Necessary and Sufficient Oracles: Toward a Computational Taxonomy for Reinforcement Learning

Dhruv Rohatgi DROHATGI@MIT.EDU
MIT

Dylan J. Foster DYLANFOSTER@MICROSOFT.COM
Microsoft Research

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Algorithms for reinforcement learning (RL) in large state spaces crucially rely on supervised learning subroutines to estimate objects such as value functions or transition probabilities. Since only the simplest supervised learning problems can be solved provably and efficiently, practical performance of an RL algorithm depends on which of these supervised learning "oracles" it assumes access to (and how they are implemented). But which oracles are better or worse? Is there a *minimal* oracle?

In this work, we clarify the impact of the choice of supervised learning oracle on the computational complexity of RL, as quantified by the oracle strength. First, for the task of reward-free exploration in Block MDPs in the standard episodic access model—a ubiquitous setting for RL with function approximation—we identify *two-context regression* as a minimal oracle, i.e. an oracle that is both necessary and sufficient (under a mild regularity assumption). Second, we identify *one-context regression* as a near-minimal oracle in the stronger *reset* access model, establishing a provable computational benefit of resets in the process. Third, we broaden our focus to *Low-Rank MDPs*, where we give cryptographic evidence that the analogous oracle from the Block MDP setting is insufficient.

Keywords: Reinforcement learning, computational complexity, oracle-efficiency

1. Introduction

An overarching paradigm in modern machine learning is to reduce a complex task of interest to a simpler supervised learning task. Instances of this paradigm range from language modeling (OpenAI, 2023) and image generation (Song and Ermon, 2019) to imitation learning (Bojarski et al., 2016). More broadly, the basic ansatz from supervised learning – that gradient descent finds good minimizers – underlies empirical progress in reinforcement learning (Mnih et al., 2015), artificial intelligence for games (Silver et al., 2018), and much more. State-of-the-art learning methods may not admit provable guarantees, but their empirical success is intuitively (to varying extents) justified by the basic ansatz.

From a theoretical perspective, this ansatz can be exploited via *oracle-efficient* algorithm design. Formally, an algorithm is oracle-efficient with respect to an oracle $\mathcal{O}$ if it is computationally efficient and provably correct when given query access to $\mathcal{O}$. Oracle-efficiency has been a cornerstone in the theory of online learning since the development of Follow-the-Perturbed-Leader (Kalai and Vempala, 2005), which solves online linear optimization by reduction to an offline linear optimization oracle over the decision set. Recently, oracle-efficiency has become ubiquitous in theoretical reinforcement learning (Dann et al., 2018; Du et al., 2019; Misra et al., 2020; Foster et al., 2021; Mhammedi et al., 2023a; Hussing et al., 2024), where end-to-end computational efficiency is often out-of-reach for settings with large observation spaces (Kane et al., 2022; Golowich et al., 2024a), yet heuristics based on deep learning (e.g., for estimation of value functions or transition dynamics) could plausibly work well on natural data. In reinforcement learning, as with online learning (Hazan and Koren, 2016; Dudík et al., 2020) and decision making (Agarwal et al., 2014; Foster and Rakhlin, 2020), the lens of oracle-efficiency has spurred numerous algorithmic improvements.

Yet, in online learning, the correct *choice of oracle* was fairly clear: to solve an *online* optimization problem, assume access to an oracle that solves the corresponding *offline* optimization problem (Kalai and Vempala, 2005; Hazan and Koren, 2016). As observed by Kalai and Vempala (2005), this assumption is essentially without loss of generality, since an online optimization algorithm must solve offline optimization as a special case. In contrast, for reinforcement learning—a more complex and structured setting, due to the interaction between the agent and environment—there is no such consensus about the "right" computational oracles, even for models that are by now well-established. Instead, oracle-efficiency has largely been used as a black-and-white prognostic, just separating "reasonable" algorithms from those requiring exhaustive enumeration (Dann et al., 2018).

In this work, we take a finer-grained view of oracle-efficiency—for example, are supervised learning oracles that perform regression onto value functions sufficient, or must we estimate more complex objects such as transition dynamics? Since access to an oracle is fundamentally an assumption, we are interested in the following question: *What are the weakest computational oracles that suffice for oracle-efficient reinforcement learning?*

In other words, while prior works have largely focused on the impact of differing structural assumptions on statistical complexity, we focus on the impact on computational complexity, as measured by the oracle strength. To begin this investigation, we study the task of exploration in *Block Markov Decision Processes* (Du et al., 2019)—one of the most well-studied families of Markov Decision Processes (MDPs) with rich observation spaces. We identify the first *minimal oracle* (Golowich et al., 2024a) for this task, under the standard episodic access model. We then consider the reset access model, and show that a strictly weaker oracle suffices. Moving beyond Block MDPs, we give cryptographic evidence of a qualitative computational separation between Block MDPs and the more general setting of *Low-Rank* MDPs.

1.1. Background: Block MDPs and Computational Oracles

A finite-horizon Markov Decision Process (MDP) is defined by a set of *states*, a set of *actions*, and an unknown transition function, which describes how the environment changes state as a result of the agent's actions. The agent learns by repeated interaction with the environment—the two most common interaction frameworks are episodic (Kearns and Singh, 2002) and resets (Weisz et al., 2021); we study both.[1] We focus on reward-free RL (Du et al., 2019; Jin et al., 2020b), where the goal is *exploration*: finding a set of *policies* (i.e. mappings from states to actions) that cover the entire state space as well as possible.[2] When the state space is extremely large, structural assumptions are needed to avoid statistical intractability: For most of this paper, we focus on the *Block MDP* (Du et al., 2019; Misra et al., 2020), a canonical setting for RL with function approximation in which the rich observed dynamics are governed by a small (unobserved) *latent state space*; just like in PAC learning, a concept class is required to model the mapping from observed states to latent states.

Definition 1.1 (Informal; see Section 2) *Let $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{A}$ be sets and let Φ be a* concept class *of functions $\phi : \mathcal{X} \to \mathcal{S}$.*[3] *An MDP with state space $\mathcal{X}$ and action space $\mathcal{A}$ is a Φ-**decodable Block***

1. We specify "episodic RL" to distinguish from "RL with resets"; we say "RL" when the distinction is unimportant.
2. See Section 6 for discussion about the technical reasons for this focus.
3. Φ is often referred to as a *decoder class* in prior work; we use *concept class* in analogy with PAC learning.

MDP *with* latent state space $\mathcal{S}$ *if there is a function* $\phi^\star \in \Phi$ *such that for any two states* $x, x' \in \mathcal{X}$ *and action* $a \in \mathcal{A}$*, the transition probability from* x *to* x' *when the agent plays action* a *depends only on* $\phi^\star(x)$*,* $\phi^\star(x')$*, and* a*; we refer to* $\phi^\star(x)$ *as the* latent state.

Henceforth, we refer to $\mathcal{X}$ as the *observed state space* or *observation space* to distinguish from the latent state space. The concept class Φ is known to the learner, but the true decoder $\phi^\star \in \Phi$ is not. The statistical complexity of reward-free exploration scales polynomially in $|\mathcal{S}|$, $|\mathcal{A}|$, and $\log|\Phi|$ (Jiang et al., 2017)—and crucially has no dependence on $|\mathcal{X}|$, which should be thought of as exponentially large or even infinite. The *computational* complexity is much more subtle. Initial algorithms required enumeration over Φ (Jiang et al., 2017); subsequent investigation identified oracle-efficient algorithms with respect to several different optimization oracles (Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b,a), but no basis for comparison between these oracles has been proposed. The first works to raise the question of which oracles are *necessary* were Golowich et al. (2024b,a), who studied the *one-context (realizable) regression* problem:

Definition 1.2 (informal; see Definition 2.2) *Fix a concept class* Φ*. Let* $(x^{(i)}, y^{(i)})_{i=1}^n$ *be i.i.d., with* $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi^\star(x^{(i)}))$ *for some unknown* $f : \mathcal{S} \to [0,1]$ *and* $\phi^\star \in \Phi$*. The goal of* ***one-context regression*** *is to compute a predictor* $\mathcal{R} : \mathcal{X} \to [0,1]$ *that approximates* $x \mapsto f(\phi^\star(x))$.

Intuitively, one-context regression can be thought of as a regression with a *well-specified model* (Tsybakov, 2009; Wainwright, 2019), as the true target function depends only on the latent state $\phi^\star(x)$ (it can also be viewed as a generalization of PAC learning with random classification noise—see Remark F.6). This oracle is well-suited for estimating objects such as value functions—which depend only on the latent state in the Block MDP—and it has been implicitly used as a subroutine in many algorithms (Foster and Rakhlin, 2020; Zhang et al., 2022; Mhammedi et al., 2023b,a).[4]

For any concept class Φ, one-context regression is *necessary* for episodic RL, i.e. there is a Cook reduction from regression to episodic RL (Golowich et al., 2024b), so as an oracle assumption, it is without loss of generality.[5] Unfortunately, one-context regression is also *insufficient*: under a standard cryptographic assumption, there exists a concept class Φ for which there is no Cook reduction from episodic RL to regression (Golowich et al., 2024a). Thus, one-context regression is not a minimal oracle for episodic RL. This motivates us to consider the problem of *two-context regression*.

Definition 1.3 (informal; see Definition 2.4) *Fix a concept class* Φ*. Let* $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$ *be i.i.d., with* $\mathbb{E}[y^{(i)} \mid x_1^{(i)}, x_2^{(i)}] = f(\phi^\star(x_1^{(i)}), \phi^\star(x_2^{(i)}))$ *for some unknown* $f : \mathcal{S} \times \mathcal{S} \to [0,1]$ *and* $\phi^\star \in \Phi$*. The goal of* ***two-context regression*** *is to compute a predictor* $\mathcal{R} : \mathcal{X} \times \mathcal{X} \to [0,1]$ *that approximates* $(x_1, x_2) \mapsto f(\phi^\star(x_1), \phi^\star(x_2))$.

Several RL algorithms use variants of this oracle (Misra et al., 2020; Mhammedi et al., 2023b)—roughly, to estimate (inverse) *transition dynamics*—and it has been suggested that these variants may be essentially minimal (Golowich et al., 2024a), but no evidence for this belief was known prior to this work. See Appendix A for a detailed discussion of prior work.

4. Information-theoretically, MSE $\lesssim \frac{|\mathcal{S}|+\log(|\Phi|\delta^{-1})}{n}$ is always possible, but not necessarily computationally efficiently.
5. We remark that this reduction assumes that the action space is small, as is common in the multi-turn theoretical RL literature (but it therefore does not apply to decision-making problems with large, structured action spaces).

Remark 1.4 (Optimization vs. learning; proper vs. improper) *Many prior works in oracle-efficient RL assume access to* optimization *oracles rather than statistical learning oracles. Informally, the former oracles require solving regression problems analogous to Definitions 1.2 and 1.3 for* arbitrary datasets *as opposed to i.i.d. and realizable datasets. This is primarily a conceptual distinction rather than technical, since in many cases a learning oracle can easily be substituted in (Misra et al., 2020; Mhammedi et al., 2023b). However, it is important from a complexity-theoretic perspective, since statistical learning can often be substantially easier (Blum et al., 1998).*

A more technically salient distinction is that our definitions above allow for improper learning*, whereas almost all prior works in oracle-efficient RL for Block MDPs require* proper *learning oracles*[6]*—an exception is the work of Misra et al. (2020), which our algorithmic results directly build on. There are many concept classes for which proper learning is* NP*-hard, but it is considered unlikely for improper learning to be* NP*-hard (Applebaum et al., 2008). Since the goal of RL is to output policies, which are fundamentally improper, it seems unlikely that a proper supervised learning task could be reduced to RL (in the manner of results such as Theorem 3.2).*

1.2. Contributions

We clarify the computational complexity of RL via the lens of oracle-efficiency, identifying (1) the first minimal oracle for episodic RL in Block MDPs, (2) provable benefits of *reset access*, and (3) computational challenges of the more general *Low-Rank* MDPs. See Section 6 for open questions.

A minimal oracle for episodic RL in Block MDPs (Section 3). We show that for *every* concept class Φ, under a mild regularity condition, two-context regression is a minimal oracle—both *sufficient* and *necessary*—for reward-free episodic RL in Φ-decodable Block MDPs. To show sufficiency, we generalize and simplify the algorithm HOMER of Misra et al. (2020), eliminating their reachability assumption and implementing their oracles with two-context regression. To show necessity, we give a novel reduction *from* two-context regression *to* reward-free RL, which simulates interaction with an appropriate MDP and "stitches together" the exploratory policies into a predictor.

A provable computational benefit for reset access in Block MDPs (Section 4). Under the same regularity condition on Φ as before, we show that *one-context regression* is a sufficient (and nearly necessary) oracle for reward-free RL in Φ-decodable Block MDPs, given the additional ability to *reset* to previously observed states (Li et al., 2021; Yin et al., 2022; Mhammedi et al., 2024). Combined with the recent work of Golowich et al. (2024a), our result gives strong evidence that reset access has computational benefits over episodic access.

Our algorithm uses a variant of the *inverse kinematics* objective (Misra et al., 2020; Mhammedi et al., 2023b), but exploits reset access to simplify the computational oracle. Previously, one-context regression was only known to be sufficient for Block MDPs with horizon 1 or with deterministic dynamics (Golowich et al., 2024a); the closest prior work is RVFS (Mhammedi et al., 2024), which solves RL with resets in general Block MDPs but requires an *agnostic*/cost-sensitive regression oracle.

A computational separation between Block MDPs and Low-Rank MDPs (Section 5). There has been recent progress on oracle-efficient algorithms for *Low-Rank* MDPs (Modi et al., 2024; Zhang et al., 2022; Mhammedi et al., 2023a), of which Block MDPs are a special case. However, these algorithms seemingly require a much more complex oracle. Do analogues of one- or two-context

6. For example, the proper learning analogue of Definition 1.2 requires computing some predictor $\mathcal{R} : \mathcal{X} \to [0, 1]$ with an explicit decomposition $\mathcal{R} = \mathcal{R}' \circ \phi$ for some $\phi \in \Phi$ and $\mathcal{R}' : \mathcal{S} \to [0, 1]$.

regression suffice for exploration in Low-Rank MDPs? We show that the analogue of one-context regression is cryptographically *insufficient* for exploration in Low-Rank MDPs under reset access, thereby separating Low-Rank MDPs from Block MDPs. Conceptually, this separation arises from the same source as cryptographic hardness of agnostic halfspace learning (Tiegel, 2023), and points to the lack of *weight function realizability* in Low-Rank MDPs as a potential computational barrier.

2. Preliminaries

To begin, we formally introduce Block MDPs, the episodic and reset access models for RL, and the computational problems: reward-free RL and one- and two-context regression. As basic notation, $[k]$ denotes the set of integers $\{1, \dots, k\}$, and $\Delta(\mathcal{Z})$ denotes the family of distributions over set $\mathcal{Z}$.

2.1. Block MDPs and Episodic RL

For a set $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$, i.e. a set of functions $\phi : \mathcal{X} \to \mathcal{S}$, a (reward-free) Φ-decodable *Block MDP* (Du et al., 2019) is a tuple $M = (H, \mathcal{S}, \mathcal{X}, \mathcal{A}, (\widetilde{\mathbb{P}}_h)_{h\in[H]}, (\widetilde{\mathbb{O}}_h)_{h\in[H]}, \phi^\star)$ where $H \in \mathbb{N}$ is the *horizon*, $\mathcal{S}$ is the *latent state space*, $\mathcal{X}$ is the *observation space*, $\mathcal{A}$ is the *action set*, $\widetilde{\mathbb{P}}_1 \in \Delta(\mathcal{S})$ is the *latent initial distribution*, $\widetilde{\mathbb{P}}_h : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the *latent transition distribution* into step h (for any $h \in \{2, \dots, H\}$), $\widetilde{\mathbb{O}}_h : \mathcal{S} \to \Delta(\mathcal{X})$ is the *observation distribution* at step h (for any $h \in [H]$), and $\phi^\star \in \Phi$ is the *decoding function*. It is required that $\phi^\star(x_h) = s_h$ with probability 1 over $x_h \sim \widetilde{\mathbb{O}}_h(\cdot \mid s_h)$, for all $h \in [H]$ and $s_h \in \mathcal{S}$ (so in particular, $\widetilde{\mathbb{O}}_h(\cdot \mid s), \widetilde{\mathbb{O}}_h(\cdot \mid s')$ have disjoint supports for all $s \neq s'$). For any $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$, we write $\mathbb{P}_h(x' \mid x, a)$ to denote $\widetilde{\mathbb{P}}_h(\phi^\star(x') \mid \phi^\star(x), a)\widetilde{\mathbb{O}}_h(x' \mid \phi^\star(x'))$. We similarly define $\mathbb{P}_1(x) = \widetilde{\mathbb{P}}_1(\phi^\star(x))\widetilde{\mathbb{O}}_1(x \mid \phi^\star(x))$. Observe that $(H, \mathcal{X}, \mathcal{A}, (\mathbb{P}_h)_h)$ is a (reward-free) MDP, with the potentially large state space $\mathcal{X}$.

Access model I: Episodic online RL. Fix a Block MDP M as specified above. We say that an algorithm `Alg` has (episodic) online access to M to mean that `Alg` is executed in the following model. First, `Alg` is given H and $\mathcal{A}$ as input. At any time, `Alg` can request a new *episode*. The model then draws $s_1 \sim \widetilde{\mathbb{P}}_1$ and $x_1 \sim \widetilde{\mathbb{O}}_1(\cdot \mid s_1)$, and sends x_1 to `Alg`. The timestep of the episode is set to $h = 1$. So long as $h \leq H$, the algorithm `Alg` can at any time play an action $a_h \in \mathcal{A}$. If $h < H$, then the model draws $s_{h+1} \sim \widetilde{\mathbb{P}}_{h+1}(\cdot \mid s_h, a_h)$, and $x_{h+1} \sim \widetilde{\mathbb{O}}_{h+1}(\cdot \mid s_{h+1})$, and sends x_{h+1} to `Alg` and increments h. Otherwise, the episode concludes. Note that `Alg` never observes the latent states $s_{1:H}$.

Access model II: (Episodic) online RL with resets. We say that an algorithm `Alg` has reset access to M to mean that `Alg` is given access to the following sampling oracles (in addition to H and $\mathcal{A}$, as before). The first sampling oracle draws $s_1 \sim \widetilde{\mathbb{P}}_1$ and $x_1 \sim \widetilde{\mathbb{O}}_1(\cdot \mid s_1)$, and outputs x_1. The second sampling oracle takes as input a step $h \in [H-1]$, any previously-seen observation $x_h \in \mathcal{X}$, and an action $a_h \in \mathcal{A}$; then, the oracle samples $s_{h+1} \sim \widetilde{\mathbb{P}}_{h+1}(\cdot \mid \phi^\star(x_h), a_h)$ and $x_{h+1} \sim \widetilde{\mathbb{O}}_{h+1}(\cdot \mid s_{h+1})$ and outputs x_{h+1}. Informally, this oracle allows the algorithm to not only sample independent episodes from M, but also to reset to any previously-seen observation.

Policies and visitations. A (randomized) *policy* $\pi = (\pi_h)_{h=1}^H$ is a collection of maps $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$. We write Π to denote the set of all policies. For $k \in [H]$, we write $d_k^{M,\pi} \in \Delta(\mathcal{S})$ to denote the distribution of s_k in an episode of interaction with M where $a_h \sim \pi_h(x_h)$ for each step h.

2.2. Computational Problems

Fix sets $\mathcal{S}, \mathcal{X}$ and a concept class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$. For the purposes of oracle reductions, an algorithm/oracle for a statistical learning problem is parametrized by its statistical efficiency (i.e. how many samples it needs in order to achieve certain accuracy) and, in the case of reward-free RL, the number of policies in the output. Throughout, we assume that the outputs of an algorithm/oracle (either policies or prediction functions) are succinctly described by circuits (and efficiently evaluatable). For an algorithm `Alg` with access to an oracle $\mathcal{O}$, the *oracle time complexity* of `Alg` is the time complexity in the computational model where each query to $\mathcal{O}$ takes linear time in the query length.

Definition 2.1 (Reward-free RL (Du et al., 2019; Jin et al., 2020b)) *Fix $N_{\mathsf{RL}}, K_{\mathsf{RL}} : (0, 1/2)^2 \times \mathbb{N}^2 \to \mathbb{N}$. An interactive algorithm* `Alg` *is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient* reward-free (episodic/reset) RL algorithm *for Φ if the following holds. Fix $\epsilon, \delta \in (0, 1/2)$, $H \in \mathbb{N}$, and a set $\mathcal{A}$. Given (episodic/reset) access to a Φ-decodable Block MDP M with horizon H and action set $\mathcal{A}$,* `Alg`$(\epsilon, \delta, H, \mathcal{A})$ *uses at most $N_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|)$ (episodes/queries), and outputs a set of policies Ψ with (1) $|\Psi| \leq K_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|)$, and (2) with probability at least $1 - \delta$, for all $s \in \mathcal{S}$ and $h \in [H]$,*

$$\max_{\pi \in \Psi} d_h^{M,\pi}(s) \geq \max_{\pi \in \Pi} d_h^{M,\pi}(s) - \epsilon. \tag{1}$$

A set Ψ that satisfies Eq. (1) is called a $(1, \epsilon)$-*policy cover*; it formalizes the notion of reaching all latent states with near-maximal probability. While the ultimate goal in many applications is *reward-directed RL*, it is straightforward to convert a reward-free RL algorithm into a reward-directed RL algorithm, and most existing oracle-efficient RL algorithms for Block MDPs (and more general classes) use some version of reward-free RL as a subroutine—see Appendix A for discussion and comparison to variants of Definition 2.1. We defer defining reward-free RL in more general settings to Section 5.

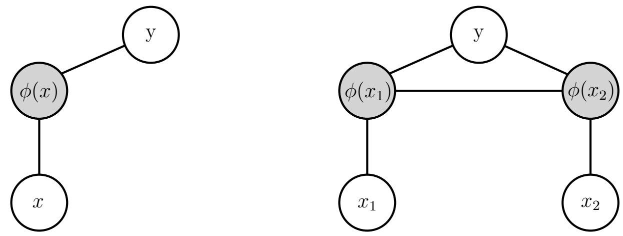
We now formally define the two notions of regression oracle we consider; versions of both (c.f. Remark 1.4) have been used extensively throughout the reinforcement learning literature.

Definition 2.2 (One-context regression) *Fix $N_{\mathsf{reg}} : (0, 1/2)^2 \to \mathbb{N}$. An algorithm* `Alg` *is an N_{reg}-efficient one-context regression algorithm for Φ if the following holds. Fix $\epsilon, \delta \in (0, 1/2)$, $n \in \mathbb{N}$, $\phi \in \Phi$, $\mathcal{D} \in \Delta(\mathcal{X})$, and $f : \mathcal{S} \to [0, 1]$. Let $(x^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $x^{(i)} \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi(x^{(i)}))$. If $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$, then with probability at least $1 - \delta$, the output of* `Alg`$((x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$ *is a circuit $\mathcal{R} : \mathcal{X} \to [0, 1]$ satisfying $\mathbb{E}_{x \sim \mathcal{D}}(\mathcal{R}(x) - f(\phi(x)))^2 \leq \epsilon$.*

See Figure 2(*a*) for the graphical model structure satisfied by each sample. One-context regression is a natural oracle for estimating value functions and Bellman backups (Ernst et al., 2005; Mhammedi et al., 2023a; Golowich et al., 2024a). In our definition, like that of Golowich et al. (2024a), the oracle is improper and only required to succeed on well-specified i.i.d. data.

Definition 2.3 *Let $\phi \in \Phi$. A distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{X})$ is ϕ-realizable if $(X_1, X_2) \sim \mathcal{D}$ satisfies $X_2 \perp X_1 \mid \phi(X_1)$ and $X_2 \perp X_1 \mid \phi(X_2)$.*

Definition 2.4 (Two-context regression) *Fix $N_{\mathsf{reg}} : (0, 1/2)^2 \to \mathbb{N}$. An algorithm* `Alg` *is an N_{reg}-efficient two-context regression algorithm for Φ if the following holds. Fix $\epsilon, \delta \in (0, 1/2)$, $n \in \mathbb{N}$, $\phi \in \Phi$, a ϕ-realizable distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{X})$, and $f : \mathcal{S} \times \mathcal{S} \to [0, 1]$. Let $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $(x_1^{(i)}, x_2^{(i)}) \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x_1^{(i)}, x_2^{(i)}] = f(\phi(x_1^{(i)}), \phi(x_2^{(i)}))$. If $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$, then with probability at least $1 - \delta$, the output of* `Alg` *on input $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$ is a circuit $\mathcal{R} : \mathcal{X} \times \mathcal{X} \to [0, 1]$ satisfying $\mathbb{E}_{(x_1, x_2) \sim \mathcal{D}}(\mathcal{R}(x_1, x_2) - f(\phi(x_1), \phi(x_2)))^2 \leq \epsilon$.*



(*a*) One-context regression (*b*) Two-context regression

Figure 1: Undirected graphical model representation for a single sample from *(a)* one-context, or *(b)* two-context regression. Note that the gray variables are unobserved.

See Figure 2(*b*) for the graphical model structure satisfied by each sample. Two-context regression is a natural oracle for estimating inverse kinematics—e.g., "given data from two policies, predict which policy the sample came from." Variants of this oracle are widely-used in RL (Misra et al., 2020; Lamb et al., 2023; Mhammedi et al., 2023b); see Appendix A for discussion.

2.3. Additional Assumptions

Our reductions *from* regression *to* RL require the following mild regularity condition on Φ; essentially, it asserts that there are two latent states $\{0, 1\}$ that are fully observable, irrespective of the decoding function $\phi^\star$. These extra states enable simulating "reward" states in the reductions.

Definition 2.5 (Regularity condition) *We say that a concept class Φ is* regular *if there are two special states $\{0, 1\} \in \mathcal{S} \cap \mathcal{X}$ that are fully observed, i.e. $\phi^{-1}(b) = \{b\}$ for all $\phi \in \Phi$ and $b \in \{0, 1\}$.*

3. A Minimal Oracle for Episodic RL in Block MDPs

In this section we show that two-context regression is a minimal oracle for reward-free episodic RL in Block MDPs. First we show that it is *sufficient* (Section 3.1), then that it is *necessary* (Section 3.2).

3.1. Sufficiency: Reducing Episodic RL to Two-Context Regression

Our first result gives a computational reduction from reward-free RL (Definition 2.1) in a Φ-decodable block MDP (with the episodic access model) to two-context regression for Φ (Definition 2.4). More precisely, we give a reward-free RL algorithm that requires access to a two-context regression oracle $\mathtt{Reg}$ (as well as episodic access to an MDP), and is oracle-efficient so long as $\mathtt{Reg}$ is sample-efficient.

Theorem 3.1 (Special case of Theorem C.1) *There is a constant $C_{3.1} > 0$ and an algorithm PCE (Algorithm 1 in Appendix C) so that the following holds. Let $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ be any concept class, let $N^\circ_{\mathsf{reg}}, C_{\mathsf{reg}} \in \mathbb{N}$, and let $\mathtt{Reg}$ be a N_{reg}-efficient two-context regression oracle for Φ with $N_{\mathsf{reg}}(\epsilon, \delta) := N^\circ_{\mathsf{reg}}/(\epsilon\delta)^{C_{\mathsf{reg}}}$. Then $\mathsf{PCE}(\mathtt{Reg}, N_{\mathsf{reg}}, |\mathcal{S}|, \cdot)$ is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free RL algorithm for Φ in the* episodic access model*, with $K_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \leq H^2|\mathcal{S}|^2$ and $N_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \leq N^\circ_{\mathsf{reg}} \cdot \left(\frac{H|\mathcal{A}||\mathcal{S}|}{\epsilon\delta}\right)^{C_{3.1}C_{\mathsf{reg}}}$. Moreover, the oracle time complexity of PCE is at most $N^\circ_{\mathsf{reg}} \cdot \left(\frac{H|\mathcal{A}||\mathcal{S}|}{\epsilon\delta}\right)^{C_{3.1}C_{\mathsf{reg}}}$.*

While Theorem 3.1 assumes a natural parametric scaling for N_{reg}, the full result (Theorem C.1) applies to any efficiency function. Note that $N^\circ_{\mathsf{reg}}, C_{\mathsf{reg}}$ will naturally be larger for more complex concept classes Φ, but there are no "hidden" dependencies on Φ. Informally, Theorem 3.1 shows that two-context regression is a *sufficient* oracle for reward-free episodic RL in block MDPs.

Proof overview. The main subroutine of PCE is EPCE, which strongly resembles the HOMER algorithm (Misra et al., 2020).[7] The basic idea of HOMER (and many other oracle-efficient RL algorithms (Du et al., 2019; Mhammedi et al., 2023a)) is to iteratively learn policy covers $\Psi_{1:H}$ for each layer of the MDP. In HOMER, given policy covers for layers $1, \dots, h$, a policy cover for layer $h+1$ is learned by applying the policy optimization method PSDP (Bagnell et al., 2003) to a set of carefully-designed internal reward functions at layer $h+1$. Ideally, the reward functions should incentivize reaching individual latent states; of course, latent states are not in general identifiable, so this criterion must be relaxed. Instead, the rewards are constructed in two steps. First, use two-context regression (with an appropriately-generated dataset, inspired by contrastive learning methods) to estimate the following *kinematics function*:

$$f_{h+1}(x_h, x_{h+1}; a_h) := \frac{\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h)}{\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h) + F_{h+1}(x_{h+1})}, \tag{2}$$

where F_{h+1} is a certain normalization function. Second, sample a large number of "cluster center" observations $(\overline{x}^{(t)}_{h+1})$, and, for each, define a reward $\mathcal{R}^{(t)}$ (derived from Eq. (2)) which is large precisely for those observations x_{h+1} that have approximately the same kinematics as $\overline{x}^{(t)}_{h+1}$.

PCE follows the same blueprint, with two modifications. First, HOMER uses an offline cost-sensitive classification oracle for PSDP. We use an alternative implementation of PSDP (Mhammedi et al., 2023b) which can be implemented with one-context regression (Lemma G.1) and hence two-context regression (via Proposition G.5). Second, the analysis of HOMER assumes that all states are reachable with non-negligible probability. We remove this assumption via truncation arguments and an iterative discovery method (Golowich et al., 2024b)—this is the reason for the outer loop in PCE. See Appendix C for the formal algorithm and analysis.

3.2. Necessity: Reducing Two-Context Regression to Episodic RL

Our second result provides a converse of Theorem 3.1. We give a two-context regression algorithm that requires access to a reward-free episodic RL *oracle* (Definition 2.1), and is oracle-efficient so long as the oracle is sample-efficient.

Theorem 3.2 *There is a constant $C_{3.2} > 0$ and an algorithm* RegToRL *(Algorithm 3 in Appendix D) so that the following holds. Let $\Phi^{\mathsf{aug}} \subseteq (\mathcal{X}^{\mathsf{aug}} \to \mathcal{S}^{\mathsf{aug}})$ be any regular concept class (Definition 2.5), let $N^{\circ}_{\mathsf{RL}}, C_{\mathsf{RL}} \in \mathbb{N}$, and let $\mathcal{O}$ be a $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free episodic RL oracle for Φ^{aug}, with $\max(N_{\mathsf{RL}}(\epsilon,\delta,H,A), K_{\mathsf{RL}}(\epsilon,\delta,H,A)) \leq N^{\circ}_{\mathsf{RL}} \cdot (AH/\epsilon\delta)^{C_{\mathsf{RL}}}$. Then* $\mathtt{RegToRL}(\mathcal{O}, \cdot)$ *is an N_{reg}-efficient two-context regression algorithm (Definition 2.4) for Φ^{aug} with $N_{\mathsf{reg}}(\epsilon,\delta) \leq N^{\circ}_{\mathsf{RL}} \left(|\mathcal{S}|/(\epsilon\delta)\right)^{C_{3.2} \cdot C_{\mathsf{RL}}}$ and with oracle time complexity at most $N^{\circ}_{\mathsf{RL}} \left(|\mathcal{S}|/(\epsilon\delta)\right)^{C_{3.2} \cdot C_{\mathsf{RL}}}$.*

Theorem 3.1 and Theorem 3.2 together show that two-context regression is a minimal oracle for reward-free episodic RL (for any regular Φ). Theorem 3.2 strengthens (Golowich et al., 2024a, Proposition B.2), which reduces *one*-context regression to RL—to our knowledge, the only prior result of this flavor. We require regularity[8] for the following technical reason: a concept class is regular if and only if it can be obtained by augmenting some base concept class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ with

7. Specifically, the "non-quantized" version of HOMER described in Appendix E of Misra et al. (2020).
8. Golowich et al. (2024a) reduce to reward-directed RL and do not use regularity; however, under regularity it is simple to adapt their reduction to reward-free RL; this adaptation is the version we sketch below.

two fully observed states $\{0,1\}$ (see Definition D.1 for the formal definition). It is easy to reduce regression with Φ^{aug} to regression with Φ (Proposition G.8), so it suffices to reduce regression with Φ to reward-free RL with Φ^{aug}; the extra states provide useful flexibility since Φ is otherwise arbitrary.

We now sketch the prior reduction before discussing how to strengthen it.

Recap: Reducing one-context regression to RL. Given a one-context regression dataset $(x^{(i)}, y^{(i)})_{i=1}^n$ with samples in $\mathcal{X} \times \{0,1\}$, consider simulating an MDP with horizon $H = 2$, where the initial observation lies in $\mathcal{X}$ and the second observation lies in $\{0,1\}$. The goal of this reduction is that a *policy* that visits state 1 at step 2 with near-optimal probability corresponds to an accurate *prediction function* for the regression. This can be achieved by defining the action space $\mathcal{A}$ to be a discretization of the interval $[0,1]$, and requiring the policy to "guess" $y^{(i)}$ after observing $x^{(i)}$.

More formally, the reduction uses a fresh datapoint $(x^{(i)}, y^{(i)})$ to simulate each new episode. It first passes observation $x_1 := x^{(i)} \in \mathcal{X}$ to the RL agent. When the agent plays an action $a_1 \in \mathcal{A} \subset [0,1]$, the reduction passes the observation $x_2 \in \{0,1\}$ sampled from $\mathrm{Ber}(1-(a-y^{(i)})^2)$. It can be checked that this procedure in fact simulates a Φ-decodable MDP, so long as the regression dataset satisfied the desideratum that $\mathbb{E}[y^{(i)} \mid x^{(i)}]$ only depends on $\phi(x^{(i)})$ for some $\phi \in \Phi$.

Reducing two-context regression to RL. Can we generalize the above construction by increasing the horizon? Given a two-context regression sample $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$, consider passing both contexts to the RL agent one-by-one and then requiring the policy to "guess" $y^{(i)}$. Unfortunately, $y^{(i)}$ may depend on both $x_1^{(i)}$ and $x_2^{(i)}$, so the simulated decision process is non-Markovian. More broadly, this points to a representational obstacle: optimal policies for a Block MDP are Markovian and hence mappings $\mathcal{X} \to \mathcal{A}$. But for two-context regression, a predictor is a function on $\mathcal{X} \times \mathcal{X}$. Thus, any successful reduction will have to "stitch together" *multiple* policies produced by the RL oracle.

To motivate our reduction, we recall why two-context regression was useful for RL in the first place: essentially, it was useful to estimate (some transformation of) transition probabilities between two consecutive states—see Eq. (2). This suggests using the regression data to simulate an MDP where *the probability of transitioning from $x_1^{(i)}$ to $x_2^{(i)}$ depends on $y^{(i)}$*.

Formally, our reduction simulates a horizon-2 MDP with first observation in $\mathcal{X}$, second observation in $\mathcal{X} \sqcup \{0\}$, and action space $\mathcal{A} \subset [0,1]$. For each sample $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$, the reduction simulates an episode of interaction with the MDP as follows: first, pass $x_1 := x_1^{(i)} \in \mathcal{X}$ to the RL oracle. Second, when the oracle plays an action $a_1 \in \mathcal{A} \subset [0,1]$, pass $x_2 \in \mathcal{X} \sqcup \{0\}$ sampled as:

$$x_2 := \begin{cases} x_2^{(i)} & \text{with probability } 1 - (a_1 - y^{(i)})^2 \\ 0 & \text{with probability } (a_1 - y^{(i)})^2 \end{cases}.$$

It can be checked that this procedure simulates a Φ-decodable block MDP under the realizability assumptions on the regression dataset. Additionally, if we fix a latent state $s \in \mathcal{S}$, then maximizing the probability of reaching state s at step 2 is equivalent to predicting $y^{(i)}$ conditioned on the observation $x_1^{(i)}$ *and* the knowledge that $\phi^\star(x_2^{(i)}) = s$. Thus, any policy $\pi_s : \mathcal{X} \to \mathcal{A}$ that visits s at step 2 with near-maximal probability must approximately minimize a *restricted* regression loss:

$$L_s(\pi_s) := \mathop{\mathbb{E}}_{x_1, x_2} \left[\mathbb{1}[\phi^\star(x_2) = s](\pi_s(x_1) - f(\phi^\star(x_1), \phi^\star(x_2)))^2 \right]$$

where $f : \mathcal{S} \times \mathcal{S} \to [0,1]$ is as in Definition 2.4. By assumption, the reward-free RL oracle will return a set of policies Ψ that contains at least one such policy π_s for each $s \in \mathcal{S}$.

The remaining challenge is how to stitch together these policies into a single predictor: given $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$, how do we use x_2 to identify the policy $\pi \in \Psi$ for which $\pi(x_1)$ is a good prediction? We accomplish this using *one-context regression*, which (as discussed above) is reducible to reward-free RL. In particular, for each policy $\pi \in \Psi$, we construct datapoints of the form $(x_2^{(i)}, (\pi(x_1^{(i)}) - y^{(i)})^2)$. Applying one-context regression yields an estimate of (an appropriate transformation of) the map $x_2 \mapsto L_{\phi^\star(x_2)}(\pi)$. After learning this map, the final predictor $\mathcal{R} : \mathcal{X} \times \mathcal{X} \to [0,1]$ is defined as follows: on input (x_1, x_2), it outputs $\widehat{\pi}^{x_2}(x_1)$, where $\widehat{\pi}^{x_2} \in \Psi$ minimizes the estimated loss. See Appendix D for the formal reduction and analysis.

4. A Simpler Oracle for RL in Block MDPs with Reset Access

We now turn to the *RL with resets* access model, which is more permissive than episodic access. In this model, we give a reward-free RL algorithm PCR that only requires access to a *one-context* regression oracle Reg, and is oracle-efficient so long as Reg is sample-efficient.

Theorem 4.1 (Special case of Theorem E.1) *There is a constant $C_{4.1} > 0$ and an algorithm* PCR *(Algorithm 5 in Appendix E) so that the following holds. Fix $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ and $N^\circ_{\mathsf{reg}}, C_{\mathsf{reg}} \in \mathbb{N}$. Let* Reg *be a N_{reg}-efficient one-context regression oracle for Φ with $N_{\mathsf{reg}}(\epsilon, \delta) := N^\circ_{\mathsf{reg}}/(\epsilon\delta)^{C_{\mathsf{reg}}}$. Then* $\mathsf{PCR}(\mathsf{Reg}, N_{\mathsf{reg}}, |\mathcal{S}|, \cdot)$ *is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free RL algorithm for Φ in the* reset access model, *with $K_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \le H^2|\mathcal{S}|^2$ and $N_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \le N^\circ_{\mathsf{reg}} \cdot (H|\mathcal{A}||\mathcal{S}|/(\epsilon\delta))^{C_{4.1}C_{\mathsf{reg}}}$. Moreover, the oracle time complexity of* PCR *is at most $N^\circ_{\mathsf{reg}} \cdot (H|\mathcal{A}||\mathcal{S}|/(\epsilon\delta))^{C_{4.1}C_{\mathsf{reg}}}$.*

Since RL in the *episodic* access model is provably harder than one-context regression (Golowich et al., 2024a), Theorem 4.1 gives a provable computational benefit of reset access—to our knowledge, the first of its kind, though statistical benefits are known in different settings (Section A.1).

Overview of algorithm design and proof. The PCR algorithm follows a similar blueprint to PCE (and HOMER): given policy covers for the first h layers, design internal reward functions for layer $h+1$ based on kinematics, and use PSDP to optimize these reward functions and obtain a policy cover for layer $h+1$ (recall that PSDP itself only requires a one-context regression oracle). The point of departure is in how the kinematics are estimated. In PCE, two-context regression is applied on a dataset consisting of some pairs $(x_h^{(i)}, x_{h+1}^{(i)})$ generated using the MDP's real transitions (and labeled $y^{(i)} = 1$), and some pairs $(x_h^{(j)}, x_{h+1}^{(j)})$ generated using "fake" transitions (and labeled $y^{(j)} = 0$). The Bayes optimal predictor $\mathbb{E}[y \mid x_1, x_2]$ turns out to be exactly the kinematics function from Eq. (2).

In the setting of Theorem 4.1, our algorithm PCR does not have access to two-context regression, but it does have the ability to "reset" the MDP to previously-seen observations. To use this, the algorithm first generates a large number of "discriminator" observations $(x_h^{(i)})_{i=1}^m$. For each fixed $i \in [m]$ and $a \in \mathcal{A}$, the algorithm then learns to predict the following predicate: "did the observation x_{h+1} come from $x_h^{(i)}$ and action a, or from some other $x_h^{(j)}$ or other action?". More precisely, using the power of resets, PCR can generate a large number of samples (x, y) where $x \sim \mathbb{P}_{h+1}(\cdot \mid x_h^{(j)}, a_h)$ for $(j, a_h) \sim \mathrm{Unif}([m] \times \mathcal{A})$, and $y := \mathbb{1}[j = i \wedge a_h = a]$. The Bayes optimal predictor $\mathbb{E}[y \mid x]$ is then:

$$w_{h+1}(x_{h+1}; i, a) := \frac{\mathbb{P}_{h+1}(x_{h+1} \mid x_h^{(i)}, a)}{\sum_{j, a_h} \mathbb{P}_{h+1}(x_{h+1} \mid x_h^{(j)}, a_h)} = \frac{\widetilde{\mathbb{P}}_{h+1}(\phi^\star(x_{h+1}) \mid \phi^\star(x_h^{(i)}), a)}{\sum_{j, a_h} \widetilde{\mathbb{P}}_{h+1}(\phi^\star(x_{h+1}) \mid \phi^\star(x_h^{(j)}), a_h)}. \quad (3)$$

This function can be estimated via one-context regression, and then used to define internal reward functions at layer $h+1$, similar to PCE. See Appendix E for the formal algorithm and analysis.

