The Role of Inherent Bellman Error in Offline Reinforcement Learning with Linear Function Approximation

Noah Golowich
nzg@mit.edu
MIT

Ankur Moitra
moitra@mit.edu
MIT

Abstract

In this paper, we study the offline RL problem with linear function approximation. Our main structural assumption is that the MDP has *low inherent Bellman error*, which stipulates that linear value functions have linear Bellman backups with respect to the greedy policy. This assumption is natural in that it is essentially the minimal assumption required for value iteration to succeed. We give a computationally efficient algorithm which succeeds under a *single-policy coverage* condition on the dataset, namely which outputs a policy whose value is at least that of any policy which is well-covered by the dataset. Even in the setting when the inherent Bellman error is 0 (termed *linear Bellman completeness*), our algorithm yields the first known guarantee under single-policy coverage. In the setting of positive inherent Bellman error $\varepsilon_{\mathsf{BE}} > 0$, we show that the suboptimality error of our algorithm scales with $\sqrt{\varepsilon_{\mathsf{BE}}}$. Furthermore, we prove that the scaling of the suboptimality with $\sqrt{\varepsilon_{\mathsf{BE}}}$ cannot be improved for *any* algorithm. Our lower bound stands in contrast to many other settings in reinforcement learning with misspecification, where one can typically obtain performance that degrades *linearly* with the misspecification error.

1 Introduction

The study of *reinforcement learning (RL)* focuses on the problem of sequential decision making in a stateful and stochastic environment, typically modeled as a *Markov Decision Process (MDP)*. An agent aims to maximize its expected reward over a finite time horizon H, also known as its *value*, by choosing a *policy*, or a mapping from states to actions. In the *offline* (or *batch*) *RL* problem, an agent's only knowledge of its environment comes from a dataset $\mathcal{D}$ consisting of samples drawn from the state transition distributions and reward functions of the MDP. Given $\mathcal{D}$, the learning algorithm aims to compute a policy $\hat{\pi}$ whose value is at least as good as that of some *reference policy* $\pi^\star$.

A key challenge in offline RL is understanding how to choose the policy $\hat{\pi}$ when the dataset $\mathcal{D}$ exhibits incomplete coverage of the environment, meaning that the transitions from many states are not represented in $\mathcal{D}$. The naive approach to this problem would proceed via some variant of *value iteration*. At each step $h = H, H-1, \ldots, 1$ of the time horizon, given a policy acting at steps $h' > h$, value iteration uses $\mathcal{D}$ to compute estimates of the *value function* of the policy, which maps each state-action pair to the policy's expected reward starting at that state-ation pair. Value iteration then greedily chooses a policy at step h which maximizes this estimated value, and proceeds to step $h-1$. Unfortunately, this approach suffers from the fact that state-action pairs which are *undercovered* by the dataset $\mathcal{D}$ may have, due to random fluctuations in $\mathcal{D}$, overly optimistic estimates of their value. The chosen policy $\hat{\pi}$ will then aim to visit such state-action pairs, which could in fact be very suboptimal.

A common approach to correcting the above issue is the principle of *pessimism* (e.g., Fujimoto et al. (2019); Xie et al. (2021a); Liu et al. (2020); Yu et al. (2020); Jin et al. (2021)), which chooses $\hat{\pi}$ so as to maximize an *underestimate* of its value, where the underestimate is chosen according to constraints that force it to be consistent with the dataset. By ensuring that $\hat{\pi}$ takes actions whose value is robust to random fluctuations in the dataset, pessimistic algorithms typically ensure a guarantee of the following form: for any policy $\pi^\star$ whose state-action pairs are well-covered by $\mathcal{D}$, the value of $\hat{\pi}$ is guaranteed to be at least that of $\pi^\star$. The assumption made on $\pi^\star$ here is typically known as a *single-policy coverage* condition (formalized in Definition 2.2); along with several variants, it has come to represent a gold standard for obtaining offline RL guarantees.

Function approximation & misspecification in offline RL. As the state and action spaces encountered in practice tend to be large or infinite, much of the theoretical work on offline RL makes *function approximation* assumptions, which introduce function classes $\mathcal{Q}_h$ for steps $h \in [H]$, whose elements can be used to approximate the value function for a policy. Due to the complexity of the offline RL problem, our understanding of the optimal guarantees attainable remains limited even for the fundamental special case in which the classes $\mathcal{Q}_h$ are *linear*, which is the focus of this paper. Concretely, letting the state and action spaces be denoted by $\mathcal{X}$ and $\mathcal{A}$, respectively, we assume the following: for some known *feature mappings* $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, $\mathcal{Q}_h$ is the class of functions $(x, a) \mapsto \langle \phi_h(x, a), w \rangle$, where $w \in \mathbb{R}^d$ belongs to some bounded set. As an added benefit of focusing on the linear setting, we will be able to obtain end-to-end computationally efficient algorithms, without reliance on a regression oracle for $\mathcal{Q}_h$.

It is unreasonable to expect that elements of the classes $\mathcal{Q}_h$ coincide exactly with the actual value functions for the underlying MDP. Therefore, it is essential to understand the price paid by having *misspecification error* in $\mathcal{Q}_h$. To quantify this misspecification error we use the *inherent Bellman error* (Munos & Szepesvári, 2008; Munos, 2005), denoted $\varepsilon_{\mathsf{BE}}$, which describes the maximum distance between the Bellman backup of any function in $\mathcal{Q}_{h+1}$ with respect to the greedy policy and a best-approximating function in $\mathcal{Q}_h$. The inherent Bellman error is particularly natural since it is exactly what quantifies the degree to which value iteration succeeds: in particular, the regression problems solved by value iteration are $\varepsilon_{\mathsf{BE}}$-approximatly well-specified. As a result, it can be shown that, if $\mathcal{D}$ covers the entire state space in an appropriate sense and the inherent Bellman error is bounded by $\varepsilon_{\mathsf{BE}}$, then value iteration produces a policy whose suboptimality may be bounded by $\mathrm{poly}(d, H) \cdot \varepsilon_{\mathsf{BE}}$ (Munos & Szepesvári, 2008). Moreover, the linear growth of suboptimality error with respect to $\varepsilon_{\mathsf{BE}}$ cannot be improved in general (Tsitsiklis & van Roy, 1996).

The results of Munos & Szepesvári (2008); Munos (2005) discussed above serve as a useful sanity check on the reasonableness of low inherent Bellman error, but do little to address the problems encountered in typical offline RL situations as a result of their stringent assumption that $\mathcal{D}$ covers the full state space. In most such settings, ranging from autonomous driving to healthcare, one should expect the offline data to be gathered in regions of the state space that result from executing reasonably good policies. (For instance, one should not expect much offline data involving states corresponding to driving a car off a cliff.) Thus, positive results under only single-policy coverage conditions are much more desirable. Our main goal is to address the following question: *is boundedness of the inherent Bellman error sufficient for computationally efficient offline RL under only a single-policy coverage condition?* Prior to our work, this question was open even for the special case of 0 inherent Bellman error, which is typically known as *linear Bellman completeness*.

1.1 Main results

Our first result is a positive answer to our main question; to state it, we need the following notation (introduced formally in Section 2). An offline RL algorithm takes as input a dataset $\mathcal{D}$ of size n, consisting of n tuples (h, x, a, r, x'), denoting a sample of the transition at step $h \in [H]$: namely, at state x, action a was taken, reward r was received, and the next state observed was x'. For each $h \in [H]$, let Σ_h denote the unnormalized covariance matrix of feature vectors in $\mathcal{D}$ at step h. Moreover, for a policy π, let $\mathscr{C}_{\mathcal{D},\pi} := \sum_{h=1}^{H} \|\mathbb{E}^\pi[\phi_h(x_h, a_h)]\|_{n\Sigma_h^{-1}}$ denote the *coverage parameter* for

π with respect to $\mathcal{D}$, which measures to what degree vectors in $\mathcal{D}$ extend in the direction of a typical feature vector drawn from π. We denote the inherent Bellman error of the MDP by $\varepsilon_{\mathsf{BE}} \geq 0$.

Theorem 1.1 (Informal version of Theorem 3.1)**.** *There is an algorithm (namely, Algorithm 1) which given the dataset $\mathcal{D}$ as input, outputs a policy $\hat{\pi}$ at random so that for any policy $\pi^\star$, we have*

$$\mathbb{E}\left[V_1^{\pi^\star}(x_1) - V_1^{\hat{\pi}}(x_1)\right] \leq \mathscr{C}_{\mathcal{D},\pi^\star} \cdot \mathrm{poly}(d, H) \cdot \left(\sqrt{\varepsilon_{\mathsf{BE}}} + \frac{1}{\sqrt{n}}\right).$$

Moreover, Algorithm 1 runs in time $\mathrm{poly}(d, H, n)$.

A notable feature of Theorem 1.1 is the fact that the suboptimality of $\hat{\pi}$ scales with $\sqrt{\varepsilon_{\mathsf{BE}}}$, which contrasts with the *linear* scaling in $\varepsilon_{\mathsf{BE}}$ seen in classic works on offline RL (Munos & Szepesvári, 2008; Munos, 2005) and recent works studying *online RL* under low inherent Bellman error (Zanette et al., 2020a;b), as well as linear dependence on the misspecification error for other types of misspecification in both online and offline RL settings (Xie et al., 2021a; Zanette et al., 2021; Nguyen-Tang & Arora, 2023; Jin et al., 2020b; Wei & Luo, 2021).[1] We show below that, perhaps surprisingly, the square-root dependence on $\varepsilon_{\mathsf{BE}}$ cannot be improved, even in a statistical sense:

Theorem 1.2 (Informal version of Theorem 4.1)**.** *Fix $\varepsilon_{\mathsf{BE}} \in (0, 1)$, $n \in \mathbb{N}$, and set $d = H = 2$. There are feature mappings $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, $h \in [H]$, where $|\mathcal{A}| = 4$, so that for any (randomized) offline RL algorithm $\mathfrak{A}$, the following holds. There is some MDP with inherent Bellman error bounded by $\varepsilon_{\mathsf{BE}}$, together with some policy $\pi^\star$ so that $\mathscr{C}_{\mathcal{D},\pi^\star} = O(1)$ yet the output policy $\hat{\pi}$ of $\mathfrak{A}$ satisfies*

$$\mathbb{E}\left[V_1^{\pi^\star}(x_1) - V_1^{\hat{\pi}}(x_1)\right] \geq \Omega\left(\sqrt{\varepsilon_{\mathsf{BE}}} + \frac{1}{\sqrt{n}}\right).$$

We emphasize that Theorem 1.2 establishes a surprising separation between offline and online RL: whereas in the online setting, as mentioned above, one can learn a policy whose suboptimality scales linearly with $\varepsilon_{\mathsf{BE}}$, the optimal scaling in the offline setting is linear in $\sqrt{\varepsilon_{\mathsf{BE}}}$. Thus, *misspecification is more expensive in the offline setting*, i.e., when one is not allowed to adaptively gather data.

Relation to prior work. Wang et al. (2021); Zanette (2021) showed exponential lower bounds for offline RL under the assumption of *all-policy realizability*, which stipulates that the Q-value function of all policies is linear (i.e., belongs to $\mathcal{Q}_h$). This lower bound is incomparable to that of Theorem 1.2: whereas the inherent Bellman error of the instances in Wang et al. (2021); Zanette (2021) satisfies $\varepsilon_{\mathsf{BE}} = \Omega(1)$ (so that lower bounds of $\sqrt{\varepsilon_{\mathsf{BE}}}$ and $\varepsilon_{\mathsf{BE}}$ are indistinguishable), the instance used to prove Theorem 1.2 does not satisfy all-policy realizability. Moreover, the lower bounds are unrelated on a technical level.

A recent line of work has investigated a strengthening of all-policy realizability under which offline RL can be achieved, known as *Bellman restricted closedness* (or often simply as *(policy) completeness*). Under this condition, there are statistically efficient algorithms for offline RL with only single-policy coverage, for general function classes $\mathcal{Q}_h$ (Zanette et al., 2021; Xie et al., 2021a; Cheng et al., 2022; Nguyen-Tang & Arora, 2023). Moreover, given a regression oracle for the class $\mathcal{Q}_h$ which implements a variant of regularized least-squares, many of these works (Xie et al., 2021a; Cheng et al., 2022; Nguyen-Tang & Arora, 2023) show that the same offline RL guarantee can be obtained in an oracle-efficient manner. Since regularized least-squares is computationally efficient when $\mathcal{Q}_h$ is linear, it follows that, under Bellman restricted closedness computationally efficient offline RL algorithms are known (Xie et al., 2021a; Cheng et al., 2022; Nguyen-Tang & Arora, 2023; Zanette et al., 2021).

For a class Π of policies, an MDP satisfies Π-Bellman restricted closedness if the Bellman backup of any function in $\mathcal{Q}_{h+1}$, *with respect to any policy in* Π, belongs to $\mathcal{Q}_h$ (see Definition F.2). Generally

[1] Note that, in Xie et al. (2021a), the realizability error $\varepsilon_{\mathcal{F}}$ and policy completeness error $\varepsilon_{\mathcal{F},\mathcal{F}}$ appear in a square root (see Theorem 3.1 of Xie et al. (2021a)), since the quantities $\varepsilon_{\mathcal{F}}, \varepsilon_{\mathcal{F},\mathcal{F}}$ actually represent the mean *squared* errors. Moreover, the cube-root dependence on $\varepsilon_{\mathcal{F}}, \varepsilon_{\mathcal{F},\mathcal{F}}$ in Xie et al. (2021a) is suboptimal and is improved in Nguyen-Tang & Arora (2023, Theorem 2).

speaking, the above papers achieving the strongest bounds assume Π-Bellman restricted closedness for the class Π^{soft} of softmax policies (Definition F.1) (Zanette et al., 2021; Nguyen-Tang & Arora, 2023). While technically speaking such an assumption is incomparable with linear Bellman completeness (i.e., 0 inherent Bellman error), the latter has the advantage of being universal in the sense that it is defined so as to allow the fundamental approach of *value iteration* to succeed. In contrast, Π-Bellman restricted closedness only enjoys a similar "universality" property when one takes Π to be the class of *all Markov policies*, in which case Π-Bellman restricted closedness allows *policy iteration* to succeed (Du et al., 2020, Theorem D.1). However, as we discuss further in Appendix A, in this case Π-Bellman restricted closendess becomes significantly stronger than linear Bellman completeness: in fact, if each state has two distinct feature vectors, then it becomes equivalent to the linear MDP assumption (Jin et al., 2020a, Proposition 5.1).

Remark 1.1 (Confluence of terminology)**.** Due to an unfortunate confluence of terminology, some prior works in the literature (e.g., Uehara & Sun (2022)) refer to the setting of Π-Bellman restricted closedness when the classes $\mathcal{Q}_h$ are linear as "linear Bellman completeness". We do not use this convention, and use "linear Bellman completeness" to refer to the setting when the inherent Bellman error is 0.

Finally, we remark that in the construction used to prove Theorem 1.2, the comparator policy $\pi^\star$ is not the optimal policy in the MDP. This observation suggests the following intriguing open problem: if one further assumes that $\pi^\star$ is the optimal policy, then can one improve the $\sqrt{\varepsilon_{\mathsf{BE}}}$ upper bound in Theorem 1.1 to be linear in $\varepsilon_{\mathsf{BE}}$, or does an analogue of Theorem 1.2 continue to hold?

Organization of the paper. In Section 2, we introduce preliminaries. In Section 3 we state and discuss our upper bound, Theorem 3.1 (the formal version of Theorem 1.1), and in Section 4 we state and discuss our lower bound, Theorem 4.1 (the formal version of Theorem 1.2). Appendix A contains a detailed discussion of related work. The full proof of our upper bound is provided in Appendices B and C, and the full proof of our lower bound is provided in Appendix D. Finally, Appendices E and F contain additional useful lemmas.

2 Preliminaries

We consider the standard setting of a *finite-horizon Markov decision process*, which consists of a tuple $M = (H, \mathcal{X}, \mathcal{A}, (P_h^M)_{h=1}^H, (r_h^M)_{h=1}^H, x_1)$, where $H \in \mathbb{N}$ denotes the *horizon*, $\mathcal{X}$ denotes the *state set*, $\mathcal{A}$ denotes the *action set*, $P_h^M(\cdot \mid x, a) \in \Delta(\mathcal{X})$ denote the *transition kernels* (for $h \in [H]$), $r_h^M : \mathcal{X} \times \mathcal{A} \to [0,1]$ denote the *reward functions* (for $h \in [H]$), and $x_1 \in \mathcal{X}$ denotes the initial state. We omit the superscript M from these notations when its value is clear.

A *Markov policy* (or simply *policy*) π is a tuple $\pi = (\pi_1, \dots, \pi_H)$, where $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$ for each $h \in [H]$. We let Π^{M} denote the set of Markov policies. A policy $\pi \in \Pi^{\mathsf{M}}$ defines a distribution over *trajectories* $(x_1, a_1, r_1, \dots, x_H, a_H, r_H) \in (\mathcal{X} \times \mathcal{A} \times [0,1])^H$, in the following manner: for each $h \in [H]$, given the state x_h, an action a_h is drawn according to $a_h \sim \pi_h(x_h)$, a reward $r_h(x_h, a_h)$ is received, and the subsequent state x_{h+1} is generated according to $x_h \sim P_h^M(\cdot \mid x_h, a_h)$. We use the notation $\mathbb{E}^{M,\pi}[\cdot]$ to denote expectaton under the draw of a trajectory from policy π in the MDP M, and we write $\mathbb{E}^{\pi}[\cdot]$ if the value of M is clear.

Fix an MDP M. The *Q-value function* and *V-value function* associated to a policy $\pi \in \Pi^{\mathsf{M}}$ in MDP M are defined as follows: for $h \in [H], x \in \mathcal{X}, a \in \mathcal{A}$,

$$Q_h^\pi(x,a) := r_h(x,a) + \mathbb{E}^\pi\left[\sum_{g=h+1}^{H} r_g(x_g, a_g) \mid (x_h, a_h) = (x, a)\right], \qquad V_h^\pi(x) := Q_h^\pi(x, \pi_h(x)),$$

where for a function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we write $Q(x, \pi_h(x)) := \mathbb{E}_{a \sim \pi_h(x)}[Q(x,a)]$. We use the convention that all rewards and value functions evaluate to 0 at step $H+1$.

Given $h \in [H]$, a function $Q_{h+1} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and a policy $\pi \in \Pi^{\mathsf{M}}$, the *Bellman backup of Q_{h+1} with respect to π* is the function $Q_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined by $Q_h(x,a) := r_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a)}[Q_{h+1}(x', \pi_{h+1}(x'))]$. It is straightforward to see that, for any $\pi \in \Pi^{\mathsf{M}}$, Q_h^π is the Bellman backup of Q_{h+1}^π with respect to π, for each $h \in [H]$.

The *optimal policy* $\pi^{\mathsf{opt}} \in \Pi^{\mathsf{M}}$ is defined as $\pi^{\mathsf{opt}} := \arg\max_{\pi \in \Pi^{\mathsf{M}}} V_1^\pi(x_1)$.

2.1 Inherent Bellman Error

MDPs encountered in practical scenarios tend to have enormous state and action spaces. To address this challenge, it is common to use *function approximation* assumptions, which consider function classes $\mathcal{Q}_h \subset \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ and posit that the value functions for the optimal policy belong to $\mathcal{Q}_h$, i.e., $Q_h^{\pi^{\mathsf{opt}}} \in \mathcal{Q}_h$ for $h \in [H]$. As our goal is to obtain provable *end-to-end computationally efficient algorithms* for offline RL, without reliance on intractable regression oracles, we focus on the setting when the classes $\mathcal{Q}_h$ are linear. In particular, for some dimension $d \in \mathbb{N}$ together with *known* feature mappings $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, we take $\mathcal{Q}_h := \{(x,a) \mapsto \langle \phi_h(x,a), w \rangle \ : \ w \in \mathbb{R}^d\}$. For simplicity of notation, we use the convention that $\phi_{H+1}(x,a)$ is the all-zeros vector for each x, a.

Generally speaking, prior work on offline RL in the linear setting (Zanette et al., 2021; Xie et al., 2021a; Cheng et al., 2022; Gabbianelli et al., 2023; Nguyen-Tang & Arora, 2023) interprets elements of the classes $\mathcal{Q}_h$ as approximations of the Q-value functions Q_h^π, for all π belonging to some subset $\Pi \subseteq \Pi^{\mathsf{M}}$ consisting of policies whose values the learning algorithm wishes to compete with. Accordingly, these works make the assumption of Π-*realizability*, which posits that for all $\pi \in \Pi$ and $h \in [H]$, $Q_h^\pi \in \mathcal{Q}_h$. This assumption is natural in that it is sufficient for correctness of the *Least-Squares Policy Iteration (LSPI)* algorithm (Lagoudakis & Parr, 2003) (which assumes knowledge of the transitions and rewards of M) for finding an optimal policy (Du et al., 2020, Theorem D.1). However, as shown in Wang et al. (2021); Amortila et al. (2020); Zanette (2021), Π-realizability alone is insufficient for offline RL to succeed with polynomial sample complexity under our desired single-policy coverage condition.[2] Thus, it is typical to make the stronger assumption of Π-*Bellman restricted closedness* (Zanette et al., 2021) (also known as *policy completeness*), namely that for all $\pi \in \Pi$ and $Q_{h+1} \in \mathcal{Q}_{h+1}$, there is some $Q_h \in \mathcal{Q}_h$ so that $Q_h(x,a) = \mathbb{E}_{x' \sim P_h(x,a)}[r_h(x,a) + Q_{h+1}(x', \pi_{h+1}(x'))]$.

Inherent Bellman error. Bellman-restricted closedness is an unwieldy assumption in that it requires quantifying over both a policy and a value function at step $h+1$. In this work, we study offline RL under the alternative assumption of *low inherent Bellman error* (Zanette et al., 2020a), as defined in Assumption 2.1 below. For $h \in [H]$, write $\mathcal{B}_h := \{w \in \mathbb{R}^d : \ |\langle \phi_h(x,a), w \rangle| \leq 1 \ \forall (x,a) \in \mathcal{X} \times \mathcal{A}\}$.

Assumption 2.1 (Low Inherent Bellman Error; Zanette et al. (2020a))**.** *We say that a MDP M has* inherent Bellman error $\varepsilon_{\mathsf{BE}}$ *if for each $h \in [H]$, there is a mapping $\mathcal{T}_h : \mathcal{B}_{h+1} \to \mathcal{B}_h$ so that*

$$\sup_{\theta \in \mathcal{B}_{h+1}} \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \left| \langle \phi_h(x,a), \mathcal{T}_h \theta \rangle - \mathbb{E}_{x' \sim P_h(x,a)} \left[r_h(x,a) + \max_{a' \in \mathcal{A}} \langle \phi_{h+1}(x', a'), \theta \rangle \right] \right| \leq \frac{\varepsilon_{\mathsf{BE}}}{2}. \quad (1)$$

If M has inherent Bellman error $\varepsilon_{\mathsf{BE}} = 0$, then we say that M is linear Bellman complete.

The assumption of linear Bellman completeness is sufficient for correctness of the *Least-Squares Value Iteration (LSVI)* algorithm (Munos & Szepesvári, 2008; Munos, 2005), which assumes knowledge of the transitions and rewards of M. This fact has made it a popular assumption under which to study *online RL*, for which algorithms typically proceed via approximate variants of LSVI (Zanette et al., 2020a;b). In contrast, as recent offline RL algorithms typically bear more resemblance to *LSPI* (Zanette et al., 2021; Xie et al., 2021a; Cheng et al., 2022; Gabbianelli et al., 2023; Nguyen-Tang & Arora, 2023), offline RL has not previously been studied under linear Bellman completeness as opposed to Bellman restricted closedness. As discused in Section 3, one of the contributions of this work is to draw connections between these two types of assumptions.

[2] In fact, these results rule out offline RL even under a stronger *all-policy* coverage condition.

It is convenient to separate the components of $\mathcal{T}_h$ capturing the rewards at step h and the transitions at step h, as follows: an immediate consequence of Assumption 2.1 (see Zanette et al. (2021, Proposition 2)) is that there are mappings $\mathcal{T}_h^\circ : \mathcal{B}_{h+1} \to \mathcal{B}_h$ for each $h \in [H-1]$ and a vector $\theta_h^{\mathsf{r}} \in \mathcal{B}_h$ for each $h \in [H]$ so that the below inequalities hold:

$$\sup_{\theta \in \mathcal{B}_{h+1}} \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \left| \langle \phi_h(x,a), \mathcal{T}_h^\circ \theta \rangle - \mathbb{E}_{x' \sim \mathbb{P}_h(x,a)} \left[\max_{a' \in \mathcal{A}} \langle \phi_{h+1}(x',a'), \theta \rangle \right] \right| \leq \varepsilon_{\mathsf{BE}} \tag{2}$$

$$\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |r_h(x,a) - \langle \phi_h(x,a), \theta_h^{\mathsf{r}} \rangle| \leq \varepsilon_{\mathsf{BE}}. \tag{3}$$

It was observed in Zanette et al. (2020a, Proposition 5) that, in general, the assumptions of linear Bellman completeness and Π^{M}-realizability are incomparable, in that neither one implies the other. Moreover, it is straightforward to see that linear Bellman completeness is a strictly weaker condition than Π^{M}-Bellman restricted closedness. For an arbitrary subset of policies $\Pi \subset \Pi^{\mathsf{M}}$, linear Bellman completeness may be incomparable to the assumption of Π-Bellman restricted closedness.

