$Q^\star$ Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison

Tengyang Xie and Nan Jiang

Department of Computer Science, University of Illinois at Urbana-Champaign
{tx10, nanjiang}@illinois.edu

Abstract

We prove performance guarantees of two algorithms for approximating $Q^\star$ in batch reinforcement learning. Compared to classical iterative methods such as Fitted Q-Iteration—whose performance loss incurs quadratic dependence on horizon—these methods estimate (some forms of) the Bellman error and enjoy linear-in-horizon error propagation, a property established for the first time for algorithms that rely solely on batch data and output stationary policies. One of the algorithms uses a novel and explicit importance-weighting correction to overcome the infamous "double sampling" difficulty in Bellman error estimation, and does not use any squared losses. Our analyses reveal its distinct characteristics and potential advantages compared to classical algorithms.

1 Introduction

We study value-function approximation for batch-mode reinforcement learning (RL), which are central to the success of modern RL as many popular off-policy deep RL algorithms find their prototypes in this literature. These algorithms are typically *iterative*, that is, they solve a series of optimization problems, aiming to mimic each step of value- or policy-iteration [Puterman, 2014].

In the setting of general function approximation, however, not only the iterative style causes instability in practice, but it also brings several theoretical issues, which have been made abundantly clear in existing analyses [e.g., Munos, 2003, 2007; Antos et al., 2008; Farahmand et al., 2010; Chen and Jiang, 2019]:

(A) Quadratic Dependence on Horizon The performance loss of most iterative methods incur quadratic dependence on the effective horizon, i.e., $\mathcal{O}(\frac{1}{(1-\gamma)^2})$, and this is tight for the popular Approximate Value/Policy Iteration (AVI/API) [Scherrer and Lesner, 2012]. One typical way this occurs in AVI analyses is through the use of (some fine-grained variants of) the following result from Singh and Yee [1994], that the performance loss of a policy greedy w.r.t. some Q is bounded by

$$\frac{2\|Q - Q^\star\|_\infty}{1-\gamma}, \tag{1}$$

and translating $\|Q - Q^\star\|$ to the quantities that the algorithm actually optimizes incurs at least another factor of horizon. Such a quadratic dependence is significantly worse than the ideal linear dependence, the best one could hope for [Scherrer, 2014].

While linear-in-horizon algorithms exist, they often require interactive access to the environment (to collect new data using policies of the algorithm's choice), or the knowledge of transition probabilities

to compute the true expectation in the Bellman operators,[1] and few of them apply to the batch learning setting.[2] *Are there batch algorithms for $Q^\star$ that incur linear-in-horizon dependence?*

(B) Characterization of Distribution Shift One of the central challenges in RL is the *distribution shift*, that the computed policy may induce a state (and action) distribution different from what it is trained on. Existing analyses characterize this effect using the *concentrability coefficients* [Munos, 2007], with a typical definition being the density ratio (or *importance weights*) between the state distribution induced at *a particular time step* by some non-stationary policy and the data distribution. These "per-step" definitions can be very loose even in the uncontrolled setting (Section 5.2) and sometimes very complicated [Farahmand et al., 2010]. *Are there algorithms whose distribution shift effects are characterized by elegantly and tightly defined quantities?*

(C) Function Approximation Assumptions Existing analyses require strong expressivity assumptions on the function classes, such as approximate closedness under Bellman update [see inherent Bellman errors; Munos and Szepesvári, 2008]. *Are there algorithms with provable guarantees under somewhat weaker conditions?*

(D) Squared-to-Average Conversion Most batch RL algorithms heavily rely on the squared loss, but bounding the performance loss (which we eventually care about) with squared-loss objectives (which we optimize) often goes through multiple relaxations, including adding point-wise absolute values and communicating between ℓ_1 and ℓ_2 norms with Jensen's inequality, reflecting a significant gap between the actual objective (maximizing return) and the surrogate squared loss. On the other hand, we know such indirectness is not necessary in RL from the policy-gradient type algorithms [Sutton et al., 2000; Williams, 1992; Kakade and Langford, 2002], but they cannot be applied in the batch setting due to on-policy roll-outs. *Are there batch algorithms whose loss functions are more directly connected to the expected return?*

In this paper we present novel analyses of two algorithms, MSBO (which has been analyzed by Chen and Jiang [2019]) and MABO (which is novel), and provide positive answers to all questions above. A simple telescoping argument (Section 4) shows that both algorithms enjoy linear-in-horizon error propagation—which immediately improves the previous bound of Chen and Jiang [2019] for MSBO—and the distribution shift effects can be characterized by simple notions of concentrability coefficients that are significantly tighter than previous per-step definitions, which address **(A)** and **(B)**. By carefully examining the difference between the two algorithms, we further show that MABO, a novel algorithm that uses explicit importance-weighting correction and plain average objectives (without squared loss) does not suffer from the looseness of squared-to-average conversion, and comes with automatically augmented expressivity for its importance-weight class, addressing **(C)** and **(D)**.

2 Preliminaries

2.1 Markov Decision Processes (MDPs)

An (infinite-horizon discounted) MDP [Puterman, 2014] is a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$: $\mathcal{S}$ and $\mathcal{A}$ are the finite state and the finite action spaces, respectively, whose cardinalities can be arbitrarily large. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function (we use $\Delta(\cdot)$ to denote the probability simplex), $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$

[1] At the minimum, two i.i.d. next-states must be drawn from the same state-action pair, known as the *double sampling* trick [Baird, 1995], which is unrealistic in non-simulator problems.

[2] Exceptions exist when we are allowed to output complex non-stationary policies; see Section 3 for details.

Table 1: Algorithms considered in this paper, all of which require $Q^\star \in \mathcal{Q}$ (definitions of approximation error differ).

Algorithm	Style	Requirement on helper class	Horizon dependence	Concentrability coefficient	Related practical algorithm
FQI	Iterative + Sq-loss	$\forall Q \in \mathcal{Q}, \mathcal{T}Q \in \mathcal{Q}$	$1/(1-\gamma)^2$	Per-step-based	DQN [Mnih et al., 2015]
MSBO	Minimax + Sq-loss	$\forall Q \in \mathcal{Q},\ \mathcal{T}Q \in \mathcal{F}$	$1/(1-\gamma)$	Occupancy-based	SBEED [Dai et al., 2018]
MABO	Minimax + Avg-loss	$\forall Q \in \mathcal{Q}$, $w_{d_{\pi_Q}/\mu} \in \mathrm{sp}(\mathcal{W})$	$1/(1-\gamma)$	$\mathcal{W}$-based	Kernel-loss [Feng et al., 2019]

is the reward function, and $\gamma \in [0, 1)$ is a parameter that characterizes how rewards are discounted over time. $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution.

A (stochastic) policy, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, induces a random trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$ with the following generative process: $s_0 \sim d_0$, $a_t \sim \pi(\cdot|s_t)$, $r_t = R(s_t, a_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, $\forall t \geq 0$. The ultimate goodness of a policy is measured by the expected discounted return (w.r.t. the initial state distribution), defined as $J(\pi) \coloneqq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 \sim d_0, \pi]$. There always exists a policy $\pi^\star$ that maximizes the expected return for any initial state distribution.

It will be useful to define the (state-)value function $V^\pi(s) \coloneqq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi]$ and the Q-function $Q^\pi(s, a) \coloneqq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a, \pi]$. Let $V^\star$ and $Q^\star$ be the shorthand for $V^{\pi^\star}$ and $Q^{\pi^\star}$. All value functions are bounded in $[0, V_{\max}]$, where $V_{\max} \coloneqq R_{\max}/(1-\gamma)$. It is also known that the greedy policy of $Q^\star$, defined as $\pi_{Q^\star}(s) = \mathrm{argmax}_{a \in \mathcal{A}} Q^\star(s, a)$,[3] is an optimal policy $\pi^\star$.

Define the Bellman optimality operator: $(\mathcal{T}Q)(s, a) \coloneqq R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_{a' \in \mathcal{A}} Q(s', a')]$ for any $Q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$. $Q^\star$ is the unique fixed point of $\mathcal{T}$, that is, $\mathcal{T}Q^\star = Q^\star$. We also use $Q(s, \pi)$ as the shorthand for $\sum_{a \in \mathcal{A}} \pi(a|s) Q(s, a)$.

Another concept crucial to this paper is the normalized discounted state occupancy:

$$d_\pi(s) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr[s_t = s | s_0 \sim d_0, \pi].$$

The state-action occupancy $d_\pi(s, a)$ is defined similarly and satisfies $d_\pi(s, a) = d_\pi(s)\pi(a|s)$.

2.2 Batch Value-Function Approximation

Setup We are concerned with approximating $Q^\star$ in the batch RL setting, where a dataset $\mathcal{D}$ consisting of n (s, a, r, s') tuples is given, and we cannot interact with the MDP to obtain new data. We adopt the following data generation protocol from Chen and Jiang [2019], that the tuples are i.i.d.[4] as $(s, a) \sim \mu$, $r = R(s, a)$, $s' \sim P(\cdot|s, a)$, and μ is fully supported on $\mathcal{S} \times \mathcal{A}$.

Function Approximation We assume access to a function class $\mathcal{Q} \subset [0, V_{\max}]^{\mathcal{S} \times \mathcal{A}}$, and focus on algorithms that approximate $Q^\star$ with some $Q \in \mathcal{Q}$ and output its greedy policy π_Q. This automatically implies a policy class $\Pi_{\mathcal{Q}} \coloneqq \{\pi_Q : Q \in \mathcal{Q}\}$, from which the output policy will be chosen. Some algorithms require additional function classes, which we introduce later. We assume all function classes have finite

[3] With a slight abuse of notations we treat deterministic policies—which are stochastic policies that put all probability mass on a single action for each state—as of type $\mathcal{S} \to \mathcal{A}$.

[4] In reality, the transition tuples extracted from the same trajectory are in general dependent, which can be handled by concentration inequalities for dependent processes with mixing assumptions [see e.g., Antos et al., 2008].

cardinalities for simplicity when analyzing statistical errors, as they are not our main focus and extension to continuous classes with e.g., finite VC-type dimensions [Natarajan, 1989] are standard.

A representative algorithm for this setting is Fitted Q-Iteration (FQI) [Ernst et al., 2005], which can be viewed as the theoretical prototype of the popular DQN algorithm [Mnih et al., 2015]: After initializing $Q_0 \in \mathcal{Q}$ arbitrarily, we iteratively compute Q_t as

$$Q_t = \operatorname*{argmin}_{Q \in \mathcal{Q}} \ell_{\mathcal{D}}(Q; Q_{t-1}), \tag{2}$$

where

$$\ell_{\mathcal{D}}(Q; Q') := \frac{1}{n} \sum_{(s,a,r,s') \in \mathcal{D}} \left(Q(s,a) - r - \gamma \max_{a' \in \mathcal{A}} Q'(s', a') \right)^2. \tag{3}$$

We will discuss the relationship between FQI (and iterative methods in general) and algorithms we analyze.

Marginalized Importance Weights We define the importance weight of any policy π to be the ratio between its normalized discounted state-action occupancy and the data distribution:

$$w_{d_\pi/\mu}(s,a) := \frac{d_\pi(s,a)}{\mu(s,a)}.$$

Such functions are of vital importance to us, as in Section 6 we model them with function approximation to explicitly correct distribution mismatch. Their norms also characterize the exploratoriness of the data distribution, which are closely related to the *concentrability coefficients* in prior analyses [Munos, 2007; Antos et al., 2008; Farahmand et al., 2010; Chen and Jiang, 2019].