Is one-context regression minimal? By a variant of (Golowich et al., 2024a, Proposition B.2), the *noiseless* version of one-context regression is necessary in the reset setting (Proposition G.4). While there are classical examples where noisy (but realizable) PAC learning is believed to be computationally harder than noiseless learning (Blum et al., 2003), in many natural settings they are comparable—see e.g. halfspace learning (Blum et al., 1998; Diakonikolas et al., 2023) and the statistical query model (Kearns, 1998). In this sense, we expect that (noisy) one-context regression is *nearly* minimal.

5. A Computational Separation for Low-Rank MDPs

We now move beyond Block MDPs. *Low-Rank* MDPs (Modi et al., 2024) are perhaps the simplest commonly-studied model class that generalizes the Block MDP. While Low-Rank MDPs admit oracle-efficient RL algorithms, the *oracles* used seem significantly more complex—both in the episodic and reset access model—than the oracles used for Block MDPs. In particular, VOX (Mhammedi et al., 2023a), which operates in the episodic access model, uses a proper min-max optimization oracle, and RVFS (Mhammedi et al., 2024), which requires reset access, uses an agnostic and cost-sensitive regression oracle. Comparing to our results for Block MDPs, it is natural to ask whether this apparent gulf in computational tractability is real, and if so, what the structural source is. In this section we make progress on this question, with a focus on the reset access model.

Preliminaries. An MDP M is Low-Rank with rank d if there are maps $\phi_h^{\mathsf{lin}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu_{h+1}^{\mathsf{lin}} : \mathcal{X} \to \mathbb{R}^d$ such that the transitions have the following factorization: $\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h) = \langle \phi_h^{\mathsf{lin}}(x_h, a_h), \mu_{h+1}^{\mathsf{lin}}(x_{h+1}) \rangle$. Prior model-free work assumes that $\phi_{1:H}^{\mathsf{lin}}$ lie in a known feature class Φ^{lin}, but the dual features $\mu_{1:H}^{\mathsf{lin}}$ are arbitrary (Modi et al., 2024; Mhammedi et al., 2023a).[9] Given implicit access to Φ^{lin} via some oracle(s), the typical goal is to design RL algorithms with time and oracle complexity polynomial in d, the horizon H, and the number of actions $\mathcal{A}$. Block MDPs with concept class Φ can be embedded in a class of Low-Rank MDPs with rank $d := |\mathcal{S}||\mathcal{A}|$: simply define $\Phi^{\mathsf{lin}} := \{\phi^{\mathsf{lin}} : \phi \in \Phi\}$ where $\phi^{\mathsf{lin}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ maps to an appropriate basis vector, i.e. $\phi^{\mathsf{lin}}(x, a) := e_{\phi(x),a}$. Analogously, there is a natural extension of one-context regression to Low-Rank MDPs.

Definition 5.1 (Informal) *The goal of* ***one-context low-rank regression*** *over Φ^{lin} is the following: given an i.i.d. dataset $(x^{(i)}, a^{(i)}, y^{(i)})_i$ satisfying the realizability assumption $\mathbb{E}[y^{(i)} \mid x^{(i)}, a^{(i)}] = \langle \phi^{\mathsf{lin}}(x^{(i)}, a^{(i)}), \theta \rangle$ for unknown $\phi^{\mathsf{lin}} \in \Phi^{\mathsf{lin}}$ and $\theta \in \mathbb{R}^d$, estimate the function $(x, a) \mapsto \mathbb{E}[y \mid x, a]$.*

This oracle suffices to implement PSDP in Low-Rank MDPs—so *given* a policy cover,[10] we can optimize a reward (Mhammedi et al., 2023a). In analogy with Theorem 4.1, we ask: is this oracle sufficient for the full task of reward-free RL with resets? We show that it is not, and shed light on why.

Generalized Block MDPs. It is well known that Block MDPs are a special case of Low-Rank MDPs (Modi et al., 2024; Zhang et al., 2022). To highlight that there are *multiple* important assumptions in the Block MDP definition, and to gain a finer understanding of how these assumptions interact with computational tractability, we introduce—and prove our hardness result in—an intermediate model class between Block MDPs and Low-Rank MDPs. Let $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ be any concept class.

9. Model-based works (Agarwal et al., 2020; Uehara et al., 2022) also assume that $\mu_{1:H}^{\mathsf{lin}}$ lie in a known dual feature class.
10. See Mhammedi et al. (2023a) for the precise notion of a policy cover in Low-Rank MDPs.

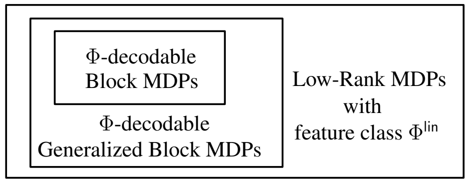


Figure 2: Containment diagram for model classes discussed in Section 5. Here, $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ is any concept class, and $\Phi^{\mathsf{lin}} := \{\phi^{\mathsf{lin}} : \phi \in \Phi\}$ where $\phi^{\mathsf{lin}}(x, a) := e_{\phi(x),a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

A Φ-decodable *Generalized* Block MDP is an MDP $M = (H, \mathcal{X}, \mathcal{A}, (\mathbb{P}_h)_h)$ with the property that there exist $\phi_1^\star, \dots, \phi_H^\star \in \Phi$ so that $\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h)$ is a function of x_{h+1}, $\phi_h^\star(x_h)$, and a_h.

Generalized Block MDPs are still Low-Rank MDPs via the same embedding discussed above (Proposition F.2)—the features are appropriate basis vectors. However, compared to standard Block MDPs, the transition probability $\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h)$ can now depend arbitrarily on x_{h+1}.[11] Moreover, in Generalized Block MDPs, Definition 5.1 simplifies back to Definition 2.2, our original definition of one-context regression (Proposition F.3). These connections are summarized in Figure 2.

Hardness for Generalized Block MDPs. Unfortunately, the arbitrary dependence of the transition probability on x_{h+1} defeats algorithms such as PCR, since the ideal kinematics (see Eq. (3)) are no longer a function of just $\phi^\star(x_{h+1})$. But could there be a more clever algorithm that avoids needing to estimate such quantities? In Theorem 5.2, we show that there is an inherent computational barrier.

For $n \in \mathbb{N}$, let $\Phi_n := \{\phi^\theta : \mathbb{R}^n \to \{0,1\} \mid \theta \in \mathbb{R}^n\}$ be the concept class of linear thresholds: $\phi^\theta(x) := \mathbb{1}[\langle x, \theta \rangle \geq 0]$. We prove that one-context regression for Φ_n is computationally tractable, but reward-free RL with resets for Φ_n-decodable Generalized Block MDPs is cryptographically hard.

Theorem 5.2 (Theorem F.5+Theorem F.8) *There is an algorithm for one-context regression with concept class Φ_n, that achieves error ϵ with probability at least $1-\delta$ and has time complexity* $\mathrm{poly}(n, 1/\epsilon, 1/\delta)$. *In contrast, suppose there exists an* $\mathtt{Alg}^M$ *that—given interactive reset access to any Φ_n-decodable Generalized Block MDP M with horizon H, observation space $\mathcal{X} = \mathbb{R}^n$, latent state space $\mathcal{S} = \{0,1\}$, and action space $\mathcal{A}$—has time complexity* $\mathrm{poly}(n, H, |\mathcal{A}|)$ *and produces a set of policies Ψ satisfying the following guarantee with probability at least* $1/2$*:*

$$\forall s \in \mathcal{S} : \max_{\pi \in \Psi} d_H^{M,\pi}(s) \geq \frac{1}{\mathrm{poly}(|\mathcal{A}|, |\mathcal{S}|, H)} \left(\max_{\pi \in \Pi} d_H^{M,\pi}(s) - \frac{1}{8} \right). \tag{4}$$

Then the Continuous Learning With Errors (cLWE) hardness assumption (Assumption F.7) is false.

Theorem 5.2 rules out any reduction from reward-free RL with resets in Φ_n-decodable Generalized Block MDPs to one-context regression, where the oracle time complexity of the reduction is allowed to scale polynomially in all relevant parameters (H, $|\mathcal{S}|$, $|\mathcal{A}|$, $1/\epsilon$, $1/\delta$, and the description length n of an observation). By Theorem 4.1, this separates standard Block MDPs from Low-Rank MDPs. We remark that the cLWE assumption can be based on classical LWE (Gupte et al., 2022).

Remark 5.3 (On the reward-free solution concept) *Comparing Eq. (4) to Definition 2.1, we have* relaxed *the notion of exploration to allow for multiplicative error in the visitation probability. In*

11. We also defined standard Block MDPs to have a single decoding function $\phi^\star$ rather than one per layer, but this is was a superficial choice made for notational simplicity; our RL algorithms from prior sections still apply if there is one function per layer, and the one-context regression oracle needed by PCR doesn't change.

other words, we are only asking for an (α, ϵ)-policy cover where $\alpha = 1/\operatorname{poly}(H, |\mathcal{A}|, |\mathcal{S}|)$. The reason is that for Generalized Block MDPs, a $(1, \epsilon)$-policy cover may not even exist. Fortunately, an (α, ϵ)-policy cover does always exist, and finding it is statistically tractable (Proposition F.17), indicating that the source of the hardness in Theorem F.8 is indeed purely computational.

Proof overview. In Theorem 5.2 (proven in Appendix F), the algorithmic result is a (simple, but not immediate) consequence of seminal work on learning halfspaces (Blum et al., 2003; Diakonikolas et al., 2023). The proof of the hardness result builds on recent work by Tiegel (2023) on the cryptographic hardness of agnostic learning of halfspaces. Fix $n \in \mathbb{N}$ and let $\mathfrak{S}^{t-1}$ denote the unit sphere in $t \approx \operatorname{polylog}(n)$ dimensions. Tiegel (2023) shows that under cLWE, there are two families $\{\nu_{w,0} : w \in \mathfrak{S}^{t-1}\}$ and $\{\nu_{w,1} : w \in \mathfrak{S}^{t-1}\}$ of distributions on $\mathbb{R}^n$ such that for any w, the distributions $\nu_{w,0}$ and $\nu_{w,1}$ are *approximately* separated by a hyperplane $\phi^{\theta(w)}$, but for unknown $w \sim \operatorname{Unif}(\mathfrak{S}^{t-1})$, distinguishing either $\nu_{w,0}$ or $\nu_{w,1}$ from a certain null distribution ν_{null} is computationally hard.

We use these distributions to construct a family of *approximate* combination lock MDPs. These MDPs are parametrized by hidden vectors $w_1, \dots, w_H \in \mathfrak{S}^{t-1}$ and hidden actions $a_1^\star, \dots, a_H^\star \in \mathcal{A}$. The transition distribution at any state $x_h \in \mathbb{R}^n$ and action a_h is either $\nu_{w_{h+1},0}$ or $\nu_{w_{h+1},1}$, depending on the current latent state $\phi^{\theta(w_h)}(x_h)$ and whether a_h equals $a_h^\star$. By a hybrid argument, we prove that these MDPs are indistinguishable from the MDP where all transitions follow ν_{null}, and hence learning $a^\star$ is impossible. However, we prove that learning $a^\star$ is necessary to solve the reward-free RL task.

Takeaway: a computational role for weight function realizability? The immediate reason why PCR fails for Low-Rank MDPs is that *weight functions*, i.e. ratios $\mathbb{P}_{h+1}(x_{h+1} \mid x, a)/\mathbb{P}_{h+1}(x_{h+1} \mid x', a')$, are not realizable as linear functions in the feature mapping. In contrast, for Block MDPs the analogous realizability does hold, and plays a central role in algorithms like HOMER and MusIK. Theorem 5.2 presents evidence that this distinction is computationally important, and perhaps suggests *regression oracles for weight functions* as a route for generalizing our theory of computational tractability beyond Block MDPs. Algorithmically, methods from Amortila et al. (2024)—which apply to MDPs with low *coverability* (subsuming Low-Rank MDPs)—offer some initial hope in this direction.

6. Discussion and Future Work

This work takes a step towards understanding the minimal computational oracles needed for reinforcement learning; however, much remains unclear, both for the specific tasks discussed in this paper, and beyond. Below, we discuss some particularly notable questions, ordered roughly from narrowest to broadest.

Pinning down a minimal oracle in the reset access model? We showed in Theorem 4.1 that one-context regression is sufficient for reward-free RL in Block MDPs with reset access. It's known to be necessary for reward-free RL in Block MDPs with *episodic* access (Golowich et al., 2024a), but for the reset access model we only show that noiseless one-context regression is necessary (Proposition G.4). Likely, the same argument shows that *one-context regression with conditional sampling queries* is necessary (which slightly strengthens noiseless one-context regression), but even this oracle does not seem sufficient for learning inverse dynamics.

Reward-free vs reward-directed? Our work focuses on reward-free RL, and in particular we mostly focus on the strongest formulation: computing a $(1, \epsilon)$-policy cover. For Block MDPs, relaxing to (α, ϵ)-policy covers or reward-directed RL does not appear to lead to simpler algorithms, but there are also technical obstacles to extending our results on necessity of two-context regression (e.g. Theorem 3.2) to these settings. Most notably, while computing a $(1, \epsilon)$-policy cover can be hard for even $H = 2$, hardness for these weaker tasks only arises when $H = \omega(1)$, and it's unclear how to use two-context regression data to usefully simulate interaction with an MDP with longer horizon. Existing hardness results that exploit long horizon include Theorem 5.2 and the main result of Golowich et al. (2024a), but both require specific concept classes with specific structure, whereas Theorem 3.2 applies to almost any concept class.

This technical difficulty leaves an open question: are there substantive computational differences between these tasks? This question may be of conceptual interest since $(1, \epsilon)$-policy covers are not meaningful beyond Block MDPs; a better understanding of weaker notions of exploration seems essential for more general characterizations.

Stronger hardness for Low-Rank MDPs? In Theorem 5.2, we show that one-context low-rank regression is insufficient for Low-Rank MDPs in the reset access model, which indicates a qualitative computational difference with Block MDPs. However, this leaves several questions:

- Is *agnostic* one-context regression necessary? The source of computational hardness in Theorem 5.2 is the same as the source for agnostic halfspace learning (Tiegel, 2023), but the result does not constitute a general-purpose reduction.

- Is two-context low-rank regression insufficient for Low-Rank MDPs in the episodic access model? In a sense, Theorem 5.2 gives weak positive evidence: the RL hardness result still holds in the episodic access model, and two-context regression for the concept class of halfspaces is likely reducible to PAC learning an intersection of two halfspaces (with random classification noise). So if two-context regression were sufficient, then it would likely imply cryptographic hardness of learning an intersection of two halfspaces under LWE, which currently seems out of reach (Tiegel, 2024).

- More broadly, is there *any* oracle-efficient algorithm for RL in Low-Rank MDPs in the episodic access model that only uses a "minimization" oracle? In stark contrast with Block MDPs, the only existing algorithm for RL in Low-Rank MDPs requires min-max optimization, and it would be interesting to understand if there is an inherent computational barrier.

A computational taxonomy for RL? Perhaps the most interesting question is whether our methods can be extended to develop a theory of computational tractability in more general interactive decision-making problems (i.e. beyond Block MDPs and Low-Rank MDPs), in analogy with the theory of statistical tractability developed by Foster et al. (2021). As we discussed in Section 5, we view *regression oracles for weight functions* as one promising direction, but a satisfying answer to this question could require developing new oracles and modeling assumptions.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International conference on machine learning*, pages 1638–1646. PMLR, 2014.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Proceedings of the 34th International*

Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Philip Amortila, Dylan J Foster, and Akshay Krishnamurthy. Scalable online exploration via coverability. In *Forty-first International Conference on Machine Learning*, 2024.

Benny Applebaum, Boaz Barak, and David Xiao. On basing lower-bounds for learning on worst-case assumptions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 211–220. IEEE, 2008.

James Bagnell, Sham M Kakade, Jeff Schneider, and Andrew Ng. Policy search by dynamic programming. *Advances in neural information processing systems*, 16, 2003.

Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22:35–52, 1998.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé. Classical hardness of learning with errors. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 575–584, 2013.

Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous lwe. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 694–707, 2021.

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

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

Ilias Diakonikolas, Christos Tzamos, and Daniel M Kane. A strongly polynomial algorithm for approximate forster transforms and its application to halfspace learning. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1741–1754, 2023.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. *Journal of the ACM (JACM)*, 67(5):1–57, 2020.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2022.

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

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable pomdps, without computationally intractable oracles. *Advances in Neural Information Processing Systems*, 35: 1458–1473, 2022.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 349–362, 2023.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1953–1967. IEEE, 2024a.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Exploring and learning in sparse linear mdps without computationally intractable oracles. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 183–193, 2024b.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1162–1173. IEEE, 2022.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. In *The mathematics of generalization*, pages 37–116. CRC Press, 2018.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 128–141, 2016.

Marcel Hussing, Michael Kearns, Aaron Roth, Sikata B Sengupta, and Jessica Sorrell. Oracle-efficient reinforcement learning for max value ensembles. *Advances in Neural Information Processing Systems*, 37:117657–117681, 2024.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33: 18530–18539, 2020a.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020b.

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

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

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Daniel Kane, Sihan Liu, Shachar Lovett, and Gaurav Mahajan. Computational-statistical gap in reinforcement learning. In *Conference on Learning Theory*, pages 1282–1302. PMLR, 2022.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J Foster, Lekan P Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=TNocbXm5MZ.

Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, and Yuting Wei. Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting. *Advances in Neural Information Processing Systems*, 34:16671–16685, 2021.

Sihan Liu, Gaurav Mahajan, Daniel Kane, Shachar Lovett, Gellért Weisz, and Csaba Szepesvári. Exponential hardness of reinforcement learning with linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1588–1617. PMLR, 2023.

Zak Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. *Advances in Neural Information Processing Systems*, 37:12334–12407, 2024.

Zakaria Mhammedi, Adam Block, Dylan J Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank mdps. *Advances in Neural Information Processing Systems*, 36:66782–66817, 2023a.

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. Representation learning with multi-step inverse kinematics: An efficient and optimal approach to rich-observation rl. In *International Conference on Machine Learning*, pages 24659–24700. PMLR, 2023b.

John Milnor. On the betti numbers of real varieties. *Proceedings of the American Mathematical Society*, 15(2):275–280, 1964.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International Conference on Machine Learning*, pages 6961–6971. PMLR, 2020.

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

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Stefan Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3029–3064. PMLR, 2023.

Stefan Tiegel. Improved hardness results for learning intersections of halfspaces. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4764–4786. PMLR, 2024.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Gellert Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, and Csaba Szepesvári. On query-efficient planning in mdps under linear realizability of the optimal state-value function. In *Conference on Learning Theory*, pages 4355–4385. PMLR, 2021.

Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory*, pages 1165–1192. PMLR, 2022.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.

Contents of Appendix

Oracle(s)	Additional assumptions	Reference
$\underset{\hat{f}:\mathcal{X}\times\mathcal{X}\to[0,1]}{\arg\min} \sum_{(x_1,x_2,y)\in\mathcal{D}} (\hat{f}(x_1,x_2)-y)^2$	Reachability	Misra et al. (2020)
$\underset{\hat{\phi}\in\Phi}{\arg\min} \underset{\substack{f:\mathcal{S}\to[0,1]\\ \phi\in\Phi}}{\max} \underset{\hat{f}:\mathcal{S}\times\mathcal{A}\to[0,1]}{\min} \sum_{(x,a)\in\mathcal{D}} (\hat{f}(\hat{\phi}(x),a)-\mathbb{E}_{x'\mid x,a} f(\phi(x')))^2$	Reachability	Modi et al. (2024)
	—	Zhang et al. (2022)
$\underset{\substack{\mu:\mathcal{S}^2\to\Delta(\mathcal{A}\times\mathcal{S})\\ \hat{\phi}\in\Phi}}{\arg\max} \sum_{(j,a,x,x')\in\mathcal{D}} \log\mu((a,j)\mid\hat{\phi}(x),\hat{\phi}(x'))$	—	Mhammedi et al. (2023b)
$\sup_{\hat{f}:\mathcal{X}\to[0,1]} \left\lvert\sum_{x'\in\mathcal{C}}\left(\hat{V}(x')-f(x')\right)\right\rvert$ s.t. $\sum_{x'\in\mathcal{D}}(\hat{V}(x')-\hat{f}(x'))^2\le\varepsilon$	**Reset access**	Mhammedi et al. (2024)

Table 1: This table, adapted from (Golowich et al., 2024a, Table 1), gives an informal overview of the oracles used in oracle-efficient RL for Φ-decodable block MDPs; the methods above the double line apply in the episodic access model. Misra et al. (2020) additionally require a cost-sensitive classification oracle for PSDP, though this can be replaced by one-context regression. The oracle in the second row can be implemented by the RepLearn algorithm (Modi et al., 2024; Zhang et al., 2022; Mhammedi et al., 2023a), but this algorithm still uses a max-min oracle. The oracle in the third row can be replaced by an analogous squared-loss minimization oracle (Mhammedi et al., 2023b, Footnote 5), equivalent to *proper* two-context regression. Mhammedi et al. (2024) additionally requires a one-context regression oracle.

Appendix A. Additional Related Work

In this section we discuss related work in more detail. Most prior work in theoretical reinforcement learning consists of algorithm design with one of three goals, reflecting varying levels of concern with computational complexity:

1. statistical efficiency with arbitrary computation (Jiang et al., 2017; Foster et al., 2021);
2. oracle-efficiency, but the choice of oracle is not the focus (Dann et al., 2018);
3. end-to-end computational efficiency (Kearns, 1998; Jin et al., 2020c; Golowich et al., 2024b).

Our work fits between the second and third levels, since we are interested in the *minimal* oracles that are necessary and sufficient for oracle-efficient RL—and we are interested in settings where end-to-end efficiency is unlikely. We therefore omit discussion of non-oracle-efficient algorithms. In Section A.1, we survey prior oracle-efficient algorithms for RL in Block MDPs, with a focus on which oracles they require. In Section A.2, we give a partial survey of oracle-efficient algorithms for RL in Low-Rank MDPs and beyond. In Section A.3, we discuss previous applications of one-context regression in theoretical RL. Finally, in Section A.4 we discuss prior works on computational hardness in RL.

A.1. Algorithms for Block MDPs

Episodic RL. In Table 1 we give an informal overview of the oracles used in oracle-efficient RL algorithms for Block MDPs, adapted from Golowich et al. (2024a). For simplicity, we omit discussion of earlier works that required stronger assumptions—beyond reachability, which we view as largely technical–such as deterministic dynamics (Dann et al., 2018), separability (Du et al., 2019), or function approximation for the observation distributions (Agarwal et al., 2020).

In Table 1, all works listed above the double line apply in the episodic access model. For our purposes, the most relevant work is that of Misra et al. (2020); their algorithm HOMER requires only a two-context regression oracle and PSDP, which can be implemented with one-context regression (Mhammedi et al., 2023b). Notably, HOMER is the only prior algorithm which uses an *improper* oracle; the others all seem to heavily rely on properness (Modi et al., 2024; Zhang et al., 2022; Mhammedi et al., 2023b). The only catch is that Misra et al. (2020) analyze HOMER under a *reachability* assumption—i.e., the sample complexity scales inverse-polynomially with the minimum visitation probability of any state $\eta := \min_{s\in\mathcal{S}} \min_{h\in[H]} \max_{\pi\in\Pi} d_h^{M,\pi}(s)$. Avoiding this dependence is the primary technical contribution of Theorem 3.1.

The conventional goal in theoretical RL is reward-directed RL, i.e. find a policy that has approximately maximal value with respect to an external reward function. The alternative goal is reward-free exploration (Jin et al., 2020b). For Block MDPs, this is typically formulated as the task of computing an (α,ϵ)-policy cover, for given $\epsilon > 0$ and any $\alpha = 1/\operatorname{poly}(H, |\mathcal{A}|, |\mathcal{S}|)$, where $\Psi \subset \Pi$ is an (α,ϵ)-policy cover if, for every layer $h \in [H]$ and latent state $s \in \mathcal{S}$,

$$\max_{\pi\in\Psi} d_h^{M,\pi}(s) \geq \alpha \cdot \max_{\pi\in\Pi} d_h^{M,\pi}(s) - \epsilon.$$

It's straightforward to see that reward-directed RL in Block MDPs is no harder than reward-free RL, since given a policy cover, PSDP can compute a near-optimal policy for any external reward function, using only a one-context regression oracle. It's unknown whether reward-directed RL is computationally *easier* than reward-free RL. However, we remark that reward-free RL is a common building block for reward-directed RL for Block MDPs (Du et al., 2019; Misra et al., 2020; Modi et al., 2024; Mhammedi et al., 2023b) and beyond (Golowich et al., 2022, 2024b; Mhammedi et al., 2023a). Moreover, the only prior algorithms for Block MDPs that do not solve reward-free RL as a byproduct (Modi et al., 2024; Zhang et al., 2022) require a seemingly *more* complicated oracle.

The particular goal that we study in this paper is computing a $(1,\epsilon)$-policy cover (Definition 2.1), which was previously studied by Du et al. (2019). This only strengthens our algorithmic results (Theorem 3.1 and Theorem 4.1), but it does leave an open question regarding our lower bound (Theorem 3.2): can it be strengthened to apply to the problem of computing an (α,ϵ)-policy cover, or the problem of reward-directed RL, or is there an inherent computational gap? See Section 6 for further discussion.

RL with resets. The reset access model augments the episodic access model by allowing the learner to revisit any previously-seen state. This access model has been extensively applied to RL settings with linear value function approximation (Weisz et al., 2021; Li et al., 2021; Yin et al., 2022), for the purpose of circumventing *sample complexity* lower bounds. These settings do not subsume the Block MDP setting, since they assume linearity of the value function(s) in a known feature mapping ϕ. More relevant to us is the work of Mhammedi et al. (2024), who give an oracle-efficient algorithm RVFS for Block MDPs (more generally, MDPs with low *pushforward-coverability*) in the reset access model. As shown in Table 1, RVFS uses a disagreement-type optimization oracle. While it is improper, and has the same solution concept as one-context regression (i.e. a real-valued function on $\mathcal{X}$), there is no obvious reduction to one-context regression, particularly one that ensures realizability. Thus, unlike our result, it's not directly apparent whether it yields a provable computational separation between the reset access model and the episodic access model (though it's certainly possible that it could imply a separation with additional work).

On the other hand, our algorithm PCR is limited to the Block MDP model class, whereas RVFS applies in significantly greater generality. Thus, the guarantees are formally incomparable.

A.2. Algorithms for Low-Rank MDPs and Beyond

Recently, Mhammedi et al. (2023a) developed an oracle-efficient algorithm for episodic RL in Low-Rank MDPs, which generalize Block MDPs. However, this algorithm relies on the analogue of the min-max oracle in Table 1. The RVFS algorithm of Mhammedi et al. (2024) solves RL in Low-Rank MDPs with reset access, but as discussed above, seems to require an agnostic optimization oracle. Beyond Low-Rank MDPs, Amortila et al. (2024) proposed an algorithm for exploring MDPs with low *pushforward coverability* (which is satisfied by Low-Rank MDPs, and hence Block MDPs as well) assuming access to a certain policy optimization oracle as well as an optimization oracle over weight functions.

A.3. Algorithms Based on One-Context Regression

One-context regression is a natural oracle for estimating value functions and Bellman backups (Ernst et al., 2005; Mhammedi et al., 2023b; Golowich et al., 2024a), though until recently, most applications were phrased in terms of the worst-case optimization oracle analogue. Prior to our work, versions of one-context regression were known to be sufficient for offline RL in Block MDPs, under all-policy concentrability (Chen and Jiang, 2019); for (online) RL in Block MDPs with deterministic dynamics (Efroni et al., 2022), and for (online) RL in horizon-one Block MDPs; these results were systematized under Definition 2.2 by Golowich et al. (2024a). A variant of one-context regression (on a "composed" function class) was shown to be sufficient for Block MDPs under a separability condition (Du et al., 2019). Recently, it was shown that (in a setting broader than Block MDPs), one-context regression is sufficient for competing with a weaker notion of optimal policy, called a "max-following policy" for a given policy ensemble (Hussing et al., 2024).

A.4. Computational Hardness in RL

Most relevant to our lower bounds are recent results by Golowich et al. (2024b, 2023), who showed that one-context regression is necessary for RL in Block MDPs (for any concept class Φ) as well as insufficient (for a particular choice of Φ), the latter under a cryptographic assumption. Other hardness results include hardness of learning Partially Observable MDPs (Papadimitriou and Tsitsiklis, 1987; Jin et al., 2020a; Golowich et al., 2023) and hardness of learning MDPs with linear $Q^\star$ and $V^\star$ (Kane et al., 2022; Liu et al., 2023), all of which hold under versions of the Exponential Time Hypothesis.

Appendix B. Additional Preliminaries

Notation in proofs. For any MDP M, any policy π induces a distribution over *trajectories* $(s_1, x_1, a_1, \ldots, s_H, x_H, a_H)$, where the trajectory is sampled via an episode of online interaction with M, and at step h the action a_h is sampled from $\pi_h(x_h)$. We write $\Pr^{M,\pi}[\cdot]$ (respectively, $\mathbb{E}^{M,\pi}[\cdot]$) to denote probability (respectively, expectation) over a trajectory sampled from M with policy π. With this notation, for $h \in [H]$ and $s \in \mathcal{S}$, we have $d_h^{M,\pi}(s) = \Pr^{M,\pi}[s_h = s]$. In a (minor) overload of notation, for $h \in [H]$ and $x \in \mathcal{X}$ we also define the visitation probability of observation x at step h as $d_h^{M,\pi}(x) := \Pr^{M,\pi}[x_h = x]$.

In the remaining appendices, we will be explicit about certain dependences on the MDP M (since we will shortly be introducing truncations of M). In particular, unless the MDP M is clear from context, we write $\widetilde{\mathbb{P}}_h^M$ to denote the latent transition distribution $\widetilde{\mathbb{P}}_h$ of M, and similarly write $\mathbb{P}_h^M$ to denote the observed transition distribution.

Notation in pseudocode. For a sampleable distribution over policies $\rho \in \Delta(\Pi)$, in the pseudocode for our algorithms we may write

$$(x_1, a_1, \ldots, x_k, a_k) \sim \rho$$

to denote sampling a policy $\pi \sim \rho$ and then a sampling (partial) trajectory via an episode of online interaction with the MDP M, using policy π. More generally, for two such distributions $\rho, \rho' \in \Delta(\Pi)$ and a step $h \in [H]$, we may write

$$(x_1, a_1, \ldots, x_k, a_k) \sim \rho \circ_h \rho'$$

to denote sampling policies $\pi \sim \rho$ and $\pi' \sim \rho'$, and then sampling a partial trajectory from M using the policy $\pi \circ_h \pi'$, i.e. playing $a_i \sim \pi_i(x_i)$ for each $i < h$ and $a_i \sim \pi_i'(x_i)$ for each $i \geq h$. In the above notation, for an action $a \in \mathcal{A}$, we may overload a to also denote the policy that deterministically plays action a.

B.1. Truncated MDPs

Fix a Block MDP $M = (H, \mathcal{S}, \mathcal{X}, \mathcal{A}, (\widetilde{\mathbb{P}}_h)_{h\in[H]}, (\widetilde{\mathbb{O}}_h)_{h\in[H]}, \phi^\star)$. For purposes of analysis, it will be useful to define *truncations* of M, as well as *truncated* policy covers defined in terms of certain truncations. To understand the material in Appendices C and E in full generality, these definitions will be important; however, if one is willing to make a reachability assumption on the MDP M (Misra et al., 2020), then one may ignore these definitions and treat $\overline{M}(\Gamma)$ and $\overline{M}(\emptyset)$ as equivalent to M itself.

In truncated Block MDPs, the latent state space and observation space are augmented with a terminal state/context $\mathfrak{t}$, and transitions that would lead to "difficult-to-reach" states in M instead lead to $\mathfrak{t}$. The following preliminaries are adapted from Golowich et al. (2024b). For notational convenience, define $\phi^\star(\mathfrak{t}) := \mathfrak{t}$.

Definition B.1 *Let $\mathcal{S}^{\mathsf{rch}} = (\mathcal{S}_1^{\mathsf{rch}}, \ldots, \mathcal{S}_H^{\mathsf{rch}})$ for some given sets $\mathcal{S}_1^{\mathsf{rch}}, \ldots, \mathcal{S}_H^{\mathsf{rch}} \subseteq \mathcal{S}$. The $\mathcal{S}^{\mathsf{rch}}$-truncation of M is the Block MDP $(H, \overline{\mathcal{S}}, \overline{\mathcal{X}}, \mathcal{A}, (\overline{\mathbb{P}}_h)_{h\in[H]}, (\overline{\mathbb{O}}_h)_{h\in[H]}, \phi^\star)$ with latent state space $\overline{\mathcal{S}} := \mathcal{S} \cup \{\mathfrak{t}\}$, observation space $\overline{\mathcal{X}} := \mathcal{X} \cup \{\mathfrak{t}\}$, observation distribution*

$$\overline{\mathbb{O}}_h(x|s) := \begin{cases} \widetilde{\mathbb{O}}_h(x|s) & \text{if } x \in \mathcal{X}, s \in \mathcal{S} \\ 1 & \text{if } x = s = \mathfrak{t} \\ 0 & \text{otherwise} \end{cases},$$

initial state distribution

$$\overline{\mathbb{P}}_1(s_1) := \begin{cases} \widetilde{\mathbb{P}}_1(s_1) & \text{if } s_1 \in \mathcal{S}_1^{\mathsf{rch}} \\ 0 & \text{if } s_1 \in \mathcal{S} \setminus \mathcal{S}_1^{\mathsf{rch}} \\ \sum_{z \in \mathcal{S} \setminus \mathcal{S}_1^{\mathsf{rch}}} \widetilde{\mathbb{P}}_1(z) & \text{if } s_1 = \mathfrak{t} \end{cases},$$

and transition distribution

$$\overline{\mathbb{P}}_h(s_h|s_{h-1},a) := \begin{cases} \widetilde{\mathbb{P}}_h(s_h|s_{h-1},a) & \text{if } s_{h-1} \in \mathcal{S}, s_h \in \mathcal{S}_h^{\mathsf{rch}} \\ 0 & \text{if } s_{h-1} \in \mathcal{S}, s_h \in \mathcal{S} \setminus \mathcal{S}_h^{\mathsf{rch}} \\ \sum_{z \in \mathcal{S} \setminus \mathcal{S}_h^{\mathsf{rch}}} \widetilde{\mathbb{P}}_h(z|s_{h-1},a) & \text{if } s_{h-1} \in \mathcal{S}, s_h = \mathfrak{t} \\ \mathbb{1}[s_h = \mathfrak{t}] & \text{if } s_{h-1} = \mathfrak{t} \end{cases}$$

for each $h \in \{2, \ldots, h\}$.

Definition of $\overline{M}(\emptyset)$ and $\overline{M}(\Gamma)$. Fix parameters $\sigma_{\mathsf{trunc}} \geq \sigma_{\mathsf{bkup}} > 0$ and a finite set of policies $\Gamma \subset \Pi$. We inductively define sets $\mathcal{S}_1^{\mathsf{rch}}(\Gamma), \ldots, \mathcal{S}_H^{\mathsf{rch}}(\Gamma)$ and truncated Block MDPs $\overline{M}_1(\Gamma), \ldots, \overline{M}_H(\Gamma)$ as follows. First, define

$$\mathcal{S}_1^{\mathsf{rch}}(\Gamma) := \mathcal{S}_1^{\mathsf{rch}}(\emptyset) := \{s \in \mathcal{S} : \widetilde{\mathbb{P}}_1(s) \geq \sigma_{\mathsf{trunc}}\}$$

and let $\overline{M}_1(\Gamma)$ be the $(\mathcal{S}_1^{\mathsf{rch}}(\Gamma), \mathcal{S}, \ldots, \mathcal{S})$-truncation of M. Next, for each $h \in \{2, \ldots, H\}$, define

$$\mathcal{S}_h^{\mathsf{rch}}(\emptyset) := \{s \in \mathcal{S} : \max_{\pi \in \Pi} d_h^{\overline{M}_{h-1},\pi}(s) \geq \sigma_{\mathsf{trunc}}\}$$

and, if $\Gamma \neq \emptyset$,

$$\mathcal{S}_h^{\mathsf{rch}}(\Gamma) := \mathcal{S}_h^{\mathsf{rch}}(\emptyset) \cup \left\{ s \in \mathcal{S} : \mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Gamma)} d_h^{M,\pi}(s) \geq \sigma_{\mathsf{bkup}} \right\}.$$

Then we let $\overline{M}_h(\Gamma)$ be the $(\mathcal{S}_2^{\mathsf{rch}}(\Gamma), \ldots, \mathcal{S}_h^{\mathsf{rch}}(\Gamma), \mathcal{S}, \ldots, \mathcal{S})$-truncation of M. Finally, define $\overline{M}(\Gamma) := \overline{M}_H(\Gamma)$. As will be evident from the final parameter settings (in Algorithms 1 and 5), one should think of $\sigma_{\mathsf{bkup}} \ll \sigma_{\mathsf{trunc}}$.

Truncated policy covers. To avoid compounding errors in the analysis of algorithms that build policy covers layer-by-layer (such as PCE and PCR), it is convenient to work with *truncated* policy covers throughout the analysis (e.g. in the inductive hypothesis at layer h), and only convert to a standard policy cover (as in Eq. (1)) at the end of the analysis. Truncated policy covers are defined below:

Definition B.2 *Let* $\alpha \in (0, 1)$. *We say that a collection of policies* $\Psi \subset \Pi$ *is an* α-truncated policy cover (for M) at step $h \in [H]$ *if for all* $x \in \mathcal{X}$,

$$\frac{1}{|\Psi|} \sum_{\pi' \in \Psi} d_h^{M,\pi'}(x) \geq \alpha \cdot \max_{\pi \in \Pi} d_h^{\overline{M}(\emptyset),\pi}(x). \tag{5}$$

Definition B.3 *Let* $\alpha \in (0, 1)$. *We say that a collection of policies* $\Psi \subset \Pi$ *is an* α-truncated max policy cover at step $h \in [H]$ *if for all* $x \in \mathcal{X}$,

$$\max_{\pi' \in \Psi} d_h^{M,\pi'}(x) \geq \alpha \cdot \max_{\pi \in \Pi} d_h^{\overline{M}(\emptyset),\pi}(x). \tag{6}$$

B.1.1. USEFUL FACTS FOR TRUNCATED MDPS

Recall that we defined $\overline{M}(\Gamma)$ as the final truncated MDP in an iterative process that produced intermediate truncated MDPs $\overline{M}_1(\Gamma), \dots, \overline{M}_H(\Gamma)$. These intermediate MDPs are only used for certain technical lemmas about $\overline{M}(\Gamma)$; nonetheless it is useful to state several facts for later use.

Recall that $\overline{M}_h(\Gamma)$ is, essentially, truncated up to and including the transitions into step h. Fact B.4 formalizes the fact that $\overline{M}_h(\Gamma)$ agrees with $\overline{M}(\Gamma)$ up to step h, and Fact B.5 formalizes the fact that it agrees with M after step h.

Fact B.4 *For any finite set $\Gamma \subset \Pi$, integers $g, h \in [H]$ with $g \le h$, policy $\pi \in \Pi$, and state $s \in \overline{\mathcal{S}}$, it holds that $d_g^{\overline{M}(\Gamma),\pi}(s) = d_g^{\overline{M}_h(\Gamma),\pi}(s)$.*

Fact B.5 *For any finite set $\Gamma \subset \Pi$, integer $h \in \{1, \dots, H-1\}$, state $s \in \overline{\mathcal{S}}$, and action $a \in \mathcal{A}$, it holds that $\widetilde{\mathbb{P}}_{h+1}^{\overline{M}_h(\Gamma)}(\cdot|s,a) = \widetilde{\mathbb{P}}_{h+1}^{M}(\cdot|s,a)$.*

The most important property of a truncated MDP is that every state is either reachable (by some policy) with probability at least σ_{trunc}, or cannot be reached by any policy:

Fact B.6 *For every $h \in \{1, \dots, H\}$ and $s \in \mathcal{S}$, we have $\max_{\pi \in \Pi} d_h^{\overline{M}(\emptyset),\pi}(s) \ge \sigma_{\mathsf{trunc}}$ if $s \in \mathcal{S}_h^{\mathsf{rch}}(\emptyset)$ and $\max_{\pi\in\Pi} d_h^{\overline{M}(\emptyset),\pi}(s) = 0$ otherwise.*

The following facts formalize that $\overline{M}_h(\Gamma)$ is "more truncated" than $\overline{M}_{h-1}(\Gamma)$, and that $\overline{M}(\Gamma)$ is "less truncated" than $\overline{M}(\emptyset)$, respectively. The truncation process only takes mass away from non-terminal states.

Fact B.7 *Fix any finite set $\Gamma \subset \Pi$ and write $\overline{M}_0(\Gamma) := M$. For any $h \in [H]$, $s \in \mathcal{S}$, and $\pi \in \Pi$, it holds that $d_h^{\overline{M}_h(\Gamma),\pi}(s) \le d_h^{\overline{M}_{h-1}(\Gamma),\pi}(s)$. Hence, $\mathbb{E}^{\overline{M}_h(\Gamma),\pi}[f(x_h)] \le \mathbb{E}^{\overline{M}_{h-1}(\Gamma),\pi}[f(x_h)]$ for any $f : \overline{\mathcal{X}} \to \mathbb{R}_{\ge 0}$ with $f(\mathfrak{t}) = 0$.*

Fact B.8 *For every $h \in [H]$, $s \in \mathcal{S}$, $\Gamma \subset \Pi$, and $\pi \in \Pi$, we have $d_h^{\overline{M}(\emptyset),\pi}(s) \le d_h^{\overline{M}(\Gamma),\pi}(s) \le d_h^{M,\pi}(s)$.*

B.1.2. FACTS ABOUT TRUNCATED POLICY COVERS

In our algorithms, we will often roll in to some step h using a uniformly random policy from $\frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma))$. The following lemma gives useful properties of the resulting visitation distribution, under the assumption that Ψ_h is an α-truncated policy cover at step h.