Finally, we make the following standard boundedness assumptions.

Assumption 2.2 (Boundedness)**.** *We assume the following:*

1. *For all $h \in [H], x \in \mathcal{X}, a \in \mathcal{A}$, we have $\|\phi_h(x,a)\|_2 \leq 1$.*
2. *For some parameter $B \in \mathbb{R}_+$: for all $w_h \in \mathcal{B}_h$, it holds that $\|w_h\|_2 \leq B$.*
3. *For all $h \in [H]$, $\|\theta_h^{\mathsf{r}}\|_2 \leq 1$ (and hence $\sup_{x,a,h} |r_h(x,a)| \leq 1$).*

The assumption that $\|\phi_h(x,a)\|_2 \leq 1$ together with the definition of $\mathcal{B}_h$ ensures that $\mathcal{B}_h$ contains a ball of radius 1 centered at the origin.

2.2 Perturbed linear policies

Given $w \in \mathbb{R}^d, h \in [H], x \in \mathcal{X}$, define $\mathcal{A}_{h,w}(x) := \arg\max_{a \in \mathcal{A}} \langle w, \phi_h(x,a) \rangle \subset \mathcal{A}$, where arg max is interpreted as the set of all actions maximizing $\langle w, \phi_h(x,a) \rangle$.

Definition 2.1 (Perturbed linear policies)**.** For $\sigma > 0$, $h \in [H]$ and $w \in \mathbb{R}^d$, define $\pi_{h,w,\sigma} : \mathcal{X} \to \Delta(\mathcal{A})$ by

$$\pi_{h,w,\sigma}(x)(a) = \Pr_{\theta \sim \mathcal{N}(w, \sigma^2 \cdot I_d)} \left(a \in \mathcal{A}_{h,\theta}(x) \right).$$

In words, to draw an action $a \sim \pi_{h,w,\sigma}(x)$, we draw $\theta \sim \mathcal{N}(w, \sigma^2 \cdot I_d)$ and then play $\arg\max_{a' \in \mathcal{A}} \langle \phi_h(x,a'), \theta \rangle$. We extend to the case that $\sigma = 0$ by taking a limit, i.e., define $\pi_{h,w,0}(x)(a) := \lim_{\sigma \downarrow 0} \pi_{h,w,\sigma}(x)(a)$ (it is straightforward to see that this limit is well-defined). Given $\sigma \geq 0$, we denote the set of all $\pi_{h,w,\sigma'}$, where $w \in \mathbb{R}^d, \sigma' \geq 0$ satisfy $\sigma'/\|w\|_2 \geq \sigma$, by $\Pi_h^{\mathsf{Plin},\sigma}$, and $\Pi^{\mathsf{Plin},\sigma} := \prod_{h=1}^H \Pi_h^{\mathsf{Plin},\sigma}$. Moreover, we write $\Pi_h^{\mathsf{Plin}} := \Pi_h^{\mathsf{Plin},0} = \bigcup_{\sigma \geq 0} \Pi_h^{\mathsf{Plin},\sigma}$ and $\Pi^{\mathsf{Plin}} := \Pi^{\mathsf{Plin},0} = \bigcup_{\sigma \geq 0} \Pi^{\mathsf{Plin},\sigma}$.

Note that, for any $c > 0$, $\pi_{h,cw,\sigma} = \pi_{h,w,\sigma/c}$. We refer to the policies in Π^{Plin} as *perturbed linear policies*. Given a (possibly randomzied) policy $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$, we use the convention that $\phi_h(x, \pi_h(x))$ refers to $\mathbb{E}_{a \sim \pi_h(x)}[\phi_h(x,a)]$.

Gaussian smoothing. For $\theta \in \mathbb{R}^d$ and $\sigma > 0$, we write

$$\mathcal{N}_\sigma(\theta) := \mathcal{N}(0, \sigma^2 \cdot I_d)(\theta) = \frac{1}{(2\pi)^{d/2}\sigma^d} \cdot \exp\left(-\frac{1}{2\sigma^2} \|\theta\|_2^2 \right).$$

Furthermore, for $f : \mathbb{R}^d \to \mathbb{R}$, we write $\mathsf{S}_\sigma f(\theta)$ to denote the convolution of f with $\mathcal{N}_\sigma$, namely

$$\mathsf{S}_\sigma f(\theta) := \int_{\mathbb{R}^d} f(z) \mathcal{N}_\sigma(\theta - z) dz = \int_{\mathbb{R}^d} f(\theta - z) \mathcal{N}_\sigma(z) dz = \mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 \cdot I_d)}[f(\theta - z)]. \tag{4}$$

2.3 The offline learning problem

In the *offline learning model*, the algorithm is not allowed to interact with the environment. Rather, it is given a dataset $\mathcal{D}$ consisting of tuples (h, x, a, r, x'), where $r = r_h(x, a)$ and $x' \sim P_h(\cdot|x, a)$. We allow the values of h, x, a in the dataset $\mathcal{D}$ to be chosen in an arbitrary adaptive manner, as formalized by the following assumption:

Assumption 2.3. *We assume the dataset $\mathcal{D} = \{(h_i, x_i, a_i, r_i, x_i')\}_{i=1}^n$ is drawn from a distribution satisfying the following conditions. For $i \in [n]$, let $\mathcal{F}_i$ denote the sigma-algebra generated by $\{(h_j, x_j, a_j, r_j, x_j')\}_{j=1}^{i-1} \cup \{(h_i, x_i, a_i)\}$. We assume that, for each $i \in [n]$, conditioned on $\mathcal{F}_i$, the reward r_i is equal to $r_{h_i}(x_i, a_i) = \langle \phi_{h_i}(x_i, a_i), \theta^{\mathsf{r}}_{h_i} \rangle$, and $x_i' \sim P_{h_i}(x_i, a_i)$.*

We remark that Assumption 2.3 is essentially the same as Assumption 1 of Zanette et al. (2021). Based on the dataset $\mathcal{D}$, the algorithm must output a policy $\hat{\pi}$ whose value, $V_1^{\hat{\pi}}(x_1)$ is as large as possible. Of course, it may be impossible to make $V_1^{\hat{\pi}}(x_1)$ very large if the dataset $\mathcal{D}$ does not include states (h_i, x_i, a_i) which explore certain directions of the feature space $\mathbb{R}^d$. A large number of conditions have been proposed in the offline RL literature which capture the degree to which $\mathcal{D}$ exhibits good "coverage" properties of the state space. Our bounds depend on one of the weakest such conditions, namely the following variant of *single-policy coverage*, which is identical to that in Zanette et al. (2021).

Definition 2.2 (Coverage parameter)**.** Given a dataset $\mathcal{D}$ as in Assumption 2.3, we define $\Sigma_h := \sum_{i:h_i=h} \phi_{h_i}(x_i, a_i)\phi_{h_i}(x_i, a_i)^\top$. For a policy $\pi \in \Pi^{\mathsf{M}}$, its *coverage parameter* for the dataset $\mathcal{D}$ is

$$\mathscr{C}_{\mathcal{D},\pi} := \sum_{h=1}^{H} \|\mathbb{E}^\pi[\phi_h(x_h, a_h)]\|_{n\Sigma_h^{-1}}.$$

In words, $\mathscr{C}_{\mathcal{D},\pi}$ measures the degree to which an average feature vector drawn from π at each step h lines up with directions spanned by feature vectors in $\mathcal{D}$. We refer to Gabbianelli et al. (2023, Section 6), Nguyen-Tang & Arora (2023, Section 4), and Jiang (2023) for a detailed comparison between $\mathscr{C}_{\mathcal{D},\pi}$ and other coverage parameters considered in prior work. As discussed there, assuming boundedness of $\mathscr{C}_{\mathcal{D},\pi}$ is essentially the *weakest* coverage assumption in the literature, as many previous works (e.g., Jin et al. (2021)) require instead boundedness of $\mathbb{E}^\pi[\|\phi_h(x_h, a_h)\|_{n\Sigma_h^{-1}}]$, where the norm is *inside* the expectation.

3 The offline actor-critic algorithm

In this section, we discuss our main upper bound, Theorem 3.1, which shows a performance guarantee for the output policy $\hat{\pi}$ of the `Actor` algorithm (Algorithm 1).

Theorem 3.1. *Consider any dataset $\mathcal{D}$ with n examples drawn according to Assumption 2.3, as well as parameters $\varepsilon_{\mathsf{final}}, \delta \in (0, 1)$. Suppose $\varepsilon_{\mathsf{BE}} \le c_0(BH)^{-2}d^{-3}$ for some sufficiently small constant c_0. Then, for η set as prescribed in Definition C.1, $\texttt{Actor}(\mathcal{D}, \varepsilon_{\mathsf{final}}, \delta, \eta)$ (Algorithm 1) returns a policy $\hat{\pi}$ so that, with probability $1 - \delta$ the following holds: for any $\pi^\star \in \Pi$,*

$$\begin{aligned} &V_1^{\pi^\star}(x_1) - V_1^{\hat{\pi}}(x_1) \\ \le &O\left(d^{3/2}BH \cdot \varepsilon_{\mathsf{BE}}^{1/2}\log(1/\varepsilon_{\mathsf{BE}}) + \frac{BHd\log^{1/2}(dnBH/(\varepsilon_{\mathsf{final}}\delta))}{\sqrt{n}}\right) \cdot (H + \mathscr{C}_{\mathcal{D},\pi^\star}) + \varepsilon_{\mathsf{final}}. \end{aligned}$$

The overall computational cost of Algorithm 1 is bounded above by $\mathrm{poly}(d, H, n, \log(B/\delta), 1/\varepsilon_{\mathsf{final}})$.

3.1 High-level proof overview

A key ingredient in the proof of Theorem 3.1 is a new structural condition (Theorem 3.2 below) proving that MDPs with low inherent Bellman error satisfy Π-Bellman restricted closedness for the

class of *perturbed linear policies*. To understand this result, we first consider the special case of linear Bellman completeness, i.e., $\varepsilon_{\mathsf{BE}} = 0$. In this case, we show (as a special case of Theorem 3.2) that Π^{lin}-Bellman restricted closedness holds, where Π^{lin} denotes the class of linear policies, namely those of the form $x \mapsto \arg\max_{a \in \mathcal{A}} \langle \phi_h(x,a), \theta \rangle$, for some $\theta \in \mathbb{R}^d$. Unfortunately, when $\varepsilon_{\mathsf{BE}}$ is positive but small, Π^{lin}-Bellman restricted closedness may no longer hold even in an approximate sense. We correct for this deficiency by modifying the policy class Π^{lin} to instead consist of perturbed linear policies (Definition 2.1).

Theorem 3.2 ($\Pi^{\mathsf{Plin},\sigma}$-Bellman restricted closedness; informal version of Corollary B.2)**.** *Suppose that $\pi \in \Pi^{\mathsf{Plin},\sigma}$ and $h \in [H-1]$. Then there is a matrix $\mathcal{T}_h^\pi : \mathbb{R}^{d \times d}$ so that, for all $w \in \mathbb{R}^d$ and $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$\left| \langle \phi_h(x,a), \mathcal{T}_h^\pi w \rangle - \mathop{\mathbb{E}}_{\substack{x' \sim P_h(x,a) \\ a' \sim \pi_{h+1}(x')}} [\langle \phi_{h+1}(x',a'), w \rangle] \right| \leq \tilde{O}\left(\|w\|_2 d^{3/2} \cdot \left(\sqrt{d} + \frac{1}{\sigma} \right) \right) \cdot \varepsilon_{\mathsf{BE}}. \tag{5}$$

The proof of Theorem 3.2 proceeds by considering, for a pair $(x,a) \in \mathcal{X} \times \mathcal{A}$, the function $w \mapsto Q_h(w;x,a) := \mathbb{E}_{x' \sim P_h(x,a)}[\max_{a' \in \mathcal{A}} \langle \phi_{h+1}(x',a'), w \rangle]$. The key observation is that if $\pi_{h+1} = \pi_{h+1,w,\sigma}$ is the perturbed linear policy specified by $w \in \mathbb{R}^d, \sigma > 0$ (see Definition 2.1), then the second term on the left-hand side of (5) may be expressed as follows in terms of the gradient with respect to w of the Gaussian smoothing of Q:

$$\nabla_w \mathsf{S}_\sigma Q(x,a,w) = \mathbb{E}_{x' \sim P_h(x,a)}[\phi_{h+1}(x', \pi_{h+1,w,\sigma}(x'))]. \tag{6}$$

The existence of the desired matrix $\mathcal{T}_h^\pi$ as claimed by Theorem 3.2 then follows by using the fact that $|Q_h(w;x,a) - \langle \phi_h(x,a), \mathcal{T}_h^\circ w \rangle| \leq \varepsilon_{\mathsf{BE}}$ (see (2)) as well as the fact that differentiating is a linear operation. This argument must overcome a few challenges in the setting that $\varepsilon_{\mathsf{BE}}, \sigma > 0$ due to the fact that $\mathsf{S}_\sigma Q(x,a,w)$ is different from $Q(x,a,w)$; full details are given in Appendix B.

We proceed to discuss the remainder of the proof of Theorem 3.1. Previous work (Zanette et al., 2021) shows that, under Bellman restricted closedness with respect to a *softmax* policy class, an actor-critic method suffices to obtain offline RL guarantees under single-policy coverage. While Bellman restricted closedness does not hold in general for the softmax policy class under the assumption of linear Bellman completeness (see Lemma F.1), we prove Theorem 3.1 by adapting the actor-critic method in Zanette et al. (2021) to work instead with the set of perturbed linear policies. To explain the requisite modifications, we briefly overview the actor-critic method: roughly speaking, the overall goal is to solve the problem $\max_\pi \min_{M \in \mathcal{M}_\mathcal{D}(\pi)} V^{M,\pi}(x_1)$, where $\mathcal{M}_\mathcal{D}(\pi)$ indicates a set of MDPs which, under trajectories drawn from π, are statistically consistent with the dataset $\mathcal{D}$. Moreover, $V^{M,\pi}(x_1)$ denotes the value of policy π in MDP M. Minimization over $M \in \mathcal{M}_\mathcal{D}(\pi)$ corresponds to the pessimism principle, and standard arguments (Xie et al., 2021a; Zanette et al., 2021) show that an exact solution to this max-min problem would solve the offline RL task.

To solve this max-min problem in a computationally efficient manner, the actor-critic method uses the "no-regret learning vs. best response" approach: a sequence of policies $\pi^{(t)}$ and MDPs $M^{(t)} \in \mathcal{M}_\mathcal{D}(\pi^{(t)})$ is generated in the following manner. At each step t, a no-regret learning algorithm, called the `Actor`, generates a policy $\pi^{(t)}$; in response, an optimization algorithm, called the `Critic`, chooses $M^{(t)} \in \mathcal{M}_\mathcal{D}(\pi^{(t)})$ so as to minimize $V^{M^{(t)},\pi^{(t)}}(x_1)$. If T is chosen sufficiently large, then, as shown in Zanette et al. (2021), a policy drawn uniformly from $\{\pi^{(t)} : t \in [T]\}$ will have sufficiently large value in the true MDP.

3.2 Algorithm description

In our setting, the `Actor` algorithm and its associated `Critic` (Algorithm 2) function similarly to the `Actor` and `Critic` algorithms in Zanette et al. (2021). For an appropriate choice of the number of iterations T, `Actor` repeats the following steps T times: at each iteration $t \in [T]$, `Actor` lets $\theta_h^{(t)}$ be the sum of vectors $w_h^{(s)}$ for iterations $s < t$, and lets $\pi_h^{(t)}$ be a perturbed linear policy with

Algorithm 1 `Actor`$(\mathcal{D}, \varepsilon_{\mathsf{final}}, \delta, \eta)$

Require: Dataset $\mathcal{D} = \{(h_i, x_i, a_i, r_i, x_i')\}_{i=1}^n$; parameters $\varepsilon_{\mathsf{final}}, \delta \in (0,1)$ and $\eta > 0$.

1: Define $T, \varepsilon_{\mathsf{apx}}, \alpha, \beta$ as a function of $n, \varepsilon_{\mathsf{final}}, \delta$ per Definition C.1.
2: **for** $1 \le t \le T+1$ **do**
3: **for** $1 \le h \le H$ **do**
4: Set $\theta_h^{(t)} = \sum_{s=1}^{t-1} w_h^{(s)}$.
5: Define $\pi_h^{(t)} := \pi_{h,\theta_h^{(t)},\eta}$. ▷ *(See Definition 2.1)*
6: Set $(w_1^{(t)}, \dots, w_H^{(t)}) =$ `Critic`$(\mathcal{D}, \pi^{(t)}, \varepsilon_{\mathsf{apx}}, \alpha, \beta, \delta/(2T))$, where $\pi^{(t)} = (\pi_1^{(t)}, \dots, \pi_H^{(t)})$. ▷ *Algorithm 2*
7: **return** a policy $\hat{\pi}$ drawn as $\hat{\pi} \sim \mathrm{Unif}(\{\pi^{(1)}, \dots, \pi^{(T)}\})$.

Algorithm 2 `Critic`$(\mathcal{D}, \pi, \varepsilon_{\mathsf{apx}}, \alpha, \beta, \delta)$

Require: Dataset $\mathcal{D} = \{(h_i, x_i, a_i, r_i, x_i')\}_{i=1}^n$; policy $\pi \in \Pi^{\mathsf{Plin}}$; parameters $\varepsilon_{\mathsf{apx}}, \alpha, \beta > 0$.

1: For each $h \in [H]$, let $\mathcal{I}_h \subset [n]$ denote the set of i so that $h_i = h$.
2: For $h \in [H]$, define $\Sigma_h = I_d + \sum_{i \in \mathcal{I}_h} \phi_h(x_i, a_i)\phi_h(x_i, a_i)^\top$.
3: For $i \in [n]$, set $\hat{\phi}_i^\pi \leftarrow$ `EstFeature`$(x_i', \pi, h_i + 1, \varepsilon_{\mathsf{apx}}, \delta/n)$. ▷ *Algorithm 3*
4: Solve the following convex program with variables $w, \xi \in \mathbb{R}^{dH}$:

$$\min \langle w_1, \phi_1(x_1, \pi_1(x_1)) \rangle \tag{7a}$$

$$\text{s.t. } w_h = \xi_h + \Sigma_h^{-1} \sum_{i \in \mathcal{I}_h} \phi_h(x_i, a_i) \cdot \left(r_i + \langle \hat{\phi}_i^\pi, w_{h+1} \rangle \right) \quad \forall h \in [H] \tag{7b}$$

$$\|\xi_h\|_{\Sigma_h}^2 \le \alpha^2 \quad \forall h \in [H] \tag{7c}$$

$$\|w_h\|_2^2 \le \beta^2 \quad \forall h \in [H]. \tag{7d}$$

5: **return** the solution $w = (w_1, \dots, w_H) \in \mathbb{R}^{dH}$ of the convex program.

Algorithm 3 `EstFeature`$(x, \pi, h, \varepsilon_{\mathsf{apx}}, \delta)$

Require: State $x \in \mathcal{X}$, policy $\pi \in \Pi^{\mathsf{Plin}}$, step $h \in [H]$, error $\varepsilon_{\mathsf{apx}} \in (0,1)$, failure probability $\delta \in (0,1)$.

1: Choose $w \in \mathbb{R}^d, \sigma > 0$ so that $\pi_h = \pi_{h,w,\sigma}$ ▷ *(This is possible by definition of* Π^{Plin}*).*
2: Choose $N \leftarrow 2\varepsilon_{\mathsf{apx}}^{-2} \log(2d/\delta)$.
3: Draw N samples $\theta_1, \dots, \theta_N \sim \mathcal{N}(w, \sigma^2 \cdot I_d)$.
4: **return** $\hat{\phi} := \frac{1}{N} \sum_{i=1}^N \phi_h(x, \pi_{h,\theta_i}(x))$.

mean vector $\theta_h^{(t)}$. The vector $\theta_h^{(t)}$ represents an aggregation of the pessimistic estimates of the value function produced by the `Critic` at previous iterations. In turn, at time step t, the `Critic` (Algorithm 2), given $\pi^{(t)} = (\pi_1^{(t)}, \dots, \pi_H^{(t)})$ as input, then produces vectors $w_h^{(t)} \in \mathbb{R}^d$ for $h \in [H]$, which constitute a solution to the optimization problem (7). The program (7) has as its objective to minimize the value at the initial state (namely, (7a)), subject to constraints that force $w_h^{(t)}$ to have Bellman backups with respect to $\pi^{(t)}$ which are consistent with $\mathcal{D}$ (Eqs. (7b) to (7d)).

The main difference between Algorithms 1 and 2 and the approach in Zanette et al. (2021) is the choice of the policies $\pi_h^{(t)}$ in the `Actor` (Algorithm 1). While we take $\pi_h^{(t)}$ to be a perturbed linear policy corresponding to $\theta_h^{(t)}$ and an appropriate choice of the noise parameter η, in Zanette et al.

(2021), $\pi_h^{(t)}$ was taken to be a softmax policy corresponding to $\theta_h^{(t)}$, with an appropriate choice of temperature. As we discuss further in Appendix C.2, our choice of $\pi_h^{(t)}$ implicitly leads $\pi_h^{(t)}(x)$ to implement the no-regret *follow-the-perturbed-leader* algorithm at each state x. In contrast, the softmax policy from Zanette et al. (2021) corresponds to the exponential weights algorithm. This difference is crucial to allow us to use Theorem 3.2 to ensure that the program (7) is feasible and outputs a pessimistic estimate of the value function $V_1^{\pi^{(t)}}(x_1)$, for each $t \in [T]$. The settings of the parameters in our algorithms are given in Definition C.1 in the appendix. We remark here that the optimal choice of the size η of the perturbation turns out to scale proportionally to $\sqrt{\varepsilon_{\mathsf{BE}}}$, which leads to the scaling of the error with $\sqrt{\varepsilon_{\mathsf{BE}}}$ in Theorem 3.1. The full details of the proof of Theorem 3.1 may be found in Appendix C.

Remark 3.1 (Relation to prior work: Bellman restricted closedness)**.** We point out that many works on offline RL consider similar actor-critic methods to ours in the setting of general function approximation (Zanette et al., 2021; Xie et al., 2021a; Cheng et al., 2022; Nguyen-Tang & Arora, 2023), generally under the assumption of Π-Bellman restricted closedness for various policy classes Π. While some of these results are sufficiently general to be instantiated in our setting (of low inherent Bellman error) using Theorem 3.2, which establishes (approximate) $\Pi^{\mathsf{Plin},\sigma}$-Bellman restricted closedness when $\varepsilon_{\mathsf{BE}}$ is small, none of the resulting implications will be as strong as Theorem 3.1, either suffering from suboptimal statistical rates or computational intractability. We defer detailed discussion to Appendix A.

4 Lower bounds

In this section we state Theorem 1.2 formally below as Theorem 4.1, and sketch its proof (which is provided in full in the appendix).

Theorem 4.1. *Let $\varepsilon_{\mathsf{BE}} \in (0,1)$ and $n \in \mathbb{N}$ be given, and set $d = H = 2$. Then there are state and action spaces $\mathcal{X}, \mathcal{A}$ with $|\mathcal{A}| = 4$ as well as feature mappings $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ ($h \in [H]$), so that the following holds, for any offline RL algorithm $\mathfrak{A}$. There is some MDP M which has inherent Bellman error with respect to the feature mappings ϕ_h bounded by $2\varepsilon_{\mathsf{BE}}$ and which satisfies Assumption 2.2, some distribution over datasets $\mathcal{D}$ satisfying Assumption 2.3, and some policy $\pi^\star \in \Pi^{\mathsf{Plin}}$ so that: $\mathscr{C}_{\mathcal{D},\pi^\star} = O(1)$ with probability 1 over the draw of $\mathcal{D}$ yet the output policy $\hat{\pi}$ of $\mathfrak{A}$ satisfies*

$$\mathbb{E}\left[V_1^{\pi^\star}(x_1) - V_1^{\hat{\pi}}(x_1)\right] \geq \Omega\left(\sqrt{\varepsilon_{\mathsf{BE}}} + \frac{1}{\sqrt{n}}\right), \tag{8}$$

where the expectation is over the draw of $\mathcal{D}$ and the randomness in $\mathfrak{A}$.

Proof overview for Theorem 4.1. To explain the idea behind Theorem 4.1, we first consider the following "toy setup": suppose that $H = d = 2$, $\mathcal{A} = \{0,1\}$, that rewards at step $h = 1$ are known to be 0, and rewards at step $h = 2$ are known to be induced by a coefficient vector $\theta_2^{\mathsf{r}} \in \{(1,1),(-1,1)\}$. Concretely, this type of uncertainty in θ_2^{r} may be implemented by having a dataset $\mathcal{D}$ for which all feature vectors at step 2 are parallel to $(0,1)$; thus, the first coordinate of θ_2^{r} may remain unknown.