Additional Notations We use the shorthand $\mathbb{E}_\mu[\cdot]$ for the population expectation of function of (s,a,r,s') drawn from the data distribution, and $\mathbb{E}_{\mathcal{D}}[\cdot]$ for its sample-based approximation. When the function only depends on (s,a), we further omit the function arguments for brevity; for example, $\mathbb{E}_\mu[Q^2] := \mathbb{E}_{(s,a)\sim\mu}[Q(s,a)^2]$. It will also be convenient to define the μ-weighted 2-norm $\|\cdot\|_{2,\mu}^2 := \mathbb{E}_\mu[(\cdot)^2]$.

3 Related Work

Linear-in-horizon Analyses As mentioned in the introduction, most of the existing linear-in-horizon results do not apply to the setting of batch learning with general function approximation. For example, Munos [2007, Section 5.2] points out that AVI enjoys linear-in-horizon error propagation if it *happens to* converge.[5] Unfortunately, AVI—and iterative methods in general—has no convergence guarantees (and known to diverge with simple linear classes) unless used with very restricted choices of function approximators [see e.g., averagers; Gordon, 1995]. As another example, linear-in-horizon error can be achieved if one can directly minimize the Bellman error [e.g., Geist et al., 2017], but computing that requires knowledge of the transition probabilities. We refer the readers to Scherrer [2014] and the references therein for further results of this kind.

The only exceptions we are aware of are the non-stationary versions of AVI/API [e.g., Scherrer and Lesner, 2012], when the algorithm is allowed to output a periodic non-stationary policies consisting of

[5] Our paper provides a novel explanation of this result: when FQI (which is a concrete instantiation of the abstract AVI procedure) happens to converge, Chen and Jiang [2019] shows that its solution coincides with that of MSBO, which we show enjoys linear-in-horizon error propagation whatsoever.

$\Omega(1/(1-\gamma))$ stationary policies. For a typical value of $\gamma = 0.99$ this translates to 100 policies, and we believe such a complexity is responsible for the clever idea not being picked up in practice despite its appealing theoretical properties. In contrast, we establish linear-in-horizon guarantees for batch algorithms that output simple stationary policies.

Clean and Tight Concentrability Coefficients The situation of concentrability coefficients is very similar. The best definition is $\|w_{d_{\pi^\star}/\mu}\|_\infty$, enjoyed by e.g., CPI [Kakade and Langford, 2002] (see also Agarwal et al. [2019]). However, concrete instantiations of these abstract algorithms (in a way that preserve their theoretical properties) typically require on-policy Monte-Carlo roll-outs, which are not available in the batch setting. The same constant has been associated with an abstract Bellman error minimization procedure [Geist et al., 2017], but the algorithm only searches over valid value-functions (instead of arbitrary functions produced by the function approximator). While our definition is worse than theirs by a maximum over policies under consideration, it is still significantly tighter and cleaner than the per-step definitions in most previous analyses of AVI/API [Szepesvári and Munos, 2005; Munos, 2007; Antos et al., 2008; Farahmand et al., 2010]. In fact, we show in Appendix B that even in a simple uncontrolled setting, our occupancy-based definition can be $1/(1-\gamma)$ multiplicatively tighter than *any* per-step definitions.

MSBO The first algorithm we analyze, MSBO, is essentially the analogy of Modified BRM [Antos et al., 2008] (which approximates Q^π) in the context of approximating $Q^\star$. To our knowledge, the algorithm is first analyzed by Chen and Jiang [2019], and we improve their loss bound by $1/(1-\gamma)$ (which translates to $1/(1-\gamma)^2$ improvement in sample complexity). It is also worth pointing out that Dai et al. [2018] has derived a closely related algorithm and demonstrated its empirical effectiveness with deep neural nets.

MABO Our second algorithm, MABO, is presented and described in such a general form for the first time. That said, the algorithmic idea can be found in several recent works: Just as MSBO is the $Q^\star$-counterpart of Modified BRM, MABO is the $Q^\star$-counterpart of the MQL algorithm for off-policy evaluation [Uehara et al., 2019]. Another closely related work is kernel loss [Feng et al., 2019], which becomes similar to MABO when the implicit maximization in the RHKS is interpreted as searching over an importance weight class (this connection is pointed out by Uehara et al. [2019]). Finally, the average Bellman error is first used by Jiang et al. [2017] for PAC-exploration with function approximation, and MABO can be viewed as the batch analogy of their OLIVE algorithm, using importance weights to mimic the data collected by different exploration policies.

4 Telescoping Performance Difference

We present the important telescoping lemmas that enable the nice guarantees of the algorithms to be introduced and analyzed later. We start with a simple telescoping lemma, which has also been used in recent off-policy evaluation literature [e.g., Uehara et al., 2019]. Unless otherwise specified, the full proofs of the results in the main text can be found in Appendix A.

Lemma 1. *For any policy π and any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

$$\mathbb{E}_{d_0}[Q(s,\pi)] - J(\pi) = \frac{\mathbb{E}_{d_\pi}\left[Q(s,a) - r - \gamma Q(s',\pi)\right]}{1-\gamma}.$$

Proof Sketch. $J(\pi) = \frac{\mathbb{E}_{d_\pi}[r]}{1-\gamma}$, so we can remove them from both sides. The remaining terms cancel out by telescoping, which is essentially the Bellman equation for d^π found in the dual linear program of MDPs. □

Using this lemma, we prove the following performance difference bound, which is central to the nice guarantees we are able to prove for MSBO and MABO. The coarse-grained, ℓ_∞ version of Theorem 2 for the specific choice of $\pi = \pi^\star$ has been given by Williams and Baird [1993], and some of the technical insights can be found in the dervations of Munos [2007]. Still, we present the results in a general and agnostic fashion, and their applications to the analyses of MSBO and MABO are also novel.

Theorem 2 (Telescoping Performance Difference). *For any policy π and any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

$$J(\pi) - J(\pi_Q) \leq \frac{\mathbb{E}_{d_\pi}\left[\mathcal{T}Q - Q\right]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_Q}}\left[Q - \mathcal{T}Q\right]}{1-\gamma}.$$

Proof Sketch. Note that $J(\pi) - J(\pi_Q) \leq J(\pi) - \mathbb{E}_{s\sim d_0}[Q(s,\pi)] + \mathbb{E}_{s\sim d_0}[Q(s,\pi_Q)] - J(\pi_Q)$, as the sum of the two terms added on the RHS is non-negative due to greediness of π_Q. Invoking Lemma 1 on Q with π and π_Q, respectively, yields $\mathbb{E}_{d_\pi}[\mathcal{T}^\pi Q - Q]$ and $\mathbb{E}_{d_{\pi_Q}}[Q - \mathcal{T}^{\pi_Q} Q]$ (up to a horizon factor). These policy-specific Bellman errors can be bounded by the optimality error using the greediness of π_Q. □

As the result shows, the difference between $J(\pi_Q)$ and that of any π is controlled by the *average* Bellman errors $\mathbb{E}_{(\cdot)}[\mathcal{T}Q - Q]$ under the distributions d_π and d_{π_Q}, with only *one factor of horizon.* This is in sharp contrast to the typical analyses for AVI sketched in the introduction (Eq.(1)), and immediately hints at a linear-in-horizon error propagation for algorithms that control (an upper bound) of the average Bellman errors, and we only need to consider d_π and d_{π_Q} when characterizing distribution shift effects. In Appendix C, we also illustrate that iterative methods (such as FQI) fail to control the Bellman error—which is in contrary to the popular folklore belief that they do—and explain in part their quadratic dependence on horizon.

In addition, the average Bellman errors $\mathbb{E}_{d_\pi}[\mathcal{T}Q - Q]$ do *not* have absolute values inside the expectation, and the errors at different (s,a) pairs with opposite signs may cancel with each other. This property is often ignored in previous works, as they add absolute values (and use Jensen's to bound ℓ_1 with ℓ_2 norms) anyway when analyzing algorithms that optimize squared-loss, just as we will do to MSBO. However, we emphasize that it is important to state this theorem in such a primitive form for the analysis of MABO, which directly estimates such average Bellman errors (allowing sign cancellations) using importance weights. Any absolute value relaxations [e.g., Williams and Baird, 1993] will immediately make the result useless for MABO.

We conclude this section with some useful corollaries of Theorem 2, which may also be of independent interest on their own.

Corollary 3 (Two-side Performance Difference Bound). *For any $Q, f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

$$|J(\pi_f) - J(\pi_Q)| \leq 2\max\left\{\frac{\mathbb{E}_{d_{\pi_f}}\left[\mathcal{T}Q - Q\right]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_Q}}\left[Q - \mathcal{T}Q\right]}{1-\gamma}, \frac{\mathbb{E}_{d_{\pi_Q}}\left[\mathcal{T}f - f\right]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_f}}\left[f - \mathcal{T}f\right]}{1-\gamma}\right\}.$$

Corollary 4 (Performance Loss w.r.t. a Class). *$\forall Q \in \mathcal{Q}$,*

$$\max_{\pi\in\Pi_\mathcal{Q}} J(\pi) - J(\pi_Q) \leq \frac{2\max_{\pi\in\Pi_\mathcal{Q}}\left|\mathbb{E}_{d_\pi}\left[\mathcal{T}Q - Q\right]\right|}{1-\gamma}.$$

5 Minimax Squared Bellman Optimality Error Minimization (MSBO)

We present the performance guarantee of the first algorithm, MSBO, which uses another helper class $\mathcal{F} \subset [0, V_{\max}]^{\mathcal{S}\times\mathcal{A}}$ to model $\mathcal{T}Q$ for any $Q \in \mathcal{Q}$, seeking to form an (approximately) unbiased estimate of

the Bellman error $\|Q - \mathcal{T}Q\|_{2,\mu}^2$:

$$\widehat{Q} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \max_{f \in \mathcal{F}} \left(\ell_{\mathcal{D}}(Q; Q) - \ell_{\mathcal{D}}(f; Q) \right), \tag{4}$$

where $\ell_{\mathcal{D}}(\cdot;\cdot)$ is defined in Eq.(3). To give some intuitions, $\ell_{\mathcal{D}}(Q;Q)$ over-estimates $\|Q - \mathcal{T}Q\|_{2,\mu}^2$ (which is why the double sampling trick was invented in the first place [Baird, 1995]), and the amount of over-estimation can be captured by $\min_{f \in \mathcal{F}} \ell_{\mathcal{D}}(f; Q)$ if $\mathcal{F}$ is a rich function class satisfying $\mathcal{T}Q \in \mathcal{F}, \forall Q \in \mathcal{Q}$; see Antos et al. [2008]; Chen and Jiang [2019] for further intuitions.

We now state the guarantee of the algorithm.