Lemma B.9 *Fix $h \in \{1, \dots, H-1\}$ and $\alpha > 0$. Let $\Psi_h, \Gamma \subset \Pi$ be finite sets of policies. Suppose that Ψ_h is an α-truncated policy cover at step h (Definition B.2) for M. Then:*

1. *For all $s_h \in \mathcal{S}_h^{\mathsf{rch}}(\Gamma)$, it holds that*

$$\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} d_h^{M,\pi}(s_h) \ge \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2}.$$

2. *For any $f : \overline{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ with $f(\mathfrak{t}) = 0$, it holds that*

$$\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}[f(x_h)] \geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2} \max_{\pi \in \Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[f(x_h)].$$

3. *For all $s_{h+1} \in \mathcal{S}$,*

$$\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))\circ_h \mathrm{Unif}(\mathcal{A})} d^{M,\pi}_{h+1}(s_{h+1}) \geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2|\mathcal{A}|} \max_{\pi \in \Pi} d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1}).$$

Proof. We start with the first claim. Pick any $s_h \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma)$. If $s_h \in \mathcal{S}^{\mathsf{rch}}_h(\emptyset)$, then $\max_{\pi \in \Pi} d^{\overline{M}(\emptyset),\pi}_h(s) \geq \sigma_{\mathsf{trunc}}$ (Fact B.6). Thus,

$$\mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Psi_h)} d^{M,\pi}_h(s_h) \geq \alpha\sigma_{\mathsf{trunc}}$$

by Definition B.2. On the other hand, if $s_h \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma) \setminus \mathcal{S}^{\mathsf{rch}}_h(\emptyset)$, then $\mathbb{E}_{\pi \sim \mathrm{Unif}(\Gamma)} d^{M,\pi}_h(s) \geq \sigma_{\mathsf{bkup}}$ by definition of $\mathcal{S}^{\mathsf{rch}}_h(\Gamma)$. In either case, we have

$$\frac{1}{2} \mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Psi_h)} d^{M,\pi}_h(s_h) + \frac{1}{2} \mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Gamma)} d^{M,\pi}_h(s_h) \geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2}$$

which proves the first claim. Next, pick any $f : \overline{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ with $f(\mathfrak{t}) = 0$. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}[f(x_h)] \\
&= \sum_{s_h \in \mathcal{S}} \mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma)} d^{M,\pi}_h(s_h) \mathop{\mathbb{E}}_{x_h \sim \mathbb{O}_h(\cdot|s_h)}[f(x_h)] \\
&\geq \sum_{s_h \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma)} \mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma)} d^{M,\pi}_h(s_h) \mathop{\mathbb{E}}_{x_h \sim \mathbb{O}_h(\cdot|s_h)}[f(x_h)] \\
&\geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2} \sum_{s_h \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma)} \max_{\pi \in \Pi} d^{\overline{M}(\Gamma),\pi}_h(s_h) \mathop{\mathbb{E}}_{x_h \sim \mathbb{O}_h(\cdot|s_h)}[f(x_h)] \\
&\geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2} \max_{\pi \in \Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[f(x_h)]
\end{aligned}$$

where the first inequality uses non-negativity of f; the second inequality uses Item 1 together with the fact that $d^{\overline{M}(\Gamma),\pi}_h(s_h) \leq 1$ for all π, s_h; and the third inequality uses the fact that $f(\mathfrak{t}) = 0$. This proves the second claim. Finally, pick any $s_{h+1} \in \mathcal{S}$. If $s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$, then for any $\pi \in \Pi$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))\circ_h \mathrm{Unif}(\mathcal{A})} d^{M,\pi}_{h+1}(s_{h+1}) \\
&= \sum_{s_h \in \mathcal{S}} \left(\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma)} d^{M,\pi}_h(s_h) \right) \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1}|s_h, a) \\
&\geq \sum_{s_h \in \mathcal{S}} \left(\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma)} d^{M,\pi}_h(s_h) \right) \frac{1}{|\mathcal{A}|} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1}|s_h, \pi(s_h))
\end{aligned}$$

$$\geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2} \sum_{s_h \in \mathcal{S}_h^{\mathsf{rch}}(\Gamma)} d_h^{\overline{M}(\Gamma),\pi}(s_h) \frac{1}{|\mathcal{A}|} \widetilde{\mathbb{P}}_{h+1}^M(s_{h+1}|s_h, \pi(s_h))$$
$$= \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2|\mathcal{A}|} d_{h+1}^{\overline{M}(\Gamma),\pi}(s_{h+1})$$

where the second inequality uses Item 1 together with the fact that $d_h^{\overline{M}(\Gamma),\pi}(s_h) \leq 1$ for all s_h. This proves the third claim whenever $s_{h+1} \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)$. Moreover, if $s_{h+1} \notin \mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)$ then $\max_{\pi \in \Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s_{h+1}) = 0$, so the inequality is vacuously true. ■

B.1.3. ADDITIONAL FACTS ABOUT TRUNCATED MDPS

The following lemma will be important in the win/win analyses for PCE and PCR—specifically, it shows that if we find a policy that visits the terminal state $\mathfrak{t}$ with reasonable probability in $\overline{M}(\Gamma)$, then it must explore some hard-to-reach state at some earlier layer of M (which is a form of progress).

Lemma B.10 *Let $\pi \in \Pi$ and $\Gamma \subset \Pi$. For any $h \in [H]$, it holds that*

$$d_h^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) \leq \sum_{k=1}^{h} \sum_{s \in \mathcal{S} \setminus \mathcal{S}_k^{\mathsf{rch}}(\Gamma)} d_k^{M,\pi}(s).$$

Proof. Observe that $d_1^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) = \sum_{s \in \mathcal{S} \setminus \mathcal{S}_1^{\mathsf{rch}}(\Gamma)} \widetilde{\mathbb{P}}_1^M(s) = \sum_{s \in \mathcal{S} \setminus \mathcal{S}_1^{\mathsf{rch}}(\Gamma)} d_1^{M,\pi}(s)$ by construction. Moreover, for any $h \in \{2, \ldots, H\}$, we have

$$d_h^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) = d_{h-1}^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) + \sum_{s \in \mathcal{S}_{h-1}^{\mathsf{rch}}(\Gamma)} d_{h-1}^{\overline{M}(\Gamma),\pi}(s) \sum_{s' \in \mathcal{S} \setminus \mathcal{S}_h^{\mathsf{rch}}(\Gamma)} \widetilde{\mathbb{P}}_h^M(s'|s, \pi(s))$$
$$\leq d_{h-1}^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) + \sum_{s \in \mathcal{S}} d_{h-1}^{M,\pi}(s) \sum_{s' \in \mathcal{S} \setminus \mathcal{S}_h^{\mathsf{rch}}(\Gamma)} \widetilde{\mathbb{P}}_h^M(s'|s, \pi(s))$$
$$= d_{h-1}^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) + \sum_{s' \in \mathcal{S} \setminus \mathcal{S}_h^{\mathsf{rch}}(\Gamma)} d_h^{M,\pi}(s')$$

where the inequality uses Fact B.8. Inducting on h completes the proof. ■

The following lemma is in the analyses of PCE and PCR, to show that a truncated max-policy cover is also a $(1, \epsilon)$-policy cover in the sense of Eq. (1).

Lemma B.11 *Let $\pi \in \Pi$. For any $h \in [H]$ and $s \in \mathcal{S}$, it holds that*

$$d_h^{\overline{M}(\emptyset),\pi}(s) \geq d_h^{M,\pi}(s) - h|\mathcal{S}|\sigma_{\mathsf{trunc}}.$$

Proof. We have $d_h^{\overline{M}(\emptyset),\pi}(\mathfrak{t}) = d_h^{\overline{M}_h(\emptyset),\pi}(\mathfrak{t})$ by Fact B.4. Moreover

$$d_h^{\overline{M}_1(\emptyset),\pi}(\mathfrak{t}) = d_1^{\overline{M}_1(\emptyset),\pi}(\mathfrak{t}) = \sum_{z \in \mathcal{S} \setminus \mathcal{S}_1^{\mathsf{rch}}(\emptyset)} \widetilde{\mathbb{P}}_1(z) \leq |\mathcal{S}|\sigma_{\mathsf{trunc}}.$$

For any $2 \leq k \leq h$, we have

$$\begin{aligned}
d_h^{\overline{M}_k(\emptyset),\pi}(\mathfrak{t}) - d_h^{\overline{M}_{k-1}(\emptyset),\pi}(\mathfrak{t}) &= d_k^{\overline{M}_k(\emptyset),\pi}(\mathfrak{t}) - d_k^{\overline{M}_{k-1}(\emptyset),\pi}(\mathfrak{t}) \\
&= \mathbb{E}^{\overline{M}_{k-1}(\emptyset),\pi}\left[\widetilde{\mathbb{P}}_k^{\overline{M}_k(\emptyset)}(\mathfrak{t} \mid s_{k-1}, a_{k-1}) - \widetilde{\mathbb{P}}_k^{\overline{M}_{k-1}(\emptyset)}(\mathfrak{t} \mid s_{k-1}, a_{k-1})\right] \\
&= \mathbb{E}^{\overline{M}_{k-1}(\emptyset),\pi}\left[\sum_{z \in \mathcal{S} \setminus \mathcal{S}_k^{\mathsf{rch}}} \widetilde{\mathbb{P}}_k^{M}(z \mid s_{k-1}, a_{k-1})\right] \\
&= \sum_{z \in \mathcal{S} \setminus \mathcal{S}_k^{\mathsf{rch}}} d_{k-1}^{\overline{M}_{k-1}(\emptyset),\pi}(z) \\
&\leq |\mathcal{S}|\sigma_{\mathsf{trunc}}.
\end{aligned}$$

Therefore $d_h^{\overline{M}(\emptyset),\pi}(\mathfrak{t}) \leq h|\mathcal{S}|\sigma_{\mathsf{trunc}}$ by telescoping. Now for any $s \in \mathcal{S}$, this means that

$$\begin{aligned}
d_h^{\overline{M}(\emptyset),\pi}(s) &= 1 - d_h^{\overline{M}(\emptyset),\pi}(\mathfrak{t}) - \sum_{s' \in \mathcal{S} \setminus \{s\}} d_h^{\overline{M}(\emptyset),\pi}(s') \\
&\geq 1 - h|\mathcal{S}|\sigma_{\mathsf{trunc}} - \sum_{s' \in \mathcal{S} \setminus \{s\}} d_h^{M,\pi}(s') \\
&= d_h^{M,\pi}(s) - h|\mathcal{S}|\sigma_{\mathsf{trunc}}
\end{aligned}$$

where the inequality also uses Fact B.8. ■

The following result, which will be used in the analysis of PSDP (Lemma G.1), is a variant of the classical Performance Difference Lemma (Kakade and Langford, 2002); see also Golowich et al. (2024b).

Lemma B.12 (Performance Difference Lemma for truncated MDPs) *Fix a finite set $\Gamma \subset \Pi$ and $k \in \{1, \dots, H-1\}$. Let $R : \overline{\mathcal{X}} \to [0,1]$ be a function with $R(\mathfrak{t}) = 0$. Then for any policies $\pi, \pi^\star \in \Pi$ it holds that*

$$\mathbb{E}^{\overline{M}(\Gamma),\pi}[R(x_{k+1})] - \mathbb{E}^{M,\pi}[R(x_{k+1})] \leq \sum_{h=1}^{k} \mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[Q_h^{M,\pi,\mathbf{r}}(x_h, a_h) - V_h^{M,\pi,\mathbf{r}}(x_h)]$$

where $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_H)$ is defined by $\mathbf{r}_{k+1}(x,a) = R(x)$ and $\mathbf{r}_h(x,a) = 0$ for all $h \neq k+1$, and we have defined $Q_h^{M,\pi,\mathbf{r}}(\mathfrak{t}, a) := 0$ and $V_h^{M,\pi,\mathbf{r}}(\mathfrak{t}) := 0$ for all $h \in [H]$ and $a \in \mathcal{A}$.

Proof. Observe that for any $h \in \{1, \dots, k\}$, $x_h \in \overline{\mathcal{X}}$, $a_h \in \mathcal{A}$, we have

$$Q_h^{M,\pi,\mathbf{r}}(x_h, a_h) = \mathop{\mathbb{E}}_{x_{h+1} \sim \mathbb{P}_{h+1}^{M}(\cdot|x_h,a_h)}[V_{h+1}^{M,\pi,\mathbf{r}}(x_{h+1})] = \mathop{\mathbb{E}}_{x_{h+1} \sim \mathbb{P}_{h+1}^{\overline{M}_h(\Gamma)}(\cdot|x_h,a_h)}[V_{h+1}^{M,\pi,\mathbf{r}}(x_{h+1})] \tag{7}$$

by Fact B.5. For notation convenience, write $\overline{M}_0(\Gamma) := M$. Now we have

$$\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[R(x_{k+1})] - \mathbb{E}^{M,\pi}[R(x_{h+1})]$$

$$
\begin{aligned}
&= \mathbb{E}^{\overline{M}_{k+1}(\Gamma),\pi^\star}[Q_{k+1}^{M,\pi,\mathbf{r}}(x_{k+1},a_{k+1})] - \mathbb{E}^{M,\pi^\star}[V_1^{M,\pi,\mathbf{r}}(x_1)] \\
&= \mathbb{E}^{\overline{M}_{k+1}(\Gamma),\pi^\star}[Q_{k+1}^{M,\pi,\mathbf{r}}(x_{k+1},a_{k+1})] - \mathbb{E}^{M,\pi^\star}[V_1^{M,\pi,\mathbf{r}}(x_1)] + \sum_{h=1}^{k} \mathbb{E}^{\overline{M}_h(\Gamma),\pi^\star}[Q_h^{M,\pi,\mathbf{r}}(x_h,a_h) - V_{h+1}^{M,\pi,\mathbf{r}}(x_{h+1})] \\
&= \sum_{h=1}^{k+1} \left(\mathbb{E}^{\overline{M}_h(\Gamma),\pi^\star}[Q_h^{M,\pi,\mathbf{r}}(x_h,a_h)] - \mathbb{E}^{\overline{M}_{h-1}(\Gamma),\pi^\star}[V_h^{M,\pi,\mathbf{r}}(x_h)] \right) \\
&\le \sum_{h=1}^{k+1} \left(\mathbb{E}^{\overline{M}_h(\Gamma),\pi^\star}[Q_h^{M,\pi,\mathbf{r}}(x_h,a_h)] - \mathbb{E}^{\overline{M}_h(\Gamma),\pi^\star}[V_h^{M,\pi,\mathbf{r}}(x_h)] \right)
\end{aligned}
$$

where the first equality is by Fact B.4 and the fact that $R(\mathfrak{t}) = 0$, the second equality is by Eq. (7), and the inequality is by Fact B.7 (along with the fact that $V_h^{M,\pi,\mathbf{r}} \ge 0$ and $V_h^{M,\pi,\mathbf{r}}(\mathfrak{t}) = 0$). Finally, observe that $Q_{k+1}^{M,\pi,\mathbf{r}}(x_{k+1},a_{k+1}) = R(x_{k+1}) = V_{k+1}^{M,\pi,\mathbf{r}}(x_{k+1})$ for any $x_{k+1} \in \overline{\mathcal{X}}$, so the final term in the summation vanishes. We conclude that

$$
\begin{aligned}
&\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[R(x_{k+1})] - \mathbb{E}^{M,\pi}[R(x_{h+1})] \\
&\le \sum_{h=1}^{k} \left(\mathbb{E}^{\overline{M}_h(\Gamma),\pi^\star}[Q_h^{M,\pi,\mathbf{r}}(x_h,a_h)] - \mathbb{E}^{\overline{M}_h(\Gamma),\pi^\star}[V_h^{M,\pi,\mathbf{r}}(x_h)] \right) \\
&= \sum_{h=1}^{k} \left(\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[Q_h^{M,\pi,\mathbf{r}}(x_h,a_h)] - \mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[V_h^{M,\pi,\mathbf{r}}(x_h)] \right)
\end{aligned}
$$

where the final equality uses Fact B.4. ■

Appendix C. Proof of Theorem 3.1

In this section we prove that for any concept class Φ, there is a reduction from reward-free RL (Definition 2.1) in the episodic access model to two-context regression (Definition 2.4). The formal statement is provided below.

Theorem C.1 (General version of Theorem 3.1) *There is a constant $C_{C.1} > 0$ and an algorithm* PCE *so that the following holds. Let $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ be any concept class, and let* Reg *be a N_{reg}-efficient two-context regression oracle for Φ. Then* $\mathsf{PCE}(\mathsf{Reg}, N_{\mathsf{reg}}, |\mathcal{S}|, \cdot)$ *is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free RL algorithm for Φ in the episodic access model, with:*

- $K_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \leq H^2|\mathcal{S}|^2$
- $N_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \leq \left(\frac{H|\mathcal{A}||\mathcal{S}|}{\epsilon\delta}\right)^{C_{C.1}} N_{\mathsf{reg}}\left(\left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{C.1}}, \left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{C.1}}\right)$.

Moreover, the oracle time complexity of PCE *is at most* $\left(\frac{H|\mathcal{A}||\mathcal{S}|}{\epsilon\delta}\right)^{C_{C.1}} N_{\mathsf{reg}}\left(\left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{C.1}}, \left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{C.1}}\right)$.

In particular, Theorem 3.1 follows from Theorem C.1 by substituting $N_{\mathsf{reg}}(\epsilon, \delta) := N^\circ_{\mathsf{reg}}/(\epsilon\delta)^{C_{\mathsf{reg}}}$ into the above bounds. Henceforth, fix a concept class Φ, a N_{reg}-efficient two-context regression oracle Reg, and a Φ-decodable block MDP M with horizon H, action set $\mathcal{A}$, and unknown decoding function $\phi^\star \in \Phi$. We also define truncations of M (see Section B.1), with the parameters $\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}} > 0$ as defined in Algorithm 1.

In Section C.1, we give pseudocode and an overview of PCE and its main subroutine EPCE. In Section C.2, we formally analyze EPCE. In Section C.3, we formally analyze PCE, completing the proof of Theorem C.1 (and hence Theorem 3.1).

C.1. PCE Pseudocode and Overview

We start by giving an overview of the algorithm PCE (Algorithm 1). The main subroutine of this algorithm is EPCE (Algorithm 2), which is used to extend a set of policy covers from layers $1, \dots, h$ to layer $h+1$. We describe EPCE first, and then explain how it fits into PCE. As discussed in Section 3, PCE is a direct extension (and, from the perspective of oracles, a simplification) of HOMER (Misra et al., 2020); we highlight relevant differences in the overview below.

C.1.1. EPCE: EXTENDING A POLICY COVER.

Algorithm overview. As shown in Algorithm 2, EPCE takes as input a two-context regression oracle Reg, a step $h \in [H]$, and a set of policy covers $\Psi_{1:h}$ (as well as several other inputs that will be discussed later as necessary—for now, consider the case $\Gamma = \emptyset$). The desired behavior of EPCE is that, if $\Psi_{1:h}$ are $(1, \epsilon)$-policy covers for layers $1, \dots, h$ of the MDP M (as defined in Eq. (1)), then the output Ψ_{h+1} should be a $(1, \epsilon)$-policy cover for layer $h+1$.

To this end, for each action $a \in \mathcal{A}$, the algorithm first uses Ψ_h to construct a two-context regression dataset $\mathcal{D}_a$, via a contrastive learning approach where datapoints (x_h, x_{h+1}) with label $y = 0$ are sampled independently, whereas datapoints with label $y = 1$ are sampled dependently according to the transition dynamics, i.e. $x_{h+1} \sim \mathbb{P}_{h+1}(\cdot \mid x_h, a)$ (in both cases, and throughout the rest of the algorithm, the algorithm rolls in to step h with a random policy from Ψ_h). The

Algorithm 1 PCE: Episodic Reward-free RL via Two-Context Regression

input: Two-context regression oracle Reg; efficiency function $N_{\mathsf{reg}} : (0,1) \times (0,1) \to \mathbb{N}$; # of latent states S; final error tolerance $\varepsilon_{\mathsf{final}} \in (0,1)$; failure parameter $\delta \in (0,1)$; horizon H; action set $\mathcal{A}$.

$N \leftarrow N_{\mathsf{reg}}\left(\left(\frac{\varepsilon_{\mathsf{final}}\delta}{H|\mathcal{A}|S}\right)^{C_{C.1}}, \left(\frac{\varepsilon_{\mathsf{final}}\delta}{H|\mathcal{A}|S}\right)^{C_{C.1}}\right)$.

$\sigma_{\mathsf{trunc}} \leftarrow \varepsilon_{\mathsf{final}}/(4+HS)$.

$R \leftarrow SH$, $\sigma_{\mathsf{bkup}} \leftarrow \frac{\sigma_{\mathsf{trunc}}^2}{RS^2H^2}$, $\alpha \leftarrow \frac{1-4\sigma_{\mathsf{trunc}}}{S}$, $m \leftarrow \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}\log(S/\delta)$, $n \leftarrow \frac{m|\mathcal{A}|}{\sigma_{\mathsf{trunc}}}$.

$\epsilon \leftarrow \min(\frac{\alpha^{32}\sigma_{\mathsf{trunc}}^{64}\sigma_{\mathsf{bkup}}^{32}}{96^{16}H^{16}S^{16}|\mathcal{A}|^{32}m^8}, \frac{\delta^4}{81n^4})$, $\gamma \leftarrow \epsilon^{1/16}$, $\gamma' \leftarrow 2\epsilon^{1/8}\sqrt{m|\mathcal{A}|}$.

$\Gamma^{(1)} \leftarrow \emptyset$.

for $1 \le r \le R$ **do**

  $\Psi_1^{(r)} := \{\pi_{\mathsf{unif}}\}$.

  for $1 \le h < H$ **do**

    $\Psi_{h+1}^{(r)} \leftarrow \mathsf{EPCE}(\mathtt{Reg}, h, \Psi_{1:h}^{(r)}, \Gamma, n, m, N, \gamma, \gamma')$. ▷ See Algorithm 2

  $\Gamma^{(r+1)} \leftarrow \Gamma^{(r)} \cup \bigcup_{h\in[H]:|\Psi_h^{(r)}|\le S} \Psi_h^{(r)}$.

return $\bigcup_{h\in[H]} \bigcup_{1\le r\le R:|\Psi_h^{(r)}|\le S} \Psi_h^{(r)}$.

algorithm then invokes Reg to compute a predictor $\widehat{f}_{h+1}(\cdot,\cdot;a)$. It can be checked that the Bayes predictor $\mathbb{E}[y \mid x_h, x_{h+1}]$ for this dataset is precisely the kinematics function from $f_{h+1}(\cdot,\cdot;a)$ from Eq. (2). By the Block MDP assumption, this function only depends on (x_h, x_{h+1}) through $\phi^\star(x_h)$ and $\phi^\star(x_{h+1})$ (and the distribution of (x_h, x_{h+1}) is $\phi^\star$-realizable), so it follows from the guarantee of Reg that $\widehat{f}_{h+1}(\cdot,\cdot;a)$ approximates the true kinematics function $f_{h+1}(\cdot,\cdot;a)$ with high probability.

Second, the algorithm samples m observations $x_h^{(1)}, \dots, x_h^{(m)}$ at step h, which will be used as "test observations" for evaluating $\widehat{f}_{h+1}$: the idea is that if

$$\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) \approx \widehat{f}_{h+1}(x_h^{(i)}, x'_{h+1}; a)$$

for some $x_{h+1}, x'_{h+1} \in \mathcal{X}$ and all $i \in [m]$ and $a \in \mathcal{A}$, then so long as these observations have appropriate coverage, in fact $\widehat{f}_{h+1}(x_h, x_{h+1}; a)$ should approximate $\widehat{f}_{h+1}(x_h, x'_{h+1}; a)$ for all $(x_h, a) \in \mathcal{X} \times \mathcal{A}$, i.e. x_{h+1} and x'_{h+1} should have approximately the same kinematics.

Third, the algorithm samples n observations $\overline{x}_{h+1}^{(1)}, \dots, \overline{x}_{h+1}^{(n)}$ at step $h+1$ (rolling in with a random policy from Ψ_h followed by a random action at step h). These observations will serve as candidate "cluster centers" for defining internal reward functions. In particular, for each $t \in [n]$, the reward function $\mathcal{R}^{(t)} : \mathcal{X} \to [0,1]$ is defined in Line 18 to be large precisely for those x_{h+1} satisfying

$$\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) \approx \widehat{f}_{h+1}(x_h^{(i)}, \overline{x}_{h+1}^{(t)}; a)$$

for all test observations $x_h^{(i)} \in \mathcal{X}$ and actions $a \in \mathcal{A}$.

Fourth, for each cluster center that has noticeably different kinematics from previous centers (measured in the same way as above), the algorithm invokes PSDP (Algorithm 7) to compute a policy $\widehat{\pi}^{(t)}$ that approximately maximizes the reward function $\mathcal{R}^{(t)}$. This policy is added to Ψ_{h+1}.

Algorithm 2 EPCE: Extend Policy Cover for Episodic RL

input: Two-context regression oracle Reg; step $h \in [H]$; policy covers $\Psi_{1:h}$; backup policy cover Γ; sample counts $n, m, N \in \mathbb{N}$; tolerances $\gamma, \gamma' \in (0,1)$.

$\mathtt{Reg}'(\cdot) \leftarrow \mathtt{OneTwo}(\mathtt{Reg}, \cdot)$ ▷ See Algorithm 10

for $a \in \mathcal{A}$ **do**

  $\mathcal{D}_a \leftarrow \emptyset$.

  for N times **do**

    Sample trajectory $(x_1, a_1, \ldots, x_h, a, x_{h+1}) \sim \frac{1}{2}\left(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)\right) \circ_h a$.

    Sample trajectory $(x'_1, a'_1, \ldots, x'_h, a'_h, x'_{h+1}) \sim \frac{1}{2}\left(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)\right) \circ_h \mathrm{Unif}(\mathcal{A})$.

    Draw $y \sim \mathrm{Ber}(1/2)$.

    If $y = 1$, update dataset: $\mathcal{D}_a \leftarrow \mathcal{D}_a \cup \{(x_h, x_{h+1}, y)\}$.

    If $y = 0$, update dataset: $\mathcal{D}_a \leftarrow \mathcal{D}_a \cup \{(x_h, x'_{h+1}, y)\}$.

  $\widehat{f}_{h+1}(\cdot, \cdot; a) \leftarrow \mathtt{Reg}(\mathcal{D}_a)$.

for $1 \le i \le m$ **do**

  Sample trajectory $(x_1, a_1, \ldots, x_h) \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma))$.

  Set $x_h^{(i)} := x_h$.

$\Psi_{h+1} \leftarrow \emptyset$, $\mathcal{T}_{\mathsf{clus}} \leftarrow \emptyset$.

for $1 \le t \le n$ **do**

  Draw $(x_1, a_1, \ldots, x_h, a_h, x_{h+1}) \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)) \circ_h \mathrm{Unif}(\mathcal{A})$ and set $\overline{x}_{h+1}^{(t)} := x_{h+1}$.

  Define $\mathcal{R}^{(t)} : \mathcal{X} \to [0,1]$ by

$$\mathcal{R}^{(t)}(x) := \max\left(0, 1 - \frac{\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{f}_{h+1}(x_h^{(i)}, \overline{x}_{h+1}^{(t)}; a) - \widehat{f}_{h+1}(x_h^{(i)}, x; a)|}{\gamma}\right).$$

  if $\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{f}_{h+1}(x_h^{(i)}, \overline{x}_{h+1}^{(t)}; a) - \widehat{f}_{h+1}(x_h^{(i)}, \overline{x}_{h+1}^{(t')}; a)| > \gamma'$ for all $t' \in \mathcal{T}_{\mathsf{clus}}$ **then**

    $\widehat{\pi}^{(t)} \leftarrow \mathtt{PSDP}(h, \mathtt{Reg}', \mathcal{R}^{(t)}, \Psi_{1:h}, \Gamma, N)$. ▷ See Algorithm 7

    Update $\Psi_{h+1} \leftarrow \Psi_{h+1} \cup \{\widehat{\pi}^{(t)}\}$.

    Update $\mathcal{T}_{\mathsf{clus}} \leftarrow \mathcal{T}_{\mathsf{clus}} \cup \{t\}$.

return Ψ_{h+1}.

Comparison with HOMER. The main differences between EPCE and the corresponding subroutine of HOMER arise in ensuring that EPCE is oracle-efficient with respect to two-context regression: in the first step, we perform an individual regression for each action a (whereas HOMER performs a single regression joint across all actions), since in our definition of two-context regression, the action is not a covariate. Also, HOMER uses an implementation of PSDP with a cost-sensitive classification oracle; it is unclear how to reduce this to two-context regression, so we instead use an implementation of

PSDP due to Mhammedi et al. (2023b)—see Algorithm 7. For this implementation, a *one*-context regression oracle suffices (Lemma G.1), and this oracle in turn can easily be implemented—via the reduction OneTwo (Algorithm 10)—with two-context regression.

Proof outline. Since EPCE is very similar (at a technical level) to the corresponding subroutine of HOMER, we do not belabor the details of the proof in this overview. The basic reason why optimizing the reward functions $\mathcal{R}^{(t)}$ is a good idea is the following. First, if two observations x_{h+1}, x'_{x+1} have the same latent state, then they have the same kinematics, i.e. $f_{h+1}(x_h, x_{h+1}; a) = f_{h+1}(x_h, x'_{h+1}; a)$ for all x, a. The converse is not necessarily true. However, something just as good is true: if two states x_{h+1}, x'_{h+1} have the same kinematics, then $d^{M,\pi}_{h+1}(x_{h+1}) = C \cdot d^{M,\pi}_{h+1}(x'_{h+1})$ for a constant C that may depend on x_{h+1} and x'_{h+1} but *does not depend on* π. In this sense, x_{h+1} and x'_{h+1} are "kinematically inseparable" (Misra et al., 2020). It follows that the policy that maximizes $d^{M,\pi}_{h+1}(x_{h+1}) + d^{M,\pi}_{h+1}(x'_{h+1})$, i.e. the probability of visiting one of these two observations, also maximizes the probability of visiting either individual observation. More generally, for any set of kinematically inseparable observations, it suffices to maximize a reward function that rewards visiting any of these observations.

Obviously, in the actual algorithm and actual reward functions there are statistical errors, but these can be handled under appropriate coverage conditions. Indeed, under a *reachability assumption* on the MDP, Misra et al. (2020) show that if $\Psi_{1:h}$ are $(1, \epsilon)$-policy covers for layers $1, \ldots, h$, then with high probability Ψ_{h+1} is a $(1, \epsilon)$-policy cover for layer $h + 1$. Thus, they can simply run EPCE iteratively from $h = 1, \ldots, H$ and produce a set of policy covers $\Psi_{1:H}$. However, we want to avoid making a reachability assumption, so we require a more sophisticated algorithm (which still uses EPCE as a subroutine)—this is precisely PCE.

We formally analyze EPCE in Section C.2; the main guarantee is Theorem C.2.

C.1.2. PCE: HANDLING REACHABILITY ISSUES VIA ITERATIVE DISCOVERY

Algorithm overview. The basic idea of PCE (Algorithm 1) is to essentially put an outer loop around the entire HOMER/EPCE algorithm; this technique was previously used by Golowich et al. (2024b) for the same reason of handling reachability issues, in the context of sparse linear MDPs. In particular, PCE proceeds in $R = |\mathcal{S}| \cdot H$ rounds. In the first round, PCE runs EPCE iteratively from $h = 1, \ldots, H$ to construct a set of candidate policy covers $\Psi^{(1)}_{1:H}$. It then adds all of these computed policies to a "backup policy cover" $\Gamma^{(2)}$, and passes $\Gamma^{(2)}$ to EPCE in the next round. The outputs of EPCE are added to $\Gamma^{(3)}$, which is the backup policy cover for round $r = 3$, and so forth. The final output of PCE is the union of all candidate policy covers (of bounded size) that were computed in all rounds.

Proof outline. We sketch why the outer loop in PCE is needed to avoid a reachability assumption (and why it works). Reachability assumptions are common in theoretical reinforcement learning (Du et al., 2019; Misra et al., 2020), but are often an artifact of the analysis. That is, they can often be avoided—without changing the algorithm—by analyzing *truncated MDPs/policies* (Golowich et al., 2022; Mhammedi et al., 2023a), which essentially avoid issues of compounding errors on hard-to-reach states by truncating away such states. This approach would work in our setting if the reward functions $\mathcal{R}^{(t)}$ were exact, and hence accurate on all states. However, in EPCE/HOMER, the reward functions are *learned*, so they could be inaccurate on hard-to-reach states (in Lemma C.8, notice that the error bound holds in expectation over the truncated MDP $\overline{M}(\Gamma)$, which doesn't include the hard to reach states from the actual MDP M). This means that the policies computed by PSDP could obtain erroneously high reward without actually being optimal for the "ideal" reward functions.

The key idea is that the above pathology only occurs if one of the policies computed by PSDP "discovers" a state that the algorithm previously was unable to cover. This is the motivation for rerunning the entire algorithm with these policies mixed into all data collection procedures via the backup policy cover Γ. By a win/win argument (Golowich et al., 2024b), after at most $R = H|\mathcal{S}|$ rounds, there will be some round r in which the algorithm does not discover any new states; it can be shown that the sets $\Psi^{(r)}_{1:H}$ constructed in this round are indeed policy covers. We discuss this win/win argument in more detail later.

We formally analyze PCE (and thereby prove Theorem C.1) in Section C.3.

C.2. Analysis of EPCE (Algorithm 2)

We now prove the following guarantee for EPCE (Algorithm 2). Recall that we have fixed a concept class Φ, a N_{reg}-efficient two-context regression oracle Reg, and a Φ-decodable block MDP M with horizon H, action set $\mathcal{A}$, and unknown decoding function $\phi^\star \in \Phi$. We have also defined truncations of M (see Section B.1), with the parameters $\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}} > 0$ as defined in Algorithm 1.

Theorem C.2 *Let $h \in \{1, \dots, H-1\}$. Let $\epsilon, \delta, \alpha > 0$ and $m, n, N \in \mathbb{N}$. Let $\Gamma \subset \Pi$ be a finite set of policies. Suppose that $\Psi_{1:h}$ are α-truncated policy covers (Definition B.2) for M at steps $1, \dots, h$. Suppose that $m \geq \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} \log(|\mathcal{S}|/\delta)$, $n \geq \frac{2|\mathcal{A}|}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})\sigma_{\mathsf{trunc}}} \log(|\mathcal{S}|/\delta)$, $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, and*

$$\epsilon^{1/16} \leq \frac{\alpha^2 \sigma_{\mathsf{trunc}}^4 \sigma_{\mathsf{bkup}}^2}{96H|\mathcal{S}||\mathcal{A}|^2\sqrt{m}}. \tag{8}$$

Set $\gamma := \epsilon^{1/16}$ and $\gamma' := 2\epsilon^{1/8}\sqrt{m|\mathcal{A}|}$ and let Ψ_{h+1} denote the output of EPCE with inputs Reg, $h, \Psi_{1:h}, \Gamma, n, m, N, \gamma, \gamma'$. Then with probability at least $1 - (2 + |\mathcal{A}| + H|\mathcal{A}|n)\delta - m|\mathcal{A}|\epsilon^{1/2} - n\epsilon^{1/4}$, the following two properties hold:

- *$|\Psi_{h+1}| \leq |\mathcal{S}|$.*
- *Either Ψ_{h+1} is a $(1 - 4\sigma_{\mathsf{trunc}})$-truncated max policy cover (Definition B.3) for M at step $h+1$, or $\max_{\pi \in \Psi_{h+1}} d^{\overline{M}(\Gamma),\pi}_{h+1}(\mathfrak{t}) \geq \sigma_{\mathsf{trunc}}^2$.*

Informally, $\overline{M}(\Gamma)$ refers to a *truncation* of the MDP M in which all latent states that (a) cannot be reached by any policy with probability at least σ_{trunc}, and (b) cannot be reached by a uniformly random policy in Γ with probability at least $\sigma_{\mathsf{bkup}} \ll \sigma_{\mathsf{trunc}}$, are "truncated away" to an artificial terminal state $\mathfrak{t}$. A truncated policy cover (Definition B.2) is essentially a set of policies Ψ so that for each latent state s that *can* be reached with probability at least σ_{trunc}, a *uniformly random* policy from Ψ covers s. A truncated max-policy cover (Definition B.3) is essentially a set of policies so that for each latent state s that *can* be reached with probability at least σ_{trunc}, the *best* policy from Ψ covers s. See Section B.1 for the formal definition of $\overline{M}(\Gamma)$ and the truncated policy covers.

With this notation, Theorem C.2 essentially asserts that if $\Psi_{1:h}$ are policy covers at steps $1, \dots, h$, then the output of EPCE is either a policy cover at step $h+1$, or else contains some policy π that visits the terminal state in $\overline{M}(\Gamma)$ with non-trivial probability. In this second case, it can be shown (via Lemma B.10 and the definition of $\mathcal{S}^{\mathsf{rch}}_k(\Gamma)$) that π visits some latent state in M that was previously nearly-unexplored by all policies in Γ. Hence, when PCE adds π to the backup policy cover $\Gamma^{(r+1)}$ in the next round, progress will have been made, and so this second case can only happen a bounded number of times—see the proof of Theorem C.1 in Section C.3.

Let us fix the inputs to EPCE: in addition to the two-context regression oracle Reg (Definition 2.4), we fix a layer $h \in [H-1]$, sets of policies $\Psi_1, \dots, \Psi_h$ and Γ, sample counts $n, m, N \in \mathbb{N}$, and tolerances $\gamma, \gamma' \in (0,1)$. As discussed above, the main idea of EPCE is to use the oracle to estimate the kinematics $f_{h+1} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ (defined informally in Eq. (2) and formally below), and then to apply the PSDP policy optimization method (Algorithm 7) on internal reward functions constructed by clustering the kinematics.

Definition C.3 *For any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, define*

$$f_{h+1}(s, s'; a) := \frac{\widetilde{\mathbb{P}}^M_{h+1}(s' \mid s, a)}{\widetilde{\mathbb{P}}^M_{h+1}(s' \mid s, a) + F_{h+1}(s')}$$

where

$$F_{h+1}(s') := \mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)) \circ_h \mathrm{Unif}(\mathcal{A})} \mathbb{E}^{M,\pi}[\widetilde{\mathbb{P}}^M_{h+1}(s' \mid s_h, a_h)].$$

The main technical lemmas in the analysis are (1) Lemma C.5, which shows that the regression problem solved in Line 11 is a realizable instance of two-context regression, and that the resulting estimator $\widehat{f}_{h+1}$ is therefore a good estimate of f_{h+1}; and (2) Lemma C.8, which shows that if $\widehat{f}_{h+1}$ is close to f_{h+1} then for any reachable latent state $s^\star$, there is some internal reward function $\mathcal{R}^{(t)}$ computed by EPCE that approximately optimizes for visiting state $s^\star$. The following notation will be useful:

Definition C.4 *Let $\mu_{h+1}(a) \in \Delta(\mathcal{X} \times \mathcal{X})$ be the marginal distribution of (x_h, x_{h+1}) where (x_h, x_{h+1}, y) is the first element of $\mathcal{D}_a$. Let $\beta_h, \beta_{h+1} \in \Delta(\mathcal{S})$ be the marginal distributions of s_h and s_{h+1} respectively, for a trajectory $(s_1, x_1, a_1, \dots, s_{h+1}, x_{h+1}) \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)) \circ_h \mathrm{Unif}(\mathcal{A})$.*

In words, β_h is the visitation distribution at step h of a uniformly random policy $\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma))$, and β_{h+1} is the visitation distribution at step $h+1$ obtained by sampling from $x_h \sim \beta_h$, $a_h \sim \mathrm{Unif}(\mathcal{A})$, and $x_{h+1} \sim \mathbb{P}^M_{h+1}(\cdot \mid x_h, a_h)$.

We now prove that $\widehat{f}_{h+1}$ approximates the true kinematics f_{h+1} with high probability. This requires checking that each dataset $\mathcal{D}_a$ constructed by EPCE satisfies the necessary realizability assumptions, specified in Definition 2.4, with respect to Φ. Item 1 is then a direct consequence of the guarantee of Reg (together with a union bound over actions). Item 2 is a useful consequence, which asserts that if we plug in the m "test observations" $x_h^{(1)}, \dots, x_h^{(m)}$ sampled by EPCE, and all $|\mathcal{A}|$ actions, the resulting $m|\mathcal{A}|$-dimensional vector $\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a)_{i,a}$ is close to the corresponding "true" vector $f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a)_{i,a}$ on average over x_{h+1}. This will be needed in Lemma C.8 since ultimately the reward functions $\mathcal{R}^{(t)}$ are constructed by clustering these vectors.

Lemma C.5 *Let $\epsilon, \delta, \delta' \in (0,1)$. Suppose that Reg is an N_{reg}-efficient two-context regression oracle for Φ, and $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$. Then:*

1. *With probability at least $1 - \delta|\mathcal{A}|$, it holds that for all $a \in \mathcal{A}$,*

$$\mathop{\mathbb{E}}_{(x_h, x_{h+1}) \sim \widetilde{\mathbb{O}}_h \beta_h \times \widetilde{\mathbb{O}}_{h+1} \beta_{h+1}} \left(\widehat{f}_{h+1}(x_h, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h), \phi^\star(x_{h+1}); a) \right)^2 \leq 2\epsilon.$$

2. *With probability at least $1 - \delta|\mathcal{A}| - \delta' \cdot m|\mathcal{A}|$, it holds that*

$$\mathop{\mathbb{E}}_{x_{h+1} \sim \widetilde{\mathbb{O}}_{h+1} \beta_{h+1}} \max_{(i,a) \in [m] \times \mathcal{A}} \left(\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a) \right)^2 \leq \frac{2\epsilon m|\mathcal{A}|}{\delta'}.$$

Proof. To prove the first claim, fix any $a \in \mathcal{A}$. The dataset $\mathcal{D}_a$ constructed by Algorithm 2 consists of N independent and identically distributed tuples $(x_h^{(i)}, x_{h+1}^{(i)}, y^{(i)})$. Fix any $i \in [N]$ and let $\mu_0 \in \Delta(\mathcal{X} \times \mathcal{X})$ be the probability density function of $(x_h^{(i)}, x_{h+1}^{(i)})$ conditioned on $y^{(i)} = 0$. Similarly, let $\mu_1 \in \Delta(\mathcal{X} \times \mathcal{X})$ be the probability density function of $(x_h^{(i)}, x_{h+1}^{(i)})$ conditioned on $y^{(i)} = 1$. For any $x, x' \in \mathcal{X}$,

$$\mu_0(x, x') = \beta_h(\phi^\star(x)) \widetilde{\mathbb{O}}_h(x \mid \phi^\star(x)) \mathop{\mathbb{E}}_{(s, \tilde{a}) \sim \beta_h \times \mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid s, \tilde{a}) \widetilde{\mathbb{O}}_{h+1}(x' \mid \phi^\star(x')) \quad (9)$$

and

$$\mu_1(x, x') = \beta_h(\phi^\star(x)) \widetilde{\mathbb{O}}_h(x \mid \phi^\star(x)) \widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid \phi^\star(x), a) \widetilde{\mathbb{O}}_{h+1}(x' \mid \phi^\star(x')).$$

The unconditional probability density function of $(x_h^{(i)}, x_{h+1}^{(i)})$ is therefore $\mu \in \Delta(\mathcal{X} \times \mathcal{X})$ defined as

$$\begin{aligned}
&\mu(x, x') \\
&= \frac{\mu_0(x, x') + \mu_1(x, x')}{2} \\
&= \frac{\beta_h(\phi^\star(x)) \widetilde{\mathbb{O}}_h(x \mid \phi^\star(x))}{2} \Bigg(\widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid \phi^\star(x), a) \widetilde{\mathbb{O}}_{h+1}(x' \mid \phi^\star(x')) \\
&\qquad\qquad + \mathop{\mathbb{E}}_{(s, \tilde{a}) \sim \beta_h \times \mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid s, \tilde{a}) \widetilde{\mathbb{O}}_{h+1}(x' \mid \phi^\star(x')) \Bigg) \\
&= \frac{\beta_h(\phi^\star(x)) \widetilde{\mathbb{O}}_h(x \mid \phi^\star(x))}{2} \left(\widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid \phi^\star(x), a) + \mathop{\mathbb{E}}_{(s, \tilde{a}) \sim \beta_h \times \mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid s, \tilde{a}) \right) \\
&\qquad\qquad \cdot \widetilde{\mathbb{O}}_{h+1}(x' \mid \phi^\star(x')).
\end{aligned}$$

From this expression it is clear that, for any $s \in \mathcal{S}$, conditioned on the event $\phi^\star(x) = s$, $x_h^{(i)}$ and $x_{h+1}^{(i)}$ are independent. Moreover, for any $s \in \mathcal{S}$, conditioned on the event $\phi^\star(x') = s$, $x_h^{(i)}$ and $x_{h+1}^{(i)}$ are independent. We conclude that μ is $\phi^\star$-realizable (Definition 2.3). Next, for any $x, x' \in \mathcal{X}$, note that

$$\begin{aligned}
\mathbb{E}[y^{(i)} \mid x_h^{(i)} = x, x_{h+1}^{(i)} = x'] &= \frac{\mu_1(x, x')}{\mu_0(x, x') + \mu_1(x, x')} \\
&= \frac{\widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid \phi^\star(x), a)}{\widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid \phi^\star(x), a) + \mathbb{E}_{(s, \tilde{a}) \sim \beta_h \times \mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}_{h+1}^M(\phi^\star(x') \mid s, \tilde{a})}
\end{aligned}$$

$$= f_{h+1}(\phi^\star(x), \phi^\star(x'); a)$$

by Definition C.3. Hence, we can apply the guarantee of Reg (Definition 2.4) with distribution μ and ground truth predictor f_{h+1}. We get that with probability at least $1 - \delta$,

$$\mathbb{E}_{(x_h, x_{h+1}) \sim \mu} \left(\widehat{f}_{h+1}(x_h, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h), \phi^\star(x_{h+1}); a) \right)^2 \leq \epsilon.$$

But $\mu(x, x') \geq \frac{1}{2}\mu_0(x, x') = \frac{1}{2}\mathbb{O}_h\beta_h \times \mathbb{O}_{h+1}\beta_{h+1}$ by Eq. (9) and definition of β_{h+1}. The first claim of the lemma statement follows.