Consider an MDP with states $x_2, x_2' \in \mathcal{X}$ so that, for each $a \in \mathcal{A}$, (x_1, a) transitions to either x_2 or x_2', but which one is unknown. Moreover suppose that feature vectors at x_2, x_2' are given as follows: for $a \in \{0,1\}$,

$$\phi_2(x_2, a) = (1-2a, 1-(1-2a)\cdot\varepsilon_{\mathsf{BE}}/2), \quad \phi_2(x_2', a) = (1-2a, 1+(1-2a)\cdot\varepsilon_{\mathsf{BE}}/2). \tag{9}$$

We may ensure that the transitions described above are consistent with the requirement that the inherent Bellman error be bounded by $\varepsilon_{\mathsf{BE}}$, since the feature vector $\phi_2(x_2, a)$ is within distance $\varepsilon_{\mathsf{BE}}$ from $\phi_2(x_2', a)$ for each a.

At a high level, our lower bound results from the following consideration: suppose the policy we wish to compete with at step 2, namely $\pi_2^\star$, takes a uniformly random action at step 2 (so that its expected feature vector at step 2 is $(0,1)$). Not knowing any information about the first coordinate of θ_2^{r}, a

natural choice for $\hat{\pi}_2$ is the "naive" policy that maximizes the reward in the one "known" direction $(0,1)$, i.e., let $\hat{\pi}_2^{\mathsf{naive}}$ be the linear policy which, at state x, chooses $\arg\max_{a\in\mathcal{A}}\langle\phi_h(x,a),(0,1)\rangle$. However, this choice suffers from the issue that, due to the $\varepsilon_{\mathsf{BE}}$ perturbation between $\phi_2(x_2,a)$ and $\phi_2(x_2',a)$, $\hat{\pi}_2^{\mathsf{naive}}$ will choose an action a at step 2, whose feature vector is either $(1,1+\varepsilon_{\mathsf{BE}}/2)$ (if x_1 transitions to x_2') or $(-1,1+\varepsilon_{\mathsf{BE}}/2)$ (if x_1 transitions to x_2). By choosing $\theta_2^{\mathsf{r}}=(-1,1)$ in the former case and $\theta_2^{\mathsf{r}}=(1,1)$ in the latter case, we can thus force this naive choice $\hat{\pi}_2^{\mathsf{naive}}$ to have suboptimality $-\Omega(1)$ compared to $\pi_2^\star$. This issue stems from the fact that the policy $\hat{\pi}_2^{\mathsf{naive}}$ is extremely brittle to small perturbations of $\phi_2(x_2,a)$: a change of size $\varepsilon_{\mathsf{BE}}$ in each of the feature vectors leads to a $\Omega(1)$-size change in the actual feature vector chosen by $\hat{\pi}_2^{\mathsf{naive}}$.

Of course, in this specific example, one can simply instead set $\hat{\pi}_2$ to be the policy which chooses each action at step 2 with probability $1/2$, which will lead to a policy $\hat{\pi}$ whose value is within $O(\varepsilon_{\mathsf{BE}})$ of $\pi^\star$. Roughly speaking, doing so corresponds to considering a perturbed linear policy, in the vein of Theorem 3.2. We can show, however, that such a perturbation must hurt the value of $\hat{\pi}$, for some alternative choice of MDP. Formally, we add transitions to states $x_2^{(\ell)}$ at step 2 with state-action feature vectors $\phi_2(x_2^{(\ell)},a)=(1-2a,1\pm(1-2a)\cdot\ell\cdot\varepsilon_{\mathsf{BE}})$, for each value of $\ell\in\{1,2,\ldots,\lfloor 1/\sqrt{\varepsilon_{\mathsf{BE}}}\rfloor\}$. A suboptimality of $\Omega(\sqrt{\varepsilon_{\mathsf{BE}}})$, with respect to some reference policy $\pi^\star$, arises because avoiding it would require $\hat{\pi}_2$ to act in a way consistent with the naive policy $\hat{\pi}_2^{\mathsf{naive}}$ (i.e., without perturbation) at states with features ϕ_ℓ, for $\ell=\Omega(1/\sqrt{\varepsilon_{\mathsf{BE}}})$. (At such states, the second component of the feature vectors deviates from 1 by enough that it cannot be "ignored".) But because the MDP has inherent Bellman error of $\varepsilon_{\mathsf{BE}}$, similar reasoning to the previous paragraph shows that the algorithm's output policy $\hat{\pi}$ cannot act significantly differently at states with feature vectors $\phi_2(x_2^{(\ell)},a),\phi_2(x_2^{(\ell-1)},a)$, for each value of ℓ and a. Since, per the previous paragraph, the algorithm must add large perturbations to $\hat{\pi}_2$ for $\ell=O(1)$, we will ultimately reach a contradiction. There are many details we have omitted from this high-level description, such as ensuring that $\mathcal{D}$ satisfies the requisite coverage property with respect to $\pi^\star$, and the fact that we wish to allow the algorithm to output arbitrary choices of $\hat{\pi}$, and not just perturbed linear policies – the full details are in Appendix D.

Acknowledgements

We thank Dylan Foster and Sham Kakade for bringing this problem to our attention and for helpful discussions. NG is supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship. AM is supported in part by a Microsoft Trustworthy AI Grant, an ONR grant and a David and Lucile Packard Fellowship.

References

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the wang-foster-kakade lower bound for the discounted setting, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. In *Proceedings of the 19th Annual Conference on Learning Theory*, COLT'06, pp. 574–588, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540352945.

Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial model for offline reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Adam Block, Ali Jadbabaie, Daniel Pfrommer, Max Simchowitz, and Russ Tedrake. Provable guarantees for generative behavior cloning: Bridging low-level stability and high-level behavior. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jonathan Daniel Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. In A. Beygelz-

imer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 09–15 Jun 2019.

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3852–3878. PMLR, 17–23 Jul 2022.

Qiwei Di, Heyang Zhao, Jiafan He, and Quanquan Gu. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning, 2023.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *CoRR*, abs/2111.10919, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 09–15 Jun 2019.

Germano Gabbianelli, Gergely Neu, Nneka Okolo, and Matteo Papini. Offline primal-dual reinforcement learning for linear MDPs. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Noah Golowich and Ankur Moitra. Linear bellman completeness suffices for efficient online reinforcement learning with few actions. In *The Thirty-Seventh Annual Conference on Learning Theory*. PMLR, 2024.

Elad Hazan. *Introduction to Online Convex Optimization*. Foundations and Trends in Optimization. Now, Boston, 2017. ISBN 978-1-68083-170-2.

Nan Jiang. What's the right notion of coverage in linear mdp? In *X (Twitter)*, pp. https://twitter.com/nanjiang_cs/status/1666994607073230848, 2023.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020a.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096. PMLR, 18–24 Jul 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *J. Mach. Learn. Res.*, 4 (null):1107–1149, dec 2003. ISSN 1532-4435.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, pp. 1006–1011. AAAI Press, 2005. ISBN 157735236x.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Asuman E. Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline RL with general function approximation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26769–26791. PMLR, 23–29 Jul 2023.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11702–11716. Curran Associates, Inc., 2021.

Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline RL with general function approximation via augmented lagrangian. In *The Eleventh International Conference on Learning Representations*, 2023.

Stéphane Ross and J. Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1905–1912, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19967–20025. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/shi22c.html`.

John N. Tsitsiklis and Benjamin van Roy. Feature-based methods for large scale dynamic programming. *Mach. Learn.*, 22(1–3):59–94, jan 1996. ISSN 0885-6125.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Offline minimax soft-q-learning under realizability and partial coverage. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Ruosong Wang, Dean Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354. PMLR, 15–19 Aug 2021.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11404–11413. PMLR, 18–24 Jul 2021.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. *Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694. Curran Associates, Inc., 2021a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021b.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent MDP and markov game. In *The Eleventh International Conference on Learning Representations*, 2023.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12287–12297. PMLR, 18–24 Jul 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10978–10989. PMLR, 13–18 Jul 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11756–11766. Curran Associates, Inc., 2020b.

Andrea Zanette, Martin Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775. PMLR, 02–05 Jul 2022.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5757–5773. PMLR, 28–30 Mar 2022.

Hanlin Zhu, Paria Rashidinejad, and Jiantao Jiao. Importance weighted actor-critic for optimal conservative offline reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

A Detailed comparison to related work

In this section, we discuss prior work on offline RL with function approximation and compare existing provable guarantees to our own.

Actor-critic methods. As mentioned above, the most closely related work is Zanette et al. (2021), which proves an analogous upper bound to ours for the special case of linear MDPs. More generally, the results of Zanette et al. (2021) apply to the broader class of MDPs satisfying *Bellman restricted closedness*, which requires that the Bellman backup of any linear function under *any policy* is (approximately) linear in the features. In contrast, the weaker condition of low inherent Bellman error (Assumption 2.1) requires that the Bellman backup of any linear function under the *single policy* induced by that linear function is approximately linear. Intuitively, the fact that Bellman restricted closedness requires that Bellman backups under all poicies be linear places it much "closer" to the assumption of linear (i.e., low-rank) MDPs than to low inherent Bellman error. This intuition is formalized in Jin et al. (2020a, Proposition 5.1), which shows that under the mild additional assumption that each state has at least two distinct feature vectors, then Bellman restricted closedness implies that the MDP is a linear MDP.

Several works have studied actor-critic methods in the setting of general function approximation, beginning with Xie et al. (2021a). In particular, Xie et al. (2021a) considers the setting of a general function class $\mathcal{F} \subset [0,1]^{\mathcal{X}\times\mathcal{A}}$ and policy class $\Pi \subset \Delta(\mathcal{A})^{\mathcal{X}}$, and assumes that $(\mathcal{F}, \Pi)$ satisfy approximate realizability and completeness.[3] Their first result, namely Xie et al. (2021a, Theorem 3.1), establishes a *computationally inefficient* algorithm for offline RL based on the principle of Bellman-consistent pessimism. One may use Corollary B.2 together with an appropriate choice of σ to instantiate their result with $\Pi = \Pi^{\mathsf{Plin},\sigma}$ and $\mathcal{F} = \{(x,a) \mapsto \langle \phi_h(x,a), w\rangle \ : \ h \in [H], w \in \mathbb{R}^d\}$

[3]Formally, to satisfy the realizability and completeness assumptions of Xie et al. (2021a), it suffices that for all $h \in [H]$, $\sup_{\pi\in\Pi, f_{h+1}\in\mathcal{F}} \inf_{f_h\in\mathcal{F}} \|f_h - \mathcal{T}_h^\pi f_{h+1}\|_\infty \le \varepsilon$.

to obtain a computationally inefficient version of Theorem 3.1.[4] Finally, Xie et al. (2021a) establish an algorithm, PSPI, which is oracle-efficient given an oracle for $\mathcal{F}$ which can solve a certain regularized least-squares problem. However, this result requires taking Π to be a softmax policy class, for which approximate Bellman completeness does not in general hold under Assumption 2.1 (see Lemma F.1). Thus, PSPI cannot be combined with Corollary B.2 to achieve an efficient algorithm under the assumption of low inherent Bellman error. (Moreover, their rate of $O(n^{-1/3})$ (Xie et al., 2021a, Corollary 5) is suboptimal.)

Subsequently, Cheng et al. (2022) considered a similar setting, with general function and policy classes $(\mathcal{F}, \Pi)$ satisfying approximate realizability and completeness. Their algorithm, ATAC, implements the actor using a generic no-regret learning algorithm, and implements the critic by solving a Lagrange relaxation of a least-squares regression problem. When combined with our main structural result, Corollary B.2, it is possible to use ATAC to obtain an end-to-end computationally efficient learning algorithm for offline RL under Assumption 2.1. In particular, one may take Π to be the class of perturbed linear policies, the no-regret learning algorithm to be expected FTPL (i.e., Algorithm 4), and one may implement the critic, which a priori appears to require minimizing a nonconvex quadratic function, using the approach in Xie et al. (2021a, Appendix D) (see also Antos et al. (2006)). However, due to the Lagrangian term in the critic's optimization problem, the resulting rate (see Cheng et al. (2022, Theorem 5)) is worse than ours, scaling with $O(n^{-1/3})$. We remark that Cheng et al. (2022) also considers the problem of *robust policy improvement*, which shows that if the data is drawn from a behavior policy, then the algorithm's output performs nearly as well as the behavior policy, with no assumptions on the hyperparameters, no dependence on any concentrability coefficient, and no assumption of Bellman completeness. Recently, Nguyen-Tang & Arora (2023) has given a refined analysis of ATAC which obtains the optimal $O(n^{-1/2})$ statistical rate, though only in the case when the policy class is the softmax policy class and Bellman completeness holds with respect to this class (which is not the case under Assumption 2.1).

Finally, Zhu et al. (2023) introduces an algorithm, A-Crab, which is similar to ATAC but incorporates the idea of *marginalized importance sampling*. As we discuss in the following paragraph, this approach cannot be instantiated in the setting of low inherent Bellman error to yield computationally or statistically efficient guarantees for offline RL in full generality.

Approaches via marginalized importance weighting. In addition to A-Crab, many other works, starting with Xie et al. (2019) have considered the approach of marginalized importance sampling (MIS), which introduces an additional bounded function class $\mathcal{W}$ consisting of *importance weights*. Elements of $\mathcal{W}$ may be interpreted as possible values for the density ratio between the state-action visitation distribution of the policy $\pi^\star$ one wishes to compete with and the data distribution. A line of work, including CORAL (Rashidinejad et al., 2023), PRO-RL (Zhan et al., 2022), PABC (Chen & Jiang, 2022), MLB-PO (Jiang & Huang, 2020) and that of Ozdaglar et al. (2023), makes the assumption that $\mathcal{W}$ contains the density ratio for $\pi^\star$, amongst other assumptions. Generally speaking, these algorithms implement pessimism for offline RL using primal-dual methods applied to the linear programming formulation of policy optimization. The introduction of $\mathcal{W}$ allows many of them to establish upper bounds even in the absence of Bellman completeness.

Several factors prevent such approaches from implying bounds similar to our own for the setting of MDPs satisfying linear Bellman completeness. First, all of the above works assume the existence of an oracle for optimizing over the class $\mathcal{W}$, which would translate into an intractable nonlinear optimization problem in the setting of linear Bellman completeness. More fundamentally, it can be impossible to satisfy the assumption of realizability with respect to $\mathcal{W}$: even if we only wish to compete with a single policy $\pi^\star$, there is a class $\mathcal{M}$ of linear Bellman complete MDPs so that there

[4] The resulting single-policy coverage parameter is somewhat larger than our own, though a subsequent observation by the authors of Xie et al. (2021a) (see Jiang (2023)) leads to a tightening of their analysis which results in a matching coverage parameter.

is no distribution μ_h for which $\sup_{M \in \mathcal{M}} \left\| \frac{d_h^{\pi^\star, M}(x,a)}{\mu_h(x,a)} \right\|_\infty$ is bounded.[5] Since the above results require that $\frac{d_h^{\pi^\star, M}(x,a)}{\mu_h(x,a)}$ belongs to $\mathcal{W}$ for any $M \in \mathcal{M}$, a bounded class $\mathcal{W}$ of importance weights, satisfying realizability for the model class $\mathcal{M}$, *does not exist*.

Uehara et al. (2023) analyses a primal-dual approach of a slightly different nature, but with the common goal of relaxing Bellman completeness at the cost of assuming realizability of a suitable class of Lagrange multipliers. Finally, Gabbianelli et al. (2023) uses a similar primal-dual method applied to the LP formulation of policy optimization as many of the above approaches, but only treats the special case of linear MDPs. Due to this additional structure, Gabbianelli et al. (2023) does not need to explicitly make any assumptions regarding a class $\mathcal{W}$.

Linear MDPs: pessimistic value iteration. In contrast to the above approaches, which phrase the problem of finding a pessimistic value function as a *global* optimization problem, a line of work, including `PEVI` (Jin et al., 2021), `VAPVI` (Yin et al., 2022), `R-LSVI` (Zhang et al., 2022), and `LinPEVI-ADV` (Xiong et al., 2023) has considered a *local* approach to implementing pessimism. In particular, these algorithms perform value iteration but subtract "penalties" at each state which are inversely proportional to how well the state is visited in the given dataset. Since the penalties do not necessarily have Bellman backups which are linear functions, these approaches do not directly generalize to the setting of linear Bellman complete MDPs.[6] The approach of pessimistic value iteration has also been extended to settings with nonlinear function approximation (Di et al., 2023).

Additional approaches for offline RL with function approximation. An older line of work (Munos & Szepesvári, 2008; Chen & Jiang, 2019) has studied offline RL under the stronger assumption of *all policy concentrability*, meaning that the data distribution covers the state-action distribution of *any* policy. These approaches proceed via variants of fitted Q-iteration, and therefore require approximate realizability and Bellman completeness in the setting of general function approximation. Xie & Jiang (2021) show that the assumption of Bellman completeness can be avoided under an even stronger concentrability assumption. Liu et al. (2020) analyzes pessimistic variants of value and policy iteration with only single-policy concentrability, under somewhat non-standard assumptions regarding completeness with respect to truncated Bellman backups. In the special case of tabular MDPs, Rashidinejad et al. (2021); Shi et al. (2022); Xie et al. (2021b) have focused on obtaining the optimal polynomial dependence on the various problem parameters, under single-policy concentrability. Finally, several works have considered model-based offline RL (Ross & Bagnell, 2012; Chang et al., 2021; Uehara & Sun, 2022; Bhardwaj et al., 2023), which construct an estimate of all of the MDP's transitions and rewards (perhaps pessimistically) as opposed to estimating the value functions.

Finally, we mention that in a distinct setting to ours (namely, that of nonlinear dynamical systems), Block et al. (2023) use the idea of injecting Gaussian noise into the learned policy to establish guarantees (see Definition 5.3 therein). This technique is analogous to our technique of using perturbed linear policies.

Lower bounds. Wang et al. (2021); Amortila et al. (2020) show an exponential lower bound for offline RL even when the value function for *any* policy is assumed to be linear in some known features, and when the distribution of the data has good coverage of *all* feature directions. Zanette (2021) shows a similar exponential lower bound, but which is stronger in that it holds no matter how the distribution of offline data is chosen. Finally, Foster et al. (2021) shows a lower bound for offline RL in a nonlinear setting when the stronger assumption of *concentrability* is made. Taken together, these results may be seen as motivating the assumption of Bellman completeness: when

[5] For instance, consider a class of MDPs for which an initial state-action pair (x_1, a_1) transitions deterministically to any of infinitely many copies of some state, denoted $x_2^1, x_2^2, \ldots$, each of which is equivalent in the sense that $\phi_2(x_2^i, a) = \phi_2(x_2^j, a)$ for all $a \in \mathcal{A}$, $i \neq j$.

[6] Moreover, a necessary truncation step in pessimistic value iteration presents another obstacle to extending this approach to our setting.

only realizability (as well as an appropriate coverage or concentrability notion) is assumed, little is possible.

B Low Inherent Bellman Error: Structural Properties

In this section, we prove Theorem 3.2 (restated below formally as Corollary B.2), which shows that the Bellman backup of any linear function at step $h+1$, with respect to any perturbed linear policy, is an approximately linear function at step h. The main ingredient in the proof of Corollary B.2 is Lemma B.1 below, which shows that the expected feature vector induced by a perturbed linear policy at step $h+1$ is a linear transformation of the state-action feature vector at step h. Lemma B.1 is a generalization of Lemma 4.3 of Golowich & Moitra (2024), which treats the special case of $\varepsilon_{\mathsf{BE}} = 0$ and used the result to develop an efficient algorithm for the related setting of *online RL* under linear Bellman completeness.

Lemma B.1. *Suppose that the MDP M has inherent bellman error bounded by $\varepsilon_{\mathsf{BE}}$, and fix $\sigma > 0$. Then for each $h \in [H]$ and $w \in \mathbb{R}^d$, there is a linear map $L_{h,w,\sigma} : \mathbb{R}^d \to \mathbb{R}^d$ so that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$\begin{aligned}&\left\|L_{h,w,\sigma}^{\top} \cdot \phi_h(x,a) - \mathbb{E}_{x' \sim \mathbb{P}_h(x,a)}[\phi_{h+1}(x', \pi_{h+1,w,\sigma}(x'))]\right\|_2 \\ \leq & C_{B.1}\varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d\log(d/(\varepsilon_{\mathsf{BE}}\sigma))} + \frac{1}{\sigma}\right),\end{aligned}$$

for some constant $C_{B.1}$. Moreover, for any w, w', σ, σ' for which $\pi_{h+1,w,\sigma}(x') = \pi_{h+1,w',\sigma'}$ for all $x' \in \mathcal{X}$, we have $L_{h,w,\sigma} = L_{h,w',\sigma'}$.

Given a perturbed linear policy $\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin}}$, so that $\pi_{h+1} = \pi_{h+1,\theta,\sigma}$ for some $\theta \in \mathbb{R}^d, \sigma > 0$, we define $\mathcal{T}_h^{\pi_{h+1}} w := \theta_h^{\mathsf{r}} + L_{h,\theta,\sigma} \cdot w$, where $L_{h,\theta,\sigma}$ is the map of Lemma B.1 and where θ_h^{r} was defined in Assumption 2.1. Note that $\mathcal{T}_h^{\pi_{h+1}}$ is well-defined in the sense that it only depends on π_{h+1} (and not on the particular choice of θ, σ), since for any $\theta, \sigma, \theta', \sigma'$ satisfying $\pi_{h+1,\theta,\sigma} = \pi_{h+1,\theta',\sigma}$, we have by Lemma B.1 that $L_{h,\theta,\sigma} = L_{h,\theta',\sigma'}$. We will at times abuse notation by writing $\mathcal{T}_h^{\pi} := \mathcal{T}_h^{\pi_{h+1}}$ for a policy $\pi \in \Pi^{\mathsf{Plin}}$. Given Lemma B.1, the proof of Corollary B.2, stated below, is straightforward.

Corollary B.2. *Suppose that $\pi \in \Pi^{\mathsf{Plin},\sigma}$. Then for all $h \in [H-1]$, $w \in \mathbb{R}^d$, and $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$\left|\langle \phi_h(x,a), \mathcal{T}_h^{\pi} w\rangle - \left(r_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a)}[\langle \phi_{h+1}(x', \pi_{h+1}(x')), w\rangle]\right)\right| \leq \|w\|_2 \cdot \zeta_\sigma, \tag{10}$$

where $\zeta_\sigma = C_{B.2}\varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d\log(d/(\varepsilon_{\mathsf{BE}}\sigma))} + \frac{1}{\sigma}\right)$, for some constant $C_{B.2}$. Moreover, if $\zeta_\sigma \leq 1$, then $\mathcal{T}_h^{\pi} w \in (1 + 2\|w\|_2) \cdot \mathcal{B}_h$.

Proof. The inequality (10) follows directly from Lemma B.1 and the definition of $\mathcal{T}_h^{\pi}$, as well as (3). To see that $\mathcal{T}_h^{\pi} w \in (1 + 2\|w\|_2) \cdot \mathcal{B}_h$, we note that, by (10), for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$|\langle \phi_h(x,a), \mathcal{T}_h^{\pi} w\rangle| \leq 1 + \|w\|_2 + \|w\|_2 \cdot \zeta_\sigma \leq 1 + 2\|w\|_2,$$

since $\zeta_\sigma \leq 1$. □

As a further corollary of Corollary B.2, the Q-function for a perturbed linear policy is linear.

Corollary B.3. *Suppose that M is linear Bellman complete, $\sigma > 0$, and that $\pi \in \Pi^{\mathsf{Plin},\sigma}$. Then for each $h \in [H]$, there is a vector $w_h^{\pi} \in 2H \cdot \mathcal{B}_h \subset \mathbb{R}^d$ so that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$|Q_h^{\pi}(x,a) - \langle w_h^{\pi}, \phi_h(x,a)\rangle| \leq 3(H+1-h)HB \cdot \zeta_\sigma,$$

where ζ_σ is as defined in Corollary B.2. Moreover, if $3HB\zeta_\sigma \leq 1$, then $w_h^{\pi} \in 2H \cdot \mathcal{B}_h$ and $\|w_h^{\pi}\|_2 \leq 2HB$.

