity of $\mathcal{Q}$. In comparison, BVFT only uses function classes of a well-controlled worst-case complexity $O(1/\epsilon^2)$.

Despite the additional assumptions, OPE has its own advantages: OPE directly estimates $J(\pi)$ instead of going through $\|f - Q^\star\|$ as a surrogate, and as a consequence, it removes the assumption that the base algorithms need to approximate $Q^\star$ and hence enjoys wider applicability. The information about $J(\pi)$ can also be valuable in certain application scenarios, which cannot be obtained by our approach (we can at the best provide an upper bound on $J(\pi^\star) - J(\pi)$). To this end, OPE-based methods and BVFT have very different characteristics when applied to model selection, and they are complementary and may be used together to provide more information and help the practitioners in making the final decision.