In the event that the first claim holds, for each $i \in [m]$ and $a \in \mathcal{A}$, since $x_h^{(i)}$ has distribution $\mathbb{O}_h\beta_h$, Markov's inequality gives that with probability at least $1 - \delta'$,

$$\mathbb{E}_{x_{h+1} \sim \mathbb{O}_{h+1}\beta_{h+1}} \left(\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a) \right)^2 \leq \frac{2\epsilon}{\delta'}.$$

By a union bound, we have with probability at least $1 - \delta' m|\mathcal{A}|$ that

$$\sum_{(i,a) \in [m] \times \mathcal{A}} \mathbb{E}_{x_{h+1} \sim \mathbb{O}_{h+1}\beta_{h+1}} \left(\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a) \right)^2 \leq \frac{2\epsilon m|\mathcal{A}|}{\delta'}.$$

Exchanging the summation and expectation completes the proof of the second claim. ■

To prove Lemma C.8, we need the following preparatory results.

Lemma C.6 *Let $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Then*

$$1 - f_{h+1}(s, s'; a) \geq \frac{\beta_h(s)}{2|\mathcal{A}|}.$$

Proof. Observe that

$$\begin{aligned}
F_{h+1}(s') &= \mathbb{E}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)) \circ_h \mathrm{Unif}(\mathcal{A})} \mathbb{E}^{M,\pi}[\widetilde{\mathbb{P}}_{h+1}^M(s' \mid s_h, a_h)] \\
&\geq \frac{1}{|\mathcal{A}|} \mathbb{E}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}[\widetilde{\mathbb{P}}_{h+1}^M(s' \mid s_h, a)] \\
&= \frac{1}{|\mathcal{A}|} \sum_{s_h \in \mathcal{S}} \beta_h(s_h) \widetilde{\mathbb{P}}_{h+1}^M(s' \mid s_h, a) \\
&\geq \frac{\beta_h(s)}{|\mathcal{A}|} \widetilde{\mathbb{P}}_{h+1}^M(s' \mid s, a).
\end{aligned}$$

It follows that

$$\begin{aligned}
1 - f_{h+1}(s, s'; a) &= \frac{F_{h+1}(s')}{\widetilde{\mathbb{P}}_{h+1}^M(s' \mid s, a) + F_{h+1}(s')} \\
&\geq \frac{\beta_h(s)}{|\mathcal{A}|} f_{h+1}(s', s; a).
\end{aligned}$$

If $f_{h+1}(s', s; a) \geq 1/2$ then the claim follows; otherwise $1 - f_{h+1}(s', s; a) \geq 1/2 \geq \beta_h(s)/(2|\mathcal{A}|)$ as well. ■

Lemma C.7 *Let $x, y \in [0,1)$. Then*

$$\left|\frac{x}{1-x} - \frac{y}{1-y}\right| \le \frac{2|x-y|}{\min\{1-x, 1-y\}^2}.$$

Proof. Define $g : [0,1) \to \mathbb{R}$ by $g(x) = x/(1-x)$. Then

$$g'(x) = \frac{1}{1-x} + \frac{x}{(1-x)^2} \le \frac{2}{(1-x)^2}.$$

Hence, $g'(w) \le 2/\min\{1-x, 1-y\}^2$ for all w in the interval between x and y. The lemma follows. ■

The following key lemma states for any reachable latent state $s^\star$, there is some reward function $\mathcal{R}^{(t)}$ so that, under any policy π, the expected value of π under this reward function is roughly proportional to the visitation probability of $s^\star$. Intuitively, this is important because optimizing with respect to $\mathcal{R}^{(t)}$ then approximately optimizes the probability of reaching $s^\star$. The key to the proof is Eq. (13), which shows that for any other state s which has roughly the same *kinematics* as $s^\star$, the probability that any policy π visits s is proportional to the probability that π visits $s^\star$ (where the constant of proportionality may depend on s and $s^\star$ but not π).

Lemma C.8 *Let $\epsilon_{\mathsf{reg}} > 0$. Condition on $\widehat{f}_{h+1}$ and $x_h^{(1)}, \dots, x_h^{(m)}$, and suppose that*

$$\mathop{\mathbb{E}}_{x_{h+1} \sim \mathbb{O}_{h+1}\beta_{h+1}} \max_{(i,a) \in [m] \times \mathcal{A}} \left(\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a)\right)^2 \le \epsilon_{\mathsf{reg}}. \tag{10}$$

Suppose that for each $s \in \mathcal{S}_h^{\mathsf{rch}}(\Gamma)$ there is some $i \in [m]$ with $\phi^\star(x^{(i)}) = s$. Suppose that $\beta_{h+1}(s) \ge \tilde{\alpha} \cdot \max_{\pi \in \Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s)$ for all $s \in \mathcal{S}$. Then for any $t \in [n]$ and $s^\star \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)$, there is some $K = K(\Gamma, s^\star) \ge 1$ such that

$$\begin{aligned}
&\max_{\pi \in \Pi} \left|\mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] - K \cdot d_{h+1}^{\overline{M}(\Gamma),\pi}(s^\star)\right| \\
&\le \frac{8\gamma|\mathcal{S}||\mathcal{A}|^2}{\min_{s \in \mathcal{S}_h^{\mathsf{rch}}(\Gamma)} \beta_h(s)^2} + \frac{\sqrt{\epsilon_{\mathsf{reg}}}}{\tilde{\alpha}\gamma} + \frac{\max_{(i,a)\in[m]\times\mathcal{A}} \left|\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), s^\star; a)\right|}{\gamma}
\end{aligned}$$

where for notational convenience we take $\mathcal{R}^{(t)}(\mathfrak{t}) := 0$.

Proof. First, observe that by the lemma assumption that $\beta_{h+1}(s) \ge \tilde{\alpha} \cdot \max_{\pi \in \Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s)$ for all $s \in \mathcal{S}$, it follows that

$$(\mathbb{O}_{h+1}\beta_{h+1})(x) \ge \tilde{\alpha} \cdot \max_{\pi \in \Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(x) \tag{11}$$

for all $x \in \mathcal{X}$. Now fix $\pi \in \Pi$. By definition of $\mathcal{R}^{(t)}$ (Line 18), we have

$$\mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] = \mathbb{E}^{\overline{M}(\Gamma),\pi}\left[g\left(\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - \widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a)\right|\right)\mathbb{1}[s_{h+1} \in \mathcal{S}]\right]$$

where $g(z) := \max(0, 1 - z/\gamma)$. For any $s, s' \in \mathcal{S}$, define $\Delta(s, s') := \max_{i,a} |f(\phi^\star(x_h^{(i)}), s; a) - f(\phi^\star(x_h^{(i)}), s'; a)|$. Then define

$$W^\pi := \mathbb{E}^{\overline{M}(\Gamma),\pi}[g(\Delta(s_{h+1}, s^\star))\mathbb{1}[s_{h+1} \in \mathcal{S}]].$$

Then

$$\begin{aligned}
&\left|\mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] - W^\pi\right| \\
&\leq \frac{1}{\gamma}\mathbb{E}^{\overline{M}(\Gamma),\pi}\left[\left|\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - \widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a)\right| - \Delta(s_{h+1}, s^\star)\right| \mathbb{1}[s_{h+1} \in \mathcal{S}]\right] \\
&\leq \frac{1}{\gamma}\mathbb{E}^{\overline{M}(\Gamma),\pi}\left[\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), s_{h+1}; a)\right| \mathbb{1}[s_{h+1} \in \mathcal{S}]\right] \\
&\quad + \frac{\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), s^\star; a)\right|}{\gamma} \\
&\leq \frac{1}{\tilde{\alpha}\gamma}\mathop{\mathbb{E}}_{x_{h+1}\sim\mathbb{O}_{h+1}\beta_{h+1}}\left[\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a)\right|\right] \\
&\quad + \frac{\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), s^\star; a)\right|}{\gamma} \\
&\leq \frac{\sqrt{\epsilon_{\mathsf{reg}}}}{\tilde{\alpha}\gamma} + \frac{\max_{i,a}\left|\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), s^\star; a)\right|}{\gamma},
\end{aligned} \tag{12}$$

where the first inequality uses the fact that g is $1/\gamma$-Lipschitz; the second inequality is by definition of $\Delta(s_{h+1}, s^\star)$ and the triangle inequality; the third inequality is by Eq. (11); and the fourth inequality is by Eq. (10).

Next, for any fixed $s_{h+1} \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)$, recall that $\widetilde{\mathbb{P}}_{h+1}^{\overline{M}(\Gamma)}(s_{h+1} \mid s_h, a_h) = \widetilde{\mathbb{P}}_{h+1}^{M}(s_{h+1} \mid s_h, a_h)$ and $\widetilde{\mathbb{P}}_{h+1}^{\overline{M}(\Gamma)}(s_{h+1} \mid \mathfrak{t}, a_h) = 0$ for all $s_h \in \mathcal{S}$, $a_h \in \mathcal{A}$ (by Definition B.1); hence,

$$\begin{aligned}
d_{h+1}^{\overline{M}(\Gamma),\pi}(s_{h+1}) &= \sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\overline{M}(\Gamma),\pi}(s_h, a_h)\widetilde{\mathbb{P}}_{h+1}^{M}(s_{h+1} \mid s_h, a_h) \\
&= \left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\overline{M}(\Gamma),\pi}(s_h, a_h)\frac{f_{h+1}(s_h, s_{h+1}; a_h)}{1 - f_{h+1}(s_h, s_{h+1}; a_h)}\right) F_{h+1}(s_{h+1})
\end{aligned} \tag{13}$$

where the second equality is by definition of f_{h+1}, F_{h+1} (Definition C.3). Substituting Eq. (13) into the definition of W^π and using the fact that $d_{h+1}^{\overline{M}(\Gamma),\pi}(s_{h+1}) = 0$ for all $s_{h+1} \in \mathcal{S} \setminus \mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)$, we get that

$$W^\pi = \sum_{s_{h+1}\in\mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)} \left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\overline{M}(\Gamma),\pi}(s_h, a_h)\frac{f_{h+1}(s_h, s_{h+1}; a_h)}{1 - f_{h+1}(s_h, s_{h+1}; a_h)}\right) F_{h+1}(s_{h+1})g(\Delta(s_{h+1}, s^\star)).$$

Define

$$\widetilde{W}^\pi := \left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\overline{M}(\Gamma),\pi}(s_h, a_h)\frac{f_{h+1}(s_h, s^\star; a_h)}{1 - f_{h+1}(s_h, s^\star; a_h)}\right) \sum_{s_{h+1}\in\mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)} F_{h+1}(s_{h+1})g(\Delta(s_{h+1}, s^\star)).$$

Then

$$
\begin{aligned}
&|W^\pi - \widetilde{W}^\pi| \\
&\le \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} \left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\overline{M}(\Gamma),\pi}(s_h,a_h) \left| \frac{f_{h+1}(s_h,s_{h+1};a_h)}{1-f_{h+1}(s_h,s_{h+1};a_h)} - \frac{f_{h+1}(s_h,s^\star;a_h)}{1-f_{h+1}(s_h,s^\star;a_h)} \right| \right) \\
&\qquad\qquad \cdot F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&\le \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} \left(\max_{(s_h,a_h)\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)\times\mathcal{A}} \left| \frac{f_{h+1}(s_h,s_{h+1};a_h)}{1-f_{h+1}(s_h,s_{h+1};a_h)} - \frac{f_{h+1}(s_h,s^\star;a_h)}{1-f_{h+1}(s_h,s^\star;a_h)} \right| \right) F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&\le 2 \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} \left(\frac{\max_{(s_h,a_h)\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)\times\mathcal{A}} |f_{h+1}(s_h,s_{h+1};a_h) - f_{h+1}(s_h,s^\star;a_h)|}{\min_{(s_h,a_h,s)\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)\times\mathcal{A}\times\mathcal{S}} (1-f_{h+1}(s_h,s;a_h))^2} \right) F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&\le 2 \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} \left(\frac{\Delta(s_{h+1},s^\star)}{\min_{(s_h,a_h,s)\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)\times\mathcal{A}\times\mathcal{S}} (1-f_{h+1}(s_h,s;a_h))^2} \right) F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&\le 8|\mathcal{A}|^2 \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} \left(\frac{\Delta(s_{h+1},s^\star)}{\min_{s\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)} \beta_h(s)^2} \right) F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&\le \frac{8\gamma|\mathcal{S}||\mathcal{A}|^2}{\min_{s\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)} \beta_h(s)^2} \qquad (14)
\end{aligned}
$$

where the second inequality uses the fact that $d_h^{\overline{M}(\Gamma),\pi}(s_h) = 0$ for all $s_h \in \mathcal{S} \setminus \mathcal{S}^{\mathsf{rch}}_h(\Gamma)$; the third inequality uses Lemma C.7; the fourth inequality uses the definition of $\Delta(s_{h+1}, s^\star)$ together with the lemma assumption that $\mathcal{S}^{\mathsf{rch}}_h(\Gamma) \subseteq \{\phi^\star(x_h^{(i)}) : i \in [m]\}$; the fifth inequality uses Lemma C.6; and the sixth inequality uses the fact that $F_{h+1}(s_{h+1}) \le 1$ for all $s_{h+1} \in \mathcal{S}$ (Definition C.3) together with the bound $z \cdot g(z) \le \gamma$ for all $z \ge 0$.

Finally, note that by definition of $\widetilde{W}^\pi$ and Eq. (13) applied to $s^\star$, we have

$$
\begin{aligned}
\widetilde{W}^\pi &= \left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\overline{M}(\Gamma),\pi}(s_h,a_h) \frac{f_{h+1}(s_h,s^\star;a_h)}{1-f_{h+1}(s_h,s^\star;a_h)} \right) \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&= \frac{d_{h+1}^{\overline{M}(\Gamma),\pi}(s^\star)}{F_{h+1}(s^\star)} \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star)) \\
&= K(\Gamma,s^\star) \cdot d_{h+1}^{\overline{M}(\Gamma),\pi}(s^\star) \qquad (15)
\end{aligned}
$$

where

$$
K(\Gamma,s^\star) := \frac{\sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} F_{h+1}(s_{h+1}) g(\Delta(s_{h+1},s^\star))}{F_{h+1}(s^\star)}.
$$

Since $s^\star \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$ and $g(\Delta(s^\star,s^\star)) = g(0) = 1$, we have $K(\Gamma,s^\star) \ge 1$. Combining Eqs. (12), (14) and (15) yields the lemma claim. ■

We now prove Theorem C.2 by combining Lemmas C.5 and C.8 with a standard guarantee for PSDP (Lemma G.1). We use the assumption that $\Psi_1, \dots, \Psi_h$ are truncateed policy covers for steps $1, \dots, h$ to show that the hypotheses for Lemma C.8 are satisfied with high probability. We remark that PSDP naturally uses a one-context regression oracle and not two; this is why we invoke it with a one-context regression oracle Reg' obtained by reduction to two-context regression (Algorithm 10; see Proposition G.5).

Proof of Theorem C.2. For any fixed $s \in \mathcal{S}_h^{\mathsf{rch}}(\Gamma)$ and $i \in [m]$, since $\phi^\star(x_h^{(i)}) \sim \beta_h$ and Ψ_h is an α-truncated policy cover for M at step h, we have by Item 1 of Lemma B.9 that

$$\Pr[\phi^\star(x_h^{(i)}) = s] \geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2}.$$

Let $\mathcal{E}_1$ be the event that for each $s \in \mathcal{S}_h^{\mathsf{rch}}(\Gamma)$ there is some $i \in [m]$ with $\phi^\star(x^{(i)}) = s$. Then

$$\Pr[\mathcal{E}_1] \geq 1 - |\mathcal{S}| \left(1 - \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2}\right)^m \geq 1 - \delta$$

by the theorem assumption that $m \geq \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} \log(|\mathcal{S}|/\delta)$. Henceforth condition on $x_h^{(1)}, \dots, x_h^{(m)}$ and suppose that $\mathcal{E}_1$ holds.

Let $\mathcal{E}_2$ be the event that

$$\mathop{\mathbb{E}}_{x_{h+1} \sim \widetilde{\mathbb{O}}_{h+1}\beta_{h+1}} \max_{(i,a) \in [m] \times \mathcal{A}} \left(\widehat{f}_{h+1}(x_h^{(i)}, x_{h+1}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(x_{h+1}); a)\right)^2 \leq 2\sqrt{\epsilon} m |\mathcal{A}|.$$

By the theorem assumptions that Reg is an N_{reg}-efficient two-context regression oracle for Φ, and $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, we may apply Item 2 of Lemma C.5 (with parameter $\delta' := \sqrt{\epsilon}$) to get that $\mathcal{E}_2$ occurs with probability at least $1 - \delta|\mathcal{A}| - \sqrt{\epsilon} m |\mathcal{A}|$ over the randomness of $(\mathcal{D}_a)_{a \in \mathcal{A}}$ and Reg. Condition on this randomness (which determines $\widehat{f}_{h+1}$) and suppose that $\mathcal{E}_2$ holds.

For each $t \in [n]$, let $\mathcal{E}_3^t$ be the event that

$$\max_{(i,a) \in [m] \times \mathcal{A}} \left(\widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t)}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), \phi^\star(\bar{x}_{h+1}^{(t)}); a)\right)^2 \leq 2\epsilon^{1/4} m |\mathcal{A}|.$$

Since $\bar{x}_{h+1}^{(t)} \sim \mathbb{O}_{h+1}\beta_{h+1}$, we have by Markov's inequality and $\mathcal{E}_2$ that $\Pr[\neg\mathcal{E}_3^t] \leq \epsilon^{1/4}$. Define $\mathcal{E}_3 := \bigcap_{t=1}^n \mathcal{E}_3^t$. By the union bound, $\mathcal{E}_3$ occurs with probability at least $1 - n\epsilon^{1/4}$ over the randomness of $\bar{x}_{h+1}^{(1)}, \dots, \bar{x}_{h+1}^{(n)}$.

Also, for each $s \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\Gamma)$, let $t(s) \in [1, n] \cup \{\infty\}$ be the infimum over t such that $\phi^\star(\bar{x}_{h+1}^{(t)}) = s$, and let $\mathcal{E}_4^s$ be the event that $t(s) < \infty$. For each $t \in [1, n]$, since $\phi^\star(\bar{x}_{h+1}^{(t)})$ has distribution β_{h+1}, we have

$$\begin{aligned}
\Pr[\phi^\star(\bar{x}_{h+1}^{(t)}) = s] &\geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2|\mathcal{A}|} \max_{\pi \in \Pi} d_{h+1}^{\overline{M}(\Gamma), \pi}(s) \\
&\geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2|\mathcal{A}|} \max_{\pi \in \Pi} d_{h+1}^{\overline{M}(\emptyset), \pi}(s) \\
&\geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})\sigma_{\mathsf{trunc}}}{2|\mathcal{A}|}
\end{aligned}$$

where the first inequality is by Item 3 of Lemma B.9; the second inequality is by Fact B.8; and the third inequality is by Fact B.6. Thus, $\Pr[\neg\mathcal{E}_4^s] \leq (1 - \frac{\min(\alpha\sigma_{\mathsf{trunc}},\sigma_{\mathsf{bkup}})\sigma_{\mathsf{trunc}}}{2|\mathcal{A}|})^n \leq \delta/|\mathcal{S}|$ for any fixed $s \in \mathcal{S}^{\mathsf{rch}}_{h+1}$, by the theorem assumption that $n \geq \frac{2|\mathcal{A}|}{\min(\alpha\sigma_{\mathsf{trunc}},\sigma_{\mathsf{bkup}})\sigma_{\mathsf{trunc}}} \log(|\mathcal{S}|/\delta)$. Define $\mathcal{E}_4 := \bigcap_{s\in\mathcal{S}^{\mathsf{rch}}_{h+1}} \mathcal{E}_4^s$. By the union bound, $\mathcal{E}_4$ occurs with probability at least $1-\delta$ over the randomness of $\bar{x}^{(1)}_{h+1}, \dots, \bar{x}^{(n)}_{h+1}$. Condition on $\bar{x}^{(1)}_{h+1}, \dots, \bar{x}^{(n)}_{h+1}$ and suppose that $\mathcal{E}_3 \cap \mathcal{E}_4$ holds.

Finally, for each $t \in \mathcal{T}_{\mathsf{clus}}$ let $\mathcal{E}_5^t$ be the event that

$$\mathbb{E}^{M,\widehat{\pi}^{(t)}}[\mathcal{R}^{(t)}(x_{h+1})] \geq \max_{\pi\in\Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] - \frac{4H\sqrt{|\mathcal{A}|\epsilon}}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}$$

where $\widehat{\pi}^{(t)}$ is defined on Line 20 of Algorithm 2. Since $\mathtt{Reg}$ is an N_{reg}-efficient two-context regression oracle, Proposition G.5 implies that $\mathtt{Reg}'$ is an N_{reg}-efficient one-context regression oracle. Hence, by the theorem assumption on $\Psi_{1:h}$, and the fact that $N \geq N_{\mathsf{reg}}(\epsilon,\delta)$, Lemma G.1 gives that $\Pr[\mathcal{E}_5^t] \geq 1 - H|\mathcal{A}|\delta$, so $\Pr[\mathcal{E}_5] \geq 1 - HAn\delta$ where $\mathcal{E}_5 := \cap_{t\in[n]}\mathcal{E}_5^t$. Condition on $\mathcal{E}_5$. We have now restricted to an event of total probability at least $1-(2+|\mathcal{A}|+H|\mathcal{A}|n)\delta - m|\mathcal{A}|\epsilon^{1/2} - n\epsilon^{1/4}$; we argue that in this event, the properties claimed in the theorem statement hold.

Size of Ψ_{h+1}. First, we argue that $|\Psi_{h+1}| \leq |\mathcal{S}|$. Indeed, suppose that there are $t, t' \in \mathcal{T}_{\mathsf{clus}}$ with $t < t'$ and $\phi^\star(\bar{x}^{(t)}_{h+1}) = \phi^\star(\bar{x}^{(t')}_{h+1})$. By Line 19 of Algorithm 2, and by choice of γ', we know that

$$\max_{(i,a)\in[m]\times\mathcal{A}} \left| \widehat{f}_{h+1}(x^{(i)}_h, \bar{x}^{(t)}_{h+1}; a) - \widehat{f}_{h+1}(x^{(i)}_h, \bar{x}^{(t')}_{h+1}; a) \right| > \gamma' = 4\epsilon^{1/8}\sqrt{m|\mathcal{A}|}.$$

But this contradicts $\mathcal{E}_3$ (in particular, the bounds implied by $\mathcal{E}_3^t$ and $\mathcal{E}_3^{t'}$ together with the triangle inequality and the fact that $\phi^\star(\bar{x}^{(t)}_{h+1}) = \phi^\star(\bar{x}^{(t')}_{h+1})$). We conclude that indeed $|\Psi_{h+1}| \leq |\mathcal{S}|$.

Coverage of Ψ_{h+1}. It remains to prove the second property of the theorem statement. Fix $s \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\emptyset)$. Define $t^\star(s) := t(s)$ if $t(s) \in \mathcal{T}_{\mathsf{clus}}$. Otherwise, let $t^\star(s)$ be the minimal $t \in \mathcal{T}_{\mathsf{clus}}$ such that $\max_{i\in[m]} |\mathcal{R}_i(\bar{x}^{(t(s))}_{h+1}) - \mathcal{R}_i(\bar{x}^{(t)}_{h+1})| \leq \gamma'$ (which exists by Line 19). In either case, we know that $t^\star(s) \in \mathcal{T}_{\mathsf{clus}}$, and

$$\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{f}_{h+1}(x^{(i)}_h, \bar{x}^{(t(s))}_{h+1}; a) - \widehat{f}_{h+1}(x^{(i)}_h, \bar{x}^{(t^\star(s))}_{h+1}; a)| \leq \gamma'. \tag{16}$$

By $\mathcal{E}_1$, $\mathcal{E}_2$, and Item 3 of Lemma B.9, we may apply Lemma C.8 with target state $s^\star := s$, index $t := t^\star(s)$, and parameters $\epsilon_{\mathsf{reg}} := 2\sqrt{\epsilon}m|\mathcal{A}|$ and $\tilde{\alpha} := \frac{\min(\alpha\sigma_{\mathsf{trunc}},\sigma_{\mathsf{bkup}})}{2|\mathcal{A}|}$. We get that there is some $K(\Gamma, s) \geq 1$ such that

$$\begin{aligned}
&\max_{\pi\in\Pi} \left| \mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - K(\Gamma,s)\cdot d^{\overline{M}(\Gamma),\pi}_{h+1}(s) \right| \\
&\leq \frac{8\gamma|\mathcal{S}||\mathcal{A}|^2}{\min_{s'\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)} \beta_h(s')^2} + \frac{\sqrt{\epsilon_{\mathsf{reg}}}}{\tilde{\alpha}\gamma} + \frac{\max_{(i,a)\in[m]\times\mathcal{A}} \left| \widehat{f}_{h+1}(x^{(i)}_h, \bar{x}^{(t^\star(s))}_{h+1}; a) - f_{h+1}(\phi^\star(x^{(i)}_h), s; a) \right|}{\gamma} \\
&\leq \frac{32\gamma|\mathcal{S}||\mathcal{A}|^2}{\min(\alpha^2\sigma^2_{\mathsf{trunc}}, \sigma^2_{\mathsf{bkup}})} + \frac{4\epsilon^{1/4}|\mathcal{A}|\sqrt{m|\mathcal{A}|}}{\gamma\cdot\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{\max_{(i,a)\in[m]\times\mathcal{A}} \left| \widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t^\star(s))}; a) - f_{h+1}(\phi^\star(x_h^{(i)}), s; a) \right|}{\gamma} \\
&\leq \frac{32\gamma|\mathcal{S}||\mathcal{A}|^2}{\min(\alpha^2\sigma_{\mathsf{trunc}}^2, \sigma_{\mathsf{bkup}}^2)} + \frac{4\epsilon^{1/4}|\mathcal{A}|\sqrt{m|\mathcal{A}|}}{\gamma \cdot \min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} \\
&\quad + \frac{\epsilon^{1/8}\sqrt{2m|\mathcal{A}|} + \max_{(i,a)\in[m]\times\mathcal{A}} \left| \widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t^\star(s))}; a) - \widehat{f}_{h+1}(x_h^{(i)}, \bar{x}_{h+1}^{(t(s))}; a) \right|}{\gamma} \\
&\leq \frac{32\gamma|\mathcal{S}||\mathcal{A}|^2}{\min(\alpha^2\sigma_{\mathsf{trunc}}^2, \sigma_{\mathsf{bkup}}^2)} + \frac{4\epsilon^{1/4}|\mathcal{A}|\sqrt{m|\mathcal{A}|}}{\gamma \cdot \min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} + \frac{\epsilon^{1/8}\sqrt{2m|\mathcal{A}|} + \gamma'}{\gamma} \\
&\leq \frac{32\epsilon^{1/16}|\mathcal{S}||\mathcal{A}|^2}{\min(\alpha^2\sigma_{\mathsf{trunc}}^2, \sigma_{\mathsf{bkup}}^2)} + \frac{4\epsilon^{3/16}|\mathcal{A}|\sqrt{m|\mathcal{A}|}}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} + 6\epsilon^{1/16}\sqrt{m|\mathcal{A}|} \\
&\leq \sigma_{\mathsf{trunc}}^2,
\end{aligned}
\tag{17}
$$

where the second inequality is by Item 1 of Lemma B.9; the third inequality is by $\mathcal{E}_3$ and the fact that $s = \phi^\star(\bar{x}_{h+1}^{(t(s))})$; the fourth inequality is by Eq. (16); the fifth inequality is by choice of γ, γ'; and the final inequality is by Eq. (8). We now distinguish two cases:

Case I. Suppose that

$$
\mathbb{E}^{M,\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] \geq \mathbb{E}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] + \sigma_{\mathsf{trunc}}^2.
$$

Then

$$
\begin{aligned}
&d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(\mathfrak{t}) \\
&= \sum_{s\in\mathcal{S}} \left(d_{h+1}^{M,\widehat{\pi}^{(t^\star(s))}}(s) - d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(s) \right) \\
&\geq \sum_{s\in\mathcal{S}} \left(d_{h+1}^{M,\widehat{\pi}^{(t^\star(s))}}(s) - d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(s) \right) \mathop{\mathbb{E}}_{x\sim\mathbb{O}_{h+1}(\cdot|s)}[\mathcal{R}^{(t^\star(s))}(x)] \\
&\geq \sigma_{\mathsf{trunc}}^2
\end{aligned}
$$

where the equality is by the fact that $d_{h+1}^{M,\widehat{\pi}^{(t^\star(s))}}(\cdot)$ is a distribution supported on $\mathcal{S}$; the first inequality uses Fact B.8 and the fact that $\mathcal{R}^{(t^\star(s))}(x) \leq 1$ for all $x \in \mathcal{X}$. Thus $\max_{\pi\in\Psi_{h+1}} d^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) \geq \sigma_{\mathsf{trunc}}^2$, so the second property of the theorem statement is satisfied.

Case II. Suppose that

$$
\mathbb{E}^{M,\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] < \mathbb{E}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] + \sigma_{\mathsf{trunc}}^2. \tag{18}
$$

Now,

$$
\begin{aligned}
K(\Gamma, s) \cdot d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{t^\star(s)}}(s) &\geq \mathbb{E}^{\overline{M}(\Gamma),\widehat{\pi}^{t^\star(s)}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - \sigma_{\mathsf{trunc}}^2 \\
&\geq \mathbb{E}^{M,\widehat{\pi}^{t^\star(s)}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - 2\sigma_{\mathsf{trunc}}^2
\end{aligned}
$$

$$
\begin{aligned}
&\geq \max_{\pi\in\Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - 3\sigma_{\mathsf{trunc}}^2 \\
&\geq \max_{\pi\in\Pi} K(\Gamma,s)\cdot d_{h+1}^{\overline{M}(\Gamma),\pi}(s) - 4\sigma_{\mathsf{trunc}}^2 \\
&\geq K(\Gamma,s)(1-4\sigma_{\mathsf{trunc}}) \max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s)
\end{aligned}
$$

where the first inequality is by Eq. (17); the second inequality is by Eq. (18); the third inequality is by $\mathcal{E}_5$ and Eq. (8); the fourth inequality is by Eq. (17); and the fifth inequality uses the fact that $\max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s) \geq \sigma_{\mathsf{trunc}}$ (Fact B.6 together with Fact B.8 and the fact that $s \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$) and the bound $K(\Gamma,s) \geq 1$ (Lemma C.8). Thus, since $\widehat{\pi}^{(t^\star(s))} \in \Psi_{h+1}$, we have

$$
\begin{aligned}
\max_{\pi\in\Psi_{h+1}} d_{h+1}^{M,\pi}(s) &\geq \max_{\pi\in\Psi_{h+1}} d_{h+1}^{\overline{M}(\Gamma),\pi}(s) \\
&\geq (1-4\sigma_{\mathsf{trunc}}) \max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s) \\
&\geq (1-4\sigma_{\mathsf{trunc}}) \max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\emptyset),\pi}(s) \qquad (19)
\end{aligned}
$$

by two applications of Fact B.8. Now recall that $s \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$ was arbitrary. Moreover, if $s \in \mathcal{S}\setminus\mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$ then $\max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\emptyset),\pi}(s) = 0$, so the inequality Eq. (19) still holds. We conclude that Ψ_{h+1} is a $(1-4\sigma_{\mathsf{trunc}})$-truncated max policy cover for M at step $h+1$ (Definition B.3), as needed. ■

C.3. Analysis of PCE (Algorithm 1)

We now complete the proof of Theorem C.1. As previously discussed, Theorem C.2 shows that in each round of PCE, either a set of policy covers was constructed, or a new state was discovered. The latter can happen at most $H|\mathcal{S}|-1$ times, so at least one round must construct a good set of policy covers. In the latter event, the union of all sets produced across all rounds is itself a good policy cover (so long as the individual sets have bounded size). Of course, since the guarantee of Theorem C.2 is probabilistic, some additional care is needed. We make the argument formal below.

Proof of Theorem C.1. Fix the remaining inputs $\varepsilon_{\mathsf{final}}, \delta > 0$ to $\mathsf{PCE}(\mathtt{Reg}, N_{\mathsf{reg}}, |\mathcal{S}|, \cdot)$. The oracle time complexity bound and bound on N_{RL} are clear from the parameter choices and pseudocode, so long as $C_{C.1}$ is a sufficiently large constant. Moreover, it is immediate from the algorithm description that $|\Psi| \leq HR|\mathcal{S}| \leq H^2|\mathcal{S}|^2$. In order to show that the algorithm is $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient, it remains to argue that with probability at least $1-\delta$, Eq. (1) holds for all $h \in [H]$ and $s \in \mathcal{S}$, with parameter $\varepsilon_{\mathsf{final}}$.

Recall that $\sigma_{\mathsf{trunc}} = \varepsilon_{\mathsf{final}}/(4+H|\mathcal{S}|)$. Fix some $1 \leq r \leq R$. For convenience, write $\alpha := \frac{1-4\sigma_{\mathsf{trunc}}}{|\mathcal{S}|}$. For each $h \in [H]$, let $\mathcal{E}_{h,r}$ be the event that $|\Psi_h^{(r)}| \leq |\mathcal{S}|$ and $\Psi_h^{(r)}$ is a $(1-4\sigma_{\mathsf{trunc}})$-truncated max policy cover for M at step h; let $\mathcal{F}_{h,r}$ be the event that $|\Psi_h^{(r)}| \leq |\mathcal{S}|$ and $\max_{\pi\in\Psi_h^{(r)}} d_h^{\overline{M}(\Gamma^{(r)}),\pi}(\mathfrak{t}) \geq \sigma_{\mathsf{trunc}}^2$. It's clear that $\Pr[\mathcal{E}_{1,r}] = 1$ (since $|\Psi_1^{(r)}| = 1$ and $d_1^{M,\pi_{\mathsf{unif}}}(s) = d_1^{M,\pi}(s)$ for all $s \in \mathcal{S}$ and $\pi \in \Pi$). Also, note that in the event $\mathcal{E}_{k,r}$, we have that $\Psi_k^{(r)}$ is an

α-truncated policy cover for M at step k. Thus, by Theorem C.2 and choice of parameters (so long as $C_{C.1}$ is a sufficiently large constant), we have for each $h \in \{2, \dots, H\}$ that

$$\Pr\left[\neg(\mathcal{E}_{h,r} \cup \mathcal{F}_{h,r}) \cap \bigcap_{1 \le k < h} \mathcal{E}_{k,r}\right] \le \Pr\left[\neg(\mathcal{E}_{h,r} \cup \mathcal{F}_{h,r}) \middle| \bigcap_{1 \le k < h} \mathcal{E}_{k,r}\right] \le \frac{\delta}{HR}. \tag{20}$$

In the event that the events $\mathcal{F}_{1,r}, \dots, \mathcal{F}_{H,r}, \bigcap_{h \in [H]} \mathcal{E}_{h,r}$ all fail, there is always some maximal $h \in [H]$ such that $\bigcap_{1 \le k \le h} \mathcal{E}_{k,r}$ holds (since $\mathcal{E}_{1,r}$ always holds); it must be that $1 \le h < H$, and $\mathcal{E}_{h+1,r}$ and $\mathcal{F}_{h+1,r}$ both fail. Thus,

$$\Pr\left[\left(\neg \bigcap_{h \in [H]} \mathcal{E}_{h,r}\right) \cap \bigcap_{h \in [H]} (\neg \mathcal{F}_{h,r})\right] \le \sum_{h=2}^{H} \Pr\left[\neg(\mathcal{E}_{h,r} \cup \mathcal{F}_{h,r}) \cap \bigcap_{1 \le k < h} \mathcal{E}_{k,r}\right] \le \frac{\delta}{R}$$

where the final inequality is by Eq. (20). Let $\mathcal{E}_r$ denote the complementary event (i.e. either $\bigcap_{h \in [H]} \mathcal{E}_{h,r}$ holds, or there is some $h \in [H]$ such that $\mathcal{F}_{h,r}$ holds), and let $\mathcal{E} := \bigcap_{1 \le r \le R} \mathcal{E}_r$; we have $\Pr[\mathcal{E}] \ge 1 - \delta$. We claim that PCE succeeds under event $\mathcal{E}$. Indeed, there are two cases to consider.

1. In the first case, there is some $r \in [R]$ such that $\bigcap_{h \in [H]} \mathcal{E}_{h,r}$ holds. Then for each $h \in [H]$, $|\Psi_h^{(r)}| \le |\mathcal{S}|$ and $\Psi_h^{(r)}$ is a $(1 - 4\sigma_{\mathsf{trunc}})$-truncated max policy cover for M at step h. Thus, $\bigcup_{1 \le r' \le R : |\Psi_h^{(r')}| \le |\mathcal{S}|} \Psi_h^{(r')} \subseteq \Psi$ is also a $(1 - 4\sigma_{\mathsf{trunc}})$-truncated max policy cover for M at step h. Therefore for each $h \in [H]$ and $s \in \mathcal{S}$, we have by Definition B.3 that

$$\begin{aligned} \max_{\pi' \in \Psi} d_h^{M,\pi'}(s) &\ge (1 - 4\sigma_{\mathsf{trunc}}) \max_{\pi \in \Pi} d_h^{\overline{M}(\emptyset),\pi}(s) \\ &\ge \max_{\pi \in \Pi} d_h^{\overline{M}(\emptyset),\pi}(s) - 4\sigma_{\mathsf{trunc}} \\ &\ge \max_{\pi \in \Pi} d_h^{M,\pi}(s) - (4 + H|\mathcal{S}|)\sigma_{\mathsf{trunc}} \end{aligned}$$

   where the final inequality is by Lemma B.11. Since $\sigma_{\mathsf{trunc}} = \varepsilon_{\mathsf{final}}/(4 + H|\mathcal{S}|)$, this bound suffices.

2. In the second case, for each $r \in [R]$, there is some $h \in [H]$ such that $\mathcal{F}_{h,r}$ holds. For each r, define

$$\mathcal{V}^{(r)} := \left\{(s,h) \in \mathcal{S} \times [H] : \max_{\pi \in \Gamma^{(r)}} d_h^{M,\pi}(s) \ge \frac{\sigma_{\mathsf{trunc}}^2}{|\mathcal{S}|H}\right\}.$$

   Fix any $r \in [R]$. By assumption, there is some $h \in [H]$ so that $\mathcal{F}_{h,r}$ holds. For this choice of h, we have by definition of $\mathcal{F}_{h,r}$ that

$$\max_{\pi \in \Psi_h^{(r)}} d_h^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) \ge \sigma_{\mathsf{trunc}}^2.$$

   By Lemma B.10, there is some $(s,k) \in (\mathcal{S} \setminus \mathcal{S}_k^{\mathsf{rch}}(\Gamma)) \times [h]$ such that

$$\max_{\pi \in \Psi_h^{(r)}} d_k^{M,\pi}(s) \ge \frac{\sigma_{\mathsf{trunc}}^2}{|\mathcal{S}|H}.$$

Thus $(s,k) \in \mathcal{V}^{(r+1)}$. Moreover, since $s \notin \mathcal{S}_k^{\mathsf{rch}}(\Gamma^{(r)})$, we have $\mathbb{E}_{\pi \sim \mathrm{Unif}(\Gamma^{(r)})} d_k^{M,\pi}(s) < \sigma_{\mathsf{bkup}}$. Using the fact that $|\Gamma^{(r)}| \leq RH|\mathcal{S}|$ and choice of σ_{bkup}, it follows that

$$\max_{\pi \in \Gamma^{(r)}} d_k^{M,\pi}(s) < RH|\mathcal{S}|\sigma_{\mathsf{bkup}} \leq \frac{\sigma_{\mathsf{trunc}}^2}{|\mathcal{S}|H}.$$

So $(s,k) \notin \mathcal{V}^{(r)}$. We conclude that $|\mathcal{V}^{(r+1)}| > |\mathcal{V}^{(r)}|$. Since this inequality holds for all $r \in [R]$ and $|\mathcal{V}^{(1)}| \geq H$, we get $|\mathcal{V}^{(R)}| \geq H + R - 1 > |\mathcal{S}|H$. Contradiction, so in fact this second case cannot occur.