Proof. We use reverse induction on h; the base case $h = H$ is immediate, so suppose the statement holds at step $h+1$. By Assumption 2.1 and Corollary B.2, we have that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned}&\left|r_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a)}[\langle \phi_{h+1}(x', \pi_{h+1}(x')), w_{h+1}^\pi\rangle - \langle \phi_h(x,a), \mathcal{T}_h^\pi w_{h+1}^\pi\rangle\right| \\ \le &\varepsilon_{\mathsf{BE}} + \zeta_\sigma \cdot \|w_{h+1}^\pi\|_2.\end{aligned}$$

Let us write $w_h^\pi := \mathcal{T}_h^\pi w_{h+1}^\pi$, Since $Q_h^\pi(x,a) = r_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a)}[Q_{h+1}^\pi(x, \pi_{h+1}(x))]$, the inductive hypothesis then gives us that

$$\begin{aligned}&|Q_h^\pi(x,a) - \langle w_h^\pi, \phi_h(x,a)\rangle| \\ \le &\mathbb{E}_{x' \sim P_h(x,a)}[|Q_{h+1}^\pi(x', \pi_{h+1}(x')) - \langle w_{h+1}^\pi, \phi_{h+1}(x', \pi_{h+1}(x'))\rangle|] \\ &+ \left|r_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a)}[\langle \phi_{h+1}(x', \pi_{h+1}(x')), w_{h+1}^\pi\rangle - \langle \phi_h(x,a), w_h^\pi\rangle\right| \\ \le &3(H-h)HB \cdot \zeta_\sigma + (\varepsilon_{\mathsf{BE}} + \zeta_\sigma \cdot \|w_{h+1}^\pi\|_2) \le 3(H+1-h)HB \cdot \zeta_\sigma, \end{aligned} \tag{11}$$

where the final inequality uses that $\varepsilon_{\mathsf{BE}} \le \zeta_\sigma$. To see the upper bound on $\|w_h^\pi\|_2$, note that, by definition of Q_h^π and (11), we have $|\langle w_h^\pi, \phi_h(x,a)\rangle| \le H + 3(H+1-h)HB\zeta_\sigma \le 2H$ for all x, a, h, where we have used that $3HB\zeta_\sigma \le 1$. Then it follows that $\|w_h^\pi\|_2 \le 2HB$ by Assumption 2.2. □

Proof of Lemma B.1. We may assume without loss of generality that $\|w\|_2 = 1$, by increasing σ by a factor of $1/\|w\|_2$. Fix $h \in [H]$. For $x' \in \mathcal{X}$ and $w \in \mathbb{R}^d$, define $V(x', w) := \max_{a' \in \mathcal{A}} \langle \phi_{h+1}(x', a'), w\rangle$. For $x \in \mathcal{X}, a \in \mathcal{A}, w \in \mathbb{R}^d$, define $Q(x,a,w) = \mathbb{E}_{x' \sim P_h(x,a)}[V(x', w)]$. Since $\mathcal{A}, \mathcal{X}$ are assumed to be countable, the mapping $w \mapsto Q(x,a,w)$ is piecewise linear with countably many pieces (and is also continuous). Next, Assumption 2.1 together with the fact that $\{w \in \mathbb{R}^d : \|w\|_2 \le 1\} \subset \mathcal{B}_{h+1}$ gives us that

$$\sup_{\|w\|_2 \le 1} \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |\langle \phi_h(x,a), \mathcal{T}_h^\circ w\rangle - Q(x,a,w)| \le \varepsilon_{\mathsf{BE}}. \tag{12}$$

Next, we may choose $k \le d$ and $(x_1, a_1), \dots, (x_k, a_k) \in \mathcal{X} \times \mathcal{A}$ so that $\{(\phi_h(x_i, a_i)\}_{i=1}^k$ forms a barycentric spanner of $\{\phi_h(x,a)\}_{(x,a) \in \mathcal{X} \times \mathcal{A}}$, and so that $\phi_h(x_i, a_i)$, $1 \le i \le k$, are linearly independent. By linear independence of $\phi_h(x_i, a_i)$, for each $w \in \mathbb{R}^d$, there is a matrix $L_{h,w,\sigma} \in \mathbb{R}^{d \times d}$ so that, for all $i \in [k]$,

$$L_{h,w,\sigma}^\top \cdot \phi_h(x_i, a_i) = \nabla \mathsf{S}_\sigma Q(x_i, a_i, w). \tag{13}$$

Next, for each $(x,a) \in \mathcal{X} \times \mathcal{A}$ and $w \in \mathbb{R}^d$, we have

$$\begin{aligned}\nabla_w \mathsf{S}_\sigma Q(x,a,w) =&\nabla_w \mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 \cdot I_d)} \mathbb{E}_{x' \sim P_h(x,a)}[Q(x', w)] \\ =&\mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 \cdot I_d)} \mathbb{E}_{x' \sim P_h(x,a)}[\nabla_w V(x', w+z)] \\ =&\mathbb{E}_{x' \sim P_h(x,a)} \mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 \cdot I_d)}[\phi_{h+1}(x', \pi_{h+1, w+z}(x'))] \\ =&\mathbb{E}_{x' \sim P_h(x,a)}[\phi_{h+1}(x', \pi_{h+1,w,\sigma}(x'))], \end{aligned} \tag{14}$$

where the second equality uses the dominated convergence theorem together with the fact that for each $x' \in \mathcal{Z}, z \in \mathbb{R}^d$, $w \mapsto V(x', w+z)$ is piecewise linear with finitely many pieces, with bounded derivative, i.e., $\|\nabla_w V(x', w+z)\| \le \max_{a' \in \mathcal{A}} \|\phi_{h+1}(x', a')\|_2 \le 1$. (Note that $\mathsf{S}_\sigma Q(x,a,w)$ is infinitely differentiable since we can write $\mathsf{S}_\sigma Q(x,a,w) = \int_{\mathbb{R}^d} Q(x,a,w) \mathcal{N}_\sigma(w-z) dz$ and since $\mathcal{N}_\sigma(w-z)$ is infinitely differentiable in w.) Using (14) with $(x,a) = (x_i, a_i)$ (for each $i \in [k]$), we see that $L_{h,w,\sigma} = L_{h,w',\sigma'}$ for all w, w', σ, σ' satisfying $\pi_{h+1,w,\sigma}(x') \pi_{h+1,w',\sigma'}(x')$ for all $x' \in \mathcal{X}$.

Fix any $(x,a) \in \mathcal{X} \times \mathcal{A}$. By the barycentric spanner property, there are coefficients $\alpha_1, \dots, \alpha_k \in [-1, 1]$ (depending on x, a) so that $\phi_h(x,a) = \sum_{i=1}^k \alpha_i \cdot \phi_h(x_i, a_i)$. It therefore follows that, for all

$w \in \mathcal{B}_{h+1}$,

$$\begin{aligned}
&\left|Q(x,a,w) - \sum_{i=1}^{k} \alpha_i Q(x_i,a_i,w)\right| \\
\leq &|Q(x,a,w) - \langle \phi_h(x,a), \mathcal{T}_h^\circ w\rangle| + \left|\langle \phi_h(x,a), \mathcal{T}_h^\circ w\rangle - \sum_{i=1}^{k} \alpha_i \langle \phi_h(x_i,a_i), \mathcal{T}_h^\circ w\rangle\right| \\
&+ \left|\sum_{i=1}^{k} \alpha_i \cdot (Q(x_i,a_i,w) - \langle \phi_h(x_i,a_i), \mathcal{T}_h^\circ w\rangle)\right| \\
\leq &|Q(x,a,w) - \langle \phi_h(x,a), \mathcal{T}_h^\circ w\rangle| + \sum_{i=1}^{k} |Q(x_i,a_i,w) - \langle \phi_h(x_i,a_i), \mathcal{T}_h^\circ w\rangle| \\
\leq &(d+1)\varepsilon_{\mathsf{BE}}, \qquad (15)
\end{aligned}$$

where the second inequality uses that $\langle \phi_h(x,a), \mathcal{T}_h^\circ w\rangle = \sum_{i=1}^{k} \alpha_i \langle \phi_h(x_i,a_i), \mathcal{T}_h^\circ w\rangle$ and the final equality uses (12) applied to each of the tuples $(x,a), (x_1,a_1), \ldots, (x_k,a_k)$.

Next we apply Lemma B.4 with

$$\begin{aligned}
f(w) =& Q(x,a,w) - \sum_{i=1}^{k} \alpha_i Q(x_i,a_i,w) \\
\epsilon =& (1 + 100\sigma\sqrt{d\log(2d/(\varepsilon_{\mathsf{BE}}\sigma))})(d+1)\varepsilon_{\mathsf{BE}},
\end{aligned}$$

$D = 2d$, $B = 1$, and $\mathcal{B} = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$. Since, for all $x \in \mathcal{X}, a \in \mathcal{A}, w \in \mathbb{R}^d$, $|Q(x,a,w)| \leq \|w\|_2$, we have $|f(w)| \leq (d+1) \cdot \|w\|_2 \leq 2d\|w\|_2$ for $w \in \mathbb{R}^d$, verifying that the choice of $D = 2d$ is admissible. Moreover, the set of z with $\mathrm{dist}(z, \mathcal{B}) \leq 100\sigma\sqrt{d\log(2d/(\epsilon\sigma))}$ is contained in $(1 + 100\sigma\sqrt{d\log(2d/\varepsilon_{\mathsf{BE}}\sigma)}) \cdot \mathcal{B}$, since $\epsilon > \varepsilon_{\mathsf{BE}}$ and $\mathcal{B}$ is a unit ball. Therefore, scaling (15) verifies that for all w with $\mathrm{dist}(w, \mathcal{B}) \leq 100\sigma\sqrt{d\log(2d/(\epsilon\sigma))}$,

$$\left|Q(x,a,w) - \sum_{i=1}^{k} \alpha_i Q(x_i,a_i,w)\right| \leq (1 + 100\sigma\sqrt{d\log(2d/(\varepsilon_{\mathsf{BE}}\sigma))}) \cdot (d+1) \cdot \varepsilon_{\mathsf{BE}} = \epsilon.$$

Then the guarantee of Lemma B.4 gives that for some constant $C > 0$, for all w with $\|w\|_2 \leq 1$,

$$\left\|\nabla \mathsf{S}_\sigma Q(x,a,w) - \sum_{i=1}^{k} \alpha_i \cdot \nabla \mathsf{S}_\sigma Q(x_i,a_i,w)\right\|_2 \leq \frac{C\epsilon\sqrt{d}}{\sigma}.$$

By our choice of α_i and (13), for all $w \in \mathbb{R}^d$,

$$\sum_{i=1}^{k} \alpha_i \cdot \nabla \mathsf{S}_\sigma Q(x_i,a_i,w) = L_{h,w,\sigma}^\top \cdot \sum_{i=1}^{k} \alpha_i \cdot \phi_h(x_i,a_i) = L_{h,w,\sigma}^\top \cdot \phi_h(x,a).$$

Combining the above with (14) and the definition of ϵ gives that, for some constants C, C', for $\|w\|_2 \leq 1$,

$$\begin{aligned}
&\left\|\mathbb{E}_{x' \sim P_h(x,a)}[\phi_{h+1}(x', \phi_{h+1,w,\sigma}(x'))] - L_{h,w,\sigma}^\top \cdot \phi_h(x,a)\right\|_2 \\
\leq &\varepsilon_{\mathsf{BE}} \cdot \frac{C \cdot (1 + 100\sigma\sqrt{d\log(2d/(\varepsilon_{\mathsf{BE}}\sigma))})(d+1) \cdot \sqrt{d}}{\sigma} \\
\leq &C'\varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d\log(d/(\varepsilon_{\mathsf{BE}}\sigma))} + \frac{1}{\sigma}\right),
\end{aligned}$$

as desired. □

Bounding the gradient of a Gaussian convolution. For a subset $\mathcal{B} \subset \mathbb{R}^d$ and $z \in \mathbb{R}^d$, we write $\mathrm{dist}(z, \mathcal{B}) := \inf\{\|w - z\|_2 \ : \ w \in \mathcal{B}\}$.

Lemma B.4. *There is a constant $C > 0$ so that the following holds. Let $\sigma, \epsilon \in (0, 1/2)$ and $B, D \geq 1$ be given, and suppose that $\mathcal{B} \subset \mathbb{R}^d$ is a set with nonempty interior $\mathcal{B}^\circ$, and for which $\max_{\theta \in \mathcal{B}} \|\theta\| \leq B$. Suppose $f : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $|f(z)| \leq \epsilon$ for all z with $\mathrm{dist}(z, \mathcal{B}) \leq 100\sigma\sqrt{d \log(BD/(\epsilon\sigma))}$, as well as $|f(z)| \leq D\|z\|_2$ for all $z \in \mathbb{R}^d$, for some $D > 0$. Then for all $z \in \mathcal{B}^\circ$,*

$$\|\nabla \mathsf{S}_\sigma f(z)\|_2 \leq \frac{C\epsilon\sqrt{d}}{\sigma}.$$

Proof. Let us write $\Delta := 100\sigma\sqrt{d \log(BD/(\epsilon\sigma))}$. Let $\mathcal{B}_\Delta := \{z \in \mathbb{R}^d \ : \ \mathrm{dist}(z, \mathcal{B}) \leq \Delta\}$, so that, by assumption, $|f(z)| \leq \epsilon$ for all $z \in \mathcal{B}_\Delta$. Then, for any $\theta \in \mathcal{B}$,

$$\begin{aligned}
\|\nabla \mathsf{S}_\sigma f(\theta)\|_2 &= \left\| \nabla \int_{\mathbb{R}^d} f(z) \mathcal{N}_\sigma(\theta - z) dz \right\|_2 \\
&= \left\| \int_{\mathbb{R}^d} f(z) \cdot \nabla \mathcal{N}_\sigma(\theta - z) dz \right\|_2 \\
&\leq \int_{\mathcal{B}_\Delta} \|f(z) \cdot \nabla \mathcal{N}_\sigma(\theta - z)\|_2 \, dz + \int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \|f(z) \cdot \nabla \mathcal{N}_\sigma(\theta - z)\|_2 \, dz. \qquad (16)
\end{aligned}$$

Note that $\nabla \mathcal{N}_\sigma(\theta) = -\frac{\theta}{\sigma^2} \cdot \mathcal{N}_\sigma(\theta)$. Then we have

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \|f(z) \cdot \nabla \mathcal{N}_\sigma(\theta - z)\|_2 \, dz &\leq \int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \frac{D\|z\|_2 \cdot \|\theta - z\|_2}{\sigma^2} \cdot \mathcal{N}_\sigma(\theta - z) dz \\
&\leq D \int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \frac{\|\theta - z\|_2^2 + B\|\theta - z\|_2}{\sigma^2} \cdot \mathcal{N}_\sigma(\theta - z) dz, \qquad (17)
\end{aligned}$$

where the second inequality uses that $\|z\| \leq \|\theta - z\|_2 + B$ since $\theta \in \mathcal{B}$.

Using the tail bound $\Pr_{z \sim \mathcal{N}(0, \sigma^2 I_d)}(\|z\|_2^2 > 2td\sigma^2) \leq e^{-td/10}$ for $t \geq 1$ (Laurent & Massart, 2000, Lemma 1)[7] it holds that

$$\begin{aligned}
\int_{\|z\|_2 \geq \Delta} \Delta \|z\|_2 \cdot \mathcal{N}_\sigma(z) dz \leq \int_{\|z\|_2 \geq \Delta} \|z\|_2^2 \cdot \mathcal{N}_\sigma(z) dz &\leq \Delta^2 \cdot e^{-\Delta^2/(20\sigma^2)} + \int_{\Delta^2}^\infty e^{-y/(20\sigma^2)} d\sigma \\
&= \Delta^2 \cdot e^{-\Delta^2/(20\sigma^2)} + \frac{1}{20\sigma^2} \cdot e^{-\Delta^2/(20\sigma^2)} \\
&\leq \left(\Delta^2 + \frac{1}{\sigma^2}\right) \cdot e^{-5d \log(BD/(\epsilon\sigma))} \\
&\leq \left(10^4 d \log BD/(\epsilon\sigma) + \frac{1}{\sigma^2}\right) \cdot (\epsilon\sigma/(BD))^{5d} \\
&\leq \frac{10^4 \epsilon\sigma^3 \epsilon}{BD} + \frac{\epsilon\sigma^3}{BD} = \frac{10001\epsilon\sigma^3}{BD}, \qquad (18)
\end{aligned}$$

where the second inequality uses the layer cake formula and the fact that $\Delta^2/\sigma^2 \geq 2d$, and the final inequality uses that $(\epsilon\sigma/(BD))^{5d} \cdot d \log(BD/\epsilon\sigma) \leq (\epsilon\sigma/BD)^{4d} \cdot d \leq \epsilon\sigma^3/(BD)$ (since $\epsilon, \sigma \leq 1/2$) and that $(\epsilon\sigma/(BD))^{5d}/\sigma^2 \leq \epsilon\sigma^3/(BD)$.

Note that for all $z \in \mathbb{R}^d \setminus \mathcal{B}_\Delta$, we have that $\|\theta - z\| \geq \Delta$ since $\theta \in \mathcal{B}$. Thus, combining Eqs. (17) and (18), we have, for some constant C,

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \|f(z) \cdot \nabla \mathcal{N}_\sigma(\theta - z)\|_2 dz &\leq \frac{D}{\sigma^2} \int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \|\theta - z\|_2^2 \cdot \mathcal{N}_\sigma(\theta - z) dz + \frac{BD}{\sigma^2} \int_{\mathbb{R}^d \setminus \mathcal{B}_\Delta} \|\theta - z\|_2 \cdot \mathcal{N}_\sigma(\theta - z) dz \\
&\leq \frac{D}{\sigma^2} \cdot \frac{C\epsilon\sigma^3}{BD} + \frac{BD}{\sigma^2} \cdot \frac{C\epsilon\sigma^3}{BD\Delta} \leq 2C\epsilon, \qquad (19)
\end{aligned}$$

[7] See also `https://math.stackexchange.com/questions/2864188/chi-squared-distribution-tail-bound`.

where we have used in the second-to-last inequality that $B \geq 1$ and $\sigma/\Delta \leq 1$.

Next, we compute

$$\begin{aligned}\int_{\mathcal{B}_\Delta} \|f(z) \cdot \nabla \mathcal{N}_\sigma(\theta - z)\|_2 dz &\leq \int_{\mathcal{B}_\Delta} |f(z)| \cdot \frac{\|\theta - z\|_2}{\sigma^2} \mathcal{N}_\sigma(\theta - z) dz \\ &\leq \frac{\epsilon}{\sigma^2} \int_{\mathcal{B}_\Delta} \|\theta - z\|_2 \mathcal{N}_\sigma(\theta - z) dz \\ &\leq \frac{\epsilon}{\sigma^2} \int_{\mathbb{R}^d} \|\theta - z\|_2 \mathcal{N}_\sigma(\theta - z) dz \leq \frac{\epsilon}{\sigma^2} \cdot \sigma\sqrt{d} = \frac{\epsilon\sqrt{d}}{\sigma}, \end{aligned} \tag{20}$$

where the final inequality uses the fact that $\mathbb{E}_{Z \sim \mathcal{N}(0, \sigma^2 \cdot I_d)}[\|Z\|_2] = \sigma\sqrt{d}$.

Combining (16), (19), and (20), we obtain that, for some constant $C' > 0$,

$$\|\nabla \mathsf{S}_\sigma f(\theta)\|_2 \leq 2C\epsilon + \epsilon\sqrt{d}/\sigma \leq \frac{C'\epsilon\sqrt{d}}{\sigma},$$

which is the desired bound. □

C Proof of Theorem 3.1

In this section we prove Theorem 3.1. We begin by defining the values for the parameters used in Algorithms 1 and 2.

Definition C.1 (Parameter settings for the algorithms)**.** Given $d, H, B > 0$ specifying the dimension, horizon, and boundedness parameters for the unknown MDP M, as well as dataset size n and error parameters $\varepsilon_{\mathsf{final}}, \delta \in (0, 1)$, we define the following parameters to be used in Algorithms 1 and 2 and its analysis:

- $\beta := 2BH$.
- $T := \frac{16\beta^2 d^{1/2}}{\varepsilon_{\mathsf{final}}^2}$.
- $\varepsilon_{\mathsf{apx}} := 1/\sqrt{n}$.
- $\eta := \beta \cdot \max\left\{T^{1/2} d^{-1/4}, T\varepsilon_{\mathsf{BE}}^{1/2}\right\}$.
- $\sigma := \frac{\eta}{T\beta}$.
- $\zeta = C_{B.2}\varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d \log(d/(\varepsilon_{\mathsf{BE}}\sigma))} + \frac{1}{\sigma}\right)$, where $C_{B.2}$ is the constant from Corollary B.2. (Note that $\zeta = \zeta_\sigma$, where ζ_σ was defined in Corollary B.2.)
- $\alpha := 4\beta\zeta\sqrt{n} + C_{C.4}\beta d \log^{1/2}(dn\beta/(\sigma\delta))$, where $C_{C.4}$ is a constant chosen sufficiently large so as to ensure Lemma C.4 holds.

C.1 Critic analysis

Consider a tuple $f = (f_1, \ldots, f_H)$, where $f_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and a policy π. We define an MDP $M^{f,\pi}$ (called the *induced MDP*, per Zanette et al. (2021)), whose transitions are identical to those of M, but whose rewards are given as follows: for $h \in [H]$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$r_h^{M^{f,\pi}}(x, a) = f_h(x, a) - \mathbb{E}_{x' \sim P_h(x,a)}[f_{h+1}(x', \pi_{h+1}(x'))]. \tag{21}$$

Lemma C.1 (Lemma 1 of Zanette et al. (2021)). *Fix any $f = (f_1, \ldots, f_H)$ with $f_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, any $\pi \in \Pi$, and write $M' := M^{f,\pi}$. Then the Q-value function of M' for the policy π, denoted $Q^{M',\pi}$, satisfies the following:*

$$\forall h \in [H],\ (x,a) \in \mathcal{X} \times \mathcal{A}, \qquad Q_h^{M',\pi}(x,a) = f_h(x,a),$$

which implies in particular that $V_h^{M',\pi}(x) = f_h(x, \pi_h(x))$.

The proof of Lemma C.1 follows a simple telescoping argument and is provided in Zanette et al. (2021). Next, we establish the following straightforward guarantee for `EstFeature`:

Lemma C.2. *For any $x \in \mathcal{X}$, $\pi \in \Pi^{\mathsf{Plin}}$, $h \in [H]$, $\varepsilon_{\mathsf{apx}} \in (0,1)$ and $\delta \in (0,1)$, `EstFeature` (Algorithm 3) runs in time $O(d\varepsilon_{\mathsf{apx}}^{-2} \log(d/\delta))$ and returns a vector $\hat{\phi}$ so that, with probability $1-\delta$, $\|\hat{\phi} - \phi_h(x, \pi_{h,w}(x))\|_2 \le \varepsilon_{\mathsf{apx}}$.*

Proof. By the definition of $\pi_{h,w,\sigma}$, it holds that for each $\theta_i \sim \mathcal{N}(w, \sigma^2 \cdot I_d)$ in Algorithm 3, we have $\mathbb{E}[\phi_h(x, \pi_{h,\theta_i}(x))] = \phi_h(x, \pi_h(x))$. Then by Hoeffding's inequality and a union bound, we have that, with probability $1-\delta$,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \phi_h(x, \pi_{h,\theta_i}(x)) - \phi_h(x, \pi_h(x)) \right\|_2 \le \sqrt{d} \cdot \left\| \frac{1}{N} \sum_{i=1}^{N} \phi_h(x, \pi_{h,\theta_i}(x)) - \phi_h(x, \pi_h(x)) \right\|_\infty \le \varepsilon_{\mathsf{apx}}.$$

□

Recall that `Critic` (Algorithm 2) computes a vector $w = (w_1, \ldots, w_H) \in \mathbb{R}^{dH}$. Given such w, we define a function $f^w = (f_1^w, \ldots, f_H^w)$, where $f_h^w(x,a) := \langle \phi_h(x,a), w_h \rangle$. We introduce the following notation, in the context of `Critic` (Algorithm 2). Given a dataset $\mathcal{D} = \{(h_i, x_i, a_i, r_i, x_i')\}_{i=1}^n$, at step $h \in [H]$, a policy $\pi \in \Pi^{\mathsf{Plin}}$, a collection of feature vectors $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_n)$ (as produced in Line 3), and a vector $w \in \mathbb{R}^d$, we define

$$\hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^{\pi} w := \Sigma_h^{-1} \sum_{i \in \mathcal{I}_h} \phi_h(x_i, a_i) \cdot (r_i + \langle \hat{\phi}_i, w \rangle), \tag{22}$$

where $\Sigma_h = I_d + \sum_{i \in \mathcal{I}_h} \phi_h(x_i,a_i)\phi_h(x_i,a_i)^\top$ and $\mathcal{I}_h = \{i :\ h_i = h\}$ are defined as in Algorithm 2. Given parameters $\alpha, \sigma > 0$, we define the following good event $\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}}$:

$$\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}} = \left\{ \sup_{\|w_{h+1}\|_2 \le \beta} \sup_{\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin},\sigma}} \sup_{\substack{(\hat{\phi}_i)_{i\in[n]} \in (\mathbb{R}^d)^n :\ \forall i \in [n], \\ \|\hat{\phi}_i - \phi_{h+1}(x_i', \pi_{h+1}(x_i'))\|_2 \le \varepsilon_{\mathsf{apx}}}} \left\| \mathcal{T}_h^{\pi_{h+1}} w_{h+1} - \hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^{\pi_{h+1}} w_{h+1} \right\|_{\Sigma_h} \le \alpha \right\}. \tag{23}$$

Lemma C.3 below establishes that under the good event $\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}}$, upon given a perturbed linear policy π as input, `Critic` produces a vector $w^\star \in \mathbb{R}^{dH}$ so that the value of π in the induced MDP $M^{f^{w^\star},\pi}$ lower bounds the value of π in the true MDP (Item 1). Moreover, for *any* policy $\pi' \in \Pi$, the value of π' in $M^{f^{w^\star},\pi}$ is close to the value of π' in M, as controlled by a concentrability coefficient depending on π' (Item 2). Recall from Corollary B.2 that we have defined $\zeta_\sigma = C_{B.2}\varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d \log(d/(\varepsilon_{\mathsf{BE}}\sigma))} + \frac{1}{\sigma}\right)$.