Theorem 5 (Improved error bound of MSBO). *Let $\widehat{Q}$ be the output of MSBO. W.p. at least $1 - \delta$,*

$$\begin{aligned}\max_{\pi \in \Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \leq & \frac{2\sqrt{2C_{\text{eff}}}}{1-\gamma} \left(\sqrt{\varepsilon_{\mathcal{Q}}^{\text{sq}}} + \sqrt{\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}} \right) \\ & + \frac{\sqrt{C_{\text{eff}}}}{1-\gamma} \mathcal{O}\left(\sqrt{\frac{V_{\max}^2 \ln \frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}} + \sqrt[4]{\frac{V_{\max}^2 \ln \frac{|\mathcal{Q}|}{\delta}}{n} \varepsilon_{\mathcal{Q}}^{\text{sq}}} + \sqrt[4]{\frac{V_{\max}^2 \ln \frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{n} \varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}} \right),\end{aligned}$$

where

$$\begin{aligned} C_{\text{eff}} &:= \max_{\pi \in \Pi_{\mathcal{Q}}} \|w_{d_\pi/\mu}\|_{2,\mu}^2. \\ \varepsilon_{\mathcal{Q}}^{\text{sq}} &:= \min_{Q \in \mathcal{Q}} \|Q - \mathcal{T}Q\|_{2,\mu}^2. \\ \varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}} &:= \max_{Q \in \mathcal{Q}} \min_{f \in \mathcal{F}} \|f - \mathcal{T}Q\|_{2,\mu}^2. \end{aligned}$$

This result improves over the bound of Chen and Jiang [2019] in several aspects, which we explain below. Furthermore, their bound for MSBO is structurally the same as that for FQI when $\mathcal{F}$ is set as $\mathcal{Q}$, and while we are able to improve the bound for MSBO, some of the improvements cannot be enjoyed by FQI (see the argument of Scherrer and Lesner [2012]), creating a gap between performance guarantees of the two algorithms.

In the rest of this section, we explain the result and discuss its significance in detail. We also include a high-level sketch of the proof at the end, deferring the full proof to Appendix A.

5.1 Errors Terms and Optimality

$\varepsilon_{\mathcal{Q}}^{\text{sq}}$ measures the violation of the realizability assumption $Q^\star \in \mathcal{Q}$, and when the assumption holds exactly we have $\varepsilon_{\mathcal{Q}}^{\text{sq}} = 0$ as $\|Q^\star - \mathcal{T}Q^\star\| = 0$. Similarly, $\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}$ measures the violation of the assumption that $\mathcal{T}Q \in \mathcal{F}, \forall Q \in \mathcal{Q}$. These definitions are directly taken from Chen and Jiang [2019] and consistent with prior literature [e.g., Antos et al., 2008]. The statistical error term within $\mathcal{O}(\cdot)$ is also the same as Chen and Jiang [2019], which consists of a $n^{-1/2}$ fast rate term and two $n^{-1/4}$ terms which vanish as the approximation errors $\varepsilon_{\mathcal{Q}}^{\text{sq}}$ and $\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}$ go to 0. The novelty of the bound is in the multiplicative constants in front of these errors.

Regarding the optimality guarantee (LHS of the bound), note that we compete with $\max_{\pi \in \Pi_{\mathcal{Q}}} J(\pi)$ as the optimal value. Slightly modifying the analyses will immediately allow us to compete with any policy π even if it is not in Π_Q (e.g., $\pi^\star$), as long as we include the policy in the definition of C_{eff}.

5.2 Concentrability Coefficient

The distribution shift effects are characterized by C_{eff} in our bound. Not only this definition is much simpler, it is also tighter than previous definitions in two ways, and we start with the minor one: we use a weighted square of $w_{d_{\pi}/\mu}$ rather than its ℓ_∞ norm, the latter of which is more common in literature [Munos, 2007; Munos and Szepesvári, 2008; Antos et al., 2008; Chen and Jiang, 2019]. It is easy to show that the squared version is tighter [Farahmand et al., 2010]: for example, consider the ℓ_∞ version of our C_{eff}, which should be defined as

$$C_\infty \coloneqq \max_{\pi\in\Pi_{\mathcal{Q}}} \|w_{d_{\pi}/\mu}\|_\infty.$$

One can easily show that C_{eff} is tighter: for any $\pi \in \Pi_{\mathcal{Q}}$,

$$\|w_{d_{\pi}/\mu}\|_{2,\mu}^2 = \mathbb{E}_\mu[w_{d_{\pi}/\mu}^2] \le \mathbb{E}_\mu[C_\infty w_{d_{\pi}/\mu}] = C_\infty.$$

The second improvement, which is much more significant, is the departure from "per-step" definitions. In all analyses of AVI/API, the concentrability coefficient takes the form of

$$C_{\text{per-step}} \coloneqq \sum_{t=0}^{\infty} \beta(t) C_t, \quad C_t \coloneqq \max_{\pi} \|w_{d_{\pi,t}/\mu}\|_\infty, \tag{5}$$

where $d_{\pi,t}$ is the marginal distribution of (s_t, a_t). $\beta(t)$ is a series of non-negative coefficients that sum up to 1. Different versions of $C_{\text{per-step}}$ differ in $\beta(t)$, the policy space considered in $\max_\pi$ (typically non-stationary policies concatenated using policies from $\Pi_{\mathcal{Q}} \bigcup \{\pi^\star\}$), and sometimes replacing $\|\cdot\|_\infty$ with $\|\cdot\|_2^2$; see Farahmand et al. [2010] for a detailed discussion. While it is difficult to directly compare this quantity to ours due to its complication, we show that in a simplest uncontrolled scenario where there is no distribution shift at all, *any* per-step definition will be at least $1/(1-\gamma)$ looser than ours. We include an intuitive but informal statement below, and defer the detailed discussions to Appendix B.

Proposition 6 (Informal). *Consider an uncontrolled deterministic problem (there is only 1 action) formed by a long chain of states. Let $\mu = d^\pi$ where π is the only policy. $C_\infty = C_{\text{eff}} = 1$, and any definition of $C_{\text{per-step}} \ge 1/(1-\gamma)$.*

5.3 Horizon Dependence

We now verify that the bound has linear dependence on horizon. Doing so can be tricky given the complicated expression, and we provide 3 verification methods following the conventions in the literature [Scherrer, 2014]: The first one is to observe that FQI has quadratic dependence on horizon and our bound for MSBO has a $1/(1-\gamma)$ net improvement over FQI [Chen and Jiang, 2019]. The second one is to read the expression, and count the explicit dependence; while the statistical error depends on $V_{\max} = R_{\max}/(1-\gamma)$, such a dependence is superficial and not produced by error accumulation over multi-stage decision-making, and is never counted in the literature.[6] The third method is to consider the fully realizable case ($\varepsilon_{\mathcal{Q}}^{\text{sq}} = \varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}} = 0$) and calculate the sample complexity. Since the statistical rate is $1/\sqrt{n}$, an algorithm with linear-in-horizon error propagation should have $\mathcal{O}(1/(1-\gamma)^2)$ sample complexity, which we show below. This contrasts the $\mathcal{O}(1/(1-\gamma)^4)$ sample complexity of FQI [Chen and Jiang, 2019].

Corollary 7 (Improved sample complexity of MSBO). *Let $\varepsilon_{\mathcal{Q}}^{\text{sq}} = \varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}} = 0$. For any $\epsilon, \delta > 0$, Eq.(4) satisfies $\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \le \varepsilon \cdot V_{\max}$ w.p. $\ge 1-\delta$, if*

$$n = \mathcal{O}\left(\frac{C_{\text{eff}} \ln \frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{\varepsilon^2(1-\gamma)^2}\right).$$

[6] See Jiang and Agarwal [2018] for a deeper discussion.

5.4 Proof Sketch

We sketch the high-level proof here, deferring the details to Appendix A; this analysis is relatively straightforward due to existing work (compared to MABO, which is novel). To bound $J(\pi) - J(\pi_{\widehat{Q}})$ for any $\pi \in \Pi_{\mathcal{Q}}$, we invoke Theorem 2, which produces two average Bellman error terms of form $|\mathbb{E}_{d_\pi}[\mathcal{T}\widehat{Q} - \widehat{Q}]|$. Then

$$|\mathbb{E}_{d_\pi}[\mathcal{T}\widehat{Q} - \widehat{Q}]| = |\mathbb{E}_\mu[w_{d_\pi/\mu} \cdot (\mathcal{T}\widehat{Q} - \widehat{Q})]| \le \sqrt{\mathbb{E}_\mu[w^2_{d_\pi/\mu}]\mathbb{E}_\mu[(\mathcal{T}\widehat{Q} - \widehat{Q})^2]} \le \sqrt{C_{\text{eff}}}\|\mathcal{T}\widehat{Q} - \widehat{Q}\|_{2,\mu}.$$

The last step follows from Cauchy-Schwarz for random variables, and the term $\|\mathcal{T}\widehat{Q} - \widehat{Q}\|_{2,\mu}$ is well-studied by Chen and Jiang [2019] and we directly use their result.

6 Minimax Average Bellman Optimality Error Minimization (MABO)

We introduce and analyze our second (and novel) algorithm, MABO, which directly estimates the average Bellman errors (allowing sign cancellations) that show up in the telescoping results from Section 4 by explicit importance-weighting correction. Doing so requires an additional function approximator $\mathcal{W}$ to model the marginalized importance weights (see Section 2.2), $\mathcal{W} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, in addition to the $\mathcal{Q}$ class that models $Q^\star$. Given $\mathcal{Q}$ and $\mathcal{W}$, the algorithm is

$$\widehat{Q} = \operatorname*{argmin}_{Q\in\mathcal{Q}} \max_{w\in\mathcal{W}} \quad |\mathcal{L}_{\mathcal{D}}(Q, w)|\,, \tag{6}$$

where

$$\mathcal{L}_{\mathcal{D}}(Q, w) \coloneqq \mathbb{E}_{\mathcal{D}}\left[w(s,a)\left(r + \gamma \max_{a'} Q(s', a') - Q(s,a)\right)\right].$$

It is important to point out that we only use the single sample estimate of Bellman error (i.e., no double sampling), but we obtain an unbiased estimate of average Bellman error thanks to not using the squared loss (unlike $\ell_{\mathcal{D}}(Q; Q)$ in MSBO, which is an over-estimation). To see how $\mathcal{L}_{\mathcal{D}}(Q, w)$ is related to the average Bellman errors, simply consider its population version:

$$\mathcal{L}_\mu(Q, w) \coloneqq \mathbb{E}_{\mathcal{D}}[\mathcal{L}_{\mathcal{D}}(Q, w)] = \mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\left((\mathcal{T}Q)(s,a) - Q(s,a)\right)\right],$$

thus $\mathcal{L}_\mu(Q, w_{d_\pi/\mu}) = \mathbb{E}_{d^\pi}[\mathcal{T}Q - Q]$. Therefore, as long as $\mathcal{W}$ realizes $w_{d_\pi/\mu}$ for all $\pi \in \Pi_{\mathcal{Q}}$ (this assumption will be relaxed), $\max_{w\in\mathcal{W}} |\mathcal{L}_\mu(Q, w)|$ will control the suboptimality gap of π_Q, which is the intuition for the algorithm.

We now state the guarantee of this algorithm. For convenience, we will use $\mathbb{E}_\mu[w \cdot (\mathcal{T}Q - Q)]$ as a shorthand for Eq.(7) in the rest of this paper.