This completes the proof. ■

Appendix D. Proof of Theorem 3.2

In this section we prove Theorem 3.2, restated below. This theorem asserts that for any *regular* (Definition 2.5) concept class Φ^{aug}, there is a reduction `RegToRL` (Algorithm 3) from two-context regression to reward-free episodic RL.

Theorem 3.2 *There is a constant $C_{3.2} > 0$ and an algorithm* `RegToRL` *(Algorithm 3 in Appendix D) so that the following holds. Let $\Phi^{\mathsf{aug}} \subseteq (\mathcal{X}^{\mathsf{aug}} \to \mathcal{S}^{\mathsf{aug}})$ be any regular concept class (Definition 2.5), let $N^{\circ}_{\mathsf{RL}}, C_{\mathsf{RL}} \in \mathbb{N}$, and let $\mathcal{O}$ be a $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free episodic RL oracle for Φ^{aug}, with $\max(N_{\mathsf{RL}}(\epsilon,\delta,H,A), K_{\mathsf{RL}}(\epsilon,\delta,H,A)) \leq N^{\circ}_{\mathsf{RL}} \cdot (AH/\epsilon\delta)^{C_{\mathsf{RL}}}$. Then* `RegToRL`$(\mathcal{O},\cdot)$ *is an N_{reg}-efficient two-context regression algorithm (Definition 2.4) for Φ^{aug} with $N_{\mathsf{reg}}(\epsilon,\delta) \leq N^{\circ}_{\mathsf{RL}}\,(|\mathcal{S}|/(\epsilon\delta))^{C_{3.2}\cdot C_{\mathsf{RL}}}$ and with oracle time complexity at most $N^{\circ}_{\mathsf{RL}}\,(|\mathcal{S}|/(\epsilon\delta))^{C_{3.2}\cdot C_{\mathsf{RL}}}$.*

Any regular concept class can be defined by augmenting some base class Φ as specified in Definition D.1. Accordingly, the main component of `RegToRL` is a reduction `TwoRed` (Algorithm 4) from two-context regression over Φ to reward-free episodic RL over Φ^{aug}. The full reduction `RegToRL` simply applies `TwoRed` in conjunction with a reduction `TwoAug` (Algorithm 12) from two-context regression over Φ^{aug} to two-context regression over Φ.

Henceforth, fix sets $\mathcal{S}, \mathcal{X}$ and a concept class $\Phi \subset (\mathcal{X} \to \mathcal{S})$. We may define an augmented concept class $\Phi^{\mathsf{aug}} \subseteq (\mathcal{X}^{\mathsf{aug}} \to \mathcal{S}^{\mathsf{aug}})$ as follows (in Section D.3 we will formally argue that any regular concept class can be expressed in this way):

Definition D.1 (Augmented concept class) *Define augmented state space $\mathcal{S}^{\mathsf{aug}} := \mathcal{S} \sqcup \{0,1\}$ and augmented observation space $\mathcal{X}^{\mathsf{aug}} := \mathcal{X} \sqcup \{0,1\}$. For each $\phi \in \Phi$ define $\operatorname{aug}(\phi) : \mathcal{X}^{\mathsf{aug}} \to \mathcal{S}^{\mathsf{aug}}$ by*

$$\operatorname{aug}(\phi)(x) := \begin{cases} \phi(x) & \text{if } x \in \mathcal{X} \\ x & \text{otherwise} \end{cases}.$$

Finally, define an extended function class $\Phi^{\mathsf{aug}} := \{\operatorname{aug}(\phi) : \phi \in \Phi\}$.

In Section D.1 we give pseudocode and an overview of `RegToRL` and its main subroutine `TwoRed`. In Section D.2 we formally analyze `TwoRed`. In Section D.3 we use this to analyze `RegToRL`, completing the proof of Theorem 3.2. We defer the analysis of `TwoAug` to Section G.2.5.

D.1. `RegToRL` Pseudocode and Overview

We start by giving a brief overview of the reduction `RegToRL` (Algorithm 3); for additional intuition, see also the discussion in Section 3.2. The main subroutine is `TwoRed` (Algorithm 4), which is used to reduce two-context regression over Φ to reward-free RL over Φ^{aug}. We start with an overview of this subroutine, and then briefly discuss `TwoAug` (Algorithm 12), the reduction from two-context regression over Φ^{aug} to two-context regression over Φ. The full reduction `RegToRL` is a direct combination of these two subroutines.

D.1.1. `TwoRed`: SIMULATING AN RL ORACLE FOR TWO-CONTEXT REGRESSION

Algorithm overview. As shown in Algorithm 4, `TwoRed` takes as input a reward-free episodic RL oracle for Φ^{aug}, a dataset $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$ for two-context regression over Φ, and accuracy

Algorithm 3 $\texttt{RegToRL}(\mathcal{O}, (x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: Two-context regression-to-RL reduction

input: Reward-free episodic RL oracle $\mathcal{O}$ for Φ; samples $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.
return:

$$\mathcal{R} \leftarrow \texttt{TwoAug}(\texttt{TwoRed}(\mathcal{O}, \cdot), (x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta).$$

Algorithm 4 $\texttt{TwoRed}(\mathcal{O}, (x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: Main subroutine in $\texttt{RegToRL}$

input: Reward-free episodic RL oracle $\mathcal{O}$ for Φ^{aug}; samples $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.
Set $\varepsilon_A := \frac{\epsilon}{2|\mathcal{S}|}$. Initialize $\mathcal{O}$ with $H = 2$ and $\mathcal{A} = \{0, \varepsilon_A, \dots, 1 - \varepsilon_A\}$ and tolerances $\frac{\epsilon^2}{4|\mathcal{S}|^2}, \delta/2$.
repeat
  Simulate episode i of interaction with $\mathcal{O}$ as follows: pass observation $x_1^{(i)}$ and receive action $a \in \mathcal{A}$. With probability $(a - y^{(i)})^2$, pass observation 0. Otherwise, pass observation $x_2^{(i)}$.
until $\mathcal{O}$ returns policy cover Ψ
for $\pi \in \Psi$ **do**
  $\mathcal{C}^\pi \leftarrow \emptyset$
  for $n/2 + 1 \le i \le n$ **do**
    Sample $z^{(i)} \sim \mathrm{Ber}((\pi(x_1^{(i)}) - y^{(i)})^2)$.
    $\mathcal{C}^\pi \leftarrow \mathcal{C}^\pi \cup \{(x_2^{(i)}, z^{(i)}\}$.
  Compute $\mathcal{R}^\pi \leftarrow \texttt{OneRed}(\mathcal{O}, \mathcal{C}^\pi)$. ▷ See Algorithm 8
return: predictor $\mathcal{R} : \mathcal{X} \times \mathcal{X} \to [0,1]$ defined as $\mathcal{R}(x_1, x_2) := (\arg\min_{\pi \in \Psi} \mathcal{R}^\pi(x_2))(x_1)$, where ties are broken in some canonical fashion.

parameters $\epsilon, \delta > 0$. The goal is to produce and estimate $\mathcal{R} : \mathcal{X} \times \mathcal{X} \to [0,1]$ of the Bayes optimal predictor $\mathbb{E}[y^{(i)} \mid x_1^{(i)}, x_2^{(i)}]$ for the dataset. To this end, the first step is to simulate the RL oracle on an MDP with horizon $H = 2$, initial observation space $\mathcal{X}$, final observation space $\mathcal{X} \sqcup \{0\}$, and action space $\mathcal{A} \subset [0,1]$. In particular, $\texttt{TwoRed}$ uses a new sample $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$ from the dataset for each episode of interaction—this is where it is crucial that the RL oracle does not have reset access. The first observation is $x_1^{(i)}$. When the oracle returns an action $a \in \mathcal{A} \subset [0,1]$, the second observation is $x_2^{(i)}$ with probability $1 - (a - y^{(i)})^2$ and 0 otherwise.

Eventually, the oracle produces a set of policies Ψ. Each policy is a map $\pi : \mathcal{X} \to [0,1]$. To "stitch" these into a single predictor on $\mathcal{X} \times \mathcal{X}$, $\texttt{TwoRed}$ estimates an error function for each. In particular, using fresh samples, the algorithm constructs a dataset consisting of samples $(x_2^{(i)}, z^{(i)})$ where $\mathbb{E}[z^{(i)} \mid x_2^{(i)}] = \mathbb{E}[(\pi(x_1^{(i)}) - y^{(i)})^2 \mid x_2^{(i)}]$ measures the error of π on $x_1^{(i)}$, conditional on $x_2^{(i)}$. Applying a one-context regression oracle to this dataset (via Algorithm 8, the reduction from one-context regression to reward-free RL) gives an estimated error function $\mathcal{R}^\pi : \mathcal{X} \to [0,1]$.

Finally, the predictor $\mathcal{R}$ output by $\texttt{TwoRed}$ is defined as follows. Given $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$, identify the policy $\pi \in \Psi$ that minimizes $\mathcal{R}^\pi(x_2)$, and return $\pi(x_1)$.

Proof outline. As discussed in Section 3.2, the basic idea for the analysis of TwoRed is as follows. Let M denote the simulated MDP, let $\phi^\star$ denote the decoding function for the dataset, and let $f : \mathcal{S} \times \mathcal{S} \to [0, 1]$ denote the true latent predictor for the dataset. It can be checked that M satisfies Φ^{aug}-decodability (Lemma D.4). For any latent state $s \in \mathcal{S}$, a policy π that approximately maximizes $d_2^{M,\pi}(s)$ must optimally "guess" $y^{(i)}$ conditioned on both $x_1^{(i)}$ and the event $\phi^\star(x_2^{(i)}) = s$; thus, as shown in Lemma D.7, such a π must approximately *minimize* the following loss:

$$L_s(\pi) := \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}} \left[\mathbb{1}[\phi^\star(x_2) = s](\pi(x_1) - f(\phi^\star(x_1), \phi^\star(x_2)))^2\right].$$

By the definition of reward-free RL, it follows that for each $s \in \mathcal{S}$, there exists some $\pi \in \Psi$ such that $L_s(\pi)$ is small (Lemma D.8), i.e. π is a good approximation of $x_1 \mapsto f(\phi^\star(x_1), s)$.

It remains to argue that the policy selection procedure in the definition of the final predictor $\mathcal{R}$ (i.e. picking the policy $\widehat{\pi}^{x_2}$ that minimizes the estimated error $\mathcal{R}^\pi(x_2)$) appropriately identifies *which* policy is good for a given covariate (x_1, x_2). Indeed, as shown in Lemma D.9, for fixed x_2, $\mathcal{R}^\pi(x_2)$ is monotonic in $L_{\phi^\star(x_2)}(\pi)$ as π varies. Thus, intuitively, the selection procedure makes sense. The remaining technical subtlety is that there is an apparent distributional mismatch. We know that the following loss is small, for each x_2 with $\phi^\star(x_2) = s$:

$$L_s(\widehat{\pi}^{x_2}) = \mathop{\mathbb{E}}_{(x_1',x_2')\sim\mathcal{D}} \left[\mathbb{1}[\phi^\star(x_2') = s](\widehat{\pi}^{x_2}(x_1') - f(\phi^\star(x_1'), \phi^\star(x_2')))^2\right].$$

However, to bound the squared error of $\mathcal{R}$, we would like to bound the following quantity:

$$\mathop{\mathbb{E}}_{(x_1',x_2')\sim\mathcal{D}} \left[\mathbb{1}[\phi^\star(x_2') = s](\widehat{\pi}^{x_2'}(x_1') - f(\phi^\star(x_1'), \phi^\star(x_2')))^2\right],$$

i.e. where context x_2' in the squared error term is the same as the context used to select the policy $\widehat{\pi}^{x_2'}$. However, it turns out that these two quantities can be related in expectation over x_2, using $\phi^\star$-realizability of the context distribution (Lemma D.10).

We formally analyze TwoRed in Section D.2; see Theorem D.2 for the formal guarantee.

D.1.2. TwoAug: AUGMENTING THE REGRESSION

Algorithm overview. As shown in Algorithm 12, TwoAug takes as input a two-context regression oracle for Φ, samples $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$, and tolerance parameters $\epsilon, \delta \in (0, 1/2)$. The goal is to solve two-context regression for Φ^{aug}. The basic idea is that since the extra states $\{0, 1\}$ are fully observed, regression onto these states is actually simpler than regression onto the hidden states. In more detail, consider all indices i for which $x_1^{(i)} = 0$ but $x_2^{(i)} \in \mathcal{X}$. We can define a reduced dataset consisting of the samples $(\overline{x}, x_2^{(i)})$ for each such i, where $\overline{x} \in \mathcal{X}$ is an arbitrary fixed observation. It's straightforward to check that this dataset is a valid two-context regression dataset for Φ, so the oracle yields a good predictor for the corresponding conditional distribution (as long as there is enough data). We can do a similar argument for all other cases (e.g. $x_1^{(i)} \in \mathcal{X}$ and $x_2^{(i)} = 1$, etc.), and it is straightforward to stitch together the predictors on subsets of $\mathcal{X}^{\mathsf{aug}} \times \mathcal{X}^{\mathsf{aug}}$ into a single predictor for the whole covariate space.

We defer the formal analysis of TwoAug to Section G.2.5; see Proposition G.8 for the formal guarantee.

D.2. Analysis of TwoRed (Algorithm 4)

The following theorem states our main guarantee for TwoRed. As discussed above, TwoRed uses the given dataset to simulate the RL oracle on a horizon-2 block MDP for which exploring a latent state s with near-maximal probability corresponds to learning the regression function on a subset of the covariate distribution determined by s. Once the oracle produces a policy cover Ψ, TwoRed performs one-context regression to learn a loss function $\mathcal{R}^\pi$ associated with each policy $\pi \in \Psi$, and then outputs a regressor obtained by stitching together the policies (according to whichever has the best loss on the given query).

Theorem D.2 *Suppose that $\mathcal{O}$ is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free episodic RL oracle (Definition 2.1) for Φ^{aug}. Then $\mathsf{TwoRed}(\mathcal{O}, \cdot)$ is a N_{reg}-efficient two-context regression algorithm (Definition 2.4) for Φ with*

$$N_{\mathsf{reg}}(\epsilon,\delta) := 2N_{\mathsf{RL}}\left(\frac{\epsilon^4}{16|\mathcal{S}|^4}, \frac{\delta}{4K'}, 2, \frac{4|\mathcal{S}|^2}{\epsilon^2}\right) + 64|\mathcal{S}|^8\epsilon^{-8}\log\left(4K'K_{\mathsf{RL}}\left(\frac{\epsilon^4}{16|\mathcal{S}|^4}, \frac{\delta}{4K'}, 2, \frac{4|\mathcal{S}|^2}{\epsilon^2}\right)/\delta\right)$$

where $K' = K_{\mathsf{RL}}(\frac{\epsilon^2}{4|\mathcal{S}|^2}, \frac{\delta}{2}, 2, \frac{2|\mathcal{S}|}{\epsilon})$.

To prove Theorem D.2, we fix $\phi^\star \in \Phi$, a $\phi^\star$-realizable distribution $\mathcal{D} \in \Delta(\mathcal{X}\times\mathcal{X})$, and a function $f : \mathcal{S}\times\mathcal{S} \to [0,1]$. For some $n \geq N_{\mathsf{reg}}(\epsilon,\delta)$, we let $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $(x_1^{(i)}, x_2^{(i)}) \sim \mathcal{D}$ and $y^{(i)} \sim \mathrm{Ber}(f(\phi^\star(x_1^{(i)}), \phi^\star(x_2^{(i)})))$. In the remainder of the section, we analyze the execution of $\mathsf{TwoRed}(\mathcal{O}, (x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$. Ultimately, we must show that with probability at least $1-\delta$, the circuit $\mathcal{R}$ produced by TwoRed satisfies

$$\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left(\mathcal{R}(x_1,x_2) - f(\phi^\star(x_1), \phi^\star(x_2))\right)^2 \leq \epsilon. \tag{21}$$

To begin, in Lemmas D.4 and D.5 we show that TwoRed is invoking the RL oracle $\mathcal{O}$ by simulating episodic access with the Φ^{aug}-decodable block MDP defined below:

Definition D.3 (Block MDP gadget for reduction) *Fix $\varepsilon_A \in (0,1)$. We define a block MDP M by*

$$M = M_{\phi^\star, f, \mathcal{D}, \varepsilon_A} := (H, \mathcal{S}^{\mathsf{aug}}, \mathcal{X}^{\mathsf{aug}}, \mathcal{A}, (\widetilde{\mathbb{P}}_h)_{h\in[2]}, (\widetilde{\mathbb{O}})_{h\in[2]}, \mathrm{aug}(\phi^\star))$$

where $H := 2$, the action space is $\mathcal{A} := \{0, \varepsilon_A, \dots, \varepsilon_A\lfloor 1/\varepsilon_A\rfloor\}$, and $\widetilde{\mathbb{P}}_1 \in \Delta(\mathcal{S}^{\mathsf{aug}})$ is the marginal distribution of $\phi^\star(x_1)$ for $(x_1,x_2)\sim\mathcal{D}$. The transition distribution $\widetilde{\mathbb{P}}_2$ is defined as follows. For initial state $s_1 \in \mathcal{S}$, next state $s_2 \in \mathcal{S}^{\mathsf{aug}}$, and action $a \in \mathcal{A}$,

$$\widetilde{\mathbb{P}}_2(s_2 \mid s_1, a) := \begin{cases} \Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2) = s_2 \mid \phi^\star(x_1) = s_1] \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2))}\left[1 - (a-y)^2\right] & \text{if } s_2 \in \mathcal{S} \\ \sum_{s_2'\in\mathcal{S}} \Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2) = s_2' \mid \phi^\star(x_1) = s_1] \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2'))}\left[(a-y)^2\right] & \text{if } s_2 = 0 \\ 0 & \text{if } s_2 = 1 \end{cases}.$$

Note that $\widetilde{\mathbb{P}}_1$ is supported on $\mathcal{S}$, so it is not necessary to define $\widetilde{\mathbb{P}}_2(s_2 \mid s_1, a)$ for $s_1 \in \{0,1\}$. Finally, for $s \in \mathcal{S}$ and $h \in [2]$, the observation distribution $\widetilde{\mathbb{O}}_h(\cdot|s)$ is defined as the conditional distribution of $x_h \mid \phi^\star(x_h) = s$ for $(x_1,x_2)\sim\mathcal{D}$. The distributions $\widetilde{\mathbb{O}}_h(\cdot \mid 0)$ are fully supported on 0, and the distributions $\widetilde{\mathbb{O}}_h(\cdot \mid 1)$ are fully supported on 1.

In order to invoke the guarantees of the RL oracle, we need to verify that M satisfies Φ^{aug}-decodability (Lemma D.4) and is in fact the MDP that is being simulated in TwoRed (Lemma D.5).

Lemma D.4 *M is a Φ^{aug}-decodable block MDP.*

Proof. First we observe that $\widetilde{\mathbb{P}}_1, \widetilde{\mathbb{P}}_2$ are well-defined: by construction $\widetilde{\mathbb{P}}_1$ is a distribution. Moreover, for any $s_1 \in \mathcal{S}$ and $a \in \mathcal{A}$, it is clear that $\widetilde{\mathbb{P}}_2(\cdot|s_1, a)$ is non-negative and

$$\sum_{s_2 \in \mathcal{S}} \widetilde{\mathbb{P}}_2(s_2 \mid s_1, a) = \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ y\sim\mathrm{Ber}(f(\phi^\star(x_1),\phi^\star(x_2)))}} \left[\left(1-(a-y)^2\right)\middle|\phi^\star(x_1) = s_1\right] = 1 - \widetilde{\mathbb{P}}_2(0 \mid s_1, a).$$

Thus, $\widetilde{\mathbb{P}}_2(\cdot \mid s_1, a)$ is a distribution. Finally, we observe that for any $h \in [2]$ and $s \in \mathcal{S}$, the observation distribution $\widetilde{\mathbb{O}}_h(\cdot \mid s)$ is fully supported on $x_h \in \mathcal{X}$ such that $\mathrm{aug}(\phi^\star)(x_h) = \phi^\star(x_h) = s$, by construction. Since $\mathrm{aug}(\phi^\star)(0) = 0$ and $\mathrm{aug}(\phi^\star)(1) = 1$, the same holds for $s \in \{0, 1\}$. ■

Lemma D.5 *Let $(x_1, x_2) \sim \mathcal{D}$ and $y \sim \mathrm{Ber}(f(\phi^\star(x_1), \phi^\star(x_2)))$. The following process simulates an episode of interaction with M:*

1. *Pass observation x_1 and receive action $a \in \mathcal{A}$.*
2. *With probability $(a-y)^2$, pass observation 0. Otherwise, pass observation x_2.*

Proof. Let $p(x_1, x_2)$ denote the density of $\mathcal{D}$, and observe that by $\phi^\star$-realizability Definition 2.3) we can write

$$p(x_1, x_2) = \widetilde{p}(\phi^\star(x_1), \phi^\star(x_2))\widetilde{q}_1(x_1 \mid \phi^\star(x_1))\widetilde{q}_2(x_2 \mid \phi^\star(x_2))$$

for some density $\widetilde{p}$ and conditional densities $\widetilde{q}_1, \widetilde{q}_2$. Also let $\widetilde{p}(s_2 \mid s_1)$ denote the conditional density of $\phi^\star(x_2) \mid \phi^\star(x_1)$ under $(x_1, x_2) \sim \mathcal{D}$. Now fix any policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ and trajectory (s_1, x_1, a, s_2, x_2) where $x_2 \neq 0$. The likelihood of this trajectory under the described process is

$$\begin{aligned}
&p(x_1, x_2)\mathbb{1}[s_1 = \phi^\star(x_1)]\mathbb{1}[s_2 = \phi^\star(x_2)]\pi(a \mid x_1) \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2))}[1-(a-y)^2] \\
&= \widetilde{p}(s_1, s_2)\widetilde{q}_1(x_1 \mid s_1)\widetilde{q}_2(x_2 \mid s_2)\pi(a \mid x_1) \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2))}[1-(a-y)^2] \\
&= \widetilde{\mathbb{P}}_1(s_1)\widetilde{p}(s_2 \mid s_1)\widetilde{\mathbb{O}}_1(x_1 \mid s_1)\widetilde{\mathbb{O}}_2(x_2 \mid s_2)\pi(a \mid x_1) \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2))}[1-(a-y)^2] \\
&= \widetilde{\mathbb{P}}_1(s_1)\widetilde{\mathbb{O}}_1(x_1 \mid s_1)\widetilde{\mathbb{O}}_2(x_2 \mid s_2)\pi(a \mid x_1)\widetilde{\mathbb{P}}_2(s_2 \mid s_1, a)
\end{aligned}$$

which is precisely the likelihood of the trajectory under M. Similarly, the likelihood of any trajectory $(s_1, x_1, a, 0, 0)$ under the described process is

$$\begin{aligned}
&\sum_{s_2\in\mathcal{S}} \widetilde{p}(s_1, s_2)\widetilde{q}_1(x_1 \mid s_1)\pi(a \mid x_1) \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2))}[(a-y)^2] \\
&= \widetilde{\mathbb{P}}_1(s_1)\widetilde{\mathbb{O}}_1(x_1 \mid s_1)\pi(a \mid x_1) \sum_{s_2\in\mathcal{S}} \widetilde{p}(s_2 \mid s_1) \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(f(s_1,s_2))}[(a-y)^2] \\
&= \widetilde{\mathbb{P}}_1(s_1)\widetilde{\mathbb{O}}_1(x_1 \mid s_1)\pi(a \mid x_1)\widetilde{\mathbb{P}}_2(s_2 \mid s_1, a)
\end{aligned}$$

as needed. ■

Next, in Lemma D.8, we show that for every latent state $s \in \mathcal{S}$, the set of policies Ψ computed by the RL oracle contains some approximate minimizer of the loss $L_s(\pi)$ defined below. To this end, the key lemma is Lemma D.7, which characterizes the loss of a policy in terms of its visitation probability for state s.

Definition D.6 *For any policy $\pi \in \Pi$ and state $s \in \mathcal{S}$, define*

$$L_s(\pi) := \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2) = s](\pi(x_1) - f(\phi^\star(x_1), \phi^\star(x_2)))^2\right],$$

and for each $s \in \mathcal{S}$ define

$$Z_s := \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ y\sim\mathrm{Ber}(f(\phi^\star(x_1),\phi^\star(x_2)))}}\left[\mathbb{1}[\phi^\star(x_2) = s]\left(y - f(\phi^\star(x_1), \phi^\star(x_2))^2\right)\right].$$

Lemma D.7 *For any policies $\pi, \pi' \in \Pi$ and state $s \in \mathcal{S}$, we have $d_2^{M,\pi}(s) - d_2^{M,\pi'}(s) = L_s(\pi') - L_s(\pi)$.*

Proof. By the characterization provided by Lemma D.5, for any policy π, the visitation distribution for a state $s_2 \in \mathcal{S}$ at step 2 is

$$\begin{aligned}
d_2^{M,\pi}(s_2) &= \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ y\sim\mathrm{Ber}(\phi^\star(x_1),\phi^\star(x_2))}}\left[\mathbb{1}[\phi^\star(x_2) = s_2]\cdot(1 - (\pi(x_1) - y)^2)\right] \\
&= \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ y\sim\mathrm{Ber}(\phi^\star(x_1),\phi^\star(x_2))}}\left[\mathbb{1}[\phi^\star(x_2) = s_2]\cdot\left(1 - (\pi(x_1) - f(\phi^\star(x_1), \phi^\star(x_2)) + f(\phi^\star(x_1), \phi^\star(x_2)) - y)^2\right)\right] \\
&= \Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2) = s_2] - L_s(\pi) - \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ y\sim\mathrm{Ber}(\phi^\star(x_1),\phi^\star(x_2))}}\left[\mathbb{1}[\phi^\star(x_2) = s_2](f(\phi^\star(x_1), \phi^\star(x_2)) - y)^2\right] \\
&\quad - 2\mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ y\sim\mathrm{Ber}(\phi^\star(x_1),\phi^\star(x_2))}}\left[\mathbb{1}[\phi^\star(x_2) = s_2](\pi(x_1) - f(\phi^\star(x_1), \phi^\star(x_2)))(f(\phi^\star(x_1), \phi^\star(x_2)) - y)\right].
\end{aligned}$$

The final term is 0 since $\mathbb{E}[y \mid x_1, x_2] = f(\phi^\star(x_1), \phi^\star(x_2))$. The remaining terms are all independent of π except for $-L_s(\pi)$; hence,

$$d_2^{M,\pi}(s_2) - d_2^{M,\pi'}(s_2) = L_s(\pi') - L_s(\pi)$$

as claimed. ■

Lemma D.8 *Suppose that $\mathcal{O}$ is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free RL oracle (Definition 2.1) for Φ^{aug}, and that $n/2 \geq N_{\mathsf{RL}}(\epsilon^2/(4|\mathcal{S}|^2), \delta/2, 2, 2|\mathcal{S}|/\epsilon)$. Then the set Ψ computed in Line 5 of Algorithm 4 satisfies $|\Psi| \leq K_{\mathsf{RL}}(\epsilon^2/(4|\mathcal{S}|^2), \delta/2, 2, 2|\mathcal{S}|/\epsilon)$. Moreover, it holds with probability at least $1 - \delta/2$ that for every $s \in \mathcal{S}$, there is some $\pi \in \Psi$ such that $L_s(\pi) \leq \epsilon^2/(2|\mathcal{S}|^2)$.*

Proof. By Lemma D.5, $\mathcal{O}$ is given interactive access to the MDP M. The claimed bound on $|\Psi|$ is immediate from Definition 2.1 together with the initialization of $\mathcal{O}$ (Line 2) and the fact that $H = 2$ and $|\mathcal{A}| = 2|\mathcal{S}|/\epsilon$.

Next, by Definition 2.1, it holds with probability at least $1 - \delta/2$ that for any $s \in \mathcal{S}^{\mathsf{aug}}$ and $h \in [2]$,

$$\max_{\pi \in \Psi} d_h^{M,\pi}(s) \geq \max_{\pi \in \Pi} d_h^{M,\pi}(s) - \frac{\epsilon^2}{4|\mathcal{S}|^2}. \tag{22}$$

Condition on this event. Fix any $s \in \mathcal{S}$ and define $\pi^\star : \mathcal{X}^{\mathsf{aug}} \to \mathcal{A}$ by $\pi^\star(x_1) := \varepsilon_A \lfloor f(\phi^\star(x_1), s)/\varepsilon_A \rfloor \in \mathcal{A}$ for $x_1 \in \mathcal{X}$ (define $\pi^\star(0), \pi^\star(1)$ arbitrarily). Note that $L_s(\pi^\star) \leq \varepsilon_A^2$. By (22), there is some $\pi \in \Psi$ such that

$$\begin{aligned}
\frac{\epsilon^2}{4|\mathcal{S}|^2} &\geq d_2^{M,\pi^\star}(s) - d_2^{M,\pi}(s) \\
&= L_s(\pi) - L_s(\pi^\star) \\
&\geq L_s(\pi) - \varepsilon_A^2
\end{aligned}$$

where the equality is by Lemma D.7. The lemma follows from the definition of ε_A. ■

The following lemma shows that with high probability, each error function $\mathcal{R}^\pi$ approximates an affine transformation of the loss $L_{\phi^\star(x_2)}(\pi)$.

Lemma D.9 *Suppose that $\mathcal{O}$ is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free RL oracle for Φ^{aug} (Definition 2.1). Set $N = N_{\mathsf{RL}}(\epsilon^4/(16|\mathcal{S}|^4), \delta/(4|\Psi|), 2, 4|\mathcal{S}|^2/\epsilon^2)$ and $K = K_{\mathsf{RL}}(\epsilon^4/(16|\mathcal{S}|^4), \delta/(4|\Psi|), 2, 4|\mathcal{S}|^2/\epsilon^2)$. If $n/2 \geq N + 64|\mathcal{S}|^8\epsilon^{-8}\log(4|\Psi|K/\delta)$ then, with probability at least $1-\delta/2$, it holds for all $\pi \in \Psi$ that*

$$\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}} \left(\mathcal{R}^\pi(x_2) - \frac{L_{\phi^\star(x_2)}(\pi) + Z_{\phi^\star(x_2)}}{\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2') = \phi^\star(x_2)]} \right)^2 \leq \frac{\epsilon^4}{4|\mathcal{S}|^4}.$$

Proof. By Proposition G.2 and the assumption on $\mathcal{O}$, we get that $\mathtt{OneRed}(\mathcal{O}, \cdot)$ is an N_{reg}-efficient one-context regression oracle for Φ (Definition 2.2) with

$$N_{\mathsf{reg}}\left(\frac{\epsilon^4}{4|\mathcal{S}|^4}, \frac{\delta}{2|\Psi|}\right) \leq N + \frac{64|\mathcal{S}|^8\log(4|\Psi|K/\delta)}{\epsilon^8} \leq n/2.$$

Fix any $\pi \in \Psi$. Observe that $\mathcal{C}^\pi$ consists of $n/2$ i.i.d. samples $(x_2^{(i)}, z^{(i)})$ with $x_2^{(i)} \in \mathcal{X}$ and $z^{(i)} \in \{0,1\}$. For any index i, by $\phi^\star$-realizability of $\mathcal{D}$ and the law of $y^{(i)}$, we have that $x_2^{(i)} \perp (\pi(x_1^{(i)}) - y^{(i)})^2 \mid \phi^\star(x_2^{(i)})$ and hence $x_2^{(i)} \perp z^{(i)} \mid \phi^\star(x_2^{(i)})$. Moreover, for any $x_2 \in \mathcal{X}$ with $s := \phi^\star(x_2)$, we have

$$\begin{aligned}
\mathbb{E}[z^{(i)} \mid x_2^{(i)} = x_2] &= \mathbb{E}[z^{(i)} \mid \phi^\star(x_2^{(i)}) = s] \\
&= \mathop{\mathbb{E}}_{\substack{(x_1',x_2')\sim\mathcal{D} \\ y\sim\mathrm{Ber}(f(\phi^\star(x_1'),\phi^\star(x_2')))}} [(\pi(x_1') - y')^2 \mid \phi^\star(x_2') = s] \\
&= \mathop{\mathbb{E}}_{(x_1',x_2')\sim\mathcal{D}} [(\pi(x_1') - f(\phi^\star(x_1'), \phi^\star(x_2')))^2 \mid \phi^\star(x_2') = s]
\end{aligned}$$

$$+ \mathop{\mathbb{E}}_{\substack{(x_1',x_2')\sim\mathcal{D} \\ y\sim\mathrm{Ber}(f(\phi^\star(x_1'),\phi^\star(x_2')))}} [(f(\phi^\star(x_1'),\phi^\star(x_2')) - y)^2 \mid \phi^\star(x_2') = s]$$
$$= \frac{L_s(\pi) + Z_s}{\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2') = s]}$$

where the penultimate equality uses that $\mathbb{E}[y \mid x_1', x_2'] = f(\phi^\star(x_1'), \phi^\star(x_2'))$. The result now follows from Definition 2.2 and a union bound over $\pi \in \Psi$. ■

The following technical lemma is needed to handle the distribution mismatch discussed in Section D.1. Essentially, it asserts that if we sample (x_1, x_2) and (x_1', x_2') independently and then condition on x_2, x_2' having the same latent state s, then x_1 and x_1' are exchangeable, i.e. (x_1, x_2) and (x_1', x_2) have the same distribution.

Lemma D.10 *For any $s \in \mathcal{S}$ and $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, it holds that*

$$\mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ (x_1',x_2')\sim\mathcal{D}}}[g(x_1,x_2) \mid \phi^\star(x_2) = \phi^\star(x_2') = s] = \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D} \\ (x_1',x_2')\sim\mathcal{D}}}[g(x_1',x_2) \mid \phi^\star(x_2) = \phi^\star(x_2') = s]$$

where the random variables (x_1, x_2) and (x_1', x_2') in the expectation are independent draws from $\mathcal{D}$.

Proof. By $\phi^\star$-realizability of $\mathcal{D}$, note that x_1, x_2 are conditionally independent under the event $\phi^\star(x_2) = \phi^\star(x_2') = s$. Additionally, for any events $\mathcal{E}, \mathcal{E}' \subseteq \mathcal{X}$,

$$\begin{aligned}
&\Pr[x_2 \in \mathcal{E} \wedge x_1' \in \mathcal{E}' \wedge \phi^\star(x_2) = s \wedge \phi^\star(x_2') = s] \\
&= \Pr[x_1 \in \mathcal{E} \wedge \phi^\star(x_2) = s] \Pr[x_2' \in \mathcal{E}' \wedge \phi^\star(x_2') = s] \\
&= \frac{\Pr[x_2 \in \mathcal{E} \wedge \phi^\star(x_2) = s \wedge \phi^\star(x_2') = s] \Pr[x_1' \in \mathcal{E}' \wedge \phi^\star(x_2) = s \wedge \phi^\star(x_2') = s]}{\Pr[\phi^\star(x_2) = s \wedge \phi^\star(x_2') = s]},
\end{aligned}$$

so

$$\begin{aligned}
&\Pr[x_2 \in \mathcal{E} \wedge x_1' \in \mathcal{E}' \mid \phi^\star(x_2) = s \wedge \phi^\star(x_2') = s] \\
&= \Pr[x_2 \in \mathcal{E} \mid \phi^\star(x_2) = s \wedge \phi^\star(x_2') = s] \Pr[x_1' \in \mathcal{E}' \mid \phi^\star(x_2) = s \wedge \phi^\star(x_2') = s].
\end{aligned}$$

Thus, x_1', x_2 are conditionally independent as well. But by symmetry, $x_1 \mid \phi^\star(x_2) = \phi^\star(x_2') = s$ and $x_1' \mid \phi^\star(x_2) = \phi^\star(x_2') = s$ have identical distributions. Thus, $(x_1, x_2) \mid \phi^\star(x_2) = \phi^\star(x_2') = s$ and $(x_1', x_2) \mid \phi^\star(x_2) = \phi^\star(x_2') = s$ have identical distributions, which implies the claimed equality. ■

With the above ingredients, we can now formally conclude the proof of Theorem D.2.

Proof of Theorem D.2. Consider the event that the claimed bounds of Lemma D.8 and Lemma D.9 both hold, which occurs with probability at least $1 - \delta$. We argue that in this event, the desired error bound (21) holds.

For each $s \in \mathcal{S}$ fix any $\pi^s \in \arg\max_{\pi\in\Psi} d_2^{M,\pi}(s)$ (i.e. π^s is optimal for reaching state s, among policies in Ψ). For each $x_2 \in \mathcal{X}$, let $\widehat{\pi}^{x_2} \in \Psi$ be the policy minimizing $\mathcal{R}^\pi(x_2)$ (breaking ties as in Algorithm 4). We have by definition of $\mathcal{R}$ that

$$\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[(\mathcal{R}(x_1,x_2) - f(\phi^\star(x_1),\phi^\star(x_2))^2\right]$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[(\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2))^2\right] \\
&= \sum_{s\in\mathcal{S}} \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s]\cdot(\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2))^2\right].
\end{aligned}
$$

Fix any $s \in \mathcal{S}$. On the one hand, observe that

$$
\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s](\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2)))^2\right] \le \Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2)=s] \tag{23}
$$

since $\widehat{\pi}^{x_2}(x_1) \in \mathcal{A} \subset [0,1]$ and $f(\phi^\star(x_1),\phi^\star(x_2)) \in [0,1]$. On the other hand,

$$
\begin{aligned}
&\Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2)=s]\cdot \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s](\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2)))^2\right] \\
&= \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D}\\(x_1',x_2')\sim\mathcal{D}}}\left[\mathbb{1}[\phi^\star(x_2)=s]\mathbb{1}[\phi^\star(x_2')=s](\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2)))^2\right] \\
&= \mathop{\mathbb{E}}_{\substack{(x_1,x_2)\sim\mathcal{D}\\(x_1',x_2')\sim\mathcal{D}}}\left[\mathbb{1}[\phi^\star(x_2)=s]\mathbb{1}[\phi^\star(x_2')=s](\widehat{\pi}^{x_2}(x_1') - f(\phi^\star(x_1'),\phi^\star(x_2)))^2\right] \\
&= \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s]\cdot \mathop{\mathbb{E}}_{(x_1',x_2')\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2')=s](\widehat{\pi}^{x_2}(x_1') - f(\phi^\star(x_1'),\phi^\star(x_2)))^2\right]\right] \\
&= \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s]\cdot \mathop{\mathbb{E}}_{(x_1',x_2')\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2')=s](\widehat{\pi}^{x_2}(x_1') - f(\phi^\star(x_1'),\phi^\star(x_2')))^2\right]\right] \\
&= \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s]\cdot L_s(\widehat{\pi}^{x_2})\right] \\
&= \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s]\cdot (L_s(\widehat{\pi}^{x_2}) - L_s(\pi^s))\right] + \Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2)=s]\cdot L_s(\pi^s)
\end{aligned} \tag{24}
$$

where the second equality is by Lemma D.10 with function $g(x_1,x_2) = (\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2)))^2$; the fourth equality is because $\phi^\star(x_2) = \phi^\star(x_2') = s$ (unless one of the indicator functions is zero); and the fifth inequality is by definition of L_s. Combining Eqs. (23) and (24), we get that

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s](\widehat{\pi}^{x_2}(x_1) - f(\phi^\star(x_1),\phi^\star(x_2)))^2\right] \\
&\le \sqrt{\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s]\cdot (L_s(\widehat{\pi}^{x_2}) - L_s(\pi^s))\right] + \Pr_{(x_1,x_2)\sim\mathcal{D}}[\phi^\star(x_2)=s]\cdot L_s(\pi^s)}.
\end{aligned} \tag{25}
$$

By the guarantee of Lemma D.8, we have $L_s(\pi^s) \le \epsilon^2/(2|\mathcal{S}|^2)$. Also,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\left[\mathbb{1}[\phi^\star(x_2)=s](L_s(\widehat{\pi}^{x_2}) - L_s(\pi^s))\right] \\
&\le \mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}}\Bigg[\mathbb{1}[\phi^\star(x_2)=s]\Big(L_s(\widehat{\pi}^{x_2}) - \mathcal{R}^{\widehat{\pi}^{x_2}}(x_2)\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2')=s] \\
&\qquad\qquad + \mathcal{R}^{\pi^s}(x_2)\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2')=s] - L_s(\pi^s)\Big)\Bigg] \\
&\le 2\max_{\pi\in\Psi} \mathop{\mathbb{E}}_{x_1,x_2\sim\mathcal{D}} \mathbb{1}[\phi^\star(x_2)=s]\left|L_s(\pi) + Z_s - \mathcal{R}^\pi(x_2)\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2')=s]\right| \\
&\le 2\max_{\pi\in\Psi} \mathop{\mathbb{E}}_{x_1,x_2\sim\mathcal{D}} \mathbb{1}[\phi^\star(x_2)=s]\left|\frac{L_s(\pi) + Z_s}{\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2')=s]} - \mathcal{R}^\pi(x_2)\right|
\end{aligned}
$$

$$\leq 2 \max_{\pi \in \Psi} \sqrt{\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}} \mathbb{1}[\phi^\star(x_2) = s] \left(\frac{L_s(\pi) + Z_s}{\Pr_{(x_1',x_2')\sim\mathcal{D}}[\phi^\star(x_2') = s]} - \mathcal{R}^\pi(x_2) \right)^2} \leq \frac{\epsilon^2}{2|\mathcal{S}|^2}$$

where the first inequality is by minimality of $\mathcal{R}^{\widehat{\pi}^{x_2}}(x_2)$ over all $\pi \in \Psi$, and the final inequality is by the guarantee of Lemma D.9. Substituting into Eq. (25) and summing over $s \in \mathcal{S}$, we get

$$\mathop{\mathbb{E}}_{(x_1,x_2)\sim\mathcal{D}} (\mathcal{R}(x_1, x_2) - f(\phi^\star(x_1), \phi^\star(x_2)))^2 \leq \epsilon$$

as needed. ■

D.3. Analysis of `RegToRL` (Algorithm 3)

The proof of Theorem 3.2 is now straightforward from the analysis of `TwoRed` (Theorem D.2) and the analysis of `TwoAug` (Proposition G.8).

Proof of Theorem 3.2. Fix a regular concept class $\Phi^{\mathsf{aug}} \subseteq (\mathcal{X}^{\mathsf{aug}} \to \mathcal{S}^{\mathsf{aug}})$. By regularity (Definition 2.5), $\{0, 1\} \subseteq \mathcal{X}^{\mathsf{aug}}, \mathcal{S}^{\mathsf{aug}}$ so we can define $\mathcal{X} := \mathcal{X}^{\mathsf{aug}} \setminus \{0, 1\}$ and $\mathcal{S} := \mathcal{S}^{\mathsf{aug}} \setminus \{0, 1\}$. Define a concept class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ by restricting each ϕ to domain $\mathcal{X}$; regularity ensures that the range of each restricted map is contained in $\mathcal{S}$, so this definition is well-defined. We now observe that the augmented concept class (Definition D.1) with base class Φ is precisely Φ^{aug}.