Lemma C.3. *Consider any $\sigma, \alpha, \varepsilon_{\mathsf{apx}}, \delta > 0$ for which $3BH\zeta_\sigma \le 1$, suppose $\beta = 2BH$, and let $\pi \in \Pi^{\mathsf{Plin},\sigma}$ be given. Consider the execution of `Critic`$(\mathcal{D}, \pi, \varepsilon_{\mathsf{apx}}, \alpha, \beta, \delta)$: there is some event $\mathcal{E}_\pi$ depending only on the randomness in the call to `EstFeature` in Line 3, with $\Pr(\mathcal{E}_\pi) > 1-\delta$, so that the following holds. Under the event $\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}} \cap \mathcal{E}_\pi$, the output of `Critic`$(\mathcal{D}, \pi, \varepsilon_{\mathsf{apx}}, \alpha, \beta, \delta)$ is a pair of vectors $w^\star, \xi^\star \in \mathbb{R}^{dH}$ satisfying the following conditions, for $f := f^{w^\star}$:*

1. $V_1^{M^{f,\pi},\pi}(x_1) \le V_1^{M,\pi}(x_1)$.

2. *For any* $\pi' \in \Pi$,

$$\left|V_1^{M^{f,\pi},\pi'}(x_1) - V_1^{M,\pi'}(x_1)\right| \le 3\beta H\zeta_\sigma + 2\alpha \sum_{h=1}^{H} \|\mathbb{E}^{M,\pi'}[\phi_h(x_h,a_h)]\|_{\Sigma_h^{-1}}.$$

Moreover, given an oracle which can solve the convex program (7), the computational cost of `Critic` *is* $\mathrm{poly}(d,n,H,\log(1/\delta)/\varepsilon_{\mathsf{apx}}^2)$.

For simplicity, we assume in the statement and proof of Lemma C.3 that the convex program (7) may be efficiently solved exactly (i.e., that an oracle provides the answer). While strictly speaking this may not be the case, it is known that for any $\epsilon > 0$, an ϵ-approximate solution to (7) may be found in time $\mathrm{poly}(\alpha,\beta,n,d,H,\log(1/\epsilon))$. The guarantee of Lemma C.3 as well as those of subsequent lemmas hold with only minor modifications if we can only solve the program (7) ϵ-approximately in this manner. In particular, our main guarantee (Theorem 3.1) may be achieved computationally efficiently, without assumption of any oracle; the necessary modifications to the proof are described in Remark C.3.

Proof of Lemma C.3. By Lemma C.2 and a union bound over $i \in \mathcal{I}_h$, there is some event $\mathcal{E}_\pi$ with $\Pr(\mathcal{E}_\pi) \ge 1-\delta$ so that, under $\mathcal{E}_\pi$, the estimates $\hat{\phi}_i^\pi$ produced on Line 3 of Algorithm 2 satisfy $\max_{i\in[n]} \|\hat{\phi}_i^\pi - \phi_{h+1}(x_i', \pi_{h+1}(x_i'))\|_2 \le \varepsilon_{\mathsf{apx}}$ for all $i \in \mathcal{I}_h$. We prove the two statements of the lemma in turn:

Proof of Item 1. Let $w_h^\pi \in 2H \cdot \mathcal{B}_h$ be defined per Corollary B.3: in particular, $w_h^\pi = \mathcal{T}_h^\pi w_{h+1}^\pi$.

For each $h \in [H]$, define $\xi_h^\pi := \mathcal{T}_h^\pi w_{h+1}^\pi - \hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^\pi w_{h+1}^\pi$. We claim that the vectors $(w_h^\pi)_{h\in[H]}, (\xi_h^\pi)_{h\in[H]}$ constitute a feasible solution to the program (7). By the definition of $\hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^\pi$ in (22), note that (7b) requires that $w_h^\pi = \xi_h^\pi + \hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^\pi w_{h+1}^\pi$. This equality holds by definition of ξ_h^π and since $w_h^\pi = \mathcal{T}_h^\pi w_{h+1}^\pi$. Next, the fact that $\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}} \cap \mathcal{E}_\pi$ holds, $\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin},\sigma}$, and $w_h^\pi \in 2H \cdot \mathcal{B}_h$ (and hence $\|w_h^\pi\|_2 \le 2BH$; here we use Corollary B.3 together with the fact that $3BH\zeta_\sigma \le 1$) gives that $\|\xi_h^\pi\|_{\Sigma_h} = \|\mathcal{T}_h^\pi w_{h+1}^\pi - \hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^\pi w_{h+1}^\pi\|_{\Sigma_h} \le \alpha$; this verifies (7c). Finally, since $\beta = 2BH$ we have that (7d) holds. It follows that (7) is feasible under the event $\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}} \cap \mathcal{E}_\pi$.

Let us denote the solution to (7) by $w^\star, \xi^\star$. Since w_h^π, ξ_h^π are feasible to (7), we must have that $\langle w_1^\star, \phi_1(x_1,\pi_1(x_1))\rangle \le \langle w_1^\pi, \phi_1(x_1,\pi_1(x_1))\rangle = V_1^\pi(x_1)$. But Lemma C.1 together with our choice of $f = f^{w^\star}$ (so that $f_h(x,a) = \langle \phi_h(x,a), w_h^\star\rangle$) guarantees that $\langle w_1^\star, \phi_1(x_1,\pi_1(x_1))\rangle = V_1^{M^f,\pi}(x_1)$, which yields $V_1^{M^f,\pi}(x_1) \le V_1^\pi(x_1)$.

Proof of Item 2. For each $h \in [H]$ and $(x,a) \in \mathcal{X}\times\mathcal{A}$, we have, by the definition (21),

$$\begin{aligned} r_h^{M^{f,\pi}}(x,a) - r_h(x,a) &= \langle w_h^\star, \phi_h(x,a)\rangle - r_h(x,a) - \mathbb{E}_{x'\sim P_h(x,a)}[\langle w_{h+1}^\star, \phi_{h+1}(x', \pi_{h+1}(x'))\rangle] \\ &= \varepsilon_h(x,a) + \langle w_h^\star - \mathcal{T}_h^\pi w_{h+1}^\star, \phi_h(x,a)\rangle \\ &= \varepsilon_h(x,a) + \langle \xi_h^\star, \phi_h(x,a)\rangle + \langle \hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^\pi w_{h+1}^\star - \mathcal{T}_h^\pi w_{h+1}^\star, \phi_h(x,a)\rangle, \end{aligned} \tag{24}$$

where $\varepsilon_h(x,a)$ satisfies $|\varepsilon_h(x,a)| \le (1 + 2\|w_{h+1}^\star\|_2)\cdot\zeta_\sigma \le 3\beta\zeta_\sigma$ (for which such a choice is possible by Corollary B.2).

Since the MDP $M^{f,\pi}$ has transitions identical to those of M, for any $\pi' \in \Pi$,

$$\begin{aligned}
\left|V_1^{M^{f,\pi},\pi'}(x_1) - V_1^{M,\pi'}(x_1)\right| &= \left|\sum_{h=1}^{H} \mathbb{E}^{M,\pi'}\left[r_h^{M^{f,\pi}}(x_h,a_h) - r_h(x_h,a_h)\right]\right| \\
&\le 3\beta H\zeta_\sigma + \sum_{h=1}^{H}\left|\mathbb{E}^{M,\pi'}[\langle \xi_h^\star, \phi_h(x_h,a_h)\rangle]\right| + \left|\mathbb{E}^{M,\pi'}\left[\langle \hat{\mathcal{T}}^{\pi}_{h,\mathcal{D},\hat{\phi}} w_{h+1}^\star - \mathcal{T}_h^\pi w_{h+1}^\star, \phi_h(x_h,a_h)\rangle\right]\right| \\
&\le 3\beta H\zeta_\sigma + \sum_{h=1}^{H}\left\|\mathbb{E}^{M,\pi'}[\phi_h(x_h,a_h)]\right\|_{\Sigma_h^{-1}} \cdot \left(\|\xi_h^\star\|_{\Sigma_h} + \|\hat{\mathcal{T}}^{\pi}_{h,\mathcal{D},\hat{\phi}} w_{h+1}^\star - \mathcal{T}_h^\pi w_{h+1}^\star\|_{\Sigma_h}\right) \\
&\le 3\beta H\zeta_\sigma + 2\alpha\sum_{h=1}^{H}\left\|\mathbb{E}^{M,\pi'}[\phi_h(x_h,a_h)]\right\|_{\Sigma_h^{-1}}, \qquad (25)
\end{aligned}$$

where the first inequality uses (24) and the triangle inequality, and the third inequality uses the fact that $(w_h^\star, \xi_h^\star)$ is feasible for (7) (in particular, (7c)) as well as the fact that $\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}} \cap \mathcal{E}_\pi$ holds and $\|w_{h+1}^\star\|_2 \le \beta = 2BH$ (using (7d)).

Finally, we analyze the computational cost of Algorithm 2. The call to `EstFeature` in Line 3 takes time $\mathrm{poly}(N,d) \le \mathrm{poly}(d, \log(1/\delta)/\varepsilon_{\mathsf{apx}}^2)$. The remaining steps take time $\mathrm{poly}(d,H)$. □

Lemma C.4. *There is a constant $C_{C.4}$ so that the following holds. Suppose the dataset $\mathcal{D}$ is drawn according to Assumption 2.3, and $\alpha, \beta, \sigma, \varepsilon_{\mathsf{apx}} > 0$ are given so that $\varepsilon_{\mathsf{apx}} \le 1/\sqrt{n}$ and $\alpha \ge 4\beta\zeta_\sigma\sqrt{n} + C_{C.4}\beta d \log^{1/2}(dn\beta/(\sigma\delta))$. Then, $\Pr(\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}}) \ge 1 - \delta/2$, where the probability is over the draw of $\mathcal{D}$.*

We remark that the only source of randomness in `Critic` is over the randomness of `EstFeature` in Line 3.

Proof of Lemma C.4. Given $h \in [H]$, $w_{h+1} \in \mathbb{R}^d$, $\pi_{h+1} \in \Pi^{\mathsf{Plin},\sigma}$, and $\hat{\phi} = (\hat{\phi}_i)_{i\in[n]} \in (\mathbb{R}^d)^n$, let us write, for $i \in \mathcal{I}_h$,

$$\begin{aligned}
\varepsilon_i(w_{h+1},\pi_{h+1}) &:= r_i + \langle \phi_{h+1}(x_i', \pi_{h+1}(x_i')), w_{h+1}\rangle - (r_h(x_i,a_i) + \mathbb{E}_{x'\sim P_h(x_i,a_i)}[\langle \phi_{h+1}(x', \pi_{h+1}(x')), w_{h+1}\rangle]) \\
\xi_i(w_{h+1},\pi_{h+1}) &:= r_h(x_i,a_i) + \mathbb{E}_{x'\sim P_h(x_i,a_i)}[\langle \phi_{h+1}(x', \pi_{h+1}(x')), w_{h+1}\rangle] - \langle \phi_h(x_i,a_i), \mathcal{T}_h^{\pi_{h+1}} w_{h+1}\rangle \\
\eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i) &:= \langle \hat{\phi}_i - \phi_{h+1}(x_i', \pi_{h+1}(x_i')), w_{h+1}\rangle.
\end{aligned}$$

Error decomposition. For any $w_{h+1} \in \mathbb{R}^d$, $\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin},\sigma}$, and $\hat{\phi} = (\hat{\phi}_i)_{i\in[n]}$, we can decompose

$$
\begin{aligned}
&\hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^{\pi_{h+1}} w_{h+1} \\
=&\Sigma_h^{-1} \sum_{i\in\mathcal{I}_h} \phi_h(x_i,a_i)\cdot(r_i + \langle \hat{\phi}_i, w_{h+1}\rangle) \\
=&\Sigma_h^{-1} \sum_{i\in\mathcal{I}_h} \phi_h(x_i,a_i)\cdot(r_i + \langle \phi_{h+1}(x_i',\pi_{h+1}(x_i')), w_{h+1}\rangle) + \Sigma_h^{-1}\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i) \\
=&\Sigma_h^{-1} \sum_{i\in\mathcal{I}_h} \phi_h(x_i,a_i)\cdot\langle \phi_h(x_i,a_i), \mathcal{T}_h^{\pi_{h+1}} w_{h+1}\rangle \\
&+ \Sigma_h^{-1}\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\Big(\varepsilon_i(w_{h+1},\pi_{h+1}) + \xi_i(w_{h+1},\pi_{h+1}) + \eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i)\Big) \\
=&-\Sigma_h^{-1}\cdot\cdot\mathcal{T}_h^{\pi} w_{h+1} + \Sigma_h^{-1}\left(+\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\phi_h(x_i,a_i)^\top\right)\cdot\mathcal{T}_h^{\pi_{h+1}} w_{h+1} \\
&+ \Sigma_h^{-1}\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot(\varepsilon_i(w_{h+1},\pi_{h+1}) + \xi_i(w_{h+1},\pi_{h+1}) + \eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i)) \\
=&\mathcal{T}_h^{\pi_{h+1}} w_{h+1} - \Sigma_h^{-1}\cdot\mathcal{T}_h^{\pi_{h+1}} w_{h+1} + \\
&+ \Sigma_h^{-1}\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot(\eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i) + \xi_i(w_{h+1},\pi_{h+1}) + \varepsilon_i(w_{h+1},\pi_{h+1})),
\end{aligned}
\tag{26}
$$

where the final equality uses the definition of Σ_h (in Line 2 of Algorithm 2).

Bounding the error terms. First, we note that for any w_{h+1} satisfying $\|w_{h+1}\|_2 \le \beta$ and any $\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin},\sigma}$,

$$
\|\Sigma_h^{-1}\cdot\mathcal{T}_h^{\pi_{h+1}} w_{h+1}\|_{\Sigma_h} = \|\mathcal{T}_h^{\pi_{h+1}} w_{h+1}\|_{\Sigma_h^{-1}} \le \|\mathcal{T}_h^{\pi_{h+1}} w_{h+1}\|_2 \le 3\cdot\beta B, \tag{27}
$$

where the first inequality uses that $\Sigma_h \succeq I_d$ and the second inequality uses that, since $\|w_{h+1}\|_2 \le \beta$, we have $\mathcal{T}_h^{\pi_{h+1}} w_{h+1} \in 3\beta\cdot\mathcal{B}_h$ (Corollary B.2), and hence $\|\mathcal{T}_h^{\pi_{h+1}} w_{h+1}\|_2 \le 3\beta B$ (Assumption 2.2).

Next, as long as $\|w_{h+1}\|_2 \le \beta$ and $\|\hat{\phi}_i - \phi_{h+1}(x_i',\pi_{h+1}(x_i'))\|_2 \le \varepsilon_{\mathsf{apx}}$ for all $i \in [n]$, we have

$$
|\eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i)| \le \|\hat{\phi}_i - \phi_{h+1}(x_i',\pi_{h+1}(x_i'))\|_2 \cdot \|w_{h+1}\|_2 \le \beta\varepsilon_{\mathsf{apx}}
$$

for all $i \in [n]$ and thus

$$
\left\|\Sigma_h^{-1}\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i)\right\|_{\Sigma_h} = \left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\eta_i(w_{h+1},\pi_{h+1},\hat{\phi}_i)\right\|_{\Sigma_h^{-1}} \le \beta\varepsilon_{\mathsf{apx}}\sqrt{n}, \tag{28}
$$

where the inequality uses Lemma E.2.

Next, by Corollary B.2, for any w_{h+1} satisfying $\|w_{h+1}\|_2 \le \beta$ and any $\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin}}$, for each $i \in \mathcal{I}_h$ we have

$$
\left|r_h(x_i,a_i) + \mathbb{E}_{x'\sim P_h(x_i,a_i)}[\langle\phi_{h+1}(x',\pi_{h+1}(x')), w_{h+1}\rangle] - \langle\phi_h(x_i,a_i), \mathcal{T}_h^{\pi_{h+1}} w_{h+1}\rangle\right| \le \beta\cdot\zeta_\sigma.
$$

Thus, by Lemma E.2,

$$
\left\|\Sigma_h^{-1}\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\xi_i(w_{h+1},\pi_{h+1})\right\|_{\Sigma_h} = \left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\xi_i(w_{h+1},\pi_{h+1})\right\|_{\Sigma_h^{-1}} \le \beta\zeta_\sigma\sqrt{n}. \tag{29}
$$

It remains to bound the final term in (26), namely the one involving $\varepsilon_i(w_{h+1}, \pi_{h+1})$. To do so, we use a covering based argument. We define the following metric on $\Pi_h^{\mathsf{Plin},\sigma}$: for $\pi_h, \pi_h' \in \Pi_h^{\mathsf{Plin},\sigma}$, define

$$\|\pi_h - \pi_h'\|_{\infty,1} := \sup_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} |\pi_h(a|x) - \pi_h'(a|x)|. \tag{30}$$

For $\epsilon > 0$, we let $\mathcal{N}_{\infty,1}(\Pi_h^{\mathsf{Plin},\sigma}, \epsilon)$ denote the minimum size of an ϵ-cover of $\Pi_h^{\mathsf{Plin},\sigma}$ with respect to $\|\cdot\|_{\infty,1}$. Similarly, let $\mathcal{G} := \{w \in \mathbb{R}^d : \|w\|_2 \le \beta\}$, and let $\mathcal{N}_2(\mathcal{G}, \epsilon)$ denote the miniimum size of an ϵ-cover of $\mathcal{G}$ with respect to $\|\cdot\|_2$. We use the following lemma:

Lemma C.5 (Covering number bounds)**.** *There is a constant $C_{C.5}$ so that for all $h \in [H]$,*

$$\begin{aligned}\log \mathcal{N}_{\infty,1}(\Pi_h^{\mathsf{Plin},\sigma}, \epsilon) &\le d \log(C_{C.5}/(\epsilon\sigma)) \\ \log \mathcal{N}_2(\mathcal{G}, \epsilon) &\le d \log(C_{C.5}\beta/\epsilon).\end{aligned}$$

Let $\widetilde{\mathcal{G}} \subset \mathcal{G}$ be an ϵ-cover of $\mathcal{G}$ with respect to $\|\cdot\|_2$ with size bounded per Lemma C.5, and $\widetilde{\Pi}_{h+1}^{\mathsf{Plin},\sigma} \subset \Pi_{h+1}^{\mathsf{Plin},\sigma}$ be an ϵ-cover of $\Pi_{h+1}^{\mathsf{Plin},\sigma}$ with respect to $\|\cdot\|_{\infty,1}$ with size bounded per Lemma C.5.

For fixed $\widetilde{w}_{h+1} \in \widetilde{\mathcal{G}}$ and $\widetilde{\pi}_{h+1} \in \widetilde{\Pi}^{\mathsf{Plin},\sigma}$, for each $i \in \mathcal{I}_h$, we have from the definition of $\varepsilon_i(\widetilde{w}_{h+1}, \widetilde{\pi}_{h+1})$ that

$$|\varepsilon_i(\widetilde{w}_{h+1}, \widetilde{\pi}_{h+1})| \le 2 + 2\|\widetilde{w}_{h+1}\|_2 \le 2 + 2\beta \le 4\beta, \tag{31}$$

where we have used Assumption 2.2 and the fact that $\beta \ge 1$. Moreover, by Assumption 2.3, if we let $\mathcal{F}_i$ denote the sigma-algebra generated by all tuples $(h_j, x_j, a_j, r_j, x_j')$ for $j \le i$ and by $(h_{i+1}, x_{i+1}, a_{i+1})$, we have $\mathbb{E}[\varepsilon_i(\widetilde{w}_{h+1}, \widetilde{\pi}_{h+1})|\mathcal{F}_{i-1}] = 0$ and $\mathbb{E}[e^{\lambda\varepsilon_i(\widetilde{w}_{h+1}, \widetilde{\pi}_{h+1})}|\mathcal{F}_{i-1}] \le e^{\lambda^2 \cdot (4\beta)^2/2}$, where we have used (31). Moreover, $\phi_h(x_i, a_i)$ is measurable with respect to $\mathcal{F}_{i-1}$. Let us write $N := |\widetilde{\mathcal{G}}| \cdot |\widetilde{\Pi}^{\mathsf{Plin},\sigma}|$, so that $\log N \le Cd \log(\beta/\epsilon\sigma)$, for some constant C. By Lemma E.1, there is some event $\mathcal{E}_{\widetilde{w},\widetilde{\pi}_{h+1}}$ which occurs with probability at least $1 - \delta/N$, so that under $\mathcal{E}_{\widetilde{w},\widetilde{\pi}_{h+1}}$, we have

$$\left\|\sum_{i \in \mathcal{I}_h} \phi_h(x_i, a_i) \cdot \varepsilon_i(\widetilde{w}_{h+1}, \widetilde{\pi}_{h+1})\right\|_{\Sigma_h^{-1}} \le 32\beta^2 \log^{1/2}\left(\frac{\det(\Sigma_h)^{1/2}}{\delta/N}\right) \le 8\beta d^{1/2} \log^{1/2}\left(N \cdot \frac{d+n}{d\delta}\right),$$

where the final inequality uses

$$\det(\Sigma_h) \le \left(\frac{1}{d}\operatorname{Tr}\Sigma_h\right)^d \le \left(\frac{1}{d} \cdot \left(d + \sum_{i \in \mathcal{I}_h} \|\phi_h(x_i, a_i)\|_2^2\right)\right)^d \le \left(\frac{d+n}{d}\right)^d.$$

By a union bound, it follows that under some event $\mathcal{E}$ that occurs with probability at least $1 - \delta$, we have

$$\sup_{\substack{\widetilde{w} \in \widetilde{\mathcal{G}} \\ \widetilde{\pi}_{h+1} \in \widetilde{\Pi}_{h+1}^{\mathsf{Plin},\sigma}}} \left\|\sum_{i \in \mathcal{I}_h} \phi_h(x_i, a_i) \cdot \varepsilon_i(\widetilde{w}_{h+1}, \widetilde{\pi}_{h+1})\right\|_{\Sigma_h^{-1}} \le 8\beta d^{1/2} \log^{1/2}\left(N \cdot \frac{d+n}{d\delta}\right). \tag{32}$$

Now consider any $w_{h+1} \in \mathcal{G}$ and $\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin},\sigma}$. Choose $\widetilde{w}_{h+1} \in \widetilde{\mathcal{G}}$ and $\widetilde{\pi}_{h+1} \in \widetilde{\Pi}_{h+1}^{\mathsf{Plin},\sigma}$ so that $\|w_{h+1} - \widetilde{w}_{h+1}\|_2 \le \epsilon$ and $\|\pi_{h+1} - \widetilde{\pi}_{h+1}\|_{\infty,1} \le \epsilon$. We have

$$\sup_{\pi_{h+1} \in \Pi_{h+1}^{\mathsf{Plin},\sigma}} |\varepsilon_i(w_{h+1}, \pi_{h+1}) - \varepsilon_i(\widetilde{w}_{h+1}, \pi_{h+1})| \le 2 \cdot \|w_{h+1} - \widetilde{w}_{h+1}\|_2 \le 2\epsilon \tag{33}$$

$$\sup_{w_{h+1} \in \mathcal{G}} |\varepsilon_i(w_{h+1}, \pi_{h+1}) - \varepsilon_i(w_{h+1}, \widetilde{\pi}_{h+1})| \le 2\beta \cdot \sup_{x' \in \mathcal{X}} \|\phi_{h+1}(x', \pi_{h+1}(x')) - \phi_{h+1}(x', \widetilde{\pi}_{h+1}(x'))\|_2 \le 2\beta\epsilon, \tag{34}$$

where (34) uses the definition of $\|\cdot\|_{\infty,1}$ in (30). Then, under the event $\mathcal{E}$,

$$\begin{aligned}
&\left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\varepsilon_i(w_{h+1},\pi_{h+1})\right\|_{\Sigma_h^{-1}}\\
\le&\left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\varepsilon_i(\widetilde{w}_{h+1},\widetilde{\pi}_{h+1})\right\|_{\Sigma_h^{-1}}+\left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot(\varepsilon_i(\widetilde{w}_{h+1},\pi_{h+1})-\varepsilon_i(w_{h+1},\pi_{h+1}))\right\|_{\Sigma_h^{-1}}\\
&+\left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot(\varepsilon_i(\widetilde{w}_{h+1},\widetilde{\pi}_{h+1})-\varepsilon_i(\widetilde{w}_{h+1},\pi_{h+1}))\right\|_{\Sigma_h^{-1}}\\
\le&\left\|\sum_{i\in\mathcal{I}_h}\phi_h(x_i,a_i)\cdot\varepsilon_i(\widetilde{w}_{h+1},\widetilde{\pi}_{h+1})\right\|_{\Sigma_h^{-1}}+4\beta\epsilon\sqrt{n}\\
\le&8\beta d^{1/2}\log^{1/2}\left(N\cdot\frac{d+n}{d\delta}\right)+4\beta\epsilon\sqrt{n},
\end{aligned}\tag{35}$$

where the first inequality uses the triangle inequality, the second inequality uses Lemma E.2 together with Eqs. (33) and (34), and the final inequality uses (32) together with the fact that $\mathcal{E}$ holds.