Theorem 8 (Error bound of MABO). *Let $\widehat{Q}$ be the output of MABO. W.p. $1-\delta$,*

$$\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \le \frac{2}{1-\gamma}\left(\varepsilon^{\text{avg}}_{\mathcal{Q}} + \varepsilon^{\text{avg}}_{\mathcal{Q},\mathcal{W}} + \varepsilon_{\text{stat},n}\right). \tag{7}$$

where

$$\varepsilon_{\mathcal{Q}}^{\text{avg}} \coloneqq \min_{Q\in\mathcal{Q}} \max_{w\in\mathcal{W}} \left|\mathbb{E}_\mu[w\cdot(\mathcal{T}Q-Q)]\right|,$$

$$\varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}} \coloneqq \max_{\pi\in\Pi_{\mathcal{Q}}} \inf_{w\in\text{sp}(\mathcal{W})} \max_{Q\in\mathcal{Q}} \left|\mathbb{E}_\mu\left[(w_{d_\pi/\mu}-w)\cdot(\mathcal{T}Q-Q)\right]\right|,$$

$$\varepsilon_{\text{stat},n} \coloneqq 2V_{\max}\sqrt{\frac{2C_{\text{eff},\mathcal{W}}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + \frac{4C_{\infty,\mathcal{W}}V_{\max}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{3n},$$

$$C_{\text{eff},\mathcal{W}} \coloneqq \max_{w\in\mathcal{W}}\|w\|_{2,\mu}^2, \quad C_{\infty,\mathcal{W}} \coloneqq \max_{w\in\mathcal{W}}\|w\|_\infty,$$

and $\text{sp}(\mathcal{W})$ *is the linear span of* $\mathcal{W}$ *using coefficients with (at most) unit* ℓ_1 *norm, i.e.,*

$$\text{sp}(\mathcal{W}) \coloneqq \left\{\textstyle\sum_{w\in\mathcal{W}}\alpha(w)w : \sum_{w\in\mathcal{W}}|\alpha(w)|\le 1\right\}.$$

In the rest of this section, we explain the bound and discuss its significance.

6.1 Error Terms and Augmented Expressivity

Similar to $\varepsilon_{\mathcal{Q}}^{\text{sq}}$ for MSBO, $\varepsilon_{\mathcal{Q}}^{\text{avg}}$ also measures the violation of $Q^\star\in\mathcal{Q}$, though in a different manner: we measure Q's worst-case average Bellman error on any $w\in\mathcal{W}$.

The situation of $\varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}}$ is a little more special. Despite that we provide intuition for MABO by requiring that $w_{d_{\pi_Q}/\mu}\in\mathcal{W}, \forall Q\in\mathcal{Q}$, it turns out we only need a much more relaxed version of this assumption (and can measure violation against the relaxed version): thanks to the linearity of $\mathcal{L}_{\mathcal{D}}(Q,\cdot)$, we are automatically approximating $w_{d_{\pi_Q}/\mu}$ from an augmented class $\text{sp}(\mathcal{W})$.[7] Moreover, the loss $\mathcal{L}_{\mathcal{D}}(Q,w)$ is "scale-free" w.r.t. w, i.e., it is completely equivalent to replace $\mathcal{W}$ with any $c\mathcal{W}\coloneqq\{cw : w\in\mathcal{W}\}$, for any $c\neq 0$. Therefore, we may rescale $\mathcal{W}$ arbitrarily in the theorem to obtain the sharpest bound.

To help develop further intuition, we illustrate the idea using a familiar tabular example: Consider the case where $|\mathcal{S}|$ and $|\mathcal{A}|$ are manageable and we use a tabular function class $\mathcal{Q}\coloneqq[0,V_{\max}]^{\mathcal{S}\times\mathcal{A}}$. It is easy to see that we can recover the standard tabular model-based algorithm (a.k.a. *certainty-equivalence*, or C-E) by using $\mathcal{W}=\{(s,a)\mapsto\mathbb{1}(s=s^*,a=a^*) : s^*\in\mathcal{S},a^*\in\mathcal{A}\}$, i.e., a set of $|\mathcal{S}\times\mathcal{A}|$ indicator functions. This is because the lowest possible value for the objective is 0, achieving which requires that $|\mathcal{L}_{\mathcal{D}}(Q,w)|=0,\forall w\in\mathcal{W}$. This set of $|\mathcal{W}|=|\mathcal{S}\times\mathcal{A}|$ equations is essentially the Bellman equation for each state-action pair in the empirical MDP, which can and can only be satisfied by the C-E solution. While the C-E solution incurs no approximation error, $\mathcal{W}$ clearly fails to realize $w_{d_{\pi_Q}/\mu}$ for all $Q\in\mathcal{Q}$. The reason, as we have already explained earlier, is because $\text{sp}(\mathcal{W})$—which now becomes the tabular function space—can model any importance weights with proper scaling.

As a final remark, given any $w\in\text{sp}(\mathcal{W})$ and the target importance weight $w_{d_{\pi_Q}/\mu}$, we measure their distance by projecting their difference using $\mathcal{T}Q-Q$ for the worst-case $Q\in\mathcal{Q}$. If we treat it as approximating distribution d_π with $(\mu\cdot w)(s,a)\coloneqq\mu(s,a)w(s,a)$, then this measure is essentially the Integral Probability Metric [Müller, 1997] between d_π and $\mu\cdot w$ using a discriminator class induced by $\mathcal{Q}$.

6.2 Concentrability Coefficients

Our $C_{\text{eff},\mathcal{W}}$ and $C_{\infty,\mathcal{W}}$ are defined in a way similar to C_{eff} and C_∞ in Section 5, except that we consider $w\in\mathcal{W}$, i.e., the functions provided by the function approximator $\mathcal{W}$ instead of the true importance weights $w_{d_{\pi_Q}/\mu}$ themselves. While these two sets of coefficients are not directly comparable, we provide some insights about their relationship.

[7] Similar properties have been recognized regarding the policy evaluation counterpart of MABO [Uehara et al., 2019].

On one hand, if we choose $\mathcal{W} = \{w_{d_{\pi_Q}/\mu} : Q \in \Pi_\mathcal{Q}\}$, which precisely satisfies the expressivity assumption, then $C_{\text{eff},\mathcal{W}} = C_{\text{eff}}$ and $C_{\infty,\mathcal{W}} = C_\infty$. Given that $\mathcal{W}$ is likely to include other functions as well, we might conclude that $C_{\text{eff},\mathcal{W}}$ and $C_{\infty,\mathcal{W}}$ are in general greater. On the other hand, to satisfy $\varepsilon^{\text{avg}}_{\mathcal{Q},\mathcal{W}} = 0$ we only need $\text{sp}(\mathcal{W})$ to be the above-mentioned class, and the actual $\mathcal{W}$ could be smaller and simpler. Also, since $C_{\text{eff},\mathcal{W}}$ and $C_{\infty,\mathcal{W}}$ only occur in the statistical error term in Theorem 8 (which is in sharp contrast to Theorem 5, where C_{eff} also amplifies approximation errors), the damage caused by $w \in \mathcal{W}$ with unnecessarily large magnitude can be mitigated by proper regularization (see e.g., Kallus [2016]; Hirshberg and Wager [2017]; Su et al. [2019] for how importance weights can be regularized in contextual bandits). Given these competing considerations, we suggest that it is reasonable to treat $C_{\text{eff},\mathcal{W}} \approx C_{\text{eff}}$, $C_{\infty,\mathcal{W}} \approx C_\infty$.

6.3 Horizon Dependence

The linear dependence on horizon of Theorem 8 can be verified in a way similar to Section 5.3, and we only include the sample complexity of MABO when all the expressivity assumptions are met exactly. The sample complexity contains two terms corresponding to the slow rate ($n^{-1/2}$) and the fast rate (n^{-1}) terms in $\varepsilon_{\text{stat},n}$, and when $C_{\infty,\mathcal{W}}$ is not too much larger than $C_{\text{eff},\mathcal{W}}$,[8] the fast rate term is dominated and the sample complexity is very similar to that of MSBO.

Corollary 9 (Sample complexity of MABO). *Suppose* $\varepsilon^{\text{avg}}_{\mathcal{Q}} = \varepsilon^{\text{avg}}_{\mathcal{Q},\mathcal{W}} = 0$. *The output of MABO Eq.*(6), *satisfies* $\max_{\pi\in\Pi_\mathcal{Q}} J(\pi) - J(\pi_{\widehat{Q}}) \le \varepsilon \cdot V_{\max}$ *w.p.* $1-\delta$, *if*

$$n = \mathcal{O}\left(\left(\frac{C_{\text{eff},\mathcal{W}}}{\varepsilon^2(1-\gamma)^2} + \frac{C_{\infty,\mathcal{W}}}{\varepsilon(1-\gamma)}\right)\ln\frac{|\mathcal{Q}||\mathcal{W}|}{\delta}\right).$$

6.4 Proof Sketch of Theorem 8

We conclude the section by a high-level proof sketch. With Theorem 2, it suffices to control $|\mathbb{E}_{d_\pi}[\mathcal{T}\widehat{Q} - \widehat{Q}]| = |\mathbb{E}_\mu[w_{d_\pi/\mu}\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]|$ for the worst-case $\pi \in \Pi_\mathcal{Q}$. Fixing any π, the first step is to peel off the approximation error of $\mathcal{W}$: for any $w \in \text{sp}(\mathcal{W})$, we have

$$\begin{aligned}
&|\mathbb{E}_\mu[w_{d_\pi/\mu}\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]| \\
\le\ &|\mathbb{E}_\mu[(w_{d_\pi/\mu} - w)(\mathcal{T}\widehat{Q} - \widehat{Q})]| + |\mathbb{E}_\mu[w\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]| \\
\le\ &\max_{Q\in\mathcal{Q}} |\mathbb{E}_\mu[(w_{d_\pi/\mu} - w)(\mathcal{T}Q - Q)]| + |\mathbb{E}_\mu[w\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]|.
\end{aligned}$$

So if we choose w as the one that achieves the infimum in the definition of $\varepsilon^{\text{avg}}_{\mathcal{Q},\mathcal{W}}$, denoted as $\widehat{w}$, then the first term is bounded by $\varepsilon^{\text{avg}}_{\mathcal{Q},\mathcal{W}}$. The second term is much closer to the loss function of MABO, and can be handled as

$$\begin{aligned}
|\mathbb{E}_\mu[\widehat{w}\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]| &\le \sup_{w\in\text{sp}(\mathcal{W})} |\mathbb{E}_\mu[w\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]| \\
&= \max_{w\in\mathcal{W}} |\mathbb{E}_\mu[w\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]|.
\end{aligned}$$

Crucially, using the linearity of $\mathbb{E}_\mu[w\cdot(\cdot)]$ in w and the norm constraints of $\text{sp}(\cdot)$, we are able to replace $\sup_{w\in\text{sp}(\mathcal{W})}$ with $\max_{w\in\mathcal{W}}$, leading to the augmented expressivity discussed in Section 6.1; see Eq.(11) in Appendix A for a detailed argument. Then with similar strategies, we peel off the approximation error of $\mathcal{Q}$ from $|\mathbb{E}_\mu[\widehat{w}\cdot(\mathcal{T}\widehat{Q} - \widehat{Q})]|$. The rest of the analysis handles statistical errors using generalization error bounds.

[8] E.g., $C_{\text{eff},\mathcal{W}} = C_{\infty,\mathcal{W}}$ when $\mathcal{W}$ only contains indicator functions (e.g., in the tabular scenario in Section 6.2).

7 Further Comparisons and Discussions

In the previous sections we have analyzed MSBO and MABO, showing that they enjoy linear-in-horizon error propagation and cleanly and tightly defined concentrability coefficients, which answers **(A)** and **(B)** in the introduction. Still, MSBO bears significant similarities to classical AVI/API algorithms[9] in the use of squared loss and the expressivity requirement on function approximation (**(C)** and **(D)**). In this section we compare its guarantee (Theorem 5) to that of MABO (Theorem 8), and discuss the potential advantages of MABO (which is novel and understudied), as well as its limitations, compared to currently popular algorithms. The recurring theme of the comparisons—as we will see below—is the pros and cons of implicit (e.g., FQI and MSBO) and explicit (MABO) distribution corrections.