Now suppose that $\mathcal{O}$ is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free episodic RL oracle for Φ^{aug} with $N_{\mathsf{RL}}, K_{\mathsf{RL}}$ bounded in terms of the parameters $N^\circ_{\mathsf{RL}}, C_{\mathsf{RL}}$ as specified in the theorem statement. By Theorem D.2 and the assumed parametric bounds on $N_{\mathsf{RL}}, K_{\mathsf{RL}}$, there is a constant $C > 0$ so that $\texttt{TwoRed}(\mathcal{O}, \cdot)$ is an N_{reg}-efficient two-context regression algorithm for Φ with

$$N_{\mathsf{reg}}(\epsilon, \delta) \leq N^\circ_{\mathsf{RL}} \left(\frac{|\mathcal{S}|}{\epsilon\delta} \right)^{C \cdot C_{\mathsf{RL}}}.$$

It follows from Proposition G.8 that $\texttt{TwoAug}(\texttt{TwoRed}(\mathcal{O}, \cdot), \cdot)$ is an N'_{reg}-efficient two-context regression algorithm for Φ^{aug} with

$$N'_{\mathsf{reg}}(\epsilon, \delta) \leq N^\circ_{\mathsf{RL}} \left(\frac{|\mathcal{S}|}{\epsilon\delta} \right)^{C_{3.2} \cdot C_{\mathsf{RL}}}$$

so long as $C_{3.2} > 0$ is a sufficiently large constant. Note that we are using the fact that $\epsilon, \delta \in (0, 1/2)$ to absorb constant factors.

To conclude the proof, we observe that the claimed oracle time complexity bound is immediate from the pseudocode of `RegToRL` and its subroutines. ■

Appendix E. Proof of Theorem 4.1

In this section we prove that for any concept class Φ, there is a reduction from reward-free RL (Definition 2.1) in the reset access model to one-context regression (Definition 2.2). The formal statement is provided below.

Theorem E.1 (General version of Theorem 4.1) *There is a constant $C_{E.1} > 0$ and an algorithm* PCR *(Algorithm 5) so that the following holds. Let $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ be any concept class, and let* $\mathtt{Reg}$ *be a N_{reg}-efficient one-context regression oracle for Φ. Then $\mathsf{PCR}(\mathtt{Reg}, N_{\mathsf{reg}}, |\mathcal{S}|, \cdot)$ is an $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free RL algorithm for Φ in the reset access model, with:*

- $K_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \leq H^2 |\mathcal{S}|^2$
- $N_{\mathsf{RL}}(\epsilon, \delta, H, |\mathcal{A}|) \leq \left(\frac{H|\mathcal{A}||\mathcal{S}|}{\epsilon\delta}\right)^{C_{E.1}} N_{\mathsf{reg}}\left(\left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{E.1}}, \left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{E.1}}\right)$.

Moreover, the oracle time complexity of PCR *is at most* $\left(\frac{H|\mathcal{A}||\mathcal{S}|}{\epsilon\delta}\right)^{C_{E.1}} N_{\mathsf{reg}}\left(\left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{E.1}}, \left(\frac{\epsilon\delta}{H|\mathcal{A}||\mathcal{S}|}\right)^{C_{E.1}}\right)$.

In particular, Theorem 4.1 follows from Theorem E.1 by substituting $N_{\mathsf{reg}}(\epsilon, \delta) := N^\circ_{\mathsf{reg}}/(\epsilon\delta)^{C_{\mathsf{reg}}}$ into the above bounds. Henceforth, fix a concept class Φ, a N_{reg}-efficient one-context regression oracle $\mathtt{Reg}$, and a Φ-decodable block MDP M with horizon H, action set $\mathcal{A}$, and unknown decoding function $\phi^\star \in \Phi$. We also define truncations of M (see Section B.1), with the parameters $\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}} > 0$ as defined in Algorithm 5.

E.1. PCR Pseudocode and Overview

We start by giving an overview of the algorithm PCR (Algorithm 5). The structure is similar to that of PCE (Algorithm 1), our algorithm from the episodic setting. In particular, the differences are entirely within the subroutine EPCR (Algorithm 6), the analogue of EPCE (Algorithm 2) from the episodic setting. Within EPCE, the difference is in how the kinematics are estimated (and what precise kinematics function is estimated)—since we no longer have access to a two-context regression oracle, but we do have the ability to reset to any previously-seen state.

Overview of EPCR. As shown in Algorithm 6, EPCR takes as input a one-context regression oracle $\mathtt{Reg}$, a step $h \in [H]$, a set of policy covers $\Psi_{1:h}$, a backup policy cover Γ, and certain sample complexity and tolerance parameters. As with EPCE, the goal is to produce a policy cover for step $h+1$, and to this end the algorithm estimates a certain kinematics function, defines internal reward functions by (implicitly) clustering together observations with similar kinematics, and uses PSDP (Algorithm 7) to find a policy that $\widehat{\pi}^{(t)}$ optimizes each reward function $\mathcal{R}^{(t)}$.

As discussed in Section 4, the main difference compared to EPCE (and HOMER) is in the estimation of kinematics. Rather than designing a dataset where the Bayes predictor must essentially distinguish between "real" and "fake" transitions (which inherently requires two contexts), EPCR samples m discriminator observations $x_h^{(1)}, \dots, x_h^{(m)}$ at step h (by rolling in with the given policy cover Ψ_h and backup policy cover Γ). For each of these observations $x_h^{(i)}$ and each action $a \in \mathcal{A}$, EPCR uses reset access to draw many conditional samples from $\mathbb{P}^M_{h+1}(\cdot \mid x_h^{(i)}, a)$, and then constructs a dataset $\mathcal{D}_{i,a}$ where the Bayes predictor must essentially predict if an observation was conditionally sampled from $(x_h^{(i)}, a)$ or from some other observation/action pair.

Algorithm 5 PCR: RL with Resets via One-Context Regression

input: One-context regression oracle Reg; efficiency function $N_{\mathsf{reg}} : (0,1) \times (0,1) \to \mathbb{N}$; # of latent states S; final error tolerance $\varepsilon_{\mathsf{final}} \in (0,1)$; failure parameter $\delta \in (0,1)$; horizon H; action set $\mathcal{A}$.

$N \leftarrow N_{\mathsf{reg}}\left(\left(\frac{\varepsilon_{\mathsf{final}}\delta}{H|\mathcal{A}|S}\right)^{C_{E.1}}, \left(\frac{\varepsilon_{\mathsf{final}}\delta}{H|\mathcal{A}|S}\right)^{C_{E.1}}\right)$.

$\sigma_{\mathsf{trunc}} \leftarrow \varepsilon_{\mathsf{final}}/(4+HS)$.

$R \leftarrow SH$, $\sigma_{\mathsf{bkup}} \leftarrow \frac{\sigma_{\mathsf{trunc}}^2}{RS^2H^2}$, $\alpha \leftarrow \frac{1-4\sigma_{\mathsf{trunc}}}{S}$, $m \leftarrow \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}},\sigma_{\mathsf{bkup}})}\log(S/\delta)$, $n \leftarrow \frac{m|\mathcal{A}|}{\sigma_{\mathsf{trunc}}}\log(S/\delta)$.

$\epsilon \leftarrow \min(\frac{\alpha^8\sigma_{\mathsf{trunc}}^{16}\sigma_{\mathsf{bkup}}^8}{6^8H^8m^{12}|\mathcal{A}|^{12}}, \frac{\delta^4}{m^2|\mathcal{A}|^2n^4})$, $\gamma \leftarrow \epsilon^{1/8}$, $\gamma' \leftarrow 2\epsilon^{1/4}$.

$\Gamma^{(1)} \leftarrow \emptyset$.

for $1 \le r \le R$ **do**

  $\Psi_1^{(r)} := \{\pi_{\mathsf{unif}}\}$.

  for $1 \le h < H$ **do**

    $\Psi_{h+1}^{(r)} \leftarrow \mathsf{EPCR}(\mathtt{Reg}, h, \Psi_{1:h}^{(r)}, \Gamma, n, m, N, \gamma, \gamma')$.

  $\Gamma^{(r+1)} \leftarrow \Gamma^{(r)} \cup \bigcup_{h\in[H]:|\Psi_h^{(r)}|\le S} \Psi_h^{(r)}$.

return: $\bigcup_{h\in[H]}\bigcup_{1\le r\le R:|\Psi_h^{(r)}|\le S}\Psi_h^{(r)}$.

More formally, as shown in Algorithm 6, for each $i \in [m]$ and $a \in \mathcal{A}$, EPCR constructs a dataset $\mathcal{D}_{i,a}$ with samples from the following procedure. First, draw $j \sim \mathrm{Unif}([m])$ and $a_h \sim \mathrm{Unif}(\mathcal{A})$. Then reset to $x_h^{(j)}$, and sample $x_{h+1} \sim \mathbb{P}_{h+1}^M(\cdot \mid x_h^{(j)}, a_h)$. If $j = i$ and $a_h = a$, then add sample $(x_{h+1}, 1)$ to the dataset; otherwise add $(x_{h+1}, 0)$. It can be checked that the Bayes predictor $\mathbb{E}[y \mid x_{h+1}]$ for this dataset is the kinematics function $w_{h+1}(\cdot; i, a)$ defined in Eq. (3).By the Block MDP assumption, $w_{h+1}(x_{h+1}; i, a)$ only depends on x_{h+1} through $\phi^\star(x_{h+1})$, so the guarantee of one-context regression applies, and invoking Reg on $\mathcal{D}_{i,a}$ gives a good approximation $\widehat{w}_{h+1}(\cdot; i, a)$ of $w_{h+1}(\cdot; i, a)$ with high probability.

This approximation is used in a similar fashion as $\widehat{f}_{h+1}$ is used in EPCE—the only difference henceforth is that there is no need to sample additional "test observations", since $x_h^{(1)}, \dots, x_h^{(m)}$ serve this purposes.

We formally analyze EPCR in Section E.2; the main guarantee is Theorem E.2.

Overview of PCR. The algorithm PCR is identical to PCE aside from (1) various parameter choices, and (2) invoking EPCR rather than EPCE. The algorithm proceeds in $R = |\mathcal{S}|H$ rounds, and in each round r, it iteratively uses EPCR to construct sets $\Psi_{1:H}^{(r)}$, which are added to the backup policy cover for the next round. The output is the union of all sets (of bounded size) that were produced by EPCR.

We formally analyze PCR (and thereby prove Theorem E.1) in Section E.3.

E.2. Analysis of EPCR (Algorithm 6)

The following theorem is our main guarantee for EPCR (Algorithm 6). Recall that we have fixed a concept class Φ, a N_{reg}-efficient one-context regression oracle Reg, and a Φ-decodable block MDP

Algorithm 6 EPCR: Extend Policy Cover in Reset Model

input: One-context regression oracle Reg; step $h \in [H]$; policy covers $\Psi_{1:h}$; backup policy cover Γ; sample counts $n, m, N \in \mathbb{N}$; tolerances $\gamma, \gamma' \in (0,1)$.
$\mathcal{C} \leftarrow \emptyset$.
for $1 \le i \le m$ **do**
  Sample trajectory $(x_1, a_1, \dots, x_h) \sim \frac{1}{2}\left(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma)\right)$.
  Set $x_h^{(i)} := x_h$.
  Update dataset: $\mathcal{C} \leftarrow \mathcal{C} \cup \{x_h\}$.
for $1 \le i \le m$ **do**
  for $a \in \mathcal{A}$ **do**
    $\mathcal{D}_{i,a} \leftarrow \emptyset$.
    for N times **do**
      Sample $j \sim \mathrm{Unif}([m])$ and $a_h \sim \mathrm{Unif}(\mathcal{A})$.
      Reset to $x_h := x_h^{(j)}$ and sample $x_{h+1} \sim \mathbb{P}^M_{h+1}(\cdot|x_h, a_h)$.
      Update dataset: $\mathcal{D}_{i,a} \leftarrow \mathcal{D}_{i,a} \cap \{(x_{h+1}, \mathbb{1}[j = i \wedge a_h = a])\}$.
    $\widehat{w}_{h+1}(\cdot; i, a) \leftarrow \mathtt{Reg}(\mathcal{D}_{i,a})$. ▷ $\widehat{w}_{h+1}(\cdot; i, a) : \mathcal{X} \to [0,1]$
$\Psi_{h+1} \leftarrow \emptyset$, $\mathcal{T}_{\mathsf{clus}} \leftarrow \emptyset$.
for $1 \le t \le n$ **do**
  Sample $j \sim \mathrm{Unif}([m])$.
  Reset to $x_h := x^{(j)}$ and sample trajectory $(x_h, a_h, x_{h+1}) \sim \mathrm{Unif}(\mathcal{A})$. Write $\overline{x}^{(t)}_{h+1} := x_{h+1}$.
  Define $\mathcal{R}^{(t)} : \mathcal{X} \to [0,1]$ by

$$\mathcal{R}^{(t)}(x) := \max\left(0, 1 - \frac{\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1}; i, a) - \widehat{w}_{h+1}(x; i, a)|}{\gamma}\right).$$

  if $\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1}; i, a) - \widehat{w}_{h+1}(\overline{x}^{(t')}_{h+1}; i, a)| > \gamma'$ for all $t' \in \mathcal{T}_{\mathsf{clus}}$ **then**
    $\widehat{\pi}^{(t)} \leftarrow \mathsf{PSDP}(h, \mathtt{Reg}, \mathcal{R}^{(t)}, \Psi_{1:h}, \Gamma, N)$. ▷ See Algorithm 7
    Update $\Psi_{h+1} \leftarrow \Psi_{h+1} \cup \{\widehat{\pi}^{(t)}\}$.
    Update $\mathcal{T}_{\mathsf{clus}} \leftarrow \mathcal{T}_{\mathsf{clus}} \cup \{t\}$.
return: Ψ_{h+1}.

M with horizon H, action set $\mathcal{A}$, and unknown decoding function $\phi^\star \in \Phi$. We have also defined truncations of M (see Section B.1), with the parameters $\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}} > 0$ as defined in Algorithm 5.

Theorem E.2 *Let $h \in \{1, \dots, H-1\}$. Let $\delta, \alpha > 0$ and $m, n, N \in \mathbb{N}$. Let $\Gamma \subset \Pi$ be a finite set of policies. Suppose that $\Psi_{1:h}$ are α-truncated policy covers (Definition B.2) for M at steps $1, \dots, h$.*

Suppose that $m \geq \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} \log(|\mathcal{S}|/\delta)$, $n \geq m|\mathcal{A}|\sigma_{\mathsf{trunc}}^{-1} \log(|\mathcal{S}|/\delta)$, $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, *and*

$$\epsilon^{1/8} \leq \frac{\alpha\sigma_{\mathsf{trunc}}^2 \sigma_{\mathsf{bkup}}}{6H(m|\mathcal{A}|)^{3/2}}. \tag{26}$$

Consider the execution of `EPCR` *with* $\gamma := \epsilon^{1/8}$ *and* $\gamma' := 2\epsilon^{1/4}$*. Then with probability at least* $1 - (2 + m|\mathcal{A}| + H|\mathcal{A}|n)\delta - \sqrt{m|\mathcal{A}|}n\epsilon^{1/4}$*, the following two properties hold:*

- $|\Psi_{h+1}| \leq |\mathcal{S}|$.
- *Either* Ψ_{h+1} *is a* $(1 - 4\sigma_{\mathsf{trunc}})$*-truncated max policy cover (Definition B.3) for* M *at step* $h+1$*, or else*

$$\max_{\pi \in \Psi_{h+1}} d_{h+1}^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) \geq \sigma_{\mathsf{trunc}}^2.$$

See Section B.1 for the definition of the truncated MDP $\overline{M}(\Gamma)$ and the truncated policy covers. Like Theorem C.2 (the analogous guarantee for `EPCE`), this result shows that either `EPCR` produces a policy cover, or one of the policies in the output reaches the terminal state $\mathfrak{t}$ in $\overline{M}(\Gamma)$, which means that it discovered a state not well-covered by policies in Γ.

Let us fix the inputs to `EPCR`: in addition to the one-context regression oracle `Reg` (Definition 2.4), we fix a layer $h \in [H-1]$, sets of policies $\Psi_1, \ldots, \Psi_h$ and Γ, sample counts $n, m, N \in \mathbb{N}$, and tolerances $\gamma, \gamma' \in (0,1)$. The first step of `EPCR` is to use one-context regression and reset access to estimate the kinematics function w_{h+1} defined informally in Eq. (3) and formally below. Lemma E.4 shows that with high probability, the estimate $\widehat{w}_{h+1}(\cdot; i, a)$ is an accurate estimate of $w_{h+1}(\cdot; \phi^\star(x_h^{(i)}), a)$ for all actions $a \in \mathcal{A}$ and "test" observations $x_h^{(1)}, \ldots, x_h^{(m)}$. Next, Lemma E.5 shows that the reward functions $\mathcal{R}^{(t)}$ designed by clustering with respect to $\widehat{w}_{h+1}$ approximately induce exploration of all latent reachable states. We then use these two lemmas to prove Theorem E.2.

Definition E.3 *For fixed* $x_h^{(1)}, \ldots, x_h^{(m)}$*, we define distributions* $\widehat{\beta}_h, \widehat{\beta}_{h+1} \in \Delta(\mathcal{S})$ *by*

$$\widehat{\beta}_h(s) := \frac{1}{m} \sum_{j=1}^m \mathbb{1}[\phi^\star(x_h^{(j)}) = s]$$

and

$$\widehat{\beta}_{h+1}(s) := \frac{1}{m} \sum_{j=1}^m \mathop{\mathbb{E}}_{a \sim \mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}_{h+1}^M(s \mid \phi^\star(x_h^{(j)}), a).$$

For any $a_h \in \mathcal{A}$ *and* $s_h, s_{h+1} \in \mathcal{S}$ *we also define*

$$w_{h+1}(s_{h+1}; s_h, a_h) := \frac{\widetilde{\mathbb{P}}_{h+1}^M(s_{h+1} \mid s_h, a_h)}{\sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}_{h+1}^M(s_{h+1} \mid \phi^\star(x_h^{(j)}), a)}.$$

Lemma E.4 *Fix a realization of* $\mathcal{C} = \{x_h^{(1)}, \ldots, x_h^{(m)}\}$*. If* $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$*, then it holds with probability at least* $1 - \delta m|\mathcal{A}|$ *that*

$$\mathop{\mathbb{E}}_{x_{h+1} \sim \widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}} \max_{(i,a) \in [m] \times \mathcal{A}} \left| \widehat{w}_{h+1}(x_{h+1}; i, a) - w_{h+1}(\phi^\star(x_{h+1}); \phi^\star(x_h^{(i)}), a) \right| \leq \sqrt{m|\mathcal{A}|\epsilon}. \tag{27}$$

Proof. Fix $i \in [m]$ and $a \in \mathcal{A}$. The dataset $\mathcal{D}_{i,a}$ consists of N i.i.d. samples. Let (x_{h+1}, y) denote the first sample, so that $x_{h+1} \sim \mathbb{P}^M_{h+1}(\cdot \mid x^{(j)}_h, a_h)$ and $y = \mathbb{1}[j = i \wedge a_h = a]$ for latent random variables $j \sim \mathrm{Unif}([m])$ and $a_h \sim \mathrm{Unif}(\mathcal{A})$. The marginal distribution of x_{h+1} is exactly $\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}$. Also,

$$
\begin{aligned}
\mathbb{E}[y \mid x_{h+1} = x] &= \Pr[j = i \wedge a_h = a \mid x_{h+1} = x] \\
&= \frac{\Pr[x_{h+1} = x \mid j = i \wedge a_h = a] \Pr[j = i \wedge a_h = a]}{\Pr[x_{h+1} = x]} \\
&= \frac{\frac{1}{m|\mathcal{A}|}\widetilde{\mathbb{O}}^M_{h+1}(x \mid \phi^\star(x))\widetilde{\mathbb{P}}^M_{h+1}(\phi^\star(x) \mid \phi^\star(x^{(i)}_h), a)}{\frac{1}{m|\mathcal{A}|}\sum_{j=1}^m \sum_{a\in\mathcal{A}} \widetilde{\mathbb{O}}^M_{h+1}(x_{h+1} \mid \phi^\star(x_{h+1}))\widetilde{\mathbb{P}}^M_{h+1}(\phi^\star(x_{h+1}) \mid \phi^\star(x^{(j)}_h), a)} \\
&= \frac{\widetilde{\mathbb{P}}^M_{h+1}(\phi^\star(x) \mid \phi^\star(x^{(i)}_h), a)}{\sum_{j=1}^m \sum_{a\in\mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(\phi^\star(x) \mid \phi^\star(x^{(j)}_h), a)} \\
&= w_{h+1}(\phi^\star(x); \phi^\star(x^{(i)}_h), a)
\end{aligned}
$$

by Definition E.3. Hence, we can apply the guarantee of $\mathtt{Reg}$ (Definition 2.2) with distribution $\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}$ and ground truth predictor $s_{h+1} \mapsto w_{h+1}(s_{h+1}; \phi^\star(x^{(i)}_h), a)$. Since $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, we get that with probability at least $1 - \delta$, the output $\widehat{w}(\cdot; i, a)$ of $\mathtt{Reg}(\mathcal{D}_{i,a})$ satisfies

$$
\mathop{\mathbb{E}}_{x_{h+1}\sim\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}} \left(\widehat{w}_{h+1}(x_{h+1}; i, a) - w_{h+1}(\phi^\star(x_{h+1}); \phi^\star(x^{(i)}_h), a)\right)^2 \leq \epsilon.
$$

Condition on the event that this bound holds for all $i \in [m]$ and $a \in \mathcal{A}$, which occurs with probability at least $1 - \delta m|\mathcal{A}|$. Then

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x_{h+1}\sim\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}} \max_{(i,a)\in[m]\times\mathcal{A}} \left|\widehat{w}_{h+1}(x_{h+1}; i, a) - w_{h+1}(\phi^\star(x_{h+1}); \phi^\star(x^{(i)}_h), a)\right| \\
&\leq \sqrt{\mathop{\mathbb{E}}_{x_{h+1}\sim\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}} \max_{(i,a)\in[m]\times\mathcal{A}} \left(\widehat{w}_{h+1}(x_{h+1}; i, a) - w_{h+1}(\phi^\star(x_{h+1}); \phi^\star(x^{(i)}_h), a)\right)^2} \\
&\leq \sqrt{\sum_{(i,a)\in[m]\times\mathcal{A}} \mathop{\mathbb{E}}_{x_{h+1}\sim\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}} \left(\widehat{w}_{h+1}(x_{h+1}; i, a) - w_{h+1}(\phi^\star(x_{h+1}); \phi^\star(x^{(i)}_h), a)\right)^2} \\
&\leq \sqrt{m|\mathcal{A}|\epsilon}
\end{aligned}
$$

as claimed. ∎

Lemma E.5 *Suppose that the event of Lemma E.4 holds. Let $\Gamma \subset \Pi$ be a finite set of policies, and let $c_{\mathsf{cov}} > 0$. Suppose that $\{\phi^\star(x^{(1)}_h), \dots, \phi^\star(x^{(m)}_h)\} \supseteq \mathcal{S}^{\mathsf{rch}}_h(\Gamma)$. Also suppose that $\widehat{\beta}_{h+1}(s) \geq c_{\mathsf{cov}} \cdot \max_{\pi\in\Pi} d^{\overline{M}(\Gamma),\pi}_{h+1}(s)$ for all $s \in \mathcal{S}$. Then for any $t \in [n]$ and $s^\star \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$, there is some $K = K(\Gamma, s^\star) \geq 1$ such that*

$$
\max_{\pi\in\Pi} \left|\mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] - K \cdot d^{\overline{M}(\Gamma),\pi}_{h+1}(s^\star)\right|
$$

$$\leq m\gamma + \frac{\sqrt{m|\mathcal{A}|\epsilon}}{c_{\mathsf{cov}}\gamma} + \frac{\max_{(i,a)\in[m]\times\mathcal{A}}\left|\widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1}; i, a) - w_{h+1}(s^\star; \phi^\star(x^{(i)}_h), a)\right|}{\gamma}$$

where $\mathcal{R}^{(t)}, \overline{x}^{(t)}_{h+1}$ are as defined in Algorithm 6, and we let $\mathcal{R}^{(t)}(\mathfrak{t}) := 0$.

Proof. Fix $\pi \in \Pi$. For $s, s' \in \mathcal{S}$, define

$$\Delta(s, s') := \max_{(i,a)\in[m]\times\mathcal{A}} |w_{h+1}(s; \phi^\star(x^{(i)}_h), a) - w_{h+1}(s'; \phi^\star(x^{(i)}_h), a)|$$

and

$$W^\pi := \mathbb{E}^{\overline{M}(\Gamma),\pi}\left[g(\Delta(s_{h+1}, s^\star))\mathbb{1}[s_{h+1} \in \mathcal{S}]\right].$$

We start by proving that $\mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})]$ is close to W^π for all $\pi \in \Pi$. Recall from Line 19 that for any $x_{h+1} \in \mathcal{X}$, we have

$$\mathcal{R}^{(t)}(x_{h+1}) = g\left(\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1}; i, a) - \widehat{w}_{h+1}(x_{h+1}; i, a)|\right)$$

where $g(z) = \max(0, 1 - z/\gamma)$. Thus, for any $s_{h+1} \in \mathcal{S}$, we have

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{x_{h+1}\sim\mathbb{O}_{h+1}(\cdot|s_{h+1})}\left[\mathcal{R}^{(t)}(x_{h+1}) - g\left(\Delta(s_{h+1}, s^\star)\right)\right]\right| \\
&\leq \frac{1}{\gamma}\mathop{\mathbb{E}}_{x_{h+1}\sim\mathbb{O}_{h+1}(\cdot|s_{h+1})}\left[\max_{(i,a)\in[m]\times\mathcal{A}}\left|\left|\widehat{w}_{h+1}(x_{h+1}; i, a) - \widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1}; i, a)\right|\right.\right. \\
&\qquad\qquad\left.\left. - \left|w_{h+1}(s_{h+1}; \phi^\star(x^{(i)}_h), a) - w_{h+1}(s^\star; \phi^\star(x^{(i)}_h), a)\right|\right|\right] \\
&\leq \frac{1}{\gamma}\underbrace{\mathop{\mathbb{E}}_{x_{h+1}\sim\mathbb{O}_{h+1}(\cdot|s_{h+1})}\left[\max_{(i,a)\in[m]\times\mathcal{A}}\left|\widehat{w}_{h+1}(x_{h+1}; i, a) - w_{h+1}(s_{h+1}; \phi^\star(x^{(i)}_h), a)\right|\right]}_{E_1(s_{h+1})} \\
&\quad + \frac{1}{\gamma}\underbrace{\max_{(i,a)\in[m]\times\mathcal{A}}\left|\widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1}; i, a) - w_{h+1}(s^\star; \phi^\star(x^{(i)}_h), a)\right|}_{E_2}
\end{aligned} \tag{28}$$

where the first inequality uses that g is $1/\gamma$-Lipschitz. It follows that

$$\begin{aligned}
\left|\mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] - W^\pi\right| &\leq \frac{1}{\gamma}\mathbb{E}^{\overline{M}(\Gamma),\pi}[E_1(s_{h+1})] + \frac{E_2}{\gamma} \\
&\leq \frac{1}{c_{\mathsf{cov}}\gamma}\mathop{\mathbb{E}}_{s_{h+1}\sim\widehat{\beta}_{h+1}}[E_1(s_{h+1})] + \frac{E_2}{\gamma} \\
&\leq \frac{\sqrt{m|\mathcal{A}|\epsilon}}{c_{\mathsf{cov}}\gamma} + \frac{E_2}{\gamma}
\end{aligned} \tag{29}$$

where the first inequality is by Eq. (28), the second inequality uses the assumption on $\widehat{\beta}_{h+1}$ and nonnegativity of E_1, and the third inequality uses Eq. (27).

Next, for any $s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$, note that $\mathbb{P}^{\overline{M}(\Gamma)}_{h+1}(s_{h+1} \mid s_h, a_h) = \mathbb{P}^{M}_{h+1}(s_{h+1} \mid s_h, a_h)$ and $\mathbb{P}^{\overline{M}(\Gamma)}_{h+1}(s_{h+1} \mid \mathfrak{t}, a_h) = 0$ for all $s_h \in \mathcal{S}$, $a_h \in \mathcal{A}$. Thus, the following equality holds for all $s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$:

$$
\begin{aligned}
d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1}) &= \sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h)\widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1} \mid s_h, a_h) \\
&= \left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h) w_{h+1}(s_{h+1}; s_h, a_h)\right) \sum_{j=1}^{m}\sum_{a\in\mathcal{A}} \widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1} \mid \phi^\star(x^{(j)}_h), a)
\end{aligned} \tag{30}
$$

where the final equality is by definition of $w_h(s_{h+1}; s_h, a_h)$ (Definition E.3). It follows from Eq. (30) and the fact that $d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1}) = 0$ for all $s_{h+1} \in \mathcal{S} \setminus \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$ that

$$
\begin{aligned}
W^\pi = \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} g(\Delta(s_{h+1}, s^\star)) &\left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h) w_{h+1}(s_{h+1}; s_h, a_h)\right) \\
&\cdot \left(\sum_{j=1}^{m}\sum_{a\in\mathcal{A}} \widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1} \mid \phi^\star(x^{(j)}_h), a)\right).
\end{aligned}
$$

Motivated by this expression, define

$$
\begin{aligned}
\widetilde{W}^\pi := \sum_{s_{h+1}\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} g(\Delta(s_{h+1}, s^\star)) &\left(\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h) w_{h+1}(s^\star; s_h, a_h)\right) \\
&\cdot \left(\sum_{j=1}^{m}\sum_{a\in\mathcal{A}} \widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1} \mid \phi^\star(x^{(j)}_h), a)\right).
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h) |w_{h+1}(s_{h+1}; s_h, a) - w_{h+1}(s^\star; s_h, a)| \\
&\leq \max_{s_h\in\mathcal{S}^{\mathsf{rch}}_{h}(\Gamma)} \max_{a\in\mathcal{A}} |w_{h+1}(s_{h+1}; s_h, a) - w_{h+1}(s^\star; s_h, a)| \\
&\leq \max_{(i,a)\in[m]\times\mathcal{A}} |w_{h+1}(s_{h+1}; \phi^\star(x^{(i)}), a) - w_{h+1}(s^\star; \phi^\star(x^{(i)}), a)| \\
&= \Delta(s_{h+1}, s^\star),
\end{aligned} \tag{31}
$$

where the first inequality uses the fact that $\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h) \leq 1$ and $d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h) = 0$ if $s \in \mathcal{S}\backslash\mathcal{S}^{\mathsf{rch}}_{h}(\Gamma)$, and the second inequality uses the assumption that $\mathcal{S}^{\mathsf{rch}}_{h}(\Gamma) \subseteq \{\phi^\star(x^{(1)}_h), \dots, \phi^\star(x^{(m)}_h)\}$. Using Eq. (31), we get

$$
\left|W^\pi - \widetilde{W}^\pi\right|
$$

$$
\begin{aligned}
&\leq \sum_{s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} g(\Delta(s_{h+1}, s^\star))\Delta(s_{h+1}, s^\star) \sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1} \mid \phi^\star(x_h^{(j)}), a) \\
&\leq \gamma \sum_{s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} \sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1} \mid \phi^\star(x_h^{(j)}), a) \\
&= \gamma m && (32)
\end{aligned}
$$

where the first inequality is by Eq. (31) and the second inequality uses the fact that $z \cdot g(z) \leq \gamma$ for all $z \geq 0$. Now from the definition of $\widetilde{W}^\pi$,

$$
\begin{aligned}
\widetilde{W}^\pi &= \left(\sum_{(s_h, a_h) \in \mathcal{S} \times \mathcal{A}} d_h^{\overline{M}(\Gamma), \pi}(s_h, a_h) w_{h+1}(s^\star; s_h, a_h) \right) \sum_{s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} g(\Delta(s_{h+1}, s^\star)) \sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1} \mid \phi^\star(x_h^{(j)}), a) \\
&= \frac{d_{h+1}^{\overline{M}(\Gamma), \pi}(s^\star)}{\sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s^\star \mid \phi^\star(x_h^{(j)}), a)} \cdot \sum_{s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} g(\Delta(s_{h+1}, s^\star)) \sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1} \mid \phi^\star(x_h^{(j)}), a) \\
&= K(\Gamma, s^\star) \cdot d_{h+1}^{\overline{M}(\Gamma), \pi}(s^\star) && (33)
\end{aligned}
$$

where the second equality uses Eq. (30) together with the assumption that $s^\star \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$, and in the final equality we have defined

$$
K(\Gamma, s^\star) := \frac{\sum_{s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)} g(\Delta(s_{h+1}, s^\star)) \sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s_{h+1} \mid \phi^\star(x_h^{(j)}), a)}{\sum_{j=1}^m \sum_{a \in \mathcal{A}} \widetilde{\mathbb{P}}^M_{h+1}(s^\star \mid \phi^\star(x_h^{(j)}), a)}.
$$

Note that $K(\Gamma, s^\star) \geq 1$, since $s^\star \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$ and $g(\Delta(s^\star, s^\star)) = g(0) = 1$. Combining Eq. (29), Eq. (32), and Eq. (33), we get that

$$
\left| \mathbb{E}^{\overline{M}(\Gamma), \pi}[\mathcal{R}^{(t)}(x_{h+1})] - K(\Gamma, s^\star) \cdot d_{h+1}^{\overline{M}(\Gamma), \pi}(s^\star) \right| \leq \frac{\sqrt{m|\mathcal{A}|\epsilon}}{c_{\mathsf{cov}}\gamma} + \frac{E_2}{\gamma} + \gamma m.
$$

Substituting in the definition of E_2 yields the result. ■

Proof of Theorem E.2. For any fixed $s \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma)$ and $i \in [m]$, we have by Item 1 of Lemma B.9 that

$$
\Pr[\phi^\star(x_h^{(i)}) = s] \geq \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2}.
$$

Let $\mathcal{E}_1$ be the event that for each $s \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma)$ there is some $i \in [m]$ with $\phi^\star(x_h^{(i)}) = s$. Then

$$
\Pr[\mathcal{E}_1] \geq 1 - |\mathcal{S}| \left(1 - \frac{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}{2}\right)^m \geq 1 - \delta
$$

by the assumption that $m \geq \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} \log(|\mathcal{S}|/\delta)$. Henceforth condition on $x_h^{(1)}, \dots, x_h^{(m)}$ and suppose that $\mathcal{E}_1$ holds.

For any $s_{h+1} \in \mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$, we have by definition of $\widehat{\beta}_{h+1}$ (Definition E.3) that

$$\begin{aligned}
\widehat{\beta}_{h+1}(s_{h+1}) &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1}|\phi^\star(x^{(j)}), a) \\
&\geq \frac{1}{m} \max_{s_h \in \mathcal{S}^{\mathsf{rch}}_h(\Gamma)} \mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})} \widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1}|s_h, a) \\
&\geq \frac{1}{m|\mathcal{A}|} \max_{(s_h,a_h)\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)\times\mathcal{A}} \widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1}|s_h, a_h) \\
&\geq \frac{1}{m|\mathcal{A}|} \max_{\pi\in\Pi} d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1})
\end{aligned} \tag{34}$$

where the first inequality uses $\mathcal{E}_1$ and the final inequality uses the fact that

$$d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1}) = \sum_{(s_h,a_h)\in\mathcal{S}^{\mathsf{rch}}_h(\Gamma)\times\mathcal{A}} d^{\overline{M}(\Gamma),\pi}_{h}(s_h,a_h)\widetilde{\mathbb{P}}^{M}_{h+1}(s_{h+1}|s_h,a_h).$$

Since $\max_{\pi\in\Pi} d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1}) = 0$ for all $s_{h+1} \in \mathcal{S}\setminus\mathcal{S}^{\mathsf{rch}}_{h+1}(\Gamma)$, in fact the bound

$$\widehat{\beta}_{h+1}(s_{h+1}) \geq \frac{1}{m|\mathcal{A}|}\max_{\pi\in\Pi} d^{\overline{M}(\Gamma),\pi}_{h+1}(s_{h+1}) \geq \frac{1}{m|\mathcal{A}|}\max_{\pi\in\Pi} d^{\overline{M}(\emptyset),\pi}_{h+1}(s_{h+1}) \tag{35}$$

holds for all $s_{h+1}\in\mathcal{S}$, where the last inequality is by Fact B.8.

Next, let $\mathcal{E}_2$ be the event that

$$\mathbb{E}_{x_{h+1}\sim\widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}} \max_{(i,a)\in[m]\times\mathcal{A}} \left|\widehat{w}_{h+1}(x_{h+1};i,a) - w_{h+1}(\phi^\star(x_{h+1});\phi^\star(x^{(i)}_h),a)\right| \leq \sqrt{m|\mathcal{A}|\epsilon}. \tag{36}$$

By Lemma E.4 and the assumption that $N \geq N_{\mathsf{reg}}(\epsilon,\delta)$, the event $\mathcal{E}_2$ occurs with probability at least $1 - m|\mathcal{A}|\delta$ over the randomness of $(\mathcal{D}_{i,a})_{i,a}$ and $\mathtt{Reg}$. Condition on $\widehat{w}_{h+1}(\cdot;\cdot,\cdot)$ and suppose that $\mathcal{E}_2$ holds.

For each $t\in[n]$, let $\mathcal{E}^t_3$ be the event that

$$\max_{(i,a)\in[m]\times\mathcal{A}} |\widehat{w}_{h+1}(\overline{x}^{(t)}_{h+1};i,a) - w_{h+1}(\phi^\star(\overline{x}^{(t)}_{h+1});\phi^\star(x^{(i)}_h),a)| \leq \epsilon^{1/4}.$$

Since $\overline{x}^{(t)}_{h+1} \sim \widetilde{\mathbb{O}}_{h+1}\widehat{\beta}_{h+1}$, we have by Markov's inequality and Eq. (36) that $\Pr[\neg\mathcal{E}^t_3] \leq \sqrt{m|\mathcal{A}|}\epsilon^{1/4}$. Define $\mathcal{E}_3 := \bigcap_{t=1}^n \mathcal{E}^t_3$. By the union bound, $\mathcal{E}_3$ occurs with probability at least $1-\sqrt{m|\mathcal{A}|}n\epsilon^{1/4}$ over the randomness of $\overline{x}^{(1)}_{h+1},\dots,\overline{x}^{(n)}_{h+1}$.

Also, for each $s\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\emptyset)$, let $t(s)\in[1,n]\cup\{\infty\}$ be the infimum over t such that $\phi^\star(\overline{x}^{(t)}_{h+1}) = s$, and let $\mathcal{E}^s_4$ be the event that $t(s) < \infty$. For each $t\in[1,n]$, since $\phi^\star(\overline{x}^{(t)}_{h+1})$ has distribution $\widehat{\beta}_{h+1}$, we have

$$\Pr[\phi^\star(\overline{x}^{(t)}_{h+1}) = s] = \widehat{\beta}(s) \geq \frac{1}{m|\mathcal{A}|}\max_{\pi\in\Pi} d^{\overline{M}(\emptyset),\pi}_{h+1}(s) \geq \frac{\sigma_{\mathsf{trunc}}}{m|\mathcal{A}|}$$

by Eq. (35) and Fact B.6. Thus, $\Pr[\neg\mathcal{E}^s_4] \leq (1-\sigma_{\mathsf{trunc}}/(m|\mathcal{A}|))^n \leq \delta/|\mathcal{S}|$ for any fixed $s\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\emptyset)$, by the assumption that $n \geq m|\mathcal{A}|\sigma^{-1}_{\mathsf{trunc}}\log(|\mathcal{S}|/\delta)$. Define $\mathcal{E}_4 := \bigcap_{s\in\mathcal{S}^{\mathsf{rch}}_{h+1}(\emptyset)} \mathcal{E}^s_4$. By the union bound, $\mathcal{E}_4$ occurs with probability at least $1-\delta$ over the randomness of $\overline{x}^{(1)}_{h+1},\dots,\overline{x}^{(n)}_{h+1}$.

Condition on $\overline{x}_{h+1}^{(1)}, \dots, \overline{x}_{h+1}^{(n)}$ and suppose that $\mathcal{E}_3 \cap \mathcal{E}_4$ holds. Note that the set $\mathcal{T}_{\mathsf{clus}}$ is now determined. For each $t \in \mathcal{T}_{\mathsf{clus}}$ let $\mathcal{E}_5^t$ be the event that

$$\mathbb{E}^{M,\widehat{\pi}^{(t)}}[\mathcal{R}^{(t)}(x_{h+1})] \geq \max_{\pi \in \Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t)}(x_{h+1})] - \frac{4H\sqrt{|\mathcal{A}|\epsilon}}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}$$

where $\widehat{\pi}^{(t)}$ is defined in Line 21. By the theorem assumptions on $\mathtt{Reg}$ and $\Psi_{1:h}$ and the fact that $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, Lemma G.1 gives $\Pr[\mathcal{E}_5^t] \geq 1 - H|\mathcal{A}|\delta$, so $\Pr[\mathcal{E}_5] \geq 1 - H|\mathcal{A}|n\delta$ where $\mathcal{E}_5 := \cap_{t \in \mathcal{T}_{\mathsf{clus}}} \mathcal{E}_5^t$. Condition on $\mathcal{E}_5$. We have now restricted to an event of total probability at least $1 - 2\delta - (m|\mathcal{A}| + H|\mathcal{A}|n)\delta - \sqrt{m|\mathcal{A}|}n\epsilon^{1/4}$; we argue that in this event, the properties claimed in the theorem statement hold.

Size of Ψ_{h+1}. First, we argue that $|\Psi_{h+1}| \leq |\mathcal{S}|$. Indeed, suppose that there are $t < t'$ with $t, t' \in \mathcal{T}_{\mathsf{clus}}$ and $\phi^\star(\overline{x}_{h+1}^{(t)}) = \phi^\star(\overline{x}_{h+1}^{(t')})$. By Line 20, we know that

$$\max_{(i,a) \in [m] \times \mathcal{A}} |\widehat{w}_{h+1}(\overline{x}_{h+1}^{(t)}; i, a) - \widehat{w}_{h+1}(\overline{x}_{h+1}^{(t')}; i, a)| > \gamma' = 2\epsilon^{1/4}.$$

But this contradicts $\mathcal{E}_3$ (in particular, the bounds implied by $\mathcal{E}_3^t$ and $\mathcal{E}_3^{t'}$ in combination with the triangle inequality and the assumption that $\phi^\star(\overline{x}_{h+1}^{(t)}) = \phi^\star(\overline{x}_{h+1}^{(t')})$). We conclude that indeed $|\Psi_{h+1}| \leq |\mathcal{S}|$.