Combining Eqs. (26) to (29) and (35), we conclude that under the event $\mathcal{E}$, for any $w_{h+1}\in\mathcal{G}$, $\pi_{h+1}\in\Pi_{h+1}^{\mathsf{Plin},\sigma}$, and $\hat{\phi}\in(\mathbb{R}^d)^n$ so that $\max_{i\in[n]}\|\hat{\phi}_i-\phi_{h+1}(x_i',\pi_{h+1}(x_i'))\|_2\le\varepsilon_{\mathsf{apx}}$,

$$\begin{aligned}
\|\hat{\mathcal{T}}_{h,\mathcal{D},\hat{\phi}}^{\pi_{h+1}}w_{h+1}-\mathcal{T}_h^{\pi_{h+1}}w_{h+1}\|_{\Sigma_h}\le&3\beta B+\beta\varepsilon_{\mathsf{apx}}\sqrt{n}+\beta\zeta_\sigma\sqrt{n}+8\beta d^{1/2}\log^{1/2}\left(N\cdot\frac{d+n}{d\delta}\right)+4\beta\epsilon\sqrt{n}\\
\le&11C\beta d\log^{1/2}\left(\frac{dn\beta}{\epsilon\delta\sigma}\right)+4\beta(\epsilon+\varepsilon_{\mathsf{apx}}+\zeta_\sigma)\sqrt{n},
\end{aligned}$$

where the second inequality uses the bound $\log N\le Cd\log(\beta/\epsilon\sigma)$. Choosing $\epsilon=1/\sqrt{n}$, and as long as $\varepsilon_{\mathsf{apx}}\le 1/\sqrt{n}$, we see that $\mathcal{E}\subset\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}}$, since we have chosen $\alpha\ge 4\beta\zeta_\sigma\sqrt{n}+C\beta d\log^{1/2}(dn\beta/(\sigma\delta))$ for a sufficiently large constant C. Thus $\Pr(\mathcal{E}_{\alpha,\beta,\sigma,\varepsilon_{\mathsf{apx}}})\ge\Pr(\mathcal{E})\ge 1-\delta$. Rescaling δ to $\delta/2$ yields the result. □

Proof of Lemma C.5. First, we note that $\log\mathcal{N}_2(\mathcal{G},\epsilon)\le d\cdot\log(3\beta/(\epsilon))$ by Wainwright (2019, Example 5.8). To bound the covering number of $\Pi_h^{\mathsf{Plin},\sigma}$, let $\mathcal{C}\subset\{w\in\mathbb{R}^d:\ \|w\|_2\le 1\}$ denote a $\epsilon\sigma$-cover of the unit ball, so that Wainwright (2019, Example 5.8) ensures that we can choose $\mathcal{C}$ with $|\mathcal{C}|\le d\log(3/(\epsilon\sigma))$. For any $w\in\mathbb{R}^d$ with $\|w\|_2\le 1$, there is some $w'\in\mathcal{C}$ with $\|w-w'\|_2\le\epsilon\sigma$. Then we have

$$D_{\mathsf{TV}}(\mathcal{N}(w,\sigma^2\cdot I_d),\mathcal{N}(w',\sigma^2\cdot I_d))\le\sqrt{\frac{1}{2}\cdot\mathrm{KL}(\mathcal{N}(w,\sigma^2\cdot I_d)||\mathcal{N}(w',\sigma^2\cdot I_d))}=\frac{1}{2\sigma}\cdot\|w-w'\|_2\le\frac{\epsilon}{2},$$

where we have used Pinsker's inequality and the formula for the KL divergence between two Gaussians. Moreover, for any $x\in\mathcal{X}$, we have

$$\sum_{a\in\mathcal{A}}|\pi_{h,w,\sigma}(a|x)-\pi_{h,w',\sigma}(a|x)|\le D_{\mathsf{TV}}(\mathcal{N}(w,\sigma^2\cdot I_d),\mathcal{N}(w',\sigma^2\cdot I_d))$$

by the data processing inequality for total variation distance (in particular, the deterministic function $W\mapsto\arg\max_{a\in\mathcal{A}}\langle\phi_h(x,a),W\rangle$ maps a random variable $W\sim\mathcal{N}(w,\sigma^2\cdot I_d)$ to an action $A\sim\pi_{h,w,\sigma}(\cdot|x)$). Combining the above displays gives that $\|\pi_{h,w,\sigma}-\pi_{h,w',\sigma}\|_{\infty,1}\le\epsilon$, as desired. □

C.2 Actor analysis

Notice that the `Actor` algorithm (Algorithm 1) is a special case of the Follow-the-Perturbed-Leader (FTPL) algorithm. This observation is central to our proof. Below we first review some basic facts pertaining to the FTPL algorithm.

Review of FTPL. For $J, L > 0$, a distribution μ on $\mathbb{R}^d$ is defined to be (J, L)-*stable* with respect to the Euclidean norm $\|\cdot\|_2$ if the following two conditions hold:

$$\mathbb{E}_{\rho\sim\mu}[\|\rho\|_2] \leq J$$
$$\forall v \in \mathbb{R}^d, \quad \int_{\rho\in\mathbb{R}^d} |\mu(\rho) - \mu(\rho - v)| d\rho \leq L \cdot \|v\|_2.$$

We consider the setting of *online linear optimization*: the algorithm is given a set of actions $\Phi \subset \mathbb{R}^d$, and operates over some number $T \in \mathbb{N}$ of rounds. At each round $t \in [T]$, an adversary chooses a *reward vector* $w^{(t)}$, which may be random and depend arbitrarily on past choices of the algorithm (i.e., we consider the case of an *adaptive adversary*). Simultaneously, the algorithm chooses a vector $\phi^{(t)} \in \Phi$, and then observes $w^{(t)}$ and receives a reward $\langle w^{(t)}, \phi^{(t)}\rangle$. The algorithm's goal is to minimize its regret with respect to the best-in-hindsight fixed choice of action in Φ. The *expected FTPL* algorithm, presented in Algorithm 4, solves the online linear optimization problem. Its choice of action at each round is given by the best action for the previous rounds, perturbed by a distribution which satisfies (J, L)-stability.

Algorithm 4 Expected Follow-the-Perturbed-Leader: $\texttt{ExpFTPL}(\Phi, \mu, T, \eta)$

Require: Action set $\Phi \subset \mathbb{R}^d$, distribution $\mu \in \Delta(\mathbb{R}^d)$, parameters $\omega > 0, T \in \mathbb{N}$.

1: **for** $1 \leq t \leq T$ **do**

2: Choose

$$\phi^{(t)} := \mathbb{E}_{\rho^{(t)}\sim\mu}\left[\arg\max_{\phi\in\mathcal{A}}\left\{\omega \cdot \sum_{s=1}^{t-1}\langle w^{(s)}, \phi\rangle + \langle \rho^{(t)}, \phi\rangle\right\}\right].$$

3: Receive reward vector $w^{(t)} \in \mathbb{R}^d$, earn reward $\langle \phi^{(t)}, w^{(t)}\rangle$.

Theorem C.6 (Hazan (2017), Theorem 5.8)**.** *Suppose that μ is (J, L)-stable, and write $D := \max_{\phi\in\Phi}\|\phi\|_2$. Then for any adaptive adversary choosing $w^{(1)}, \ldots, w^{(T)}$ satisfying $\|w^{(t)}\|_2 \leq G$ for all t, the iterates $\phi^{(t)}$ produced by $\texttt{ExpFTPL}(\Phi, \mu, T, \omega)$ (Algorithm 4) satisfy*

$$\max_{\phi^\star\in\Phi}\left\{\sum_{t=1}^{T}\langle \phi^\star, w^{(t)}\rangle - \sum_{t=1}^{T}\langle \phi^{(t)}, w^{(t)}\rangle\right\} \leq \omega LDG^2T + \frac{JD}{\omega}.$$

We remark that since the iterates $\phi^{(t)}$ of Algorithm 4 are deterministic, in the setting of no-regret learning (i.e., that of Theorem C.6), adaptive and oblivious adversaries are equivalent.

Analysis of `Actor`. Recall that for $w = (w_1, \ldots, w_H) \in \mathbb{R}^{dH}$, we have defined $f^w := (f_1^w, \ldots, f_H^w)$, where $f_h^w(x, a) := \langle \phi_h(x, a), w_h\rangle$. To simplify notation, we write $f^{(t)} := f^{w^{(t)}}$, where $w^{(t)} = (w_1^{(t)}, \ldots, w_H^{(t)})$ is the vector returned by `Critic` on Line 6 of `Actor` (Algorithm 1).

Lemma C.7. *For any $\mathcal{D}, \varepsilon_{\mathsf{final}}, \delta$, the algorithm $\texttt{Actor}(\mathcal{D}, \varepsilon_{\mathsf{final}}, \delta, \eta)$ (Algorithm 1) satisfies the following: for any $\pi^\star \in \Pi$, $h \in [H]$, and $x \in \mathcal{X}$, we have*

$$\sum_{t=1}^{T} f_h^{(t)}(x, \pi_h^\star(x)) - \sum_{t=1}^{T} f_h^{(t)}(x, \pi_h^{(t)}(x)) \leq \eta^{-1}(2BH)^2T + \eta\sqrt{d}.$$

We emphasize that the policy $\pi^\star$ in Lemma C.7 is not required to be a (perturbed) linear policy.

Proof of Lemma C.7. Fix $\pi^\star \in \Pi$ and $x \in \mathcal{X}$. Note that, by definition of $f_h^{(t)}$,

$$\sum_{t=1}^{T} f_h^{(t)}(x, \pi_h^\star(x)) = \sum_{t=1}^{T}\langle w_h^{(t)}, \phi_h(x, \pi_h^\star(x))\rangle,$$

and

$$\sum_{t=1}^{T} f_h^{(t)}(x, \pi^{(t)}(x)) = \sum_{t=1}^{T} \langle w_h^{(t)}, \phi_h(x, \pi_h^{(t)}(x)) \rangle.$$

Moreover, the definition of $\pi_h^{(t)}$ in `Actor` (Algorithm 1) ensures that

$$\begin{aligned}\phi_h(x, \pi_h^{(t)}(x)) =& \mathbb{E}_{\rho_h^{(t)} \sim \mathcal{N}(0, \eta^2 \cdot I_d)} \left[\arg\max_{a \in \mathcal{A}} \left\{ \left\langle \phi_h(x, a), \theta_h^{(t)} + \rho_h^{(t)} \right\rangle \right\} \right] \\ =& \mathbb{E}_{\rho_h^{(t)} \sim \mathcal{N}(0, \eta^2 \cdot I_d)} \left[\arg\max_{a \in \mathcal{A}} \left\{ \sum_{s=1}^{t-1} \langle \phi_h(x, a), w_h^{(s)} \rangle + \langle \phi_h(x, a), \rho_h^{(t)} \rangle \right\} \right],\end{aligned}$$

where we have written $\theta_h^{(t)} = \sum_{s=1}^{t-1} w_h^{(s)}$, as in Algorithm 1. In particular, $\phi_h(x, \pi_h^{(t)}(x))$ is exactly the choice of $\texttt{ExpFTPL}(\Phi, \mu, T, \omega)$ (Algorithm 4) at round t with action set $\Phi = \bar{\Phi}_h(x) = \mathrm{co}(\{\phi_h(x, a) : a \in \mathcal{A}\})$, distribution $\mu = \mathcal{N}(0, \eta^2 \cdot I_d)$, and $\omega = 1$. It follows from Theorem C.6 that

$$\begin{aligned}\sum_{t=1}^{T} f_h^{(t)}(x, \pi_h^\star(x)) - \sum_{t=1}^{T} f_h^{(t)}(x, \pi_h^{(t)}(x)) =& \sum_{t=1}^{T} \langle w_h^{(t)}, \phi_h(x, \pi_h^\star(x)) \rangle - \sum_{t=1}^{T} \langle w_h^{(t)}, \phi_h(x, \pi_h^{(t)}(x)) \rangle \\ \le& \eta^{-1} (2BH)^2 T + \eta \sqrt{d},\end{aligned}$$

where we have used that $\|\phi_h(x, a)\| \le 1$ for all x, a, h, that $\mathcal{N}(0, \eta^2 \cdot I_d)$ is $(\eta\sqrt{d}, \eta^{-1})$-stable (Lemma C.8), and that $\|w_h^{(t)}\|_2 \le \beta = 2BH$, using the constraint (7d) in `Critic` (and where the value of β is set on Line 1 of Algorithm 1 per Definition C.1). □

Lemma C.8. *For any $\eta > 0$, $\mathcal{N}(0, \eta^2 \cdot I_d)$ is $(\eta\sqrt{d}, \eta^{-1})$-stable.*

Proof. We first note that $\mathbb{E}_{Z \sim \mathcal{N}(0, \eta^2 I_d)}[\|Z\|_2] \le \eta \sqrt{\mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}[Z^2]} = \eta\sqrt{d}$.

Let $\mu_\eta : \mathbb{R}^d \to \mathbb{R}$ denote the probability density function of $\mathcal{N}(0, \eta^2 I_d)$. To verify the second condition of stability, we compute, for any $v \in \mathbb{R}^d$,

$$\begin{aligned}\int_{\rho \in \mathbb{R}^d} |\mu_\eta(\rho) - \mu_\eta(\rho - v)| d\rho =& 2 D_{\mathsf{TV}}(\mathcal{N}(0, \eta^2 I_d), \mathcal{N}(v, \eta^2 I_d)) \\ \le& \sqrt{2 \,\mathrm{KL}(\mathcal{N}(0, \eta^2 I_d) || \mathcal{N}(v, \eta^2 I_d))} = \|v\|_2 / \eta,\end{aligned}$$

where we have used Pinsker's inequality and the formula for KL divergence between Gaussians. □

Finally, we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. Let the parameters $T, \eta, \varepsilon_{\mathsf{apx}}, \alpha, \beta, \sigma$ be chosen as in Definition C.1. Note that these parameter settings ensure that $1/\sigma \le 1/\sqrt{\varepsilon_{\mathsf{BE}}}$, which in turn ensures that

$$3BH\zeta \le 3BH \cdot C_{B.2} \varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d \log(d/\varepsilon_{\mathsf{BE}}^2)} + 1/\sqrt{\varepsilon_{\mathsf{BE}}} \right) \le 1, \tag{36}$$

by our assumption that $\varepsilon_{\mathsf{BE}} \le c_0 (BH)^{-2} d^{-2}$ and as long as c_0 is sufficiently small.

Fix an arbitrary policy $\pi^\star \in \Pi$. Recall that $\pi^{(t)}$ denotes the policy chosen by `Actor` (Algorithm 1) in step $t \in [T]$, and $f^{(t)} = f^{w^{(t)}}$. Moreover, let us write $M^{(t)} := M^{f^{(t)}, \pi^{(t)}}$. The choices of $\alpha, \beta, \varepsilon_{\mathsf{apx}}$ in Definition C.1 ensure that, by Lemma C.4, we have that, over the draw of $\mathcal{D}$, $\Pr(\mathcal{E}_{\alpha, \beta, \sigma, \varepsilon_{\mathsf{apx}}}) \ge 1 - \delta/2$. We next wish to use Lemma C.3 with $\pi = \pi^{(t)}$, for each $t \in [T]$. To do so, we need to check that $\pi^{(t)} \in \Pi^{\mathsf{Plin}, \sigma}$: indeed, $\pi_h^{(t)} = \pi_{h, \theta_h^{(t)}, \eta}$, where $\theta_h^{(t)}$ satisfies $\|\theta_h^{(t)}\|_2 \le t \cdot \max_{s \le t} \|w_h^{(s)}\|_2 \le T\beta$ (by

Corollary B.3, which is applicable because of (36)). Then $\frac{\eta}{\|\theta_h^{(t)}\|_2} \geq \frac{\eta}{T\beta} = \sigma$, where we have used the definition of η, σ in Definition C.1.

If we let $\mathcal{F}^{(t)}$ denote the sigma-algebra generated by $\mathcal{D}, w^{(1)}, \ldots, w^{(t)}$, then Lemma C.3 ensures that, for each t, $\Pr(\mathcal{E}_{\pi^{(t)}} | \mathcal{F}^{(t-1)}) \geq 1 - \delta/(2T)$. (In particular, we use here that the randomness in the call to `EstFeature` in `Critic` is chosen independently at each step t.) By a union bound, it follows that $\Pr\left(\mathcal{E}_{\alpha,\sigma,\varepsilon_{\mathsf{apx}}} \cap \bigcap_{t\in[T]} \mathcal{E}_{\pi^{(t)}}\right) \geq 1 - \delta$. We write $\mathcal{E}^\star := \mathcal{E}_{\alpha,\sigma,\varepsilon_{\mathsf{apx}}} \cap \bigcap_{t\in[T]} \mathcal{E}_{\pi^{(t)}}$.

By Lemma C.3, under the event $\mathcal{E}^\star$, for each $t \in [T]$, we have

$$V_1^{M,\pi^\star}(x_1) - V_1^{M,\pi^{(t)}}(x_1) \leq V_1^{M^{(t)},\pi^\star}(x_1) - V_1^{M^{(t)},\pi^{(t)}}(x_1) + 3\beta H\zeta + 2\alpha \sum_{h=1}^H \|\mathbb{E}^{M\,\pi^\star}[\phi_h(x_h, a_h)]\|_{\Sigma_h^{-1}}. \tag{37}$$

Next, using the performance difference lemma applied (Lemma E.3) to the MDP $M^{(t)}$, we have

$$\begin{aligned}
&\sum_{t=1}^T V_1^{M^{(t)},\pi^\star}(x_1) - V_1^{M^{(t)},\pi^{(t)}}(x_1) \\
=&\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M^{(t)},\pi^\star}\left[Q_h^{M^{(t)},\pi^{(t)}}(x_h, \pi_h^\star(x_h)) - Q_h^{M^{(t)},\pi^{(t)}}(x_h, \pi_h^{(t)}(x_h))\right] \\
=&\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M^{(t)},\pi^\star}\left[f_h^{(t)}(x_h, \pi_h^\star(x_h)) - f_h^{(t)}(x_h, \pi_h^{(t)}(x_h))\right] \\
=&\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M,\pi^\star}\left[f_h^{(t)}(x_h, \pi_h^\star(x_h)) - f_h^{(t)}(x_h, \pi_h^{(t)}(x_h))\right] \\
=&\sum_{h=1}^H \mathbb{E}^{M,\pi^\star}\left[\sum_{t=1}^T f_h^{(t)}(x_h, \pi_h^\star(x_h)) - f_h^{(t)}(x_h, \pi_h^{(t)}(x_h))\right],
\end{aligned} \tag{38}$$

where the second equality uses Lemma C.1, the third equality uses the fact that the dynamics of $M^{(t)}$ are the same as those of M, and the final equality rearranges. Next, Lemma C.7 guarantees that for each possible choice of x_h, we have

$$\sum_{t=1}^T f_h^{(t)}(x_h, \pi_h^\star(x_h)) - \sum_{t=1}^T f_h^{(t)}(x_h, \pi_h^{(t)}(x_h)) \leq \eta^{-1}(2BH)^2 T + \eta\sqrt{d}. \tag{39}$$

Combining (38), (39), and (37) yields that, under $\mathcal{E}^\star$,

$$\frac{1}{T}\sum_{t=1}^T \left(V_1^{M,\pi^\star} - V_1^{M,\pi^{(t)}}(x_1)\right) \leq 2\alpha \sum_{h=1}^H \|\mathbb{E}^{M,\pi^\star}[\phi_h(x_h, a_h)]\|_{\Sigma_h^{-1}} + \eta^{-1}(2BH)^2 + \frac{\eta\sqrt{d}}{T} + 3\beta H\zeta. \tag{40}$$

Note that the choices of η, σ, ζ in Definition C.1 gives that $1/\sigma \leq \varepsilon_{\mathsf{BE}}^{-1/2}$, meaning that

$$\zeta \leq C_{B.2}\varepsilon_{\mathsf{BE}} d^{3/2} \cdot \left(\sqrt{d\log(d/(\varepsilon_{\mathsf{BE}}\sigma))} + \varepsilon_{\mathsf{BE}}^{-1/2}\right) \leq C\varepsilon_{\mathsf{BE}}^{1/2} d^{3/2}\log(1/\varepsilon_{\mathsf{BE}}), \tag{41}$$

for some constant $C > 0$; note that we have used that $\varepsilon_{\mathsf{BE}} \leq d^{-1}$ and thus $\sqrt{d\log d} \leq O(\varepsilon_{\mathsf{BE}}^{-1/2}\log^{1/2}(1/\varepsilon_{\mathsf{BE}}))$ in the above display. Combining Eqs. (40) and (41) gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \left(V_1^{M,\pi^\star} - V_1^{M,\pi^{(t)}}(x_1)\right) \\
\leq& 2\alpha \sum_{h=1}^H \|\mathbb{E}^{M,\pi^\star}[\phi_h(x_h, a_h)]\|_{\Sigma_h^{-1}} + \frac{4BHd^{1/4}}{\sqrt{T}} + 2BH\sqrt{d}\varepsilon_{\mathsf{BE}}^{1/2} + 6CBH^2 d^{3/2}\cdot\varepsilon_{\mathsf{BE}}^{1/2}\log(1/\varepsilon_{\mathsf{BE}}).
\end{aligned}$$

Thus, by our choice of $T = \frac{16B^2H^2d^{1/2}}{\varepsilon_{\mathsf{final}}^2}$ (Definition C.1), it follows that the policy $\hat{\pi} := \frac{1}{T}\sum_{t=1}^{T}\pi^{(t)}$ satisfies

$$\begin{aligned}
&V_1^{M,\pi^\star}(x_1) - V_1^{M,\hat{\pi}}(x_1) \\
\leq& 2\alpha \sum_{h=1}^{H} \|\mathbb{E}^{M,\pi^\star}[\phi_h(x_h,a_h)]\|_{\Sigma_h^{-1}} + \varepsilon_{\mathsf{final}} + O\left(BH^2d^{3/2}\cdot\varepsilon_{\mathsf{BE}}^{1/2}\log(1/\varepsilon_{\mathsf{BE}})\right) \\
\leq& O\left(d^{3/2}BH\varepsilon_{\mathsf{BE}}^{1/2}\log(1/\varepsilon_{\mathsf{BE}})\sqrt{n} + BHd\log^{1/2}(dnBH/(\varepsilon_{\mathsf{final}}\delta))\right)\cdot\left(\frac{H}{\sqrt{n}} + \sum_{h=1}^{H}\|\mathbb{E}^{M,\pi^\star}[\phi_h(x_h,a_h)]\|_{\Sigma_h^{-1}}\right) + \varepsilon_{\mathsf{final}},
\end{aligned}$$

which is equal to the desired bound.

Finally, we analyze the computational cost of Algorithm 1. The only nontrivial computation to be performed is the call to `Critic` in Line 6, which is made T times. Thus, by Lemma C.3, given an oracle which can solve the convex program (7), the overall computational cost of Algorithm 1 is $\mathrm{poly}(T,d,n,H,\log(1/\delta)/\varepsilon_{\mathsf{apx}}^2) \leq \mathrm{poly}(d,H,n,\log(1/\delta),1/\varepsilon_{\mathsf{final}})$. The same guarantee holds without existence of such an oracle; the necessary modifications to the proof are described below in Remark C.3. □

Remark C.2. Notice that in the case $\varepsilon_{\mathsf{BE}} = 0$, the final steps of the proof above yield the (slightly stronger) bound

$$\begin{aligned}
&V_1^{M,\pi^\star}(x_1) - V_1^{M,\hat{\pi}}(x_1) \\
\leq& O\left(\frac{BHd\log^{1/2}(dnBH/(\varepsilon_{\mathsf{final}}\delta))}{\sqrt{n}}\right)\cdot\sum_{h=1}^{H}\|\mathbb{E}^{M,\pi^\star}[\phi_h(x_h,a_h)]\|_{n\Sigma_h^{-1}} + \varepsilon_{\mathsf{final}}.
\end{aligned}$$

Remark C.3 (Solving the convex program)**.** In this remark, we discuss the minor modifications necessary to the proof of Theorem 3.1 to establish its guarantee without assumption of an oracle which can solve the program (7).

Note that the program (7) is a convex program in $O(dH)$ variables consisting of linear equalities and ℓ_2 norm constraints, for which all coefficients of the variables can be specified with $\log\mathrm{poly}(\alpha,\beta,d,n) \leq \log\mathrm{poly}(n,B,d,H,\log(1/\delta))$ bits. Thus, for any $\epsilon > 0$, the ellipsoid algorithm returns vectors $w,\xi \in \mathbb{R}^{dH}$ which satisfy the constraints up to ϵ error and ϵ-approximately minimize the objective (7a) in time $\mathrm{poly}(d,H,\log(nB/(\delta\epsilon)))$. Then it is immediate that Item 1 of Lemma C.3 gives only the weaker guarantee $V_1^{M^{\hat{f},\pi}}(x_1) \leq V_1^{M,\pi}(x_1) + \epsilon$. Moreover, in the proof of Item 2 of Lemma C.3, the right-hand side of (25) has an additional $O(\epsilon\cdot\alpha H)$ term (as (7c) holds up to additive ϵ), which may be absorbed in the first term (namely, $3\beta H\zeta_\sigma$) by increasing the constants, as long as ϵ is sufficiently small. In turn, Lemma C.3 is used to establish (37): thus this equation gains an additional term of ϵ on the right-hand side as well as additional constant factor. This additional terms propagate to degrade the bound of Theorem 3.1 by a constant factor.