7.1 Robustness Against Misspecified $\mathcal{Q}$

We compare the robustness of the two algorithms against misspecified $\mathcal{Q}$, that is, how much we pay when $Q^\star \notin \mathcal{Q}$. Omitting the common horizon factor, MSBO pays $\mathcal{O}(\sqrt{C_{\text{eff}} \cdot \varepsilon_{\mathcal{Q}}^{\text{sq}}})$ and MABO pays $\mathcal{O}(\varepsilon_{\mathcal{Q}}^{\text{avg}})$. Again, they are not directly comparable, but we can still offer some useful insights. Imagine the scenario of $\mathcal{W} = \{w_{d_{\pi_Q}/\mu} : Q \in \mathcal{Q}\}$ (as we did in Section 6.2), then

$$\varepsilon_{\mathcal{Q}}^{\text{avg}} = \min_{Q\in\mathcal{Q}} \max_{\pi\in\Pi_{\mathcal{Q}}} \left|\mathbb{E}_\mu[w_{d_\pi/\mu} \cdot (\mathcal{T}Q - Q)]\right| \tag{8}$$
$$\leq \min_{Q\in\mathcal{Q}} \max_{\pi\in\Pi_{\mathcal{Q}}} \sqrt{\mathbb{E}_\mu[w_{d_\pi/\mu}^2] \cdot \mathbb{E}_\mu[(\mathcal{T}Q - Q)^2]} = \sqrt{C_{\text{eff}} \cdot \varepsilon_{\mathcal{Q}}^{\text{sq}}}.$$

Here the second step follows from Cauchy-Schwarz, which we also used in Section 5.4. As we can see, if $\mathcal{W}$ is specified "just right", MABO's guarantee never suffers more than that of MSBO on misspecified $\mathcal{Q}$, and any looseness from Cauchy-Schwarz[10] enters the gap. On the other hand, such an advantage of MABO may be weakened if $\mathcal{W}$ includes additional functions that do not correspond to real importance weights.

Another difference between MSBO and MABO is that MSBO pays $\sqrt{C_{\text{eff}}}$ in front of $\sqrt{\varepsilon_{\mathcal{Q}}^{\text{sq}}}$, whereas MABO does not pay any concentrability coefficients in its approximation error terms, thanks to explicit distribution correction. While Eq.(8) might leave the impression that the difference is superficial, the inequality only relaxes $\varepsilon_{\mathcal{Q}}^{\text{avg}}$ (apart from the nice choice of $\mathcal{W}$) hence unfairly favors MSBO, and there are scenarios where the $\sqrt{C_{\text{eff}}}$ difference is real: for example, consider the scenario where Q has uniformly low error across all distributions, and Q' has small Bellman error on μ but (up to $\sqrt{C_{\text{eff}}}$ times) higher errors on e.g., $d_{\pi'_Q}$. In this case, MABO clearly prefers Q over Q' due to explicit distribution correction, whereas MSBO is indifferent between them and can suffer the poor performance of Q'.

7.2 Statistical Rates

The $n^{-1/2}$ terms in Theorems 5 and 8 match each other if we treat $C_{\text{eff}} \approx C_{\text{eff},\mathcal{W}}$ (see Section 6.2). MABO suffers another $C_{\infty,\mathcal{W}}/n$ term, whereas C_∞ does not enter the guarantee of MSBO; this is an (unfortunately) inevitable consequence of explicit importance weighting and concentration inequalities. On the other hand, the term fades away quickly with n and will be of minor issue with sufficient data. Finally, MSBO suffers two $n^{-1/4}$ terms, and although they can be absorbed by the worse between the fast rate and the approximation error terms in Big-Oh notations [Chen and Jiang, 2019, Appendix C], doing so worsens the constant.

[9] Recall that FQI coincides with MSBO using $\mathcal{F} = \mathcal{Q}$ when FQI converges [Chen and Jiang, 2019], and in this sense MSBO can be viewed as a best-case scenario for FQI.

[10] See **(D)** in the introduction.

7.3 Assumptions on the Helper Classes

A characteristic shared by MSBO and MABO is the use of a helper class ($\mathcal{F}$ for MSBO and $\mathcal{W}$ for MABO) to assist the estimation of the Bellman error. These helper classes also take the heaviest expressivity burdens in their corresponding algorithms: while $\mathcal{Q}$ is only required to capture $Q^\star$, $\mathcal{F}$ and $\mathcal{W}$ are required to capture $\mathcal{T}Q$ and $w_{d_{\pi_Q}/\mu}$, respectively, *for all* $Q \in \mathcal{Q}$.

While $\mathcal{F}$ and $\mathcal{W}$ model completely different objects, we note that $\mathcal{W}$ enjoys a superior property that $\mathcal{F}$ does not have, that is we essentially approximate the importance weights from $\mathrm{sp}(\mathcal{W})$, allowing simple $\mathcal{W}$ to have high expressivity. This property crucially comes from the linearity of the average Bellman error loss, which is another advantage of the average loss over the squared loss.

To further illustrate the representation power of $\mathrm{sp}(\mathcal{W})$, we provide the following result, showing that in MDPs with low-rank dynamics (which are often sufficient conditions that allow an exploratory[11] μ to exist in the first place [Chen and Jiang, 2019]), there exists very simple (in the sense of low statistical complexity) $\mathcal{W}$ that satisfies $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}} = 0$.

Proposition 10. *Suppose the rank of the MDP's transition matrix is k. Then,*

1. *For any choice of $\mathcal{Q}$, there exists $\mathcal{W}$ with cardinality $|\mathcal{W}| \le (k+1)|\mathcal{Q}|$, such that $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}} = 0$.*
2. *Let the transition matrix $P = \Phi P'$, where $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times k}$ and let $\phi(s,a)^\top$ denote its (s,a)-th row. For the choice of $\mathcal{Q} = \{(s,a) \mapsto R(s,a) + \gamma\phi(s,a)^\top\theta : \theta \in \mathbb{R}^k\}$, there exists $\mathcal{W}$ with cardinality $|\mathcal{W}| \le k+1$ such that $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}} = 0$.*

The formal definitions and proofs are deferred to Appendix D. In the first claim (general case), $\mathcal{W}$ has low statistical capacity despite scaling with $|\mathcal{Q}|$, as we need to pay $\ln|\mathcal{Q}|$ anyway by using the $\mathcal{Q}$ class, and the dependence of $|\mathcal{W}|$ on $|\mathcal{Q}|$ is not a significant burden. In the second claim, which is the more restricted "linear MDP" setting recently studied by e.g., Yang and Wang [2019], we are able to bring $|\mathcal{W}|$ down to as low as $k+1$; it is also interesting to point out that we cannot guarantee $w_{d_{\pi_Q}/\mu} \in \mathrm{sp}(\mathcal{W})$, but using the linear structure of $\mathcal{Q}$ we can still prove that $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}} = 0$. Finally, we emphasize that the existence of such a simple $\mathcal{W}$ does not imply that we are guaranteed to find it for every problem, as the design of function approximation always requires appropriate prior knowledge and inductive biases.

8 Conclusions

We analyze two algorithms, MSBO and MABO, which enjoy linear-in-horizon error propagation, a property established for the first time for batch algorithms outputting stationary policies. MABO uses a novel importance-weight correction to handle the difficulty of Bellman error estimation, and our analyses reveal its distinct properties and potential advantages compared to classical squared-loss-based algorithms.

Acknowledgement

The authors thank Aditya Modi for providing the references to some important related works.

[11] Technically, a small C_{eff} or C_∞.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

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

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

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051, 2019.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1133–1142, 2018.

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

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pages 568–576, 2010.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. In *Advances in Neural Information Processing Systems*, pages 15430–15441, 2019.

Matthieu Geist, Bilal Piot, and Olivier Pietquin. Is the bellman residual a bad proxy? In *Advances in Neural Information Processing Systems*, pages 3205–3214, 2017.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings 1995*, pages 261–268. Elsevier, 1995.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org, 2017.

Sham Kakade and John Langford. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pages 267–274, 2002.

Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pages 560–567, 2003.

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

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

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

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

Bruno Scherrer. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pages 1314–1322, 2014.

Bruno Scherrer and Boris Lesner. On the use of non-stationary policies for stationary infinite-horizon markov decision processes. In *Advances in Neural Information Processing Systems*, pages 1826–1834, 2012.

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

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. *arXiv preprint arXiv:1907.09623*, 2019.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pages 880–887. ACM, 2005.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Ronald J Williams and Leemon C Baird. Tight performance bounds on greedy policies based on imperfect value functions. Technical report, Citeseer, 1993.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

Appendix

A Detailed Proofs

Lemma 1 (Evaluation error lemma, restated). *For any policy π and any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

$$\mathbb{E}_{d_0}[Q(s,\pi)] - J(\pi) = \frac{\mathbb{E}_{d_\pi}\left[Q(s,a) - r - \gamma Q(s',\pi)\right]}{1-\gamma}.$$

Proof of Lemma 1. Since $J(\pi) = \frac{\mathbb{E}_{d_\pi}[r]}{1-\gamma}$, we remove these terms from both sides, and prove the rest of the identity.

$$\begin{aligned}
&\frac{\mathbb{E}_{(s,a,r,s')\sim d_\pi}\left[Q(s,a) - \gamma Q(s',\pi(s'))\right]}{1-\gamma} \\
&= \sum_{s,a}\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s, a_t = a | s_0 \sim d_0, \pi) Q(s,a) - \sum_{s,a}\sum_{t=1}^{\infty}\gamma^t \Pr(s_t = s | s_0 \sim d_0, \pi) Q(s,\pi(s)) \\
&= \sum_{s,a}\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s, a_t = a | s_0 \sim d_0, \pi) Q(s,a) - \sum_{s,a}\sum_{t=1}^{\infty}\gamma^t \Pr(s_t = s, a_t = a | s_0 \sim d_0, \pi) Q(s,a) \\
&= \sum_{s,a}\Pr(s_0 = s, a_0 = a | s_0 \sim d_0, \pi) Q(s,a) = \mathbb{E}_{s\sim d_0}[Q(s,\pi(s))],
\end{aligned}$$

where the first equation follows from the definition of d_π, the second equation follows from the definition of $Q(s,\pi(s))$. □

Theorem 2 (Telescoping Performance Difference, restated). *For any policy π and any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

$$J(\pi) - J(\pi_Q) \leq \frac{\mathbb{E}_{d_\pi}\left[\mathcal{T}Q - Q\right]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_Q}}\left[Q - \mathcal{T}Q\right]}{1-\gamma}.$$

Proof of Theorem 2.

$$\begin{aligned}
J(\pi) - J(\pi_Q) = &\underbrace{J(\pi) - \mathbb{E}_{s\sim d_0}[Q(s,\pi(s))]}_{\text{(I)}} + \underbrace{\mathbb{E}_{s\sim d_0}[Q(s,\pi(s))] - \mathbb{E}_{s\sim d_0}[Q(s,\pi_Q(s))]}_{\text{(II)}} \\
&+ \underbrace{\mathbb{E}_{s\sim d_0}[Q(s,\pi_Q(s))] - J(\pi_Q)}_{\text{(III)}}.
\end{aligned}$$

These three terms can be bound separately as follows.

$$\begin{aligned}
\text{(I)} =& J(\pi) - \mathbb{E}_{s\sim d_0}[Q(s,\pi)] \\
=& \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\left[r + \gamma Q(s',\pi) - Q(s,a)\right] \\
\leq& \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\left[r + \gamma \max_{a'\in\mathcal{A}} Q(s',a') - Q(s,a)\right] \\
=& \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\left[\mathcal{T}Q - Q\right].
\end{aligned}$$

The second equation follows from Lemma 1, and the last step follows from marginalizing out r and s' by conditioning on (s,a) using law of total expectation.