Coverage of Ψ_{h+1}. It remains to prove the second property of the theorem statement. Fix $s \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$. Let $t^\star(s)$ be the minimal $t \in \mathcal{T}_{\mathsf{clus}}$ such that

$$\max_{(i,a) \in [m] \times \mathcal{A}} |\widehat{w}_{h+1}(\overline{x}_{h+1}^{(t(s))}; i, a) - \widehat{w}(\overline{x}_{h+1}^{(t)}; i, a)| \leq \gamma'. \tag{37}$$

Note that $t^\star(s)$ necessarily exists, because either $t(s) \in \mathcal{T}_{\mathsf{clus}}$ or, if not, Line 20 implies that some other $t \in \mathcal{T}_{\mathsf{clus}}$ satisfies the above bound. Next, by $\mathcal{E}_1$, Eq. (35), and Eq. (36), we can apply Lemma E.5 with coverage parameter $c_{\mathsf{cov}} := 1/(m|\mathcal{A}|)$, target state $s^\star = s$, and index $t := t^\star(s)$. We get that

$$\begin{aligned}
&\max_{\pi \in \Pi} \left| \mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - K(\Gamma, s) \cdot d_{h+1}^{\overline{M}(\Gamma),\pi}(s) \right| \\
&\leq m\gamma + \frac{(m|\mathcal{A}|)^{3/2}\sqrt{\epsilon}}{\gamma} + \frac{\max_{(i,a) \in [m] \times \mathcal{A}} \left| \widehat{w}_{h+1}(\overline{x}_{h+1}^{(t^\star(s))}; i, a) - w_{h+1}(s; \phi^\star(x_h^{(i)}), a) \right|}{\gamma} \\
&\leq m\gamma + \frac{(m|\mathcal{A}|)^{3/2}\sqrt{\epsilon}}{\gamma} + \frac{\epsilon^{1/4} + \max_{(i,a) \in [m] \times \mathcal{A}} \left| \widehat{w}_{h+1}(\overline{x}_{h+1}^{(t^\star(s))}; i, a) - \widehat{w}_{h+1}(\overline{x}_{h+1}^{(t(s))}; i, a) \right|}{\gamma} \\
&\leq m\gamma + \frac{(m|\mathcal{A}|)^{3/2}\sqrt{\epsilon}}{\gamma} + \frac{\epsilon^{1/4} + \gamma'}{\gamma} \\
&\leq \sigma_{\mathsf{trunc}}^2,
\end{aligned} \tag{38}$$

where the second inequality is by $\mathcal{E}_3$ and the fact that $s = \phi^\star(\overline{x}_{h+1}^{(t(s))})$, the third inequality is by Eq. (37), and the final inequality is by choice of γ, γ' and Eq. (26). We now distinguish two cases.

Case I. Suppose that

$$\mathbb{E}^{M,\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] \geq \mathbb{E}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] + \sigma_{\mathsf{trunc}}^2.$$

Then

$$\begin{aligned}
&d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(\mathfrak{t}) \\
&= \sum_{s\in\mathcal{S}} \left(d_{h+1}^{M,\widehat{\pi}^{(t^\star(s))}}(s) - d^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(s) \right) \\
&\geq \sum_{s\in\mathcal{S}} \left(d_{h+1}^{M,\widehat{\pi}^{(t^\star(s))}}(s) - d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(s) \right) \mathop{\mathbb{E}}_{x\sim\mathbb{O}_{h+1}(\cdot|s)}[\mathcal{R}^{(t^\star(s))}(x)] \\
&\geq \sigma_{\mathsf{trunc}}^2
\end{aligned}$$

where the equality is by the fact that $d_{h+1}^{M,\widehat{\pi}^{(t^\star(s))}}(\cdot)$ is a distribution supported on $\mathcal{S}$; the first inequality uses Fact B.8 and the fact that $\mathcal{R}^{(t^\star(s))}(x) \leq 1$ for all $x \in \mathcal{X}$. Thus $\max_{\pi\in\Psi_{h+1}} d^{\overline{M}(\Gamma),\pi}(\mathfrak{t}) \geq \sigma_{\mathsf{trunc}}^2$, so the second property of the theorem statement is satisfied.

Case II. Suppose that

$$\mathbb{E}^{M,\widehat{\pi}^{(t^\star(s))}\circ_h \mathrm{Unif}(\mathcal{A})}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] < \mathbb{E}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}\circ_h \mathrm{Unif}(\mathcal{A})}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] + \sigma_{\mathsf{trunc}}^2. \tag{39}$$

Now,

$$\begin{aligned}
K(\Gamma,s)\cdot d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(s) &\geq \mathbb{E}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - \sigma_{\mathsf{trunc}}^2 \\
&\geq \mathbb{E}^{M,\widehat{\pi}^{(t^\star(s))}}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - 2\sigma_{\mathsf{trunc}}^2 \\
&\geq \max_{\pi\in\Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[\mathcal{R}^{(t^\star(s))}(x_{h+1})] - 3\sigma_{\mathsf{trunc}}^2 \\
&\geq \max_{\pi\in\Pi} K(\Gamma,s)\cdot d_{h+1}^{\overline{M}(\Gamma),\pi}(s) - 4\sigma_{\mathsf{trunc}}^2 \\
&\geq K(\Gamma,s)(1-4\sigma_{\mathsf{trunc}}) \max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s)
\end{aligned}$$

where the first inequality is by Eq. (38), the second inequality is by Eq. (39), the third inequality is by $\mathcal{E}_5$ and Eq. (26), the fourth inequality is by Eq. (38), and the fifth inequality uses the fact that $\max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s) \geq \sigma_{\mathsf{trunc}}$ (Fact B.6 together with Fact B.8 and the fact that $s \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$) and $K(\Gamma,s) \geq 1$. Thus, since $\widehat{\pi}^{(t^\star(s))} \in \Psi_{h+1}$, we have

$$\begin{aligned}
\max_{\pi\in\Psi_{h+1}} d_{h+1}^{M,\pi}(s) &\geq \max_{\pi\in\Psi_{h+1}} d_{h+1}^{\overline{M}(\Gamma),\widehat{\pi}^{(t^\star(s))}}(s) \\
&\geq (1-4\sigma_{\mathsf{trunc}}) \max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\Gamma),\pi}(s) \\
&\geq (1-4\sigma_{\mathsf{trunc}}) \max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\emptyset),\pi}(s)
\end{aligned} \tag{40}$$

by two applications of Fact B.8. Now recall that $s \in \mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$ was arbitrary. Moreover, if $s \in \mathcal{S} \setminus \mathcal{S}_{h+1}^{\mathsf{rch}}(\emptyset)$ then $\max_{\pi\in\Pi} d_{h+1}^{\overline{M}(\emptyset),\pi}(s) = 0$, so Eq. (40) still holds. We conclude that Ψ_{h+1} is a $(1-4\sigma_{\mathsf{trunc}})$-truncated max policy cover for M at step $h+1$ (Definition B.3) as needed. ■

E.3. Analysis of PCR (Algorithm 5)

We can now use Theorem E.2 to complete the analysis of PCR, proving Theorem E.1 (and thus Theorem 4.1). After modularizing out the analysis of EPCR/EPCE, the analyses of PCR/PCE are essentially identical, so we omit the details here for brevity.

Proof of Theorem E.1. Fix the remaining inputs $\varepsilon_{\mathsf{final}}, \delta > 0$ to $\mathsf{PCE}(\mathtt{Reg}, N_{\mathsf{reg}}, |\mathcal{S}|, \cdot)$. The oracle time complexity bound and bound on N_{RL} are clear from the parameter choices and pseudocode, so long as $C_{E.1}$ is a sufficiently large constant. Moreover, it is immediate from the algorithm description that $|\Psi| \leq HR|\mathcal{S}| \leq H^2|\mathcal{S}|^2$. In order to show that the algorithm is $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient, it remains to argue that with probability at least $1-\delta$, Eq. (1) holds for all $h \in [H]$ and $s \in \mathcal{S}$, with parameter $\varepsilon_{\mathsf{final}}$.

Recall that $\sigma_{\mathsf{trunc}} = \varepsilon_{\mathsf{final}}/(4 + H|\mathcal{S}|)$. Fix some $1 \leq r \leq R$. For convenience, write $\alpha := \frac{1-4\sigma_{\mathsf{trunc}}}{|\mathcal{S}|}$. For each $h \in [H]$, let $\mathcal{E}_{h,r}$ be the event that $|\Psi_h^{(r)}| \leq |\mathcal{S}|$ and $\Psi_h^{(r)}$ is a $(1-4\sigma_{\mathsf{trunc}})$-truncated max policy cover for M at step h; let $\mathcal{F}_{h,r}$ be the event that $|\Psi_h^{(r)}| \leq |\mathcal{S}|$ and $\max_{\pi \in \Psi_h^{(r)}} d_h^{\overline{M}(\Gamma^{(r)}),\pi}(\mathfrak{t}) \geq \sigma_{\mathsf{trunc}}^2$. It's clear that $\Pr[\mathcal{E}_{1,r}] = 1$ (since $|\Psi_1^{(r)}| = 1$ and $d_1^{M,\pi_{\mathsf{unif}}}(s) = d_1^{M,\pi}(s)$ for all $s \in \mathcal{S}$ and $\pi \in \Pi$). Also, note that in the event $\mathcal{E}_{k,r}$, we have that $\Psi_k^{(r)}$ is an α-truncated policy cover for M at step k. Thus, by Theorem E.2 and choice of parameters (so long as C is a sufficiently large constant), we have for each $h \in \{2, \dots, H\}$ that

$$\Pr\left[\neg(\mathcal{E}_{h,r} \cup \mathcal{F}_{h,r}) \cap \bigcap_{1 \leq k < h} \mathcal{E}_{k,r}\right] \leq \Pr\left[\neg(\mathcal{E}_{h,r} \cup \mathcal{F}_{h,r}) \middle| \bigcap_{1 \leq k < h} \mathcal{E}_{k,r}\right] \leq \frac{\delta}{HR}. \tag{41}$$

The remainder of the proof is identical to that of Theorem C.1. ∎

Appendix F. Proofs from Section 5

In this section, we re-introduce Generalized Block MDPs, and verify (Proposition F.2) that generalized block MDPs are a special case of low-rank MDPs. Then, in Section F.2, we prove Theorem F.5, which asserts that one-context regression is computationally tractable for the concept class Φ_n defined by halfspaces. In Section F.3, we prove Theorem F.8, which asserts that reward-free RL for Φ_n-decodable generalized block MDPs is computationally *hard*. Together, Theorems F.5 and F.8 immediately imply Theorem 5.2. Finally, in Section F.4, we prove Proposition F.17, which asserts that reward-free RL in the family of Generalized Block MDPs (and hence Low-Rank MDPs) $\mathcal{M}_n$ is *statistically* tractable—and hence the preceding hardness result is purely a *computational* phenomenon.

F.1. Preliminaries

One-context low-rank regression. We will not actually need a formal definition of one-context low-rank regression in our results, but we introduce it formally for the sake of discussion:

Definition F.1 *Fix $N_{\mathsf{reg}} : (0, 1/2) \to \mathbb{N}$. An algorithm $\mathtt{Alg}$ is an N_{reg}-efficient* one-context low-rank regression *algorithm for a feature class $\Phi^{\mathsf{lin}} \subseteq (\mathcal{X} \times \mathcal{A} \to \mathbb{R}^d)$ if the following holds. Fix $\epsilon, \delta \in (0, 1/2)$, $n \in \mathbb{N}$, $\phi^{\mathsf{lin}} \in \Phi^{\mathsf{lin}}$, $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{A})$, and $\theta \in \mathbb{R}^d$. Let $(x^{(i)}, a^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $(x^{(i)}, a^{(i)}) \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x^{(i)}, a^{(i)}] = \langle \phi^{\mathsf{lin}}(x^{(i)}, a^{(i)}), \theta \rangle$. If $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$, then with probability at least $1 - \delta$, the output of $\mathtt{Alg}((x^{(i)}, a^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$ is a circuit $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to [0, 1]$ satisfying*

$$\mathop{\mathbb{E}}_{(x,a)\sim\mathcal{D}} (\mathcal{R}(x, a) - \langle \phi^{\mathsf{lin}}(x, a), \theta \rangle)^2 \leq \epsilon.$$

Generalized Block MDPs. Recall from Section 5 that for sets $\mathcal{S}, \mathcal{X}$ and a concept class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$, a *Generalized* Φ-decodable Block MDP is an MDP $M = (H, \mathcal{X}, \mathcal{A}, (\mathbb{P}_h)_h)$ with the property that there exists some $\phi_1^\star, \dots, \phi_H^\star \in \Phi$ so that $\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h)$ is a function of x_{h+1}, $\phi_h^\star(x_h)$, and a_h.

The following proposition makes formal the observation from Section 5 that Generalized Block MDPs (and, as a special case, Block MDPs themselves) are Low-Rank MDPs with an appropriate feature class:

Proposition F.2 *Set $d = |\mathcal{S}||\mathcal{A}|$ and identify $[d] \equiv \mathcal{S} \times \mathcal{A}$. Define $\Phi^{\mathsf{lin}} := \{\phi^{\mathsf{lin}} : \phi \in \Phi\}$, where $\phi^{\mathsf{lin}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is defined by*

$$\phi^{\mathsf{lin}}(x, a) := e_{\phi(x),a}.$$

Then any Generalized Φ-decodable Block MDP M is a Low-Rank MDP (Agarwal et al., 2020; Mhammedi et al., 2023a) with features $(\phi_h^\star)_h \subset \Phi^{\mathsf{lin}}$ and dual features $(\mu_h^\star)_h$ where $(\mu_h^\star)_{s,a} \in \Delta(\mathcal{X})$ for all $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$.

Proof. Let $\phi_1^\star, \dots, \phi_H^\star$ be the decoding functions for M. For each $h \in [H-1]$ and $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, we fix any $\overline{x}_h \in \mathcal{X}$ with $\phi_h^\star(\overline{x}_h) = s_h$ and define $\mu_{h+1}^\star(x_{h+1})_{s_h,a_h} = \mathbb{P}_{h+1}(x_{h+1} \mid \overline{x}_h, a_h)$. Then for any $x_h, x_{h+1} \in \mathcal{X}$ with $\phi_h^\star(x_h) = s_h$, we have

$$\mathbb{P}_{h+1}(x_{h+1} \mid x_h, a_h) = \mathbb{P}_{h+1}(x_{h+1} \mid \overline{x}_h, a_h)$$

$$= \langle e_{s_h,a_h}, \mu^\star_{h+1}(x_{h+1})\rangle$$
$$= \langle \phi^{\mathsf{lin}}_h(x_h, a_h), \mu^\star_{h+1}(x_{h+1})\rangle$$

where $\phi^{\mathsf{lin}}_h \in \Phi^{\mathsf{lin}}$ is defined by $\phi^{\mathsf{lin}}_h(x, a) := e_{\phi^\star_h(x),a}$. The first equality above used the Generalized Block MDP assumption. ■

The following proposition asserts that for Generalized Block MDPs, when viewed as a special case of Low-Rank MDPs, the one-context *low-rank* regression problem reduces to one-context regression. Since this fact is only important for purposes of motivation and discussion, we omit the formal statement (i.e. parameters of the reduction), and simply remark that the reduction proceeds by independently solving a one-context regression problem for each action. The proof of correctness would proceed similarly to e.g. that of Proposition G.8.

Proposition F.3 (Informal) *For any concept class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ and action space $\mathcal{A}$, let $\Phi^{\mathsf{lin}} \subseteq (\mathcal{X} \times \mathcal{A} \to \mathbb{R}^d)$ be the feature class defined in terms of Φ as specified in Proposition F.2. Then there is a polynomial-time reduction from one-context low-rank regression for Φ^{lin} (Definition F.1) to one-context regression for Φ (Definition 5.1).*

For the results in Section F.3, it is convenient to introduce the following class of Φ_n-decodable Generalized Block MDPs, where Φ_n is the class of linear threshold functions, as previously defined in Section 5.

Definition F.4 *Fix $n \in \mathbb{N}$. We define $\mathcal{M}_n$ to be the family of generalized Φ_n-decodable block MDPs with horizon $H := (\log n)^{\log \log n}$, observation space $\mathcal{X} = \mathbb{R}^n$, latent state space $\mathcal{S} = \{0, 1\}$, action space $\mathcal{A} = \{0, 1\}$, and feature class $\Phi_n = \{\phi^\theta : \theta \in \mathbb{R}^n\}$ consisting of linear threshold functions, i.e. where*

$$\phi^\theta(x) := \mathbb{1}[\langle x, \theta\rangle \geq 0].$$

The hardness result in Theorem 5.2 will be against (a subset of) $\mathcal{M}_n$. Observe that $H \leq n$ and $|\mathcal{A}| = 2$, so to rule out an RL algorithm for Φ_n-decodable Generalized Block MDPs with time complexity $\mathrm{poly}(n, H, |\mathcal{A}|)$, it suffices to rule out an RL algorithm for $\mathcal{M}_n$ with time complexity $\mathrm{poly}(n)$.

F.2. Tractability of One-Context Regression

We start by showing that one-context regression with concept class Φ_n can be solved with time complexity $\mathrm{poly}(n)$. Note that n is the description length of an element of the observation space $\mathcal{X}$ (ignoring issues of finite precision arithmetic). To prove Theorem F.5, recall that PAC learning of halfspaces with random classification noise is computationally tractable (Blum et al., 1998; Diakonikolas et al., 2023). The only difference between this problem and one-context regression for Φ_n is that in the latter setting, the noise levels for the two classes may be different. To reduce the latter to the former, we apply a careful symmetrization step; a priori this seems to require knowing the noise levels, but this can be avoided by gridding over all possibilities.

Theorem F.5 *There is a constant $C_{F.5} > 0$ so that the following holds. There is an N_{reg}-efficient algorithm for one-context regression with concept class Φ_n, where*

$$N_{\mathsf{reg}}(\epsilon, \delta) = (n/\epsilon)^{C_{F.5}} \log^2(1/\delta).$$

Moreover, the time complexity of the algorithm with error tolerance ϵ and failure probability δ is $\mathrm{poly}(n, 1/\epsilon, 1/\delta)$.

Proof. We define a one-context regression algorithm `Alg` as follows. We are given, as input, samples $(x^{(i)}, y^{(i)})_{i=1}^m$ with $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \{0,1\}$, and parameters $\epsilon, \delta \in (0, 1/2)$. Set $\varepsilon_{\mathsf{disc}} := c_{\mathsf{disc}}\delta/m$ for a universal constant $c_{\mathsf{disc}} > 0$ that will be determined in the analysis. Let $\mathcal{G}$ be a set of functions $g : \{0,1\} \to [0,1]$ such that $|\mathcal{G}| \leq O(1/\varepsilon_{\mathsf{disc}}^2)$ and for any $(a,b) \in [0,1]^2$ there is some $g \in \mathcal{G}$ with $|a - g(0)| + |b - g(1)| \leq \varepsilon_{\mathsf{disc}}$.

For each $g \in \mathcal{G}$, we compute a predictor $\mathcal{R}_g : \mathbb{R}^n \to [0,1]$ as follows. First, we define $(y_g^{(i)})_{i=1}^{m/2}$ as follows:

- If $g(0) + g(1) \geq 1$, then draw

$$y_g^{(i)} \sim \begin{cases} \mathrm{Ber}\left(\frac{g(0)+g(1)-1}{g(0)+g(1)}\right) & \text{if } y^{(i)} = 0 \\ \mathrm{Ber}(1) & \text{if } y^{(i)} = 1 \end{cases}.$$

- If $g(0) + g(1) < 1$, then draw

$$y_g^{(i)} \sim \begin{cases} \mathrm{Ber}(0) & \text{if } y^{(i)} = 0 \\ \mathrm{Ber}\left(\frac{1-g(0)-g(1)}{2-g(0)+g(1)}\right) & \text{if } y^{(i)} = 1 \end{cases}.$$

Let `HalfspaceLearn` denote the halfspace learning algorithm guaranteed by (Diakonikolas et al., 2023, Theorem 1.8). Invoke `HalfspaceLearn` on dataset $(x^{(i)}, y_g^{(i)})_{i=1}^{m/2}$, which returns a hypothesis $h_g^+ : \mathbb{R}^n \to \{0,1\}$. Define $\mathcal{R}_g^+ : \mathbb{R}^n \to [0,1]$ by $\mathcal{R}_g^-(x) = g(h_g^+(x))$. Similarly define h_g^- by invoking `HalfspaceLearn` on $(x^{(i)}, 1 - y_g^{(i)})_{i=1}^{m/2}$, and define $\mathcal{R}_g^- : \mathbb{R}^n \to [0,1]$ by $\mathcal{R}_g^-(x) = g(h_g^-(x))$.

Finally, output $\mathcal{R}_{\widehat{g}}^{\widehat{b}}$ with $\widehat{g}, \widehat{b}$ defined by

$$(\widehat{g}, \widehat{b}) := \operatorname*{arg\,min}_{g \in \mathcal{G}, b \in \{-,+\}} \sum_{i=m/2+1}^{m} (\mathcal{R}_g^b(x^{(i)}) - y^{(i)})^2.$$

Analysis. Let $m \geq (n/\epsilon)^{C_{F.5}} \log^2(1/\delta)$. Suppose that $(x^{(i)}, y^{(i)})_{i=1}^m$ are i.i.d. samples with $\mathbb{E}[y^{(i)}|x^{(i)}] = f(\phi(x^{(i)}))$ for some $\phi \in \Phi$ and $f : \{0,1\} \to [0,1]$. Suppose that $f(0) + f(1) \geq 1$; the argument in the other case is similar. By construction, there is some $g \in \mathcal{G}$ such that $g(0) + g(1) \geq 1$ and $|f(0) - g(0)| + |f(1) - g(1)| \leq O(\varepsilon_{\mathsf{disc}})$. We distinguish two cases.

1. First, if $|f(0) - f(1)| \leq \epsilon$, then $|g(0) - g(1)| \leq O(\epsilon)$ (since $\varepsilon_{\mathsf{disc}} \leq \epsilon$), and thus

$$|\mathcal{R}_g^+(x) - f(\phi(x))| \leq \max_{b,b' \in \{0,1\}} |g(b) - f(b')| \leq O(\epsilon)$$

for all x.

2. Second, suppose that $|f(0) - f(1)| \geq \epsilon$. Further suppose that $f(0) > f(1)$. Fix $i \in [m/2]$ and condition on $x^{(i)}$. If $\phi(x^{(i)}) = 0$, then

$$\Pr[y_g^{(i)} = 0] = \frac{1}{g(0) + g(1)} \Pr[y^{(i)} = 1]$$

$$= \frac{f(0)}{g(0)+g(1)}.$$

If $\phi(x^{(i)}) = 1$, then

$$\begin{aligned}\Pr[y_g^{(i)} = 1] &= \frac{g(0)+g(1)-1}{g(0)+g(1)} \Pr[y^{(i)} = 1] + \Pr[y^{(i)} = 0] \\ &= \frac{g(0)+g(1)-1}{g(0)+g(1)} f(1) + 1 - f(1) \\ &= \frac{g(0)+g(1)-f(1)}{g(0)+g(1)}.\end{aligned}$$

In particular, for each $b \in \{0,1\}$, if $\phi(x^{(i)}) = b$, then

$$\left| \Pr[y_g^{(i)} = b] - \frac{f(0)}{f(0)+f(1)} \right| \le O(\varepsilon_{\mathsf{disc}}).$$

Consider the alternative (idealized) dataset $(x^{(i)}, \widetilde{y}^{(i)})_{i=1}^{m/2}$ where

$$\Pr[\widetilde{y}^{(i)} = \phi(x^{(i)}) \mid x^{(i)}] = \frac{f(0)}{f(0)+f(1)}.$$

Then the total variation distance between $\mathrm{Law}((x^{(i)}, y_g^{(i)})_{i=1}^{m/2})$ and $\mathrm{Law}((x^{(i)}, \widetilde{y}^{(i)})_{i=1}^{m/2})$ is at most $O(\varepsilon_{\mathsf{disc}} m) \le \delta$, where the inequality holds by definition of $\varepsilon_{\mathsf{disc}}$, so long as $c_{\mathsf{disc}} > 0$ is chosen sufficiently small. Moreover, $(x^{(i)}, \widetilde{y}^{(i)})_{i=1}^{m/2}$ is an instance of the halfspace learning problem with noise level $\eta := \frac{f(1)}{f(0)+f(1)}$. Let $\widetilde{h}^+ : \mathbb{R}^n \to \{0,1\}$ denote the output of `HalfspaceLearn` on $(x^{(i)}, \widetilde{y}^{(i)})_{i=1}^{m/2}$. By the guarantee of (Diakonikolas et al., 2023, Theorem 1.8), so long as $C_{F.5}$ is a sufficiently large constant, we have that with probability at least $1-\delta$,

$$\mathrm{Pr}_{x,\widetilde{y}}[\widetilde{h}(x) \neq \widetilde{y}] \le \eta + \frac{\epsilon^2}{4},$$

where $(x, \widetilde{y})$ is a fresh sample from the same distribution. Now

$$\begin{aligned}\mathrm{Pr}_{x,\widetilde{y}}[\widetilde{h}(x) \neq \widetilde{y}] &= (1-\eta)\,\mathrm{Pr}_x[\widetilde{h}(x) \neq \phi(x)] + \eta(1 - \mathrm{Pr}_x[\widetilde{h}(x) \neq \phi(x)]) \\ &= \eta + (1-2\eta)\,\mathrm{Pr}_x[\widetilde{h}(x) \neq \phi(x)].\end{aligned}$$

Observe that $1 - 2\eta \ge \epsilon/2$ (since $f(0) - f(1) \ge \epsilon$). Therefore it holds with probability at least $1-\delta$ that

$$\mathrm{Pr}_x[\widetilde{h}(x) \neq \phi(x)] \le \frac{\mathrm{Pr}_{x,\widetilde{y}}[\widetilde{h}(x) \neq \widetilde{y}] - \eta}{1-2\eta} \le \epsilon.$$

It follows from the preceding total variation bound and the data processing inequality that in an event $\mathcal{E}_g^+$ that occurs with probability at least $1-2\delta$, the estimator h_g^+ (which our algorithm actually computes) satisfies

$$\mathrm{Pr}_x[h_g^+(x) \neq \phi(x)] \le \epsilon.$$

Moreover, in event $\mathcal{E}_g^+$, we have

$$\mathop{\mathbb{E}}_x(\mathcal{R}_g^+(x) - f(\phi(x)))^2 = \mathop{\mathbb{E}}_x(g(h_g^+(x)) - f(\phi(x)))^2 \le O(\varepsilon_{\mathsf{disc}}^2) + \epsilon \le O(\epsilon).$$

Recall that this holds under the assumption that $f(0) > f(1)$. If $f(0) < f(1)$, then by a similar argument there is an event $\mathcal{E}_g^-$ that occurs with probability at least $1 - 2\delta$, in which

$$\mathop{\mathbb{E}}_x(\mathcal{R}_g^-(x) - f(\phi(x)))^2 \le O(\epsilon).$$

In either case, there is an event $\mathcal{E}$ that holds with probability at least $1 - 2\delta$, in which

$$\min_{g\in\mathcal{G}, b\in\{-,+\}} \mathop{\mathbb{E}}_x(\mathcal{R}_g^b(x) - f(\phi(x)))^2 \le O(\epsilon).$$

Next, notice that

$$m \ge \epsilon^{-2}\log(4m^2/(c_{\mathsf{disc}}^2\delta^3)) = \epsilon^{-2}\log(4|\mathcal{G}|/\delta)$$

so long as $C_{F.5}$ is a sufficiently large constant. Therefore by Hoeffding's inequality and a union bound over $\mathcal{G} \times \{-,+\}$, there is an event $\mathcal{E}'$ that occurs with probability at least $1 - \delta$ in which, for all $g \in \mathcal{G}$ and $b \in \{-,+\}$,

$$\left| \frac{2}{m} \sum_{i=m/2+1}^{m} (\mathcal{R}_g^b(x^{(i)} - y^{(i)}))^2 - \mathop{\mathbb{E}}_{x,y}(\mathcal{R}_g^b(x) - y)^2 \right| \le \epsilon.$$

Condition on the event $\mathcal{E} \cap \mathcal{E}'$. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_x(\mathcal{R}_{\widehat{g}}^{\widehat{b}}(x) - f(\phi(x)))^2 &= \mathop{\mathbb{E}}_{x,y}(\mathcal{R}_{\widehat{g}}^{\widehat{b}}(x) - y)^2 - \mathop{\mathbb{E}}_{x,y}(f(\phi(x)) - y)^2 \\
&\le \epsilon + \frac{2}{m}\sum_{i=m/2+1}^{m} (\mathcal{R}_{\widehat{g}}^{\widehat{b}}(x^{(i)} - y^{(i)}))^2 - \mathop{\mathbb{E}}_{x,y}(f(\phi(x)) - y)^2 \\
&\le \epsilon + \min_{g\in\mathcal{G}, b\in\{-,+\}} \frac{2}{m}\sum_{i=m/2+1}^{m} (\mathcal{R}_g^b(x^{(i)} - y^{(i)}))^2 - \mathop{\mathbb{E}}_{x,y}(f(\phi(x)) - y)^2 \\
&\le 2\epsilon + \min_{g\in\mathcal{G}, b\in\{-,+\}} \mathop{\mathbb{E}}_{x,y}(\mathcal{R}_g^b(x) - y)^2 - \mathop{\mathbb{E}}_{x,y}(f(\phi(x)) - y)^2 \\
&= 2\epsilon + \min_{g\in\mathcal{G}, b\in\{-,+\}} \mathop{\mathbb{E}}_{x,y}(\mathcal{R}_g^b(x) - f(\phi(x)))^2 \\
&\le O(\epsilon).
\end{aligned}$$

Rescaling ϵ, δ by the appropriate constants yields the desired bound. ∎

Remark F.6 *In fact, the reduction in Theorem F.5 does not use any special properties of halfspaces. Thus, it shows that for* any *binary concept class Φ, one-context regression is polynomial-time reducible to PAC learning Φ with random classification noise.*

F.3. Hardness of Reward-Free RL with Resets

Next, we prove Theorem F.8, which asserts that reward-free RL in $\mathcal{M}_n$ with the reset access model is computationally intractable under the *Continuous LWE* assumption formally defined below.

Together with Theorem F.5, this completes the proof of Theorem 5.2. Throughout this section, we write $\mathfrak{S}^{t-1}$ to denote the set of Euclidean unit vectors in $\mathbb{R}^t$.

Assumption F.7 (Continuous LWE (Bruna et al., 2021)) *Let $\delta \in (0,1)$. For $\beta = \beta(t) := 1/t$ and $\gamma = \gamma(t) := 2\sqrt{t}$, for any time-$2^{t^\delta}$ algorithm $\mathtt{Alg}^{\mathcal{O}}$ with sampling oracle $\mathcal{O}$ and outputs in $[0,1]$, it holds that*

$$\left| \mathop{\mathbb{E}}_{w \sim \mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\mathtt{Alg}^{C_{w,\beta,\gamma}}] - \mathbb{E}[\mathtt{Alg}^{N(0,\frac{1}{2\pi}I_t)}] \right| \leq t^{-\omega(1)},$$

where $C_{w,\beta,\gamma}$ is the continuous LWE distribution *with secret $w \in \mathfrak{S}^{t-1}$ and parameters $\beta, \gamma > 0$.*

Recent work has shown that this assumption is well-founded—in particular, it can be based on hardness of discrete LWE (Gupte et al., 2022) and therefore on worst-case hardness of the approximate shortest vector problem (Brakerski et al., 2013).

Theorem F.8 (One-context regression is insufficient for Low-Rank MDPs with resets) *Suppose that $\mathtt{Alg}^M$ is an algorithm that, given interactive reset access to any MDP $M \in \mathcal{M}_n$, has time complexity $\mathrm{poly}(n)$ and produces a set of policies Ψ satisfying the following guarantee with probability at least $1/2$:*

$$\forall s \in \mathcal{S} : \max_{\pi \in \Psi} d_H^{M,\pi}(s) \geq \frac{1}{\mathrm{poly}(|\mathcal{A}|, |\mathcal{S}|, H)} \left(\max_{\pi \in \Pi} d_H^{M,\pi}(s) - \frac{1}{8} \right). \tag{42}$$

Then the Continuous LWE hardness assumption (Assumption F.7) is false.

Since the generalized Block MDPs in $\mathcal{M}_n$ have horizon $H \leq n$ and action space of size $O(1)$, this result rules out algorithms with time complexity $\mathrm{poly}(n, H, |\mathcal{A}|)$, and therefore proves the claimed hardness result in Theorem 5.2. Also, the description length of each $x \in \mathcal{X} = \mathbb{R}^n$ is $O(n)$ (again, ignoring issues of bit complexity). Thus, Theorem F.8, in conjunction with Theorem F.5, rules out any reduction from reward-free RL with resets (in Generalized Block MDPs) to one-context regression that is efficient, i.e. that has oracle time complexity $\mathrm{poly}(H, |\mathcal{S}|, |\mathcal{A}|, \epsilon^{-1}, \delta^{-1}, n)$.

Proof overview. To prove Theorem F.8, we need the following result, which we prove (in Section F.3.1) by unpacking various proofs from Tiegel (2023). Essentially, it states that there are two parametric families of distributions $\{\nu_{w,0} : w \in \mathfrak{S}^{t-1}\}$ and $\{\nu_{w,1} : w \in \mathfrak{S}^{t-1}\}$ that are each computationally indistinguishable from some null distribution under Continuous LWE (Item 1), and yet for each w, the two corresponding distributions are approximately separated by a halfspace (Item 2).

Theorem F.9 *Suppose Assumption F.7 holds. Let $n \in \mathbb{N}$ and define $t := \frac{\log^2 n}{\log^5(\log n)}$. There is a distribution ν_{null} and families $\{\theta(w) : w \in \mathfrak{S}^{t-1}\}$, $\{\nu_{w,0} : w \in \mathfrak{S}^{t-1}\}$, $\{\nu_{w,1} : w \in \mathfrak{S}^{t-1}\}$ where $\theta(w) \in \mathbb{R}^n$, $\nu_{w,0}, \nu_{w,1} \in \Delta(\mathbb{R}^n)$ are distributions, and the following properties hold:*

1. *For every algorithm* $\texttt{Alg}^{\mathcal{O}_0,\mathcal{O}_1}$ *with time complexity* $\mathrm{poly}(n)$*, access to sampling oracles* $\mathcal{O}_0, \mathcal{O}_1$*, and outputs in* $[0,1]$*,*

$$\left| \mathop{\mathbb{E}}_{w \sim \mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\texttt{Alg}^{\nu_{w,0},\nu_{w,1}}] - \mathbb{E}[\texttt{Alg}^{\nu_{\mathsf{null}},\nu_{\mathsf{null}}}] \right| \le n^{-\omega(1)}.$$

2. *There is* $\gamma(n) \le (\log n)^{-\Omega(\log^2 \log n)}$ *so that for every* $w \in \mathfrak{S}^{t-1}$*,*

$$\mathrm{Pr}_{x \sim \nu_{w,1}}[\langle x, \theta(w) \rangle < 0] \le \gamma(n)$$

and

$$\mathrm{Pr}_{x \sim \nu_{w,0}}[\langle x, \theta(w) \rangle \ge 0] \le \gamma(n)$$

and

$$\mathrm{Pr}_{x \sim \nu_{\mathsf{null}}}[\langle x, \theta(w) \rangle \ge 0] \ge 1/2.$$

3. *There is a* $\mathrm{poly}(n)$*-time algorithm that takes as input* $w \in \mathfrak{S}^{t-1}$ *and* $x \in \mathbb{R}^n$*, and outputs* $\mathrm{sgn}(\langle x, \theta(w) \rangle)$.

4. *There is a* $\mathrm{poly}(n)$*-time algorithm that takes as input* $w \in \mathfrak{S}^{t-1}$ *and* $a \in \{0,1\}$*, and outputs* $x \in \mathbb{R}^n$ *where* $\mathrm{Law}(x)$ *is* $n^{-\omega(1)}$*-close to* $\nu_{w,a}$ *in total variation distance.*

5. *There is a* $\mathrm{poly}(n)$*-time algorithm that samples from* ν_{null}.

We exploit these distribution families by designing a family of *approximate* combination lock MDPs within $\mathcal{M}_n$. Concretely, for each sequence of vectors $(w_1, \dots, w_H)$ and action sequence $(a_1^\star, \dots, a_H^\star)$ we design a Φ_n-decodable MDP $M^{:H}_{\mathbf{w},\mathbf{a}^\star}$, defined formally below, for which playing an action sequence similar to $\mathbf{a}^\star$ is necessary in order to reach the latent state $\phi^{\theta(w_H)}(x_H) = 1$ with decent probability. This idea is formalized in Lemmas F.12 and F.13. We then use the indistinguishability property of the distributions families (Item 1) together with a careful hybrid argument to prove that any computationally efficient RL algorithm has similar behavior on $M^{:H}_{\mathbf{w},\mathbf{a}^\star}$ as on a "null MDP" that is independent of $\mathbf{w}$ and $\mathbf{a}^\star$. In particular, Lemma F.15 shows that the transitions of $M^{:H}_{\mathbf{w},\mathbf{a}^\star}$ can be replaced with "null" transitions layer-by-layer, starting with the last layer, and no efficient algorithm can detect the difference under Assumption F.7. We then complete the proof of Theorem F.8 by observing that an RL algorithm which is independent of $\mathbf{a}^\star$ cannot always play a policy similar to $\mathbf{a}^\star$.

Definition F.10 (MDPs for hardness construction) *Set* $H := (\log n)^{\frac{1}{3} \log\log n}$*. For any* $\mathbf{w} = (w_1, \dots, w_H) \in (\mathfrak{S}^{t-1})^H$ *and* $\mathbf{a}^\star = (a_1^\star, \dots, a_H^\star) \in \mathcal{A}^H$ *and* $k \in [H]$*, we define an MDP* $M^{:k}_{\mathbf{w},\mathbf{a}}$ *with observation space* $\mathcal{X} = \mathbb{R}^n$*, action space* $\mathcal{A} = \{0,1\}$*, and horizon* $H = (\log n)^{\log\log n}$ *as follows. The initial distribution is* ν_{null}*. For each* $x_h \in \mathcal{X}$ *and* $a_h \in \mathcal{A}$ *and* $h \in [H-1]$*, the transition distribution is*

$$\mathbb{P}^{M^{:k}_{\mathbf{w},\mathbf{a}}}_{h+1}(x_{h+1}|x_h, a_h) := \begin{cases} \nu_{w_{h+1},1}(x_{h+1}) & \text{if } \langle x_h, \theta(w_h) \rangle \ge 0 \text{ and } a_h = a_h^\star \text{ and } h < k \\ \nu_{w_{h+1},0}(x_{h+1}) & \text{if } (\langle x_h, \theta(w_h) \rangle < 0 \text{ or } a_h \ne a_h^\star) \text{ and } h < k \\ \nu_{\mathsf{null}} & \text{if } h \ge k \end{cases}.$$

Fact F.11 *For any* $\mathbf{w} = w_{1:H} \in (\mathfrak{S}^{t-1})^H$, $\mathbf{a}^\star = a^\star_{1:H} \in \mathcal{A}^H$, *and* $k \in [H]$, *it holds that* $M^{:k}_{\mathbf{w},\mathbf{a}} \in \mathcal{M}_n$.

The following lemma shows that the fixed action sequence $\mathbf{a}^\star$ is a policy that reaches the latent state $\phi^{\theta(w_H)}(x_H) = 1$ with constant probability.

Lemma F.12 *For any* $\mathbf{w} = w_{1:H} \in (\mathfrak{S}^{t-1})^H$ *and* $\mathbf{a}^\star = a^\star_{1:H} \in \mathcal{A}^H$,

$$\max_{\pi \in \Pi} \Pr^{M^{:H}_{\mathbf{w},\mathbf{a}},\pi}[\phi^{\theta(w_H)}(x_H) = 1] \geq \frac{1}{4}$$

so long as n *is sufficiently large.*

Proof. For notational convenience, write $M := M^{:H}_{\mathbf{w},\mathbf{a}^\star}$. Consider the policy π defined by $\pi_h(x_h) := a^\star_h$ for all $h \in [H]$ and $x_h \in \mathcal{X}$. The initial distribution of M is ν_{null}, so $\Pr^{M,\pi}[\phi^{\theta(w_1)}(x_1) = 1] \geq 1/2$ by Item 2 of Theorem F.9. For each $h < H$, conditioned on $x_h \in \mathcal{X}$ with $\phi^{\theta(w_h)}(x_h) = 1$, the distribution of x_{h+1} under policy π is $\nu_{w_{h+1},1}$, so

$$\Pr^{M,\pi}[\phi^{\theta(w_{h+1})}(x_{h+1}) = 1 \mid \phi^{\theta(w_h)}(x_h) = 1] \geq 1 - \gamma(n).$$

Therefore by induction,

$$\Pr^{M,\pi}[\phi^{\theta(w_H)}(x_H) = 1] \geq \frac{1}{2}(1 - \gamma(n))^{H-1} \geq \frac{1}{4}$$

where the final inequality holds for all sufficiently large n, since $H = H(n) = o(\gamma(n))$. ■

The following lemma shows that any policy that reaches $\phi^{\theta(w_H)}(x_H) = 1$ with decent probability must also often play the action sequence $\mathbf{a}^\star$.

Lemma F.13 *For any* $\mathbf{w} = w_{1:H} \in (\mathfrak{S}^{t-1})^H$, $\mathbf{a}^\star = a^\star_{1:H} \in \mathcal{A}^H$, *and policy* $\pi \in \Pi$,

$$\Pr^{M^{:H}_{\mathbf{w},\mathbf{a}^\star},\pi}[\forall h \in [H-1] : \pi_h(x_h) = a^\star_h] \geq \Pr^{M^{:H}_{\mathbf{w},\mathbf{a}^\star},\pi}\left[\phi^{\theta(w_H)}(x_H) = 1\right] - H^2\gamma(n).$$

Proof. For notational convenience, write $M = M^{:H}_{\mathbf{w},\mathbf{a}^\star}$. For each $h \in [H-1]$ let $\mathcal{E}_h$ be the event that $\pi_h(x_h) \neq a^\star_h$. Then

$$\Pr\left[\left(\phi^{\theta(w_H)}(x_H) = 1\right) \wedge (\pi_h(x_h) \neq a^\star_h)\right] \leq \Pr\left[\phi^{\theta(w_H)}(x_H) = 1 \middle| \pi_h(x_h) \neq a^\star_h\right] \leq H\gamma(n)$$

since conditioned on any $x_h \in \mathcal{X}$ with $\pi_h(x_h) \neq a^\star_h$, x_{h+1} is distributed according to $\nu_{w_{h+1},0}$, and thus $\Pr^{M,\pi}[\phi^{\theta(w_{h+1})}(x_{h+1}) = 1 \mid x_h] \leq \gamma(n)$ (Item 2 of Theorem F.9), and at each subsequent step $k \geq h+1$, conditioned on any $x_k \in \mathcal{X}$ with $\phi^{\theta(w_k)}(x_k) = 0$, x_{k+1} is distributed according ot $\nu_{w_k,0}$ so once again $\Pr^{M,\pi}[\phi^{\theta(w_{k+1})}(x_{k+1}) = 1 \mid x_k] \leq \gamma(n)$.