D Proof of Theorem 4.1

In this section we prove Theorem 4.1. First, in Appendix D.1 we introduce the family of MDP instances used to prove the theorem, and in Appendix D.2 we analyze the performance of any algorithm on this family.

D.1 Construction of the family $\mathcal{M}_{\varepsilon_{\mathsf{BE}}}$.

Fix $\varepsilon_{\mathsf{BE}} > 0$. For simplicity we assume that $L := 1/\sqrt{\varepsilon_{\mathsf{BE}}}$ is an even integer. Given bits $b_{\mathsf{rew}}, b_{\mathsf{init}} \in \{0,1\}$ and $(b_{\ell,e})_{\ell\in[L],e\in\{0,1\}} \in \{0,1\}^{2L}$, we write $\mathbf{b} = (b_{\mathsf{rew}}, b_{\mathsf{init}}, (b_{\ell,e})_{\ell\in[L],e\in\{0,1\}}) \in \{0,1\}^{2L+2}$ to denote the collection of all these bits. We construct a class $\mathcal{M}_{\varepsilon_{\mathsf{BE}}} = \{M^{\mathbf{b}}\}_{\mathbf{b}\in\{0,1\}^{2L+2}}$ of MDPs $M^{\mathbf{b}}$, each of which has inherent Bellman error bounded above by $\varepsilon_{\mathsf{BE}}$ with respect to some fixed feature

mappings. All MDPs $M \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$ have horizon $H = 2$ and feature dimension $d = 2$. The state and action spaces of each $M \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$ are given as follows:

$$\mathcal{X} = \{\mathfrak{s}_1, \mathfrak{t}_1, \mathfrak{s}_2, \bar{\mathfrak{s}}_2, \mathfrak{q}_2\} \cup \{\mathfrak{s}_2^\zeta, \mathfrak{t}_{2,0}^\zeta, \mathfrak{t}_{2,1}^\zeta\}_{\zeta \in [0,1]} \qquad \mathcal{A} = \{0, 1, 2, 3\}, \tag{42}$$

where ζ ranges over $[0, 1]$. (We remark that we will only need to use the states $\mathfrak{s}_2^\zeta, \mathfrak{t}_{2,0}^\zeta, \mathfrak{t}_{2,1}^\zeta$ for values of ζ which are in $\{0, \varepsilon_{\mathsf{BE}}, 2\varepsilon_{\mathsf{BE}}, \ldots, \sqrt{\varepsilon_{\mathsf{BE}}}\}$, but to simplify notation we opt to define states corresponding to all $\zeta \geq 0$.)

Feature vectors. The feature vectors corresponding to $\mathcal{X}, \mathcal{A}$ are defined below. We use the convention that if we specify fewer than 4 actions at a state, then all remaining actions at the state have equal behavior (i.e., feature vectors and transition) to action 0 at that state. For normalizing constant $\alpha := (\sqrt{2})^{-1}$, we define, for each $h \in [2]$:

- $\phi_h(\mathfrak{s}_1, 0) = \alpha \cdot (1, 0)$, $\phi_h(\mathfrak{s}_1, 1) = \alpha \cdot (0, 1)$.
- $\phi_h(\mathfrak{t}_1, 0) = \alpha \cdot (1, 1)$ and $\phi_h(\mathfrak{t}_1, 1) = \alpha \cdot (1, -1)$.
- $\phi_h(\bar{\mathfrak{s}}_2, 0) = \alpha \cdot (0, 0)$.
- $\phi_h(\mathfrak{s}_2, 0) = \alpha \cdot (0, 1)$, $\phi_h(\mathfrak{s}_2, 1) = \alpha \cdot (0, -1)$.
- For each $\zeta \in [0, 1]$, $\phi_h(\mathfrak{s}_2^\zeta, 0) = \alpha \cdot (0, \zeta)$, $\phi_h(\mathfrak{s}_2^\zeta, 1) = \alpha \cdot (1, 0)$, $\phi_h(\mathfrak{s}_2^\zeta, 2) = \alpha \cdot (0, -\zeta)$, $\phi_h(\mathfrak{s}_2^\zeta, 3) = \alpha \cdot (-1, 0)$.
- For each $\zeta \in [0, 1]$ and $b \in \{0, 1\}$, $\phi_h(\mathfrak{t}_{2,b}^\zeta, 0) = \alpha \cdot (1, (1-2b)\zeta)$ and $\phi_h(\mathfrak{t}_{2,b}^\zeta, 1) = \alpha \cdot (-1, -(1-2b)\zeta)$. Via slight abuse of notation, we identify $\mathfrak{t}_{2,0}^0, \mathfrak{t}_{2,1}^0$, i.e., $\mathfrak{t}_{2,0}^0 = \mathfrak{t}_{2,1}^0$.

Transitions and rewards. Consider any $M := M^{\mathbf{b}} \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$, and write $\mathbf{b} = (b_{\mathsf{rew}}, b_{\mathsf{init}}, \{b_{\ell,e}\}_{\ell \in [L], e \in \{0,1\}}) \in \{0, 1\}^{2L+2}$. We proceed to define the initial state distribution, transitions and rewards of M. The initial state distribution has all its mass concentrated on $\mathfrak{t}_1$. The transitions are defined as follows:

- $(\mathfrak{s}_1, 1)$ transitions to $\bar{\mathfrak{s}}_2$ with probability 1.
- $(\mathfrak{s}_1, 0)$ and $(\mathfrak{t}_1, b_{\mathsf{init}})$ each transition to the distribution that puts mass $1/L$ on each of the states $\mathfrak{s}_2^{\ell \cdot \varepsilon_{\mathsf{BE}}}$, for $\ell \in [L]$.
- $(\mathfrak{t}_1, 1 - b_{\mathsf{init}})$ transitions to the distribution which:
  - Puts mass $1/(2L)$ on $\mathfrak{t}_{2,0}^{(\ell - b_{\ell,0}) \cdot \varepsilon_{\mathsf{BE}}}$ for each $\ell \in [L]$;
  - Puts mass $1/(2L)$ on $\mathfrak{t}_{2,1}^{(\ell - b_{\ell,1}) \cdot \varepsilon_{\mathsf{BE}}}$ for each $\ell \in [L]$.

The rewards are defined as follows:

- $r_1(x, a) = 0$ for all $x \in \mathcal{X}, a \in \mathcal{A}$, i.e., rewards at step 1 are linear with $\theta_1^{\mathsf{r}} = (0, 0)$.
- The rewards at step 2 are linear with respect to the coefficient vector $\theta_2^{\mathsf{r}} = (1 - 2b_{\mathsf{rew}}, 1/R)$, where $R := 16$.

Verifying low inherent bellman error. Next, we verify that each MDP $M^{\mathbf{b}} \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$ has low inherent Bellman error and satisfies Assumption 2.2 with respect to the feature mappings ϕ_h defined above.

Lemma D.1 (Low inherent Bellman error and boundedness)**.** *For any $M = M^{\mathbf{b}} \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$, M has inherent Bellman error $2\varepsilon_{\mathsf{BE}}$ and satisfies Assumption 2.2 with $B = \sqrt{2}$.*

Proof. It is straightforward to see that

$$\mathcal{B}_1 = \mathcal{B}_2 = \left\{w \in \mathbb{R}^2 \ : \ \alpha|w_1| + \alpha|w_2| \le 1\right\},$$

meaning that the second item of Assumption 2.2 is satisfied with $B = 1/\alpha = \sqrt{2}$. Moreover, by choice of α the first item is immediate, and the third item holds since for all (x, a), $|r_2(x, a)| \le \alpha \cdot \|\theta^{\mathsf{r}}_2\|_1 = \alpha \cdot (1 + 1/R) \le 1$.

We must verify (1) for $h \in \{1, 2\}$. We begin with the case $h = 1$. For $\theta \in \mathbb{R}^2$, define

$$\mathcal{T}_1\theta := \left(\frac{1}{L}\sum_{\ell=1}^{L}\left(\max_{a\in\mathcal{A}}\langle\theta, \phi_2(\mathfrak{s}_2^{\ell\cdot\varepsilon_{\mathsf{BE}}}, a)\rangle\right), 0\right),$$

which belongs to $\mathcal{B}_1$ since its first coordinate is bounded above in absolute value by $\alpha \cdot \max\{|\theta_1|, L\varepsilon_{\mathsf{BE}}|\theta_2|\} \le \alpha \cdot (|\theta_1| + |\theta_2|) \le 1$.

Since $r_1(x, a) = 0$ for all x, a, we certainly have that, for each $(x, a) \in (\{\mathfrak{s}_1\} \times \mathcal{A}) \cup (\mathfrak{t}_1, b_{\mathsf{init}})$, $(\phi_1(x, a))_1 = 1$ and thus

$$\langle\phi_h(x, a), \mathcal{T}_1\theta\rangle = \mathbb{E}_{x'\sim P_h^M(x,a)}\left[r_1(x, a) + \max_{a'\in\mathcal{A}}\langle\phi_2(x', a'), \theta\rangle\right], \tag{43}$$

meaning that (1) holds for these (x, a). Now consider the state-action pair $(x, a) = (\mathfrak{t}_1, 1 - b_{\mathsf{init}})$. For each $\ell \in [L]$ and $\theta \in \mathbb{R}^2$, we have

$$\begin{aligned}
&\left|\max_{a'\in\mathcal{A}}\langle\phi_2(\mathfrak{s}_2^{\ell\varepsilon_{\mathsf{BE}}}, a'), \theta\rangle - \frac{1}{2}\max_{a'\in\mathcal{A}}\langle\phi_2(\mathfrak{t}_{2,0}^{(\ell-b_{\ell,0})\varepsilon_{\mathsf{BE}}}, a'), \theta\rangle - \frac{1}{2}\max_{a'\in\mathcal{A}}\langle\phi_2(\mathfrak{t}_{2,1}^{(\ell-b_{\ell,1})\varepsilon_{\mathsf{BE}}}, a'), \theta\rangle\right| \\
\le&\alpha|\theta_2|\varepsilon_{\mathsf{BE}} + \left|\max_{a'\in\mathcal{A}}\langle\phi_2(\mathfrak{s}_2^{\ell\varepsilon_{\mathsf{BE}}}, a'), \theta\rangle - \frac{1}{2}\max_{a'\in\mathcal{A}}\langle\phi_2(\mathfrak{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}, a'), \theta\rangle - \frac{1}{2}\max_{a'\in\mathcal{A}}\langle\phi_2(\mathfrak{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}}, a'), \theta\rangle\right| = \alpha|\theta_2|\varepsilon_{\mathsf{BE}} \le \varepsilon_{\mathsf{BE}}.
\end{aligned}$$

It follows that

$$\left|\langle\phi_h(\mathfrak{t}_1, 1 - b_{\mathsf{init}}), \mathcal{T}_1\theta\rangle - \mathbb{E}_{x'\sim P_h^M(\mathfrak{t}_1, b_{\mathsf{init}})}\left[r_1(\mathfrak{t}_1, b_{\mathsf{init}}) + \max_{a'\in\mathcal{A}}\langle\phi_2(x', a'), \theta\rangle\right]\right| \le \varepsilon_{\mathsf{BE}},$$

verifying (1) holds for $(x, a) = (\mathfrak{t}_1, b_{\mathsf{init}})$. Finally, the validity of (1) for $(x, a) = (\mathfrak{s}_1, 1)$ is immediate since $(\mathcal{T}_1\theta)_2 = 0$.

Next we verify (1) for $h = 2$. Since all feature vectors are identically 0 at step $h = H + 1 = 3$ (by convention), we take $\mathcal{T}_2\theta = \theta^{\mathsf{r}}_2$ (which is in $\mathcal{B}_2$ since $\alpha \cdot (1 + 1/R) \le 1$), and satisfies $r_2(x, a) = \langle\phi_2(x, a), \theta^{\mathsf{r}}_2\rangle$ for all (x, a). □

D.2 Proof of Theorem 4.1

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. First, note that if $1/\sqrt{n} \ge \sqrt{\varepsilon_{\mathsf{BE}}}$, then the lower bound is straightforward and well-known. In particular, consider $\mathcal{X} := \{\mathfrak{s}_1, \mathfrak{s}_2\}, \mathcal{A} := \{0, 1\}$ with $\phi_1(x, 0) = (1, 0), \phi_1(x, 1) = (0, 1)$ for each $x \in \mathcal{X}$, and $\phi_2(\mathfrak{s}_1, a) = (1, 0), \phi_2(\mathfrak{s}_2, a) = (0, 1)$ for each $a \in \mathcal{A}$. We let the rewards be linear with respect to some vectors $\theta^{\mathsf{r}}_1, \theta^{\mathsf{r}}_2 \in \mathbb{R}^2$ with $\theta^{\mathsf{r}}_1 = (0, 0)$ some choice of $\theta^{\mathsf{r}}_2 \in \{(0, 1), (1, 0)\}$. The initial state is $\mathfrak{s}_1$; $(\mathfrak{s}_1, 0)$ transitions to $\mathfrak{s}_1$ with probability $1/2 + (10\sqrt{n})^{-1}$ (and to $\mathfrak{s}_2$ with the remaining probability), and $(\mathfrak{s}_1, 1)$ transitions to $\mathfrak{s}_1$ with probability $1/2 - (10\sqrt{n})^{-1}$ (and to $\mathfrak{s}_2$ with the remaining probability). Finally, the dataset $\mathcal{D}$ consists of $n/4$ transitions from each of the tuples $(x, a) \in \{(\mathfrak{s}_1, 0), (\mathfrak{s}_1, 1), (\mathfrak{s}_2, 0), (\mathfrak{s}_2, 0)\}$. It is straightforward to see that the inherent Bellman error is 0 and that Assumptions 2.2 and 2.3 are satisfied. Moreover, it follows from well-known arguments (Lattimore & Szepesvári, 2020, Chapter 15) that for any algorithm $\mathfrak{A}$, there is some choice of θ^{r}_2 as above so that the optimal policy $\pi^\star$ (i.e., defined by $\pi^\star_1(\mathfrak{s}_1) = 0$ if $\theta^{\mathsf{r}}_2 = (1, 0)$, and $\pi^\star_1(\mathfrak{s}_1) = 1$ if $\theta^{\mathsf{r}}_2 = (0, 1)$) and the output policy $\hat{\pi}$ of $\mathfrak{A}$ satisfy (8).

For the remainder of the proof we may therefore assume that $\sqrt{\varepsilon_{\mathsf{BE}}} > 1/\sqrt{n}$. Moreover, by decreasing $\varepsilon_{\mathsf{BE}}$ by a constant factor, we may assume that $L = 1/\sqrt{\varepsilon_{\mathsf{BE}}}$ is an even integer. We use the state and action spaces defined in (42) and the feature mappings ϕ_h defined in Appendix D.1 above. Fix some randomized offline RL algorithm $\mathfrak{A}$. We will choose some MDP $M \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$ and define a distribution over datasets $\mathcal{D}$ satisfying Assumption 2.3 so that (8) holds for some $\pi^\star$.

Consider some $\mathbf{b} = (b_{\mathsf{rew}}, b_{\mathsf{init}}, (b_{\ell,e})_{\ell\in[L],e\in\{0,1\}}) \in \{0,1\}^{2L+2}$, to be specified below, and set $M = M^{\mathbf{b}}$. By Lemma D.1, M has inherent Bellman error $2\varepsilon_{\mathsf{BE}}$ and satisfies Assumption 2.2 with $B = \sqrt{2}$. The dataset $\mathcal{D}$ consists of tuples $(h_i, x_i, a_i, r_i, x_i')$, $i \in [n]$, drawn as follows:

- There are $n/3$ points of the form $(1, \mathfrak{s}_1, 1, 0, \bar{\mathfrak{s}}_2)$.
- There are $n/3$ points of the form $(1, \mathfrak{s}_1, 0, 0, x_i)$, where $x_i \sim P_1^M(\cdot \mid \mathfrak{s}_1, 0)$ for $i \in [n/3]$.
- There are $n/3$ points of the form $(2, \mathfrak{s}_2^{L\varepsilon_{\mathsf{BE}}}, 0, L\varepsilon_{\mathsf{BE}}\alpha/R, \perp)$.

It is immediate that the distribution of $\mathcal{D}$ satisfies Assumption 2.3.

Controlling the coverage coefficient. Note that we have

$$\Sigma_1 = \frac{n}{3} \cdot \phi_1(\mathfrak{s}_1, 1)\phi_1(\mathfrak{s}_1, 1)^\top + \frac{n}{3} \cdot \phi_1(\mathfrak{s}_1, 0)\phi_1(\mathfrak{s}_1, 0)^\top = \frac{\alpha^2 n}{3} \cdot I_2$$
$$\Sigma_2 = \frac{n}{3} \cdot \phi_2(\mathfrak{s}_2, 0)\phi_2(\mathfrak{s}_2, 0)^\top = \frac{n \cdot (\alpha L \varepsilon_{\mathsf{BE}})^2}{3} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Define $w_1^\star := (1, 1-2b_{\mathsf{init}})$, $w_2^\star := (0, 1)$, and $\pi_h^\star := \pi_{h,w_h^\star,0}$ for $h \in [2]$, and write $\pi^\star = (\pi_1^\star, \pi_2^\star) \in \Pi^{\mathsf{Plin}}$. Note that

$$\mathbb{E}^{M,\pi^\star}[\phi_1(x_1, a_1)] = \alpha \cdot (1, 1 - 2b_{\mathsf{init}}), \qquad \mathbb{E}^{M,\pi^\star}[\phi_2(x_2, a_2)] = \alpha \cdot (0, L\varepsilon_{\mathsf{BE}}),$$

so that

$$\|\mathbb{E}^{M,\pi^\star}[\phi_1(x_1, a_1)]\|_{n\Sigma_1^{-1}} = \sqrt{6}, \qquad \|\mathbb{E}^{M,\pi^\star}[\phi_2(x_2, a_2)]\|_{n\Sigma_2^{-1}} = \sqrt{3}.$$

The value function of $\pi^\star$ for $M = M^{\mathbf{b}} \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$ does not depend on the choice of $\mathbf{b}$ and may be computed as follows: first, note that $V_2^{M,\pi^\star}(\mathfrak{s}_2^\zeta) = \langle \phi_2(\mathfrak{s}_2^\zeta, 0), \theta_2^{\mathsf{r}} \rangle = \alpha\zeta/R$, for each $\zeta \in [0, 1]$, which implies that $V_1^{M,\pi^\star}(\mathfrak{t}_1) = Q_1^{M,\pi^\star}(\mathfrak{t}_1, b_1) = V_2^{M,\pi^\star}(\mathfrak{s}_2) = \frac{\alpha}{R} \cdot \frac{(1+\cdots+L)\varepsilon_{\mathsf{BE}}}{L} = \frac{\alpha}{R} \cdot (L+1)\varepsilon_{\mathsf{BE}}/2$.

Controlling the performance of $\mathfrak{A}$. Let $\hat{\pi}$ denote the (random) output of the algorithm $\mathfrak{A}$, given the random dataset $\mathcal{D}$. Note that the distribution $P_1^M(\cdot \mid \mathfrak{s}_1, 0)$ does not depend on the choice of $\mathbf{b}$. Thus, the distribution of $\hat{\pi}$, which we denote by $\mathcal{D}_0 \in \Delta(\Pi^{\mathsf{M}})$, is the same for all possible choices of $\mathbf{b}$. Lemma D.2 below, which is the main technical component of the proof, yields that there is some $\mathbf{b}$ so that (8) holds (as we have assumed $1/\sqrt{n} < \sqrt{\varepsilon_{\mathsf{BE}}}$), which completes the proof of Theorem 4.1.

Lemma D.2. *Let $\mathcal{D}_0 \in \Delta(\Pi^{\mathsf{M}})$ be an arbitrary distribution over Markov policies. Then there is some choice of $\mathbf{b}$ so that, for $M := M^{\mathbf{b}} \in \mathcal{M}_{\varepsilon_{\mathsf{BE}}}$, we have*

$$\mathbb{E}_{\pi\sim\mathcal{D}_0}\left[V_1^{M,\pi}(\mathfrak{t}_1)\right] \le V_1^{M,\pi^\star}(\mathfrak{t}_1) - \Omega\left(\sqrt{\varepsilon_{\mathsf{BE}}}\right). \tag{44}$$

The proof of Lemma D.2 is provided below. □

Proof of Lemma D.2. Let $\mathcal{D}_0 \in \Delta(\Pi^{\mathsf{M}})$ be given. Choose $b_{\mathsf{init}} \in \{0, 1\}$ so that $Z_0 := \mathbb{E}_{\pi\sim\mathcal{D}_0}[\pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1)] \ge 1/2$. For $0 \le \ell \le L$, define

$$\bar{\eta}(\ell) := \mathbb{E}_{\pi\sim\mathcal{D}_0}\left[\pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot \left(\pi(0 \mid \mathfrak{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(1 \mid \mathfrak{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}})\right)\right]$$
$$\bar{\gamma}(\ell) := \mathbb{E}_{\pi\sim\mathcal{D}_0}\left[\pi(b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot \left(\pi(1 \mid \mathfrak{s}_2^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(3 \mid \mathfrak{s}_2^{\ell\varepsilon_{\mathsf{BE}}})\right)\right].$$

Let $\mathcal{D} \in \Delta(\Pi)$ be defined by

$$\mathcal{D}(\pi) := \frac{\mathcal{D}_0(\pi) \cdot \pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1)}{\mathbb{E}_{\pi' \sim \mathcal{D}_0}[\pi'(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1)]} = \frac{\mathcal{D}_0(\pi) \cdot \pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1)}{Z_0}.$$

For $0 \le \ell \le L$, define

$$\rho(\ell) := \mathbb{E}_{\pi \sim \mathcal{D}} \left[\pi(0 \mid \mathfrak{t}_{2,0}^{\ell \varepsilon_{\mathsf{BE}}}) + \pi(1 \mid \mathfrak{t}_{2,1}^{\ell \varepsilon_{\mathsf{BE}}}) \right].$$

Consider any choice of $b_{\mathsf{rew}} \in \{0,1\}$ and $b_{\ell,e} \in \{0,1\}$ for each $\ell \in [L], e \in \{0,1\}$, and write $\mathbf{b} = (b_{\mathsf{rew}}, b_{\mathsf{init}}, (b_{\ell,e})_{\ell \in [L], e \in \{0,1\}})$. Then for any policy π, $\zeta > 0$, and $b \in \{0,1\}$, we have

$$\begin{aligned}
V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,b}^{\zeta}) =& \pi(b \mid \mathfrak{t}_{2,b}^{\zeta}) \cdot \langle \phi_2(\mathfrak{t}_{2,b}^{\zeta}, b), \theta_2^{\mathsf{r}} \rangle + \pi(1-b \mid \mathfrak{t}_{2,b}^{\zeta}) \cdot \langle \phi_2(\mathfrak{t}_{2,b}^{\zeta}, 1-b), \theta_2^{\mathsf{r}} \rangle \\
=& \alpha \cdot \pi(b \mid \mathfrak{t}_{2,b}^{\zeta}) \cdot \langle (1-2b, \zeta), (1 - 2b_{\mathsf{rew}}, 1/R) \rangle + \alpha \cdot (1 - \pi(b \mid \mathfrak{t}_{2,b}^{\zeta})) \cdot \langle (2b-1, -\zeta), (1 - 2b_{\mathsf{rew}}, 1/R) \rangle \\
=& 2\alpha \cdot \pi(b \mid \mathfrak{t}_{2,b}^{\zeta}) \cdot ((1-2b)(1-2b_{\mathsf{rew}}) + \zeta/R) - \alpha \cdot ((1-2b)(1-2b_{\mathsf{rew}}) + \zeta/R). \qquad (45)
\end{aligned}$$

Hence

$$V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,0}^{\zeta}) + V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,1}^{\zeta}) = 2\alpha(1-2b_{\mathsf{rew}}) \cdot \left(\pi(0 \mid \mathfrak{t}_{2,0}^{\zeta}) - \pi(1 \mid \mathfrak{t}_{2,1}^{\zeta}) \right) + 2\zeta\alpha R^{-1} \left(\pi(0 \mid \mathfrak{t}_{2,0}^{\zeta}) + \pi(1 \mid \mathfrak{t}_{2,1}^{\zeta}) - 1 \right). \qquad (46)$$

Moreover,

$$\begin{aligned}
V_2^{M^{\mathbf{b}},\pi}(\mathfrak{s}_2^{\zeta}) =& \alpha \cdot (1 - 2b_{\mathsf{rew}}) \cdot \left(\pi(1 \mid \mathfrak{s}_2^{\zeta}) - \pi(3 \mid \mathfrak{s}_2^{\zeta}) \right) + \alpha R^{-1} \zeta \cdot \left(\pi(0 \mid \mathfrak{s}_2^{\zeta}) - \pi(2 \mid \mathfrak{s}_2^{\zeta}) \right) \\
\le& \alpha \cdot (1 - 2b_{\mathsf{rew}}) \cdot \left(\pi(1 \mid \mathfrak{s}_2^{\zeta}) - \pi(3 \mid \mathfrak{s}_2^{\zeta}) \right) + \alpha R^{-1} \zeta. \qquad (47)
\end{aligned}$$