For (II),

$$\text{(II)} \leq \mathbb{E}_{s\sim d_0}\left[Q(s,\pi(s))\right] - \mathbb{E}_{s\sim d_0}\left[Q(s,\pi_Q(s))\right] = \mathbb{E}_{s\sim d_0}\left[Q(s,\pi(s)) - \max_a Q(s,a)\right] \leq 0.$$

Finally, (III), which is handled similarly to (I).

$$\begin{aligned}
\text{(III)} =& \mathbb{E}_{s\sim d_0}[Q(s,\pi_Q)] - J(\pi_Q) \\
=& \frac{1}{1-\gamma}\mathbb{E}_{d_{\pi_Q}}\left[Q(s,a) - r - \gamma Q(s',\pi_Q)\right] \\
=& \frac{1}{1-\gamma}\mathbb{E}_{d_{\pi_Q}}\left[Q(s,a) - r - \gamma \max_{a'} Q(s',a')\right] \\
=& \frac{1}{1-\gamma}\mathbb{E}_{(s,a,s')\sim d_{\pi_Q}}\left[Q - \mathcal{T}Q\right],
\end{aligned}$$

where the third equation follows from the definition of π_Q being greedy w.r.t. Q. The result follows by putting all three parts together. □

Theorem 5 (Improved error bound of MSBO, restated). *Let $\widehat{Q}$ be the output of MSBO. W.p. at least $1-\delta$,*

$$\begin{aligned}
\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \leq& \frac{2\sqrt{2C_{\text{eff}}}}{1-\gamma}\left(\sqrt{\varepsilon^{\text{sq}}_{\mathcal{Q}}} + \sqrt{\varepsilon^{\text{sq}}_{\mathcal{Q},\mathcal{F}}}\right) \\
&+ \frac{\sqrt{C_{\text{eff}}}}{1-\gamma}\mathcal{O}\left(\sqrt{\frac{V_{\max}^2 \ln\frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}} + \sqrt[4]{\frac{V_{\max}^2 \ln\frac{|\mathcal{Q}|}{\delta}}{n}\varepsilon^{\text{sq}}_{\mathcal{Q}}} + \sqrt[4]{\frac{V_{\max}^2 \ln\frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}\varepsilon^{\text{sq}}_{\mathcal{Q},\mathcal{F}}}\right),
\end{aligned}$$

where

$$\begin{aligned}
C_{\text{eff}} &:= \max_{\pi\in\Pi_{\mathcal{Q}}} \|w_{d_\pi/\mu}\|^2_{2,\mu}. \\
\varepsilon^{\text{sq}}_{\mathcal{Q}} &:= \inf_{Q\in\mathcal{Q}} \|Q - \mathcal{T}Q\|^2_{2,\mu}. \\
\varepsilon^{\text{sq}}_{\mathcal{Q},\mathcal{F}} &:= \sup_{Q\in\mathcal{Q}} \inf_{f\in\mathcal{F}} \|f - \mathcal{T}Q\|^2_{2,\mu}.
\end{aligned}$$

Proof of Theorem 5. We use $\pi^\star$ to denote $\operatorname{argmax}_{\pi\in\Pi_{\mathcal{Q}}} J(\pi)$. By applying Theorem 2, we can obtain

$$\begin{aligned}
\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \leq& \frac{\mathbb{E}_{d_{\pi^\star}}\left[\mathcal{T}\widehat{Q} - \widehat{Q}\right]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_{\widehat{Q}}}}\left[\widehat{Q} - \mathcal{T}\widehat{Q}\right]}{1-\gamma} \\
=& \frac{\mathbb{E}_{\mu}\left[w_{d_{\pi^\star}/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]}{1-\gamma} + \frac{\mathbb{E}_{\mu}\left[w_{d_{\pi_{\widehat{Q}}}/\mu}\cdot\left(\widehat{Q} - \mathcal{T}\widehat{Q}\right)\right]}{1-\gamma} \\
\overset{\text{(a)}}{\leq}& \frac{\sqrt{\mathbb{E}_{(s,a)\sim\mu}\left[\left(w_{d_{\pi^\star}/\mu}(s,a)\right)^2\right]\mathbb{E}_{(s,a)\sim\mu}\left[\left((\mathcal{T}\widehat{Q})(s,a) - \widehat{Q}(s,a)\right)^2\right]}}{1-\gamma} \\
&+ \frac{\sqrt{\mathbb{E}_{(s,a)\sim\mu}\left[\left(w_{d_{\pi_{\widehat{Q}}}/\mu}(s,a)\right)^2\right]\mathbb{E}_{(s,a)\sim\mu}\left[\left((\mathcal{T}\widehat{Q})(s,a) - \widehat{Q}(s,a)\right)^2\right]}}{1-\gamma} \\
\overset{\text{(b)}}{\leq}& \frac{2\sqrt{C_{\text{eff}}}}{1-\gamma}\|Q - \mathcal{T}Q\|_{2,\mu}.
\end{aligned} \tag{9}$$

where (a) follows from the Cauchy-Schwarz inequality for random variables ($|\mathbb{E}XY| \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$) and (b) follows from the definition of C_{eff}.

We then directly adopt the upper bound on $\left\|\widehat{Q} - \mathcal{T}\widehat{Q}\right\|_{2,\mu}$ from Chen and Jiang [2019]:

$$\left\|\widehat{Q} - \mathcal{T}\widehat{Q}\right\|_{2,\mu}^2 \le \frac{16V_{\max}^2 \ln\frac{2|\mathcal{Q}|}{\delta}}{3n} + 2\varepsilon_2 + \varepsilon_3 + \sqrt{\frac{8V_{\max}^2 \ln\frac{2|\mathcal{Q}|}{\delta}}{n}\left(\frac{10V_{\max}^2 \ln\frac{2|\mathcal{Q}|}{\delta}}{3n} + 2\varepsilon_2 + \varepsilon_3\right)},$$

$$\text{where,} \quad \varepsilon_2 = \frac{43V_{\max}^2 \ln\frac{8|\mathcal{Q}||\mathcal{F}|}{\delta}}{n} + \sqrt{\frac{239V_{\max}^2 \ln\frac{8|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}} + \varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}},$$

$$\text{and,} \quad \varepsilon_3 = \varepsilon_{\mathcal{Q}}^{\text{sq}} + \sqrt{\frac{8V_{\max}^2 \ln\frac{2|\mathcal{F}|}{\delta}}{n}\varepsilon_{\mathcal{Q}}^{\text{sq}}} + \frac{4V_{\max}^2 \ln\frac{2|\mathcal{Q}|}{\delta}}{3n}. \tag{10}$$

By substitute Eq.(9) into Eq.(10) and adapt the the proof of Theorem 17 in Chen and Jiang [2019], we have

$$\begin{aligned}
\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \le& \frac{2\sqrt{C_{\text{eff}}}}{1-\gamma}\left\|\widehat{Q} - \mathcal{T}\widehat{Q}\right\|_{2,\mu} \\
\le& \frac{2\sqrt{C_{\text{eff}}}}{1-\gamma}\left(\sqrt{2\varepsilon_{\mathcal{Q}}^{\text{sq}}} + \sqrt{2\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}}\right) + \frac{2\sqrt{C_{\text{eff}}}}{1-\gamma}\left(\sqrt{\frac{24V_{\max}^2 \ln\frac{2|\mathcal{Q}|}{\delta}}{n}} + \sqrt{\frac{172V_{\max}^2 \ln\frac{8|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}}\right) \\
&+ \frac{2\sqrt{C_{\text{eff}}}}{1-\gamma}\left(\sqrt[4]{\frac{32V_{\max}^2 \ln\frac{2|\mathcal{Q}|}{\delta}}{n}\varepsilon_{\mathcal{Q}}^{\text{sq}}} + \sqrt[4]{\frac{3824V_{\max}^2 \ln\frac{8|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}}\right). \qquad \square
\end{aligned}$$

Theorem 8 (Error bound of MABO, restated). *Let $\widehat{Q}$ be the output of MABO. W.p. $1-\delta$,*

$$\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \le \frac{2}{1-\gamma}\left(\varepsilon_{\mathcal{Q}}^{\text{avg}} + \varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}} + \varepsilon_{\text{stat},n}\right).$$

where

$$\begin{aligned}
\varepsilon_{\mathcal{Q}}^{\text{avg}} &:= \min_{Q\in\mathcal{Q}}\max_{w\in\mathcal{W}}\left|\mathbb{E}_{\mu}[w\cdot(\mathcal{T}Q - Q)]\right|, \\
\varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}} &:= \max_{\pi\in\Pi_{\mathcal{Q}}}\inf_{w\in\text{sp}(\mathcal{W})}\max_{Q\in\mathcal{Q}}\left|\mathbb{E}_{\mu}\left[(w_{d_\pi/\mu} - w)\cdot(\mathcal{T}Q - Q)\right]\right|, \\
\varepsilon_{\text{stat},n} &:= 2V_{\max}\sqrt{\frac{2C_{\text{eff},\mathcal{W}}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + \frac{4C_{\infty,\mathcal{W}}V_{\max}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{3n}, \\
C_{\text{eff},\mathcal{W}} &:= \max_{w\in\mathcal{W}}\|w\|_{2,\mu}^2, \quad C_{\infty,\mathcal{W}} := \max_{w\in\mathcal{W}}\|w\|_{\infty},
\end{aligned}$$

and $\text{sp}(\mathcal{W})$ is the linear span of $\mathcal{W}$ using coefficients with (at most) unit ℓ_1 norm, i.e.,

$$\text{sp}(\mathcal{W}) := \left\{\textstyle\sum_{w\in\mathcal{W}}\alpha(w)w : \sum_{w\in\mathcal{W}}|\alpha(w)| \le 1\right\}.$$

Proof of Theorem 8. Let $\pi_{\mathcal{Q}}^{\star} := \text{argmax}_{\pi\in\Pi_{\mathcal{Q}}} J(\pi)$. By Theorem 2, we have

$$\begin{aligned}
\max_{\pi\in\Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \le& \frac{\mathbb{E}_{d_{\pi_{\mathcal{Q}}^{\star}}}\left[\mathcal{T}\widehat{Q} - \widehat{Q}\right]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_{\widehat{Q}}}}\left[\widehat{Q} - \mathcal{T}\widehat{Q}\right]}{1-\gamma} \\
\le& \frac{2\max_{\pi\in\Pi_{\mathcal{Q}}}\left|\mathcal{L}_{\mu}(\widehat{Q}, w_{d_\pi/\mu})\right|}{1-\gamma}.
\end{aligned}$$

We now bound $\left|\mathcal{L}_\mu(\widehat{Q}, w_{d_\pi/\mu})\right|$ for any policy $\pi \in \Pi_\mathcal{Q}$. Let

$$\widehat{w}_{d_\pi/\mu} \coloneqq \operatorname*{argmin}_{w\in \mathrm{sp}(\mathcal{W})} \max_{Q\in\mathcal{Q}} \left|\mathbb{E}_\mu\left[(w_{d_\pi/\mu} - w)\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right|,$$

and we obtain

$$\begin{aligned}\left|\mathcal{L}_\mu(\widehat{Q}, w_{d_\pi/\mu})\right| &= \left|\mathbb{E}_\mu\left[(w_{d_\pi/\mu} - \widehat{w}_{d_\pi/\mu})\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right] + \mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right| \\ &\le \left|\mathbb{E}_\mu\left[(w_{d_\pi/\mu} - \widehat{w}_{d_\pi/\mu})\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right| + \left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right| \\ &= \varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}} + \left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right|,\end{aligned}$$

where the last equation follows from the definition of $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}}$.

To bound the remaining term, we first need a helper lemma that $\sup_{w\in\mathrm{sp}(\mathcal{W})}|f(\cdot)| = \max_{w\in\mathcal{W}}|f(\cdot)|$ for any linear function $f(\cdot)$: consider any $w \in \mathrm{sp}(\mathcal{W})$, which can be written as $w = \sum_i \alpha_i w_i$, where $w_i \in \mathcal{W}, \forall i$ and $\sum_i |\alpha_i| \le 1$. For linear $f(\cdot)$ and any $w \in \mathrm{sp}(\mathcal{W})$ we have

$$|f(w)| = \left|f\left(\sum_i \alpha_i w_i\right)\right| = \left|\sum_i \alpha_i f(w_i)\right| \le \sum_i |\alpha_i|\,|f(w_i)| \le \sup_{w'\in\mathcal{W}}|f(w')|. \tag{11}$$

So $\sup_{w\in\mathrm{sp}(\mathcal{W})}|f(\cdot)| \le \max_{w\in\mathcal{W}}|f(\cdot)|$. On the other hand, since $\mathcal{W} \subset \mathrm{sp}(\mathcal{W})$, we conclude that $\sup_{w\in\mathrm{sp}(\mathcal{W})}|f(\cdot)| = \max_{w\in\mathcal{W}}|f(\cdot)|$ for linear $f(\cdot)$.

With this preparation, now we are ready to bound $\left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right|$. Note that

$$\varepsilon^{\mathrm{avg}}_{\mathcal{Q}} \coloneqq \min_{Q\in\mathcal{Q}}\max_{w\in\mathcal{W}}|\mathbb{E}_\mu[w\cdot(\mathcal{T}Q - Q)]| = \min_{Q\in\mathcal{Q}}\sup_{w\in\mathrm{sp}(\mathcal{W})}|\mathbb{E}_\mu[w\cdot(\mathcal{T}Q - Q)]|,$$

so we have

$$\left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right| = \left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right| - \min_{Q\in\mathcal{Q}}\sup_{w\in\mathrm{sp}(\mathcal{W})}|\mathbb{E}_\mu\left[w\cdot(\mathcal{T}Q - Q)\right]| + \varepsilon^{\mathrm{avg}}_{\mathcal{Q}},$$

At this point, we peeled off all the approximation errors from $\left|\mathcal{L}_\mu(\widehat{Q}, w_{d_\pi/\mu})\right|$, and it remains to bound the estimation error

$$\left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q} - \widehat{Q}\right)\right]\right| - \inf_{Q\in\mathcal{Q}}\sup_{w\in\mathrm{sp}(\mathcal{W})}|\mathbb{E}_\mu\left[w\cdot(\mathcal{T}Q - Q)\right]|.$$

Let $\widetilde{Q} \coloneqq \operatorname{argmin}_{Q\in\mathcal{Q}} \sup_{w\in\mathrm{sp}(\mathcal{W})} |\mathbb{E}_\mu [w\cdot(\mathcal{T}Q - Q)]|$ and $\mathcal{W}_1 \coloneqq \{aw : a\in[-1,1], w\in\mathcal{W}\}$.

$$
\begin{aligned}
&\left|\mathbb{E}_\mu\left[\widehat{w}_{d_\pi/\mu}\cdot\left(\mathcal{T}\widehat{Q}-\widehat{Q}\right)\right]\right| - \inf_{Q\in\mathcal{Q}}\sup_{w\in\mathrm{sp}(\mathcal{W})}|\mathbb{E}_\mu[w\cdot(\mathcal{T}Q-Q)]| \\
&\le \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widehat{Q}-\widehat{Q}\right)\right]\right| - \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&= \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widehat{Q}-\widehat{Q}\right)\right]\right| - \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| \\
&\quad + \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| - \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&\overset{(a)}{\le} \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widehat{Q}-\widehat{Q}\right)\right] - \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| \\
&\quad + \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| - \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&\overset{(b)}{\le} \underbrace{\sup_{w\in\mathcal{W}}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widehat{Q}-\widehat{Q}\right)\right] - \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right|}_{\text{(I)}} \\
&\quad + \underbrace{\sup_{w\in\mathcal{W}}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widetilde{Q}(s',a')-\widetilde{Q}(s,a)\right)\right] - \mathbb{E}_\mu\left[w(s,a)\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right|}_{\text{(II)}}.
\end{aligned}
\tag{12}
$$

where (a) follows form $\sup_x|f(x)| - \sup_x|g(x)| \le \sup_x|f(x)-g(x)|$ and (b) follows from Eq.(11) and the following argument:

$$
\begin{aligned}
&\sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| - \sup_{w\in\mathrm{sp}(\mathcal{W})}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&\le \sup_{w\in\mathcal{W}}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| - \sup_{w\in\mathcal{W}_1}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&\le \sup_{w\in\mathcal{W}}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widehat{Q}(s',a')-\widehat{Q}(s,a)\right)\right]\right| - \sup_{w\in\mathcal{W}}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&\le \sup_{w\in\mathcal{W}}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widetilde{Q}(s',a')-\widetilde{Q}(s,a)\right)\right]\right| - \sup_{w\in\mathcal{W}}\left|\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right| \\
&\le \sup_{w\in\mathcal{W}}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(r+\max_{a'}\widetilde{Q}(s',a')-\widetilde{Q}(s,a)\right)\right] - \mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]\right|,
\end{aligned}
$$

where the first inequality follows from Eq.(11) and the fact that $\mathcal{W}_1 \subseteq \mathrm{sp}(\mathcal{W})$, the second inequality follows from the linearity of $\mathbb{E}_\mu\left[w\cdot\left(\mathcal{T}\widetilde{Q}-\widetilde{Q}\right)\right]$, the third inequality follows from the fact that $\widehat{Q}$ optimizes $\max_{w\in\mathcal{W}}|\mathcal{L}_\mathcal{D}(\cdot,w)|$, and the last inequality follows from $\sup_x|f(x)| - \sup_x|g(x)| \le \sup_x|f(x)-g(x)|$.

Now, since the only difference between term (I) and term (II) is the choice of Q and w, it suffices to provide a uniform deviation bound that applies to all $w\in\mathcal{W}$ and $Q\in\mathcal{Q}$. Before applying concentration bounds, it will be useful to first verify the boundedness of the random variables: $w(s,a)\in[-C,C]$, and $r+\gamma\max_{a'}Q(s',a')-Q(s,a)\in[-V_{\max},V_{\max}]$ (recall that we assumed $Q\in[0,V_{\max}]$). Therefore, by

Bernstein's inequality and the union bound, w.p. at least $1-\delta$ we have that for any $w \in \mathcal{W}$ and $Q \in \mathcal{Q}$,

$$\begin{aligned}
&\left|\mathbb{E}_\mu\left[w \cdot (\mathcal{T}Q - Q)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[w(s_i,a_i)\left(r_i + \gamma \max_{a'} Q(s_i',a') - Q(s_i,a_i)\right)\right]\right| \\
\leq &\sqrt{\frac{2\mathrm{Var}_\mu\left[w(s,a)\left(r+\gamma\max_{a'}Q(s',a')-Q(s,a)\right)\right]\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + \frac{2C_{\infty,\mathcal{W}}V_{\max}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{3n} \\
\overset{(a)}{\leq} &V_{\max}\sqrt{\frac{2C_{\text{eff},\mathcal{W}}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + \frac{2C_{\infty,\mathcal{W}}V_{\max}\ln\frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{3n} = \frac{\varepsilon_{\text{stat},n}}{2},
\end{aligned} \tag{13}$$

where (a) is obtained by the following argument:

$$\begin{aligned}
&\mathrm{Var}_\mu\left[w(s,a)\left(r+\gamma\max_{a'}Q(s',a')-Q(s,a)\right)\right] \\
\leq &\mathbb{E}_\mu\left[w(s,a)^2\left(r+\gamma\max_{a'}Q(s',a')-Q(s,a)\right)^2\right] \\
\leq &V_{\max}^2\mathbb{E}_\mu\left[w(s,a)^2\right] \leq V_{\max}^2 C_{\text{eff},\mathcal{W}}.
\end{aligned}$$

Substituting Eq.(13) into Eq.(12), we obtain that the both of term (I) and term (II) in Eq.(12) can be simultaneously bounded by $\varepsilon_{\text{stat},n}/2$ w.p. $1-\delta$. Therefore, we bound $\max_{\pi\in\Pi_\mathcal{Q}} J(\pi) - J(\pi_{\widehat{Q}})$ w.p. $1-\delta$ as follows

$$\max_{\pi\in\Pi_\mathcal{Q}} J(\pi) - J(\pi_{\widehat{Q}}) \leq \frac{2}{1-\gamma}\left(\varepsilon_\mathcal{Q}^{\text{avg}} + \varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}} + \varepsilon_{\text{stat},n}\right). \qquad \square$$

B Comparison between Per-step vs. Occupancy-based Concentrability Coefficients

We provide an example to illustrate the limitation of the per-step concentrability coefficients (Proposition 6). Consider a deterministic chain MDP, where there are $L+1$ states, $\{s_0, s_1, s_2, \ldots, s_L\}$. There is only one action, which we omit in the notations. s_0 is the deterministic initial state, and each s_l transitions to s_{l+1} under the only action for $0 \leq l < L$. s_L is an absorbing state (i.e., it transitions to itself). The reward function is inconsequential.

There is only one possible policy π for this MDP, and we let the data distribution $\mu = d^\pi$. The occupancy-based concentrability coefficient is always 1 (either C_∞ or C_{eff}), which agrees with the intuition that there is no distribution shift. Since the per-step definitions (Eq.(5)) are always the convex combinations of $C_t = \max_\pi \|w_{d_\pi,t/\mu}\|_\infty$ for $t \geq 0$, we can assert that it is never lower than $\min_t C_t$ however the combination coefficients are chosen.

Now we calculate C_t for this MDP:

$$C_t = \begin{cases} \frac{1}{\mu(s_t)} = \frac{1}{(1-\gamma)\gamma^t}, & 0 \leq t < L \\ \frac{1}{\mu(s_L)} = \frac{1}{\gamma^L}, & t \geq L \end{cases}$$

Replacing $\|\cdot\|_\infty$ with $\|\cdot\|_{2,\mu}^2$ gives exactly the same results. (When the distribution on the enumerator is a point mass, $\|\cdot\|_{2,\mu}^2$ of the importance weight is the same as $\|\cdot\|_\infty$.) Therefore, as long as L is sufficiently large so that $\frac{1}{\gamma^L} \geq \frac{1}{(1-\gamma)}$, we have $C_t \geq 1/1-\gamma$ for all t, and the per-step concentrability coefficient is at least $1/1-\gamma$. As a final remark, since the MDP only has 1 policy, the result has no dependence on the choice of policy class in $\max_\pi$ in the definition of concentrability coefficient, so we have virtually covered all existing definitions in the AVI/API literature.