Case 1: $\left| \frac{1}{L} \sum_{\ell=1}^{L} (\bar{\eta}(\ell) + \bar{\gamma}(\ell)) \right| > \frac{\sqrt{\varepsilon_{\mathsf{BE}}}}{10}$**.** Recall our choice b_{init} from above. Set $b_{\mathsf{rew}} = 0$ if $\sum_{\ell=1}^{\ell} (\bar{\eta}(\ell) + \bar{\gamma}(\ell)) < 0$, and otherwise $b_{\mathsf{rew}} = 1$. Finally set $b_{\ell,e} = 0$ for all $\ell \in [L], e \in \{0,1\}$, and write $\mathbf{b} = (b_{\mathsf{rew}}, b_{\mathsf{init}}, (b_{\ell,e})_{\ell,e})$. Therefore,

$$\begin{aligned}
& \mathbb{E}_{\pi \sim \mathcal{D}_0} \left[V_1^{M^{\mathbf{b}},\pi}(\mathfrak{t}_1) \right] \\
=& \frac{1}{2L} \sum_{\ell=1}^{L} \mathbb{E}_{\pi \sim \mathcal{D}_0} \left[\pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot \left(V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,0}^{\ell \varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,1}^{\ell \varepsilon_{\mathsf{BE}}}) \right) + 2\pi(b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot V_2^{M^{\mathbf{b}},\pi}(\mathfrak{s}_2^{\ell \varepsilon_{\mathsf{BE}}}) \right] \qquad (48) \\
\le& \frac{1}{2L} \sum_{\ell=1}^{L} 2\alpha(1-2b_{\mathsf{rew}}) \cdot \mathbb{E}_{\pi \sim \mathcal{D}_0} \left[\pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot (\pi(0 \mid \mathfrak{t}_{2,0}^{\ell \varepsilon_{\mathsf{BE}}}) - \pi(1 \mid \mathfrak{t}_{2,1}^{\ell \varepsilon_{\mathsf{BE}}})) \right] \\
& + \frac{1}{2L} \sum_{\ell=1}^{L} 2\alpha R^{-1} \ell \varepsilon_{\mathsf{BE}} \mathbb{E}_{\pi \sim \mathcal{D}_0} \left[\pi(1 - b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot (\pi(0 \mid \mathfrak{t}_{2,0}^{\ell \varepsilon_{\mathsf{BE}}}) + \pi(1 \mid \mathfrak{t}_{2,1}^{\ell \varepsilon_{\mathsf{BE}}}) - 1) \right] \\
& + \frac{1}{L} \sum_{\ell=1}^{L} \left(\alpha(1-2b_{\mathsf{rew}}) \cdot \mathbb{E}_{\pi \sim \mathcal{D}_0} \left[\pi(b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot (\pi(1 \mid \mathfrak{s}_2^{\ell \varepsilon_{\mathsf{BE}}}) - \pi(3 \mid \mathfrak{s}_2^{\ell \varepsilon_{\mathsf{BE}}})) \right] + \alpha R^{-1} \mathbb{E}_{\pi \sim \mathcal{D}_0} \left[\pi(b_{\mathsf{init}} \mid \mathfrak{t}_1) \cdot \ell \varepsilon_{\mathsf{BE}} \right] \right) \\
=& \alpha(1-2b_{\mathsf{rew}}) \frac{1}{L} \sum_{\ell=1}^{L} (\bar{\eta}(\ell) + \bar{\gamma}(\ell)) + \frac{1}{L} \sum_{\ell=1}^{L} Z_0 \cdot \alpha R^{-1} \ell \varepsilon_{\mathsf{BE}} \cdot (\rho(\ell) - 1) + \frac{1}{L} \sum_{\ell=1}^{L} \alpha R^{-1} (1 - Z_0) \cdot \ell \varepsilon_{\mathsf{BE}} \qquad (49) \\
\le& - \frac{\alpha}{L} \left| \sum_{\ell=1}^{L} (\bar{\eta}(\ell) + \bar{\gamma}(\ell)) \right| + \alpha R^{-1} (L+1) \varepsilon_{\mathsf{BE}}/2 \\
=& - \frac{\alpha}{L} \left| \sum_{\ell=1}^{L} (\bar{\eta}(\ell) + \bar{\gamma}(\ell)) \right| + V_1^{M^{\mathbf{b}},\pi^\star}(\mathfrak{t}_1) \le - \frac{\alpha \sqrt{\varepsilon_{\mathsf{BE}}}}{10} + V_1^{M^{\mathbf{b}},\pi^\star}(\mathfrak{t}_1),
\end{aligned}$$

where the first inequality uses Eqs. (46) and (47), and the second inequality uses $\rho(\ell) \le 2$ for each $\ell \in [L]$ as well as our choice of b_{rew}. The above chain of inequalities thus verifies (44) in this case.

From here on, we assume that Case 1 does not hold. Since $\mathrm{t}^0_{2,0} = \mathrm{t}^0_{2,1}$, we have $\rho(0) - 1 = 0$. Therefore, either $\sum_{\ell=L/2}^{L}(\rho(\ell)-1) \le \frac{L}{2}\cdot 1/2$ or $\frac{1}{L}\sum_{\ell=1}^{L}[\rho(\ell)-\rho(\ell-1)]_+ \ge 1/(2L) = \sqrt{\varepsilon_{\mathsf{BE}}}/2$. Thus, it suffices to consider the (exhaustive) Cases 2 and 3 below:

Case 2: $\sum_{\ell=L/2}^{L}(\rho(\ell)-1) \le \frac{L}{2}\cdot 1/2$**.** We make the same choice of $\mathbf{b}$ as in Case 1. Then using (49), we have

$$\begin{aligned}
\mathbb{E}_{\pi\sim\mathcal{D}_0}\left[V_1^{M^{\mathbf{b}},\pi}(\mathrm{t}_1)\right] =& \alpha(1-2b_{\mathsf{rew}})\frac{1}{L}\sum_{\ell=1}^{L}(\bar{\eta}(\ell)+\bar{\gamma}(\ell)) + \frac{\alpha Z_0}{RL}\sum_{\ell=1}^{L}(\rho(\ell)-1)\cdot\ell\varepsilon_{\mathsf{BE}} + \frac{\alpha(1-Z_0)}{RL}\sum_{\ell=1}^{L}\ell\varepsilon_{\mathsf{BE}} \\
\le& \frac{\alpha Z_0}{RL}\sum_{\ell=1}^{L}(\rho(\ell)-1)\cdot\ell\varepsilon_{\mathsf{BE}} + \alpha R^{-1}(1-Z_0)\cdot(L+1)\varepsilon_{\mathsf{BE}}/2 \\
\le& \frac{\alpha(L+1)\varepsilon_{\mathsf{BE}}}{2R}\cdot(1-Z_0) + \frac{\alpha Z_0}{R}\cdot\left(\frac{(L+1)\varepsilon_{\mathsf{BE}}}{2} - \frac{L}{4}\cdot\frac{L\varepsilon_{\mathsf{BE}}/2}{L}\right) \\
=& \frac{\alpha(L+1)\varepsilon_{\mathsf{BE}}}{2R} - \frac{\alpha Z_0\cdot L\varepsilon_{\mathsf{BE}}}{8R} \\
\le& V_1^{M^{\mathbf{b}},\pi^\star}(\mathrm{t}_1) - \alpha R^{-1}L\varepsilon_{\mathsf{BE}}/16 = V_1^{M^{\mathbf{b}},\pi^\star}(\mathrm{t}_1) - \alpha R^{-1}\sqrt{\varepsilon_{\mathsf{BE}}}/16,
\end{aligned}$$

where the second inequality uses our assumption that $\sum_{\ell=L/2}^{L}(\rho(\ell)-1) \le L/4$, and the final inequality uses the fact that $Z_0 \ge 1/2$.

Case 3: $\sum_{\ell=1}^{L}[\rho(\ell)-\rho(\ell-1)]_+ \ge \sqrt{\varepsilon_{\mathsf{BE}}}/2$**.** For each $\ell \in [L]$ and $b \in \{0,1\}$, define $\rho_b(\ell) := \mathbb{E}_{\pi\sim\mathcal{D}}\left[\pi(b \mid \mathrm{t}_{2,b}^{\ell\varepsilon_{\mathsf{BE}}})\right]$, so that $\rho(\ell) = \rho_0(\ell) + \rho_1(\ell)$. We may choose some $e^\star \in \{0,1\}$ so that $\sum_{\ell=1}^{L}[\rho_{e^\star}(\ell) - \rho_{e^\star}(\ell-1)]_+ \ge \sqrt{\varepsilon_{\mathsf{BE}}}/4$. Choose $b_{\mathsf{rew}} = e^\star$ and define the values $b_{\ell,e}$ as follows:

$$b_{\ell,e} := \begin{cases} 1 & : \rho_{e^\star}(\ell) - \rho_{e^\star}(\ell-1) \ge 0,\ e = e^\star \\ 0 & : \text{otherwise.} \end{cases} \tag{50}$$

Write $\mathbf{b} = (b_{\mathsf{init}}, b_{\mathsf{rew}}, (b_{\ell,e})_{\ell,e})$. Using (45), for each $\ell \in [L]$, we have

$$\begin{aligned}
& V_2^{M^{\mathbf{b}},\pi}(\mathrm{t}_{2,0}^{(\ell-b_{\ell,0})\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathrm{t}_{2,1}^{(\ell-b_{\ell,1})\varepsilon_{\mathsf{BE}}}) \\
=& 2\alpha\pi(1-e^\star \mid \mathrm{t}_{2,1-e^\star}^{\ell\varepsilon_{\mathsf{BE}}})\cdot((2e^\star-1)(1-2b_{\mathsf{rew}}) + \ell\varepsilon_{\mathsf{BE}}/R) - \alpha((2e^\star-1)(1-2b_{\mathsf{rew}}) + \ell\varepsilon_{\mathsf{BE}}/R) \\
& + 2\alpha\pi(e^\star \mid \mathrm{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}})\cdot((1-2e^\star)(1-2b_{\mathsf{rew}}) + (\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}/R) - \alpha((1-2e^\star)(1-2b_{\mathsf{rew}}) + (\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}/R) \\
=& 2\alpha\pi(1-e^\star \mid \mathrm{t}_{2,1-e^\star}^{\ell\varepsilon_{\mathsf{BE}}})\cdot((2e^\star-1)(1-2b_{\mathsf{rew}}) + \ell\varepsilon_{\mathsf{BE}}/R) - \alpha(2\ell - b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}/R \\
& - 2\alpha\pi(e^\star \mid \mathrm{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}})\cdot((2e^\star-1)(1-2b_{\mathsf{rew}}) - (\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}/R) \\
\le& 2\alpha(2e^\star-1)(1-2b_{\mathsf{rew}})\left(\pi(1-e^\star \mid \mathrm{t}_{2,1-e^\star}^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(e^\star \mid \mathrm{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}})\right) + 2\alpha\ell\cdot\varepsilon_{\mathsf{BE}}/R.
\end{aligned}$$

Combining the above display with (46), we obtain that

$$\begin{aligned}
& V_2^{M^{\mathbf{b}},\pi}(\mathrm{t}_{2,0}^{(\ell-b_{\ell,0})\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathrm{t}_{2,1}^{(\ell-b_{\ell,1})\varepsilon_{\mathsf{BE}}}) - \left(V_2^{M^{\mathbf{b}},\pi}(\mathrm{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathrm{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}})\right) \\
\le& \left(2\alpha(2e^\star-1)(1-2b_{\mathsf{rew}})\left(\pi(1-e^\star \mid \mathrm{t}_{2,1-e^\star}^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(e^\star \mid \mathrm{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}})\right) + 2\alpha R^{-1}\ell\cdot\varepsilon_{\mathsf{BE}}\right) \\
& - \left(2\alpha(1-2b_{\mathsf{rew}})\cdot\left(\pi(0 \mid \mathrm{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(1 \mid \mathrm{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}})\right) + 2\alpha R^{-1}\ell\varepsilon_{\mathsf{BE}}\left(\pi(0 \mid \mathrm{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}) + \pi(1 \mid \mathrm{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}}) - 1\right)\right) \\
\le& 2\alpha(1-2e^\star)(1-2b_{\mathsf{rew}})\cdot\pi(e^\star \mid \mathrm{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}}) - 2\alpha(1-2b_{\mathsf{rew}})(1-2e^\star)\cdot\pi(e^\star \mid \mathrm{t}_{2,e^\star}^{\ell\varepsilon_{\mathsf{BE}}}) + 4\alpha R^{-1}\ell\varepsilon_{\mathsf{BE}} \\
=& -2\alpha\cdot\left(\pi(e^\star \mid \mathrm{t}_{2,e^\star}^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(e^\star \mid \mathrm{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}})\right) + 4\alpha R^{-1}\ell\varepsilon_{\mathsf{BE}},
\end{aligned} \tag{51}$$

where the final equality uses our choice of $b_{\mathsf{rew}} = e^\star$. Define $b'_{\ell,e} = 0$ for all $\ell \in [L], e \in \{0,1\}$, and set $\mathbf{b}' := (b_{\mathsf{rew}}, b_{\mathsf{init}}, (b_{\ell,e})'_{\ell,e})$. Then, using (48) as well as the fact that $V_2^{M^{\mathbf{b}},\pi}(\mathfrak{s}_2^\zeta) = V_2^{M^{\mathbf{b}'},\pi}(\mathfrak{s}_2^\zeta)$ for all π and ζ,

$$
\begin{aligned}
&\frac{1}{Z_0}\mathbb{E}_{\pi\sim\mathcal{D}_0}\left[V_1^{M^{\mathbf{b}},\pi}(\mathfrak{t}_1) - V_1^{M^{\mathbf{b}'},\pi}(\mathfrak{t}_1)\right] \\
=&\frac{1}{2LZ_0}\sum_{\ell=1}^{L}\mathbb{E}_{\pi\sim\mathcal{D}_0}\left[\pi(1-b_{\mathsf{init}}\mid \mathfrak{t}_1)\cdot\left(V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,0}^{(\ell-b_{\ell,0})\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,1}^{(\ell-b_{\ell,1})\varepsilon_{\mathsf{BE}}}) - \left(V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}})\right)\right)\right] \\
=&\frac{1}{2L}\sum_{\ell=1}^{L}\mathbb{E}_{\pi\sim\mathcal{D}}\left[V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,0}^{(\ell-b_{\ell,0})\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,1}^{(\ell-b_{\ell,1})\varepsilon_{\mathsf{BE}}}) - \left(V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,0}^{\ell\varepsilon_{\mathsf{BE}}}) + V_2^{M^{\mathbf{b}},\pi}(\mathfrak{t}_{2,1}^{\ell\varepsilon_{\mathsf{BE}}})\right)\right] \\
\le&2\alpha R^{-1}(L+1)\varepsilon_{\mathsf{BE}} - \alpha\sum_{\ell=1}^{L}\mathbb{E}_{\pi\sim\mathcal{D}}\left[\pi(e^\star\mid \mathfrak{t}_{2,e^\star}^{\ell\varepsilon_{\mathsf{BE}}}) - \pi(e^\star\mid \mathfrak{t}_{2,e^\star}^{(\ell-b_{\ell,e^\star})\varepsilon_{\mathsf{BE}}})\right] \\
=&2\alpha R^{-1}(L+1)\varepsilon_{\mathsf{BE}} - \alpha\sum_{\ell=1}^{L}\left(\rho_{e^\star}(\ell) - \rho_{e^\star}(\ell-b_{\ell,e^\star})\right) \\
=&2\alpha R^{-1}(L+1)\varepsilon_{\mathsf{BE}} - \alpha\sum_{\ell=1}^{L}[\rho_{e^\star}(\ell) - \rho_{e^\star}(\ell-b_{\ell,e^\star})]_+ \\
\le&2\alpha R^{-1}(L+1)\varepsilon_{\mathsf{BE}} - \alpha\sqrt{\varepsilon_{\mathsf{BE}}}/2 \le -\alpha\sqrt{\varepsilon_{\mathsf{BE}}}/4,
\end{aligned}
\tag{52}
$$

where the first inequality uses (51), the final equality uses the definition of $b_{\ell,e^\star}$ in (50), and the final inequality uses the fact that $R = 16$. Also note that, from (49),

$$
\begin{aligned}
\mathbb{E}_{\pi\sim\mathcal{D}_0}\left[V_1^{M^{\mathbf{b}'},\pi}(\mathfrak{t}_1)\right] =&\alpha(1-2b_{\mathsf{rew}})\frac{1}{L}\sum_{\ell=1}^{L}(\bar{\eta}(\ell)+\bar{\gamma}(\ell)) + \frac{\alpha}{RL}\sum_{\ell=1}^{L}Z_0\cdot\ell\varepsilon_{\mathsf{BE}}\cdot(\rho(\ell)-1) + \frac{\alpha}{RL}\sum_{\ell=1}^{L}(1-Z_0)\cdot\ell\varepsilon_{\mathsf{BE}} \\
\le&\alpha\cdot\frac{\sqrt{\varepsilon_{\mathsf{BE}}}}{10} + \frac{\alpha(L+1)\varepsilon_{\mathsf{BE}}}{2R} = V_1^{M^{\mathbf{b}},\pi^\star}(\mathfrak{t}_1) + \alpha\cdot\sqrt{\varepsilon_{\mathsf{BE}}}/10.
\end{aligned}
\tag{53}
$$

Combining Eqs. (52) and (53) and using the fact that $Z_0 \ge 1/2$ gives that

$$
\mathbb{E}_{\pi\sim\mathcal{D}_0}\left[V_1^{M^{\mathbf{b}},\pi}(\mathfrak{t}_1)\right] - V_1^{M^{\mathbf{b}},\pi^\star}(\mathfrak{t}_1) \le \alpha\sqrt{\varepsilon_{\mathsf{BE}}}/10 - \alpha\sqrt{\varepsilon_{\mathsf{BE}}}/8 = -\alpha\sqrt{\varepsilon_{\mathsf{BE}}}/40,
$$

as desired. □

E Useful lemmas

E.1 Concentration

Lemma E.1 (Concentration for self-normalized process; e.g., Theorem D.3 of Jin et al. (2020a)). *Fix $n \in \mathbb{N}$ and let $\varepsilon_1, \dots, \varepsilon_n$ be random variables which are adapted to a filtration $(\mathcal{F}_i)_{0\le i\le n}$. Suppose that for each $i \in [n]$, $\mathbb{E}[\varepsilon_i|\mathcal{F}_{i-1}] = 0$ and $\mathbb{E}[e^{\lambda\varepsilon_i}|\mathcal{F}_{i-1}] \le e^{\lambda^2\sigma^2/2}$. Suppose that $\phi_1, \dots, \phi_n$ is a sequence which is predictable with respect to $(\mathcal{F}_i)_{0\le i\le n}$, i.e., ϕ_i is measurable with respect to $\mathcal{F}_{i-1}$ for all $i \in [n]$. Suppose that $\Gamma_0 \in \mathbb{R}^{d\times d}$ is positive definite, and let $\Gamma_i = \Gamma_0 + \sum_{j=1}^{i}\phi_j\phi_j^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$,*

$$
\left\|\sum_{i=1}^{n}\phi_i\varepsilon_i\right\|_{\Gamma_i^{-1}}^2 \le 2\sigma^2\log\left(\frac{\det(\Gamma_t)^{1/2}\det(\Gamma_0)^{-1/2}}{\delta}\right).
$$

E.2 Projection bound

Lemma E.2. *Consider any sequence of vectors $\phi_1, \dots, \phi_n \in \mathbb{R}^d$ and a sequence of real numbers $b_1, \dots, b_n \in \mathbb{R}$, so that, for some $\epsilon > 0$, $|b_i| \le \epsilon$ for all $i \in [n]$. Then for any $\lambda \ge 0$,*

$$\left\| \sum_{i=1}^n b_i \phi_i \right\|^2_{\left(\lambda I + \sum_{i=1}^n \phi_i \phi_i^\top\right)^{-1}} \le n\epsilon^2.$$

E.3 Performance difference lemma

Lemma E.3 (Performance difference lemma; Kakade & Langford (2002))**.** *For any MDP M, policies $\pi, \pi' \in \Pi$, it holds that*

$$\mathbb{E}^{M,\pi}\left[\sum_{h=1}^H r_h(x_h, a_h)\right] - \mathbb{E}^{M,\pi'}\left[\sum_{h=1}^H r_h(x_h, a_h)\right] = \sum_{h=1}^H \mathbb{E}^{M,\pi'}\left[V_h^{M,\pi}(x_h) - Q_h^{M,\pi}(x_h, a_h)\right].$$

F Bellman restricted closedness

In this section, we show that linear Bellman completeness does not, in general, imply that Bellman restricted closedness holds, even when the policy class is restricted to be softmax policies. We first make the requisite definitions.

Definition F.1 (Softmax policy class)**.** Given feature mappings $(\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d)_{h \in [H]}$, the associated *softmax policy class* Π^{soft} consists of the set of all policies $\pi = (\pi_1, \dots, \pi_H)$, for which there is some $\eta > 0$ and $w_1, \dots, w_H \in \mathbb{R}^d$ so that for all $h \in [H]$ and $x \in \mathcal{X}$,

$$\pi_h(a|x) = \frac{\exp(\eta \cdot \langle \phi_h(x, a), w_h \rangle)}{\sum_{a' \in \mathcal{A}} \exp(\eta \cdot \langle \phi_h(x, a'), w_h \rangle)}.$$

If π_h satisfies the above display, we write $\pi_h = \pi_h^{\mathsf{soft}}[w_h, \eta]$.

Definition F.2 (Bellman restricted closedness)**.** An MDP M is said to satisfy *Bellman restricted closendess* with respect to d-dimensional feature mappings $(\phi_h)_{h \in [H]}$ for a policy class Π' if for each $\pi \in \Pi'$, there are mappings $\mathcal{T}_h^\pi : \mathbb{R}^d \to \mathbb{R}^d$ so that the following holds for each $h \in [H]$:

$$\sup_{w \in \mathbb{R}^d} \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \left| \langle \phi_h(x, a), \mathcal{T}_h^\pi w \rangle - \mathbb{E}_{x' \sim P_h(x,a)} \left[r_h(x, a) + \langle \phi_{h+1}(x', \pi_{h+1}(x')), w \rangle \right] \right| = 0.$$

Lemma F.1. *There is an MDP with $H = 2$, $A = 3$ together with feature mappings in $d = 1$ dimension which satisfies linear Bellman completeness (i.e., Assumption 2.1 with $\varepsilon_{\mathsf{BE}} = 0$) but not Bellman restricted closedness (Definition F.2) with respect to the softmax policy class Π^{soft}.*

Proof. Consider the MDP with $H = 2, \mathcal{A} = \{1, 2, 3\}, d = 1$, $\mathcal{X} = \{\mathfrak{s}_1, \mathfrak{s}_2\}$, and

$$\begin{aligned} \phi_1(x, a) &= 1 \quad \forall x \in \mathcal{X}, a \in \mathcal{A} \\ \phi_2(\mathfrak{s}_1, 1) = 1,\ \phi_2(\mathfrak{s}_1, 2) &= 0,\ \phi_2(\mathfrak{s}_1, 3) = -1 \\ \phi_2(\mathfrak{s}_2, 1) = 1,\ \phi_2(\mathfrak{s}_2, 2) &= 1, \phi_2(\mathfrak{s}_2, 3) = -1. \end{aligned}$$

All rewards are 0. The transitions are as follows: for any $a \in \mathcal{A}$, $x \in \mathcal{X}$, (x, a) transitions to x at step 1. By defining $\mathcal{T}_1 w = |w|$, we may ensure that linear Bellman completeness holds with respect to the above feature mappings: indeed, for each $x \in \mathcal{X}$, $\max_{a \in \mathcal{A}} w \cdot \phi_2(x, a) = |w| = \langle \phi_1(x, a'), \mathcal{T}_1 w \rangle$ for all $a' \in \mathcal{X}$.

Now let $\pi_2 := \pi_2^{\mathsf{soft}}[1, 1]$ (see Definition F.1) and $w = 1$. Then

$$\begin{aligned} \langle \phi_2(\mathfrak{s}_1, \pi_2(\mathfrak{s}_1)), w \rangle &= \frac{e - e^{-1}}{e + 1 + e^{-1}} \\ \langle \phi_2(\mathfrak{s}_2, \pi_2(\mathfrak{s}_2)), w \rangle &= \frac{2e}{2e + e^{-1}}. \end{aligned}$$

Since $\frac{e-e^{-1}}{e+1+e^{-1}} \neq \frac{2e}{2e+e^{-1}}$, and $\phi_1(\mathfrak{s}_1, a) = \phi_1(\mathfrak{s}_2, a)$ for all a, Bellman restricted closedness cannot hold (even up to constant approximation error). □