C On Iterative Methods' Lack of Control of Bellman Errors

We demonstrate that iterative methods fail to directly control the Bellman error on the data distribution μ. Consider a two-state deterministic MDP with just 1 action, where s_1 transitions to s_2, and s_2 is absorbing. The reward is always 0.

We use the tabular representation for this MDP, where $Q = [Q(s_1, a), Q(s_2, a)]^\top$. Assume our batch data $\mathcal{D}$ only contains transition tuples of form $(s_1, a, 0, s_2)$, and no data points from (s_2, a_2) are present. We first show how FQI behave on this example. Given the update rule of FQI (Eq.(2)),

$$Q_t \in \operatorname*{argmin}_Q \ell_{\mathcal{D}}(Q; Q_{t-1}) = \{[Q(s_1, a), Q(s_2, a)]^\top : Q(s_1, a) = \gamma Q_{t-1}(s_2, a)\}.$$

Therefore, with very update, $Q(s_1, a)$ will obtain the old value of $\gamma Q(s_2, a)$ from the previous iteration, whereas the new value of $Q(s_2, a)$ will be set arbitrarily. Since the mean square Bellman error is $\|Q_t - \mathcal{T}Q_t\|^2_{2,\mu} = (Q_t(s_1, a) - \gamma Q_t(s_2, a))^2$, its value can be arbitrarily away from 0 and do not become smaller over iterations. In comparison, it is easy to verify that MSBO and MABO do not suffer from this issue: although there is also arbitrariness in their outputs due to insufficient data coverage, their outputs will always satisfy $Q(s_1, a) = \gamma Q(s_2, a)$ and hence imply zero Bellman error on μ.

As a final remark, it should be noted that the counterexample holds because μ is non-exploratory and $C_{\text{eff}} = C_\infty = \infty$, which breaks the assumption for all algorithms considered in this paper. Although $\|Q - \mathcal{T}Q\|^2_{\mu,2}$ will be controlled by FQI when μ is exploratory, this is an indirect consequence of FQI finding $Q \approx Q^\star$, and our example illustrates that these iterative methods do not *directly* control the Bellman error on the data distribution.

D Existence of Simple $\mathcal{W}$ in Low-rank MDPs (Proposition 10)

Claim 1: General Low-rank Case Consider an MDP whose transition matrix $P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ satisfies $\operatorname{rank}(P) = k$. Let there be a total of N policies in $\Pi_{\mathcal{Q}}$, and we stack $\nu_\pi \in \mathbb{R}^{|\mathcal{S}|}$ for all $\pi \in \Pi_{\mathcal{Q}}$ as a matrix: $M_\nu \coloneqq \begin{bmatrix}\nu_{\pi_1} & \cdots & \nu_{\pi_N}\end{bmatrix}^\top$; all vectors in this proof are treated as column vectors. We first argue that $\operatorname{rank}(M_\nu) \le k + 1$.

Let $\nu_{\pi,t}(s)$ be the marginal distribution of s_t under π. Also let $\Pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}\mathcal{A}|}$ be the standard matrix representation of a policy π, that is, $\Pi_{s',(s,a)} \coloneqq \mathbb{1}(s = s', a = \pi(s))$. It is known that $\nu^\top_{\pi,t} = d_0^\top(\Pi P)^t$, which shows that $\nu^\top_{\pi,t}$ is in the row-space of $\begin{bmatrix} P \\ d_0^\top \end{bmatrix}$ for any π and t. Since $\nu_\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t \nu_{\pi,t}$, the same holds for ν_π. Therefore, we have $\operatorname{rank}(M_\nu) \le \operatorname{rank}(\begin{bmatrix} P \\ d_0^\top \end{bmatrix}) \le k+1$. For convenience, let $k' \coloneqq k+1$.

Then, following a determinant(volume)-maximization argument similar to Chen and Jiang [2019, Proposition 10], we can find k' rows from M_ν, denoted as $\eta_1, \ldots, \eta_{k'}$, which satisfies the following: for any $i = 1, \ldots, N$, there exists $\alpha_1, \ldots, \alpha_{k'}$, such that $\nu_{\pi_i} = \sum_{j=1}^{k'} \alpha_j \cdot k' \cdot \eta_j$, and $|\alpha_j| \le 1/k'$ for $j = 1, \ldots, k'$. This implies that $\{\nu_{\pi_1}, \ldots, \nu_{\pi_N}\} \subseteq \operatorname{sp}(\{\eta'_1, \ldots, \eta'_{k'}\})$, where $\eta'_i \coloneqq k'\eta_i$. Now consider $\operatorname{sp}(\{\eta'_1, \ldots, \eta'_{k'}\}) \times \Pi_{\mathcal{Q}}$, where the Cartesian product produces $k'|\Pi_{\mathcal{Q}}|$ pairs of state-action functions, defined as $\eta' \times \pi \coloneqq ((s,a) \mapsto \eta'(s) \cdot \mathbb{1}(a = \pi(s)))$. We claim that $\{d_{\pi_1}, \ldots, d_{\pi_N}\} \subset \operatorname{sp}(\{\eta'_1, \ldots, \eta'_{k'}\} \times \Pi_{\mathcal{Q}}$: for any π_i, since ν_{π_i} can be expressed as the linear combination of $\{\eta'_1, \ldots, \eta'_{k'}\}$ with coefficients satisfying the norm constraints, $d_{\pi_i} = \nu_{\pi_i} \times \pi_i$ is also the combination of $\{\eta'_1 \times \pi_i, \ldots, \eta'_{k'} \times \pi_i\}$ with exactly the same coefficients.

Since μ is supported on the entire $\mathcal{S} \times \mathcal{A}$, we have $w_{d_\pi/\mu} = \operatorname{diag}(\mu)^{-1} d_\pi$. Putting all results together, it suffices to choose $\mathcal{W} = \{\operatorname{diag}(\mu)^{-1}(\eta'_i \times \pi_Q) : i \in [k'], Q \in \mathcal{Q}\}$, and $|\mathcal{W}| \le (k+1)|\Pi_{\mathcal{Q}}|$.

Remark on the $|\mathcal{Q}|$ Dependence in the General Case The annoying dependence on $|\mathcal{Q}|$ comes from the fact that we hope the *state-action occupancy* vectors of different policies to have low-rank factorization

(which is satisfied in the more restricted case; see Claim 2). In general low-rank MDPs, however, only state occupancy factorizes and the state-action one does not; a counter-example can be easily shown in contextual bandits:

Consider an MDP with 2 actions per state. d_0 is uniform among $|\mathcal{S}| - 1$ states, all of which transition deterministically to the last state, which is absorbing. This MDP essentially emulates a contextual bandit. Since all states share exactly the same next-state distribution, the rank of the transition matrix is 1 regardless of how large $|\mathcal{S}|$ is. Now consider a policy space $\Pi_{\mathcal{Q}}$, where each policy takes action a_1 in one of the $|\mathcal{S}| - 1$ states, and takes a_2 in all other states; there are $|\mathcal{S}| - 1$ such policies. It is easy to show that the matrix consisting of state-action occupancy d_π for all policies in $\Pi_{\mathcal{Q}}$ has full-rank $|\mathcal{S}| - 1$, which cannot be bounded by the rank of the transition matrix when $|\mathcal{S}|$ is large.

Given this difficulty, our strategy is to first find the policies whose state occupancy vectors span the entire low-dimensional space, and take their Cartesian product with $\Pi_{\mathcal{Q}}$ to handle the actions, which results in the $|\mathcal{Q}|$ dependence. As we will see below, we can avoid paying $|\mathcal{Q}|$ when the $\mathcal{Q}$ class is more structured.

Claim 2: Restricted Case of Knowing the Left Factorization Matrix as Features [Yang and Wang, 2019] Here we consider the setting of $P = \Phi P'$, where $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times k}$ and $\phi(s,a)^\top$ denotes its (s,a)-th row. For the choice of $\mathcal{Q} = \{(s,a) \mapsto R(s,a) + \gamma\phi(s,a)^\top\theta : \theta \in \mathbb{R}^k\}$, note that any $Q \in \mathcal{Q}$ is in the column space of $\Phi^+ := [\Phi\ R]$, where the reward function R is treated as an $|\mathcal{S}\times\mathcal{A}| \times 1$ vector. Yang and Wang [2019, Proposition 2] shows that it is realizable and closed under Bellman update, i.e., $\mathcal{T}Q \in \mathcal{Q}, \forall Q \in \mathcal{Q}$. Therefore, the Bellman error $Q - \mathcal{T}Q$ is also in the column space of Φ^+. Let $\phi^+(s,a)^\top$ be the (s,a)-th row of Φ^+, and θ^+_Q and $\theta^+_{\mathcal{T}Q}$ be the coefficients such that $Q = \phi^+(s,a)^\top\theta^+_Q$ and $\mathcal{T}Q = \phi^+(s,a)^\top\theta^+_{\mathcal{T}Q}$.

Fixing any π, consider

$$\begin{aligned}
&\mathbb{E}_\mu[(w - w_{d_\pi/\mu}) \cdot (\mathcal{T}Q - Q)] \\
&= \mathbb{E}_\mu[(w - w_{d_\pi/\mu}) \cdot (\phi^+(s,a)^\top(\theta^+_Q - \theta^+_{\mathcal{T}Q}))] \\
&= (w - w_{d_\pi/\mu})^\top \mathrm{diag}(\mu)\Phi^+(\theta^+_Q - \theta^+_{\mathcal{T}Q}).
\end{aligned}$$

According to the definition of $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}}$, to achieve $\varepsilon^{\mathrm{avg}}_{\mathcal{Q},\mathcal{W}} = 0$ it suffices to have the following: for every $\pi \in \Pi_{\mathcal{Q}}$, there exists $w \in \mathrm{sp}(\mathcal{W})$, such that $\mathbb{E}_\mu[(w - w_{d_\pi/\mu}) \cdot (Q - \mathcal{T}Q)] = 0$ for any $Q \in \mathcal{Q}$. Given the linear structure of Q and $\mathcal{T}Q$, we can relax the last statement to its sufficient condition:

$$(w - w_{d_\pi/\mu})^\top \mathrm{diag}(\mu)\Phi^+ = \mathbf{0}^\top_{k+1},$$

where $\mathbf{0}$ is the all-zero vector. The rest of the proof is very similar to Claim 1: we simply stack $w^\top_{d_\pi/\mu}\mathrm{diag}(\mu)\Phi^+ \in \mathbb{R}^{1\times(k+1)}$ together into a $|\Pi_{\mathcal{Q}}| \times (k+1)$ matrix, use the determinant-maximization argument to select its rows, and form $\mathcal{W}$ with the corresponding $w_{d_\pi/\mu}$ after proper rescaling.

Remark Since $\mathcal{Q}$ is closed under Bellman update in this setting, one may also use $\mathcal{Q}$ as the helper class $\mathcal{F}$ for MSBO. However, the complexity of $\mathcal{F}$ in this case only matches that of $\mathcal{W}$ in the more general case (Claim 1) and is significant worse than what we can achieve here ($|\mathcal{W}| \le k + 1$).