By the union bound,

$$\Pr\left[\left(\phi^{\theta(w_H)}(x_H) = 1\right) \wedge (\exists h \in [H-1] : \pi_h(x_h) \neq a^\star_h)\right] \leq H^2\gamma(n).$$

Therefore

$$\begin{aligned}&\Pr^{M,\pi}[\forall h \in [H-1] : \pi_h(x_h) = a^\star_h]\\ &\geq \Pr^{M,\pi}\left[\left(\phi^{\theta(w_H)}(x_H) = 1\right) \wedge (\forall h \in [H-1] : \pi_h(x_h) = a^\star_h)\right]\\ &\geq \Pr^{M,\pi}\left[\phi^{\theta(w_H)}(x_H) = 1\right] - H^2\gamma(n)\end{aligned}$$

as needed. ■

We now implement a hybrid argument to show that any efficient RL algorithm has similar behavior on $M^{:1}_{\mathbf{w},\mathbf{a}^\star}, \dots, M^{:H}_{\mathbf{w},\mathbf{a}^\star}$. In particular, we need that the *value* of its output is similar on these different MDPs, where we define value as the maximum agreement probability (over all policies output by the algorithm) with the true action sequence $\mathbf{a}^\star$:

Definition F.14 *For an MDP M, set of policies Ψ, and action sequence $\mathbf{a}^\star$, define*

$$\mathrm{Val}^M_{\mathbf{a}^\star}(\Psi) = \max_{\pi \in \Psi} \Pr^{M,\pi}[\forall h \in [H-1] : \pi_h(x_h) = a^\star_h].$$

Lemma F.15 *Let $\texttt{Alg}^M$ be a $\mathrm{poly}(n)$-time algorithm that uses interactive reset access to an MDP $M \in \mathcal{M}_n$ and outputs a set of policies. Fix $\mathbf{a}^\star \in \mathcal{A}^H$ and $k \in [H-1]$. Then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}} \mathbb{E}[\mathrm{Val}^{M_\mathbf{w}}_{\mathbf{a}^\star}(\texttt{Alg}^{M_\mathbf{w}})] - \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}} \mathbb{E}[\mathrm{Val}^{M'_\mathbf{w}}_{\mathbf{a}^\star}(\texttt{Alg}^{M'_\mathbf{w}})] \right| \leq O(n^{-5}).$$

where $M_\mathbf{w} = M^{:k}_{\mathbf{w},\mathbf{a}^\star}$ and $M'_\mathbf{w} = M^{:k+1}_{\mathbf{w},\mathbf{a}^\star}$.

Proof. We define an algorithm $\overline{\texttt{Alg}}^{\mathcal{O}_0,\mathcal{O}_1}$ with access to sampling oracles $\mathcal{O}_0, \mathcal{O}_1$ as follows. Sample $\mathbf{w} = (w_1, \dots, w_H)$ from $\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}$. Then simulate $\texttt{Alg}$ with RL-with-resets access to an MDP $M^{\mathcal{O}_0,\mathcal{O}_1}_{w_{1:k}}$ defined as follows. First, define an initial sampling subroutine that samples from ν_{null} (this can be implemented in time $\mathrm{poly}(n)$ by Item 5). Next, define a conditional sampling subroutine that, given h, x_h, and a_h, samples x_{h+1} according to the following distribution:

$$\mathbb{P}(x_{h+1}|x_h, a_h) = \begin{cases} \nu_{w_{h+1},1}(x_{h+1}) & \text{if } \phi^{\theta(w_h)}(x_h) = 1 \text{ and } a_h = a^\star_h \text{ and } h < k \\ \nu_{w_{h+1},0}(x_{h+1}) & \text{if } (\phi^{\theta(w_h)}(x_h) = 0 \text{ or } a \neq a^\star_h) \text{ and } h < k \\ \mathcal{O}_0(x_{h+1}) & \text{if } \phi^{\theta(w_h)}(x_h) = 1 \text{ and } a_h = a^\star_h \text{ and } h = k \\ \mathcal{O}_1(x_{h+1}) & \text{if } (\phi^{\theta(w_h)}(x_h) = 0 \text{ or } a_h \neq a^\star_h) \text{ and } h = k \\ \nu_{\mathsf{null}} & \text{if } h > k \end{cases}.$$

This sampler can be implemented in time $\mathrm{poly}(n)$, up to total variation error $n^{-\omega(1)}$: in particular, for any $x_h \in \mathcal{X}$ and $w_h \in \mathfrak{S}^{t-1}$, we can compute $\mathrm{sgn}(\langle x_h, \theta(w_h)\rangle)$ in time $\mathrm{poly}(n)$ (Item 3) and hence evaluate $\phi^{\theta(w_h)}(x_h) = \mathbb{1}[\langle x_h, \theta(w_h)\rangle \geq 0]$; moreover, we can sample from $\nu_{w_{h+1},0}$ and $\nu_{w_{h+1},1}$ for any given $w_{h+1} \in \mathfrak{S}^{t-1}$ (Item 4) up to total variation error $n^{-\omega(1)}$, we can sample from ν_{null} (Item 5), and we are given access to $\mathcal{O}_0, \mathcal{O}_1$.

Let Ψ be the output of $\texttt{Alg}$ after interaction with this MDP. Next, $\overline{\texttt{Alg}}^{\mathcal{O}_0,\mathcal{O}_1}$ computes an estimate $\widehat{\mathrm{Val}}$ of $\mathrm{Val}^{M^{\mathcal{O}_0,\mathcal{O}_1}_{w_{1:k}}}_{\mathbf{a}^\star}(\Psi)$ by enumerating over Ψ and performing Monte Carlo estimation: note that we can draw trajectories from this MDP, and we know $\mathbf{a}^\star$. By drawing $N = n^{10}$ trajectories, we can guarantee that conditioned on all prior randomness,

$$\mathbb{E}\left|\widehat{\mathrm{Val}} - \mathrm{Val}_{\mathbf{a}^\star}^{M_{w_{1:k}}^{\mathcal{O}_0,\mathcal{O}_1}}(\Psi)\right| \leq O(1/\sqrt{N}).$$

Finally, $\overline{\texttt{Alg}}^{\mathcal{O}_0,\mathcal{O}_1}$ output $\widehat{\mathrm{Val}}$.

Analysis. Suppose $(\mathcal{O}_0,\mathcal{O}_1) = (\nu_{\mathsf{null}},\nu_{\mathsf{null}})$. If the conditional sampler defined above had no error, then it would hold for any choice of $\mathbf{w} = w_{1:H}$ that $M_{w_{1:k}}^{\mathcal{O}_0,\mathcal{O}_1}$ is exactly $M_{\mathbf{w},\mathbf{a}^\star}^{:k}$. Since the error is $n^{-\omega(1)}$ per sample and $\texttt{Alg}$ has time complexity $\mathrm{poly}(n)$, this error affects $\mathbb{E}[\overline{\texttt{Alg}}^{\nu_{\mathsf{null}},\nu_{\mathsf{null}}}]$ by at most $n^{-\omega(1)}$. Therefore

$$\left|\mathbb{E}[\overline{\texttt{Alg}}^{\nu_{\mathsf{null}},\nu_{\mathsf{null}}}] - \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}} \mathbb{E}[\mathrm{Val}_{\mathbf{a}^\star}^{M_{\mathbf{w},\mathbf{a}^\star}^{:k}}(\texttt{Alg}^{M_{\mathbf{w},\mathbf{a}^\star}^{:k}})]\right| \leq O(1/\sqrt{N}),$$

where the error due to Monte Carlo estimation dominates the error in the conditional sampler.

Next, for any $w_{k+1}^\star \in \mathfrak{S}^{t-1}$, suppose that $(\mathcal{O}_0,\mathcal{O}_1) = (\nu_{w_{k+1}^\star,0},\nu_{w_{k+1}^\star,1})$. If the conditional sampler had no error, then it would hold for any choice of $\mathbf{w}$ that $M_{w_{1:k}}^{\mathcal{O}_0,\mathcal{O}_1}$ is the same as $M_{\mathbf{w}^\star,\mathbf{a}^\star}^{:k+1}$ where $\mathbf{w}^\star = (w_1,\dots,w_k,w_{k+1}^\star,w_{k+2},\dots,w_H)$. Again the error is at most $n^{-\omega(1)}$ per sample, and hence affects $\mathbb{E}[\overline{\texttt{Alg}}^{\nu_{w_{k+1}^\star,0},\nu_{w_{k+1}^\star,1}}]$ by at most $n^{-\omega(1)}$. Therefore

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{w_{k+1}^\star\sim\mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\overline{\texttt{Alg}}^{\nu_{w_{k+1}^\star,0},\nu_{w_{k+1}^\star,1}}] - \mathop{\mathbb{E}}_{w_{k+1}^\star,w_1,\dots,w_H\sim\mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\mathrm{Val}_{\mathbf{a}^\star}^{M_{\mathbf{w}^\star,\mathbf{a}^\star}^{:k+1}}(\texttt{Alg}^{M_{\mathbf{w}^\star,\mathbf{a}^\star}^{:k+1}})]\right| \\
&\leq \mathop{\mathbb{E}}_{w_{k+1}^\star\sim\mathrm{Unif}(\mathfrak{S}^{t-1})}\left|\mathbb{E}[\overline{\texttt{Alg}}^{\nu_{w_{k+1}^\star,0},\nu_{w_{k+1}^\star,1}}] - \mathop{\mathbb{E}}_{w_1,\dots,w_H\sim\mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\mathrm{Val}_{\mathbf{a}^\star}^{M_{\mathbf{w}^\star,\mathbf{a}^\star}^{:k+1}}(\texttt{Alg}^{M_{\mathbf{w}^\star,\mathbf{a}^\star}^{:k+1}})]\right| \\
&\leq O(1/\sqrt{N}).
\end{aligned}$$

But $\overline{\texttt{Alg}}$ has time complexity $\mathrm{poly}(n)$, so by Item 1,

$$\left|\mathop{\mathbb{E}}_{w_{k+1}^\star\sim\mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\overline{\texttt{Alg}}^{\nu_{w_{k+1}^\star,0},\nu_{w_{k+1}^\star,1}}] - \mathbb{E}[\overline{\texttt{Alg}}^{\nu_{\mathsf{null}},\nu_{\mathsf{null}}}]\right| \leq n^{-\omega(1)}.$$

The result follows by the triangle inequality. ■

We can now put together Lemma F.15 (applied for all $1 \leq k \leq H-1$) with Lemmas F.12 and F.13 and the coverage guarantee Eq. (42) assumed in the theorem statement.

Proof of Theorem F.8. For any $\mathbf{w} \in (\mathfrak{S}^{t-1})^H$ and $\mathbf{a} \in \mathcal{A}^H$, let $\Psi_{\mathbf{w},\mathbf{a}^\star}^{:k}$ denote the (random) output of $\texttt{Alg}^{M_{\mathbf{w},\mathbf{a}^\star}^{:k}}$. By the theorem assumption and the fact that $|\mathcal{A}| = |\mathcal{S}| = 2$, we know that for any $\mathbf{w},\mathbf{a}^\star$, with probability at least $1/2$ over the execution of $\texttt{Alg}^{M_{\mathbf{w},\mathbf{a}^\star}^{:H}}$, there is some $\widehat{\pi} \in \Psi_{\mathbf{w},\mathbf{a}^\star}^{:H}$ such that

$$\Pr^{M_{\mathbf{w},\mathbf{a}^\star}^{:H},\widehat{\pi}}[\phi^{\theta(w_H)}(x_H) = 1] \geq \frac{1}{\mathrm{poly}(H)}\left(\max_{\pi\in\Pi}\Pr^{M_{\mathbf{w},\mathbf{a}}^{:H},\pi}[\phi^{\theta(w_H)}(x_H) = 1] - \frac{1}{8}\right) \geq \frac{1}{\mathrm{poly}(H)}$$

where the second inequality is by Lemma F.12. In this event, by Lemma F.13, the policy $\widehat{\pi}$ satisfies

$$\Pr^{M^{:H}_{\mathbf{w},\mathbf{a}^\star},\widehat{\pi}}[\forall h \in [H-1] : \widehat{\pi}_h(x_h) = a^\star_h] \geq \frac{1}{\operatorname{poly}(H)} - H^2\gamma(n) \geq \frac{1}{\operatorname{poly}(H)}$$

where the second inequality holds for sufficiently large n, and uses the fact that $H = (\log n)^{\log\log n}$ whereas $\gamma(n) = (\log n)^{-\Omega(\log^2 \log n)}$. Therefore we have

$$\mathbb{E}[\mathrm{Val}_{\mathbf{a}^\star}^{M^{:H}_{\mathbf{w},\mathbf{a}^\star}}(\Psi^{:H}_{\mathbf{w},\mathbf{a}^\star})] \geq \frac{1}{\operatorname{poly}(H)}$$

where the expectation is over the execution of $\mathtt{Alg}^{M^{:H}_{\mathbf{w},\mathbf{a}^\star}}$. Since this bound holds for any fixed $\mathbf{w}$, it also holds in expectation over $\mathbf{w} \sim \mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}$. By Lemma F.15, we get that for any $\mathbf{a}^\star \in \mathcal{A}^H$,

$$\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}} \mathbb{E}[\mathrm{Val}_{\mathbf{a}^\star}^{M^{:1}_{\mathbf{w},\mathbf{a}^\star}}(\Psi^{:1}_{\mathbf{w},\mathbf{a}^\star})] \geq \frac{1}{\operatorname{poly}(H)} - O(H \cdot n^{-5}) \geq \frac{1}{\operatorname{poly}(H)} \tag{43}$$

since $H = (\log n)^{\log\log n} = n^{o(1)}$. But the initial distribution and transition distributions of $M^{:1}_{\mathbf{w},\mathbf{a}^\star}$ are independent of $\mathbf{a}^\star$ (and $\mathbf{w}$), i.e. we can write $M^{:1}_{\mathbf{w},\mathbf{a}^\star} = M$ for a fixed MDP M. It follows that the random variables $\{\Psi^{:1}_{\mathbf{w},\mathbf{a}} : \mathbf{a} \in \mathcal{A}^H\}$ are identically distributed. So for any $\mathbf{w} \in (\mathfrak{S}^{t-1})^H$ and $\mathbf{a}^\star \in \mathcal{A}^H$,

$$\begin{aligned}
\sum_{\mathbf{a}\in\mathcal{A}^H} \mathbb{E}[\mathrm{Val}_{\mathbf{a}}^{M^{:1}_{\mathbf{w},\mathbf{a}}}(\Psi^{:1}_{\mathbf{w},\mathbf{a}})] &= \sum_{\mathbf{a}\in\mathcal{A}^H} \mathbb{E}[\mathrm{Val}_{\mathbf{a}}^{M^{:1}_{\mathbf{w},\mathbf{a}}}(\Psi^{:1}_{\mathbf{w},\mathbf{a}^\star})] \\
&= \sum_{\mathbf{a}\in\mathcal{A}^H} \mathbb{E}[\mathrm{Val}_{\mathbf{a}}^{M}(\Psi^{:1}_{\mathbf{w},\mathbf{a}^\star})] \\
&= \mathbb{E}\left[\sum_{\mathbf{a}\in\mathcal{A}^H} \max_{\pi\in\Psi^{:1}_{\mathbf{w},\mathbf{a}^\star}} \Pr^{M,\pi}[\forall h \in [H-1] : \pi_h(x_h) = a_h]\right] \\
&\leq \mathbb{E}\left[\sum_{\mathbf{a}\in\mathcal{A}^H} \sum_{\pi\in\Psi^{:1}_{\mathbf{w},\mathbf{a}^\star}} \Pr^{M,\pi}[\forall h \in [H-1] : \pi_h(x_h) = a_h]\right] \\
&\leq \mathbb{E}[|\Psi^{:1}_{\mathbf{w},\mathbf{a}^\star}|] \\
&\leq \operatorname{poly}(n)
\end{aligned}$$

since the size of the output of $\mathtt{Alg}$ is bounded by its runtime. It follows that

$$\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}} \sum_{\mathbf{a}\in\mathcal{A}^H} \mathbb{E}[\mathrm{Val}_{\mathbf{a}}^{M^{:1}_{\mathbf{w},\mathbf{a}}}(\Psi^{:1}_{\mathbf{w},\mathbf{a}})] \leq \operatorname{poly}(n)$$

and thus there is some $\mathbf{a} \in \mathcal{A}^H$ with

$$\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathrm{Unif}(\mathfrak{S}^{t-1})^{\otimes H}} \mathbb{E}[\mathrm{Val}_{\mathbf{a}}^{M^{:1}_{\mathbf{w},\mathbf{a}}}(\Psi^{:1}_{\mathbf{w},\mathbf{a}})] \leq \frac{\operatorname{poly}(n)}{2^H} \leq n^{-\omega(1)}$$

where the final inequality uses that $H = \omega(\log n)$. But this contradicts Eq. (43). ■

F.3.1. TECHNICAL RESULTS REGARDING CLWE AND PTFS

To prove Theorem F.9, we follow the strategy used by Tiegel (2023): we first prove an analogous result for polynomial threshold functions (Lemma F.16, below), and then lift to linear threshold functions (albeit suffering a blowup in the dimension) via the Veronese mapping. To reiterate, these results essentially follow from inspecting various statements and proofs from Tiegel (2023).

Lemma F.16 *Let $t, \ell \in \mathbb{N}$ with $\ell \geq 2\sqrt{t}$, and $\delta \in (0,1)$. Define $\mathcal{D}^\circ := N(0, \frac{1}{2\pi} I_t)$. There is a family of degree-ℓ PTFs $\{f_w : \mathbb{R}^t \to \{-1,1\}\}$ and two families of distributions $\{\mathcal{D}_w^+\}$ and $\{\mathcal{D}_w^-\}$ over $\mathbb{R}^t$, all indexed by $w \in \mathbb{R}^t$, with the following properties:*

1. *Under (Tiegel, 2023, Assumption 3.4), for every algorithm $\mathtt{Alg}^{\mathcal{O}_0,\mathcal{O}_1}$ with time complexity 2^{t^δ}, access to sampling oracles $\mathcal{O}_0, \mathcal{O}_1$, and outputs in $[0,1]$,*

$$\left| \mathop{\mathbb{E}}_{w \sim \mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\mathtt{Alg}^{\mathcal{D}_w^+, \mathcal{D}_w^-}] - \mathbb{E}[\mathtt{Alg}^{\mathcal{D}^\circ, \mathcal{D}^\circ}] \right| \leq 2^{-t^\delta}.$$

2. *For every $w \in \mathfrak{S}^{t-1}$,*

$$\Pr_{x \sim \mathcal{D}_w^+}[f_w(x) \neq 1] \leq \exp(-\Omega(\ell^2/t)) \tag{44}$$

and

$$\Pr_{x \sim \mathcal{D}_w^-}[f_w(x) \neq -1] \leq \exp(-\Omega(\ell^2/t)) \tag{45}$$

and

$$\Pr_{x \sim \mathcal{D}^\circ}[f_w(x) = 1] \geq 1/2. \tag{46}$$

3. *There is a $\mathrm{poly}(t)$-time algorithm that takes as input $w \in \mathfrak{S}^{t-1}$ and $a \in \{-,+\}$, and outputs $(x, f_w(x))$ with $\mathsf{TV}(\mathrm{Law}(x), \mathcal{D}_w^a) \leq 2^{-\Omega(t)}$.*

Proof. Set $\beta = 1/t$ and $\gamma = 2\sqrt{t}$. For each $w \in \mathfrak{S}^{t-1}$, let $\mathcal{D}_w^+$ be the distribution $\mathrm{NH}_{w,\beta,\gamma,0}$, and let $\mathcal{D}_w^-$ be the distribution $\mathrm{NH}_{w,\beta,\gamma,1/2}$, where both of these *non-overlapping homogeneous CLWE* distributions are defined in (Tiegel, 2023, Definition 3.3). By (Tiegel, 2023, Lemma 4.3), for each $w \in \mathfrak{S}^{t-1}$ there is some degree-ℓ polynomial threshold function $f_w : \mathbb{R}^t \to \{-1,1\}$ such that

$$\frac{1}{2} \Pr_{x \sim \mathcal{D}_w^+}[f_w(x) \neq 1] + \frac{1}{2} \Pr_{x \sim \mathcal{D}_w^-}[f_w(x) \neq 1] \leq \exp(-\Omega(\ell^2/t)),$$

which proves Eqs. (44) and (45). Moreover, inspecting the proof of (Tiegel, 2023, Lemma 4.3) we see that $f_w(x) = \mathrm{sgn}(p(\langle w, x\rangle))$ for a fixed degree-ℓ polynomial $p : \mathbb{R} \to \mathbb{R}$ (i.e. depending only on t, ℓ, β, γ). Since the distribution $\mathcal{D}^\circ$ is radially symmetric, the distribution of $\langle w, x\rangle$ under $x \sim \mathcal{D}^\circ$ does not depend on w, so $\Pr_{x \sim \mathcal{D}^\circ}[f_w(x) = 1] = r$ for some constant $r \in [0,1]$. If $r \geq 1/2$ then Eq. (46) holds for all $w \in \mathfrak{S}^{t-1}$. If $r < 1/2$, then we can simply swap the definitions of $\mathcal{D}_w^+$ and $\mathcal{D}_w^-$ for all w, and replace f_w with $-f_w$ for all w. This preserves Eqs. (44) and (45), and after this swap Eq. (46) is now satisfied for all $w \in \mathfrak{S}^{t-1}$.

Since the sign pattern of p is explicit (in terms of t, ℓ, β, and γ), for any $w \in \mathfrak{S}^{t-1}$ and $x \in \mathbb{R}^t$ we can efficiently compute $\mathrm{sgn}(p(\langle w, x\rangle)) = f_w(x)$. To prove Item 3 it remains to show that given $w \in \mathfrak{S}^{t-1}$ we can approximately sample from $\mathrm{NH}_{w,\beta,\gamma,0}$ and $\mathrm{NH}_{w,\beta,\gamma,1/2}$. By (Tiegel, 2023, Lemma 6.2), for any $c \in [0,1)$, given a sampling oracle for the CLWE distribution $C_{w,\beta,\gamma}$ there is a $\mathrm{poly}(t)$-time algorithm for sampling from a distribution that is 2^{-t}-close to $\mathrm{H}_{w,\beta,\gamma,c}$

in total variation distance, where $\mathrm{H}_{w,\beta,\gamma,c}$ is the homogeneous CLWE distribution ((Tiegel, 2023, Definition 3.2)) and $C_{w,\beta,\gamma}$ is the CLWE distribution ((Tiegel, 2023, Definition 3.1)). But by definition, there is an explicit, $\mathrm{poly}(t)$-time algorithm for sampling from $C_{w,\beta,\gamma}$ (assuming the ability to sample from a standard Gaussian distribution). Moreover, (Tiegel, 2023, Lemma A.1) shows that $\mathsf{TV}(\mathrm{NH}_{w,\beta,\gamma,c}, \mathrm{H}_{w,\beta,\gamma,c}) \le 4 \cdot \exp(-\frac{1}{100\beta^2}) \le 2^{-\Omega(t)}$ by choice of β. Thus, given $w \in \mathfrak{S}^{t-1}$, we can sample from a distribution that is $2^{-\Omega(t)}$-close to $\mathrm{NH}_{w,\beta,\gamma,c}$ in total variation distance, in time $\mathrm{poly}(t)$. This proves Item 3.

By (Tiegel, 2023, Theorem 6.1), it holds under (Tiegel, 2023, Assumption 3.4) that there is no constant $c > 0$ and 2^{t^δ}-time algorithm $\mathtt{Alg}^{\mathcal{O}}$ with outputs in $[0,1]$ and

$$\left| \mathop{\mathbb{E}}_{w\sim \mathrm{Unif}(\mathfrak{S}^{t-1})}[\mathtt{Alg}^{\frac{1}{2}(D_w^+,1)+\frac{1}{2}(D_w^-,-1)}] - \mathbb{E}[\mathtt{Alg}^{N(0,\frac{1}{2\pi}I_t)\times \mathrm{Ber}(1/2)}] \right| \ge \Omega(n^{-c}).$$

In fact, by a standard boosting argument, the advantage can be driven down to 2^{-t^δ}. Moreover, since samples from D_w^+ and D_w^- can be individually extracted using a sampling oracle for $\frac{1}{2}(D_w^+,1) + \frac{1}{2}(D_w^-,-1)$, it follows that there is no 2^{t^δ}-time algorithm $\mathtt{Alg}^{\mathcal{O}_0,\mathcal{O}_1}$ with outputs in $[0,1]$ and

$$\left| \mathop{\mathbb{E}}_{w\sim \mathrm{Unif}(\mathfrak{S}^{t-1})}[\mathtt{Alg}^{D_w^+,D_w^-}] - \mathbb{E}[\mathtt{Alg}^{N(0,\frac{1}{2\pi}I_t),N(0,\frac{1}{2\pi}I_t)}] \right| \ge \Omega(2^{-t^\delta}).$$

This proves Item 1. ■

Proof of Theorem F.9. Invoke Lemma F.16 with $t = \frac{\log^2 n}{\log^4(\log n)}$ and $\ell = \frac{\log n}{3\log\log n} \ge 2\sqrt{t}$, where the inequality holds for sufficiently large n. For each $w \in \mathfrak{S}^{t-1}$, let $\mathcal{D}_w^+, \mathcal{D}_w^- \in \Delta(\mathbb{R}^t)$ be the distributions and let $f_w : \mathbb{R}^t \to \{-1,1\}$ be the $2\sqrt{t}$-PTF given by the lemma. Let $\tau : \mathbb{R}^t \to \mathbb{R}^n$ be defined as follows. Identify the first $t^0 + t^1 + \cdots + t^\ell$ coordinates of $[n]$ with sequences $\alpha = (\alpha_1, \dots, \alpha_k) \in [t]^k$ of length at most ℓ, and for $x \in \mathbb{R}^t$ and any such sequence α, define

$$\tau(x)_\alpha := \prod_{i=1}^{|\alpha|} x_i.$$

Define the remaining $n - (t^0 + \cdots + t^\ell)$ coordinates of $\tau(x)$ to be 0. Note that

$$t^0 + \cdots + t^\ell \le 2t^\ell \le 2(\log^2 n)^{\log(n)/(3\log\log n)} \le 2n^{2/3}$$

so this map is well-defined for sufficiently large n. Next, observe that for any polynomial $q : \mathbb{R}^t \to \mathbb{R}$ of degree at most ℓ, there is some $\theta \in \mathbb{R}^n$ such that $q(x) = \langle \tau(x), \theta\rangle$ for all $x \in \mathbb{R}^t$. Accordingly, for each $w \in \mathfrak{S}^{t-1}$, define $\theta(w) \in \mathbb{R}^n$ so that $f_w(x) = \mathrm{sgn}(\langle \tau(x), \theta(w)\rangle)$ for all $x \in \mathbb{R}^t$. Let $\nu_{w,1}$ be the distribution of $\tau(x)$ under $x \sim \mathcal{D}_w^+$, and let $\nu_{w,0}$ be the distribution of $\tau(x)$ under $x \sim \mathcal{D}_w^-$. Let ν_{null} be the distribution of $\tau(x)$ under $x \sim \mathcal{D}^\circ$. We now check the theorem claims:

1. Let $\mathtt{Alg}^{\mathcal{O}_0,\mathcal{O}_1}$ be an $\mathrm{poly}(n)$-time algorithm with access to sampling oracles $\mathcal{O}_0, \mathcal{O}_1 \in \Delta(\mathbb{R}^n)$. We define an algorithm $\overline{\mathtt{Alg}}^{\overline{\mathcal{O}}_0,\overline{\mathcal{O}}_1}$ with access to sampling oracles $\overline{\mathcal{O}}_0, \overline{\mathcal{O}}_1 \in \Delta(\mathbb{R}^t)$ as follows. Simulate $\mathtt{Alg}$; when it queries $\mathcal{O}_0$, draw $x \sim \overline{\mathcal{O}}_0$ and pass $\tau(x)$; when it queries $\mathcal{O}_1$, draw $x \sim \overline{\mathcal{O}}_1$ and pass $\tau(x)$. Finally, output the final output of $\mathtt{Alg}$.

Since τ can be evaluated in time $\mathrm{poly}(n)$, the time complexity of $\overline{\texttt{Alg}}$ is $\mathrm{poly}(n) = 2^{O(\log n)} \leq 2^{O(t^{2/3})}$. Moreover, the distribution of $\overline{\texttt{Alg}}^{\mathcal{D}^\circ,\mathcal{D}^\circ}$ is exactly the distribution of $\texttt{Alg}^{\nu_{\mathsf{null}},\nu_{\mathsf{null}}}$, and for each $w \in \mathfrak{S}^{t-1}$, the distribution of $\overline{\texttt{Alg}}^{\mathcal{D}_w^+,\mathcal{D}_w^-}$ is exactly the distribution of $\texttt{Alg}^{\nu_{w,1},\nu_{w,0}}$. Thus, Item 1 of Lemma F.16, applied to $\overline{\texttt{Alg}}$, implies that

$$\left| \mathop{\mathbb{E}}_{w \sim \mathrm{Unif}(\mathfrak{S}^{t-1})} \mathbb{E}[\texttt{Alg}^{\nu_{w,0},\nu_{w,1}}] - \mathbb{E}[\texttt{Alg}^{\nu_{\mathsf{null}},\nu_{\mathsf{null}}}] \right| \leq 2^{-t^{2/3}}.$$

By definition of t, this bound decays super-polynomially in n.

2. Fix $w \in \mathfrak{S}^{t-1}$. We have

$$\begin{aligned}
\Pr_{x \sim \nu_{w,1}}[\langle x, \theta(w) \rangle < 0] &= \Pr_{x \sim \mathcal{D}_w^+}[\langle \tau(x), \theta(w) \rangle < 0] \\
&= \Pr_{x \sim \mathcal{D}_w^+}[f_w(x) = -1] \\
&\leq \exp(-\Omega(\ell^2/t)) \\
&\leq \exp(-\Omega(\log^2 \log n)) \\
&\leq (\log n)^{-\Omega(\log \log n)}
\end{aligned}$$

by Eq. (44) and the definitions of ℓ and t. A symmetric argument with $\mathcal{D}_w^-$ and Eq. (45) shows that

$$\Pr_{x \sim \nu_{w,0}}[\langle x, \theta(w) \rangle \geq 0] \leq (\log n)^{-\Omega(\log \log n)}.$$

Finally,

$$\begin{aligned}
\Pr_{x \sim \nu_{\mathsf{null}}}[\langle x, \theta(w) \rangle \geq 0] &= \Pr_{x \sim \mathcal{D}^\circ}[\langle \tau(x), \theta(w) \rangle \geq 0] \\
&= \Pr_{x \sim \mathcal{D}^\circ}[f_w(x) = 1] \\
&\geq 1/2
\end{aligned}$$

by Eq. (46).

3. Given $w \in \mathfrak{S}^{t-1}$, we can sample $x \in \mathbb{R}^t$ with $\mathsf{TV}(\mathrm{Law}(x), \mathcal{D}_w^+) \leq 2^{-\Omega(t)}$ in time $\mathrm{poly}(t)$ (by Item 3), and we can then compute $\tau(x)$ in time $\mathrm{poly}(n)$. By the data processing inequality, $\mathsf{TV}(\mathrm{Law}(\tau(x)), \nu_{w,1}) \leq 2^{-\Omega(t)} \leq n^{-\omega(1)}$. Moreover, we can compute $f_w(x) = \mathrm{sgn}(\langle \tau(x), \theta(w) \rangle)$ in time $\mathrm{poly}(t)$ by Item 3. The same argument works for $\nu_{w,0}$.

4. Since $\mathcal{D}^\circ = N(0, \frac{1}{2\pi} I_t)$, we can efficiently sample $x \sim \mathcal{D}^\circ$ and then compute $\tau(x)$, which has distribution ν_{null} by definition.

This completes the proof. ■

F.4. Statistical Tractability of Generalized Block MDPs

In this section we prove Proposition F.17, which asserts that reward-free RL in the family of Generalized Block MDPs (and hence Low-Rank MDPs) $\mathcal{M}_n$ is *statistically* tractable—and hence the hardness result from the preceding section is purely a *computational* phenomenon.

Proposition F.17 *There is a (computationally* inefficient*) algorithm* $\texttt{Alg}^M$ *that, given episodic access to any MDP* $M \in \mathcal{M}_n$*, has sample complexity* $\mathrm{poly}(n)$ *and produces a set of policies* Ψ *satisfying the following guarantee with probability at least* $1 - o(1)$*:*

$$\forall s \in \mathcal{S} : \max_{\pi \in \Psi} d_H^{M,\pi}(s) \geq \frac{1}{64} \cdot \left(\max_{\pi \in \Pi} d_H^{M,\pi}(s) - \frac{1}{8} \right).$$

Since Φ_n has infinite cardinality, this does not follow from a black-box application of prior results on learning Low-Rank MDPs. However, as we verify in Theorem F.19, the main result of Mhammedi et al. (2023a) can be extended to this setting so long as an appropriate function class based on Φ_n has bounded *pseudo-dimension* (Definition F.18). We then verify the requisite pseudo-dimension bound (Lemma F.20). From these, the proof of Proposition F.17 is straightforward.

Definition F.18 (Pseudo-dimension (Haussler, 2018)) *For any set* $\mathcal{X}$ *and function class* $\mathcal{F} \subseteq \{\mathcal{X} \to \mathbb{R}\}$*, the pseudo-dimension of* $\mathcal{F}$ *is defined as the VC dimension of* $\mathcal{F}^+$*, where* $\mathcal{F}^+ \subseteq \{\mathcal{X} \times \mathbb{R} \to \{0, 1\}\}$ *is defined as* $\mathcal{F}^+ := \{(x, \xi) \mapsto \mathbb{1}[f(x) > \xi] : f \in \mathcal{F}\}$*.*

We prove the following straightforward extension of a result by Mhammedi et al. (2023a):

Theorem F.19 (Extension of (Mhammedi et al., 2023a, Theorem 3.2)) *Let* $\mathcal{S}, \mathcal{X}, \mathcal{A}$ *be sets, let* $H \in \mathbb{N}$*, and let* Φ *be a set of functions* $\phi : \mathcal{X} \to \mathcal{S}$*. Let* $\mathfrak{d}$ *denote the pseudo-dimension (Definition F.18) of the function class* $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R})$ *defined by*

$$\mathcal{F} := \left\{ (x, a, x') \mapsto f(\phi^1(x), \phi^2(x), \phi^3(x'), \phi^4(x'), a) \mid f : \mathcal{S}^4 \times \mathcal{A} \to [0, 1], \phi^1, \phi^2, \phi^3, \phi^4 \in \Phi \right\}. \tag{47}$$

There is an algorithm $\texttt{Alg}^M$ *that takes input* $\epsilon, \delta \in (0, 1/2)$ *and episodic access to an MDP* M *with observation space* $\mathcal{X}$*, action space* $\mathcal{A}$*, and horizon* H*, and has the following property. If* M *is a generalized* Φ*-decodable block MDP, then* $\texttt{Alg}^M(\epsilon, \delta)$*, with probability at least* $1 - \delta$*, produces sets* $\Psi_{1:H}$ *such that*

$$\begin{aligned} &\mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Psi_h)} d_h^{M,\pi}(x) \geq \frac{1}{8|\mathcal{A}|^2|\mathcal{S}|} \max_{\pi \in \Pi} d_h^{M,\pi}(x) \\ &\forall h \in \{2, \dots, H\}, x \in \mathcal{X} : \max_{\pi \in \Pi} d_h^{M,\pi}(x) \geq \epsilon \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{P}_h^M(x|s, a). \end{aligned} \tag{48}$$

Moreover, the algorithm has sample complexity $\mathrm{poly}(H, |\mathcal{S}|, |\mathcal{A}|, 1/\epsilon, \mathfrak{d}, \log(1/\delta))$*.*

Proof. First suppose that $|\Phi| < \infty$ and we replace the pseudo-dimension term $\mathfrak{d}$ in the sample complexity with $\log|\Phi|$. Then we claim that this theorem is essentially immediate from (Mhammedi et al., 2023a, Theorem 3.2): by Proposition F.2, any generalized Φ-decodable block MDP is low-rank with rank $d := |\mathcal{S}||\mathcal{A}|$ and function class Φ^{lin} of size exactly $|\Phi|$. It can be checked that the norm bounds required by Mhammedi et al. (2023a) are satisfied: $\left\|\phi^{\mathsf{lin}}(x, a)\right\|_2 \leq 1$ for all $x \in \mathcal{X}, a \in \mathcal{A}$, and $\phi^{\mathsf{lin}} \in \Phi^{\mathsf{lin}}$, and

$$\left\| \int_{\mathcal{X}} g(x) d\mu_{h+1}^{\star} \right\|_2 \leq \sqrt{d}$$

for any $h \in [H - 1]$ and $g : \mathcal{X} \to [0, 1]$, since $(\mu_{h+1}^{\star})_{s,a}$ is a distribution over $\mathcal{X}$ for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Thus, (Mhammedi et al., 2023a, Theorem 3.2) shows that with probability at least $1-\delta$, the output $\Psi_{1:H}$ is a $(1/(8|\mathcal{A}|^2||\mathcal{S}|), \epsilon)$-policy cover. However, inspecting the proof, in fact the stronger guarantee is shown that the output is a $(1/(8|\mathcal{A}|^2|\mathcal{S}|), \epsilon)$-*randomized* policy cover – see (Mhammedi et al., 2023a, Definition 2.2). This gives Eq. (48), using the fact that $\left\|\mu^\star_{h+1}(x)\right\|_2 \leq \sum_{s,a} \mathbb{P}_{h+1}(x \mid s,a)$ for any $x \in \mathcal{X}$.

It remains to argue that the result can be generalized to infinite classes Φ using pseudo-dimension. The dependence on $\log|\Phi|$ in the sample complexity bound in (Mhammedi et al., 2023a, Theorem 3.2) arises due to the need to prove uniform concentration over Φ. In particular, it arises in two places:

1. Analysis of PSDP: (Mhammedi et al., 2023a, Lemma D.2) incurs dependence on $\log|\Phi|$ while proving that the regression estimate $\hat{g}^{(t)}$ computed on line 7 of (Mhammedi et al., 2023a, Algorithm 3) is close to the Q-function $Q_t^{\hat{\pi}^{t+1}}$. However, we can remove this dependence as follows. With notation as in line 7 of (Mhammedi et al., 2023a, Algorithm 3), it suffices to prove that so long as $|\mathcal{D}^{(t)}| \geq \mathrm{poly}(\mathfrak{d}, 1/\epsilon, \log(1/\delta))$, we have with probability at least $1-\delta$ that
$$\sup_{g \in \mathcal{G}_t} \left| \frac{1}{|\mathcal{D}^{(t)}|} \sum_{(x,a,R) \in \mathcal{D}^{(t)}} (g(x,a) - R)^2 - \mathbb{E}[(g(x,a)-R)^2] \right| \leq \epsilon.$$
In particular, we need this to hold for $\mathcal{G}_t := \{(x,a) \mapsto f(\phi(x),a) \mid f : \mathcal{S} \times \mathcal{A} \to [0,1], \phi \in \Phi\}$. By (Modi et al., 2024, Corollary 4.3) and the triangle inequality, it suffices to bound the pseudo-dimension of the function classes
$$\mathcal{F}_1 := \{(x,a) \mapsto f(\phi(x),a)^2 \mid f : \mathcal{S} \times \mathcal{A} \to [0,1], \phi \in \Phi\},$$
$$\mathcal{F}_2 := \{(x,a,R) \mapsto f(\phi(x),a) \cdot R \mid f : \mathcal{S} \times \mathcal{A} \to [0,1], \phi \in \Phi\},$$
and
$$\mathcal{F}_3 := \{R \mapsto R^2\}.$$
Note that $\mathcal{F}_3$ has constant size and hence constant pseudo-dimension; each function in $\mathcal{F}_2$ can be expressed as a product of some function in $\mathcal{G}_t$ with the function $R \mapsto R$; and each function in $\mathcal{F}_1$ can be expressed as a product of two functions in $\mathcal{G}_t$. Thus, by (Modi et al., 2024, Lemma 50), it suffices to bound the pseudo-dimension of $\mathcal{G}_t$. Since $\mathcal{G}_t$ can be embedded in $\mathcal{F}$, the pseudo-dimension of $\mathcal{G}_t$ is bounded by $\mathfrak{d}$.

2. Analysis of RepLearn: (Mhammedi et al., 2023a, Lemma F.1) incurs dependence on $\log|\Phi|$ through application of (Modi et al., 2024, Lemma 14), which in turn uses (Modi et al., 2024, Lemma 17), which in turn uses (Modi et al., 2024, Lemma 34). In the proof of (Modi et al., 2024, Lemma 34), a factor of $\log|\Phi|$ is incurred to bound the log covering number of a particular function class $\mathcal{H}$, which consists of certain functions $h : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$. For our application, the function classes Φ, Φ' in the lemma statement are both Φ^{lin}, and the reward function class $\mathcal{R}$ is $\{(x,a) \mapsto f(\phi(x),a) : f : \mathcal{S} \times \mathcal{A} \to [0,1], \phi \in \Phi\}$. From this, we can check that each function $h \in \mathcal{H}$ depends on its input (x,a,x') only through a and $\phi^1(x), \phi^2(x), \phi^3(x'), \phi^4(x')$ for four functions $\phi^1, \phi^2, \phi^3, \phi^4 \in \Phi$. Thus, $\mathcal{H}$ can be embedded in $\mathcal{F}$, and so the pseudo-dimension of $\mathcal{H}$ can be bounded by $\mathfrak{d}$. This means that in the proof of (Modi et al., 2024, Lemma 34), instead of bounding the log covering number, we can apply (Modi et al., 2024, Corollary 43) using this bound on the pseudo-dimension of $\mathcal{H}$. The rest of the proof of the lemma is unchanged.

This completes the proof. ∎

Lemma F.20 *Fix $n \in \mathbb{N}$ and let Φ_n denote the class of linear threshold functions (Definition F.4). The function class $\mathcal{F}_n \subseteq (\mathbb{R}^n \times \{0,1\} \times \mathbb{R}^n \to \mathbb{R})$ defined by*

$$\mathcal{F}_n := \left\{(x, a, x') \mapsto f(\phi^1(x), \phi^2(x), \phi^3(x'), \phi^4(x'), a) \mid f : \{0,1\}^5 \to [0,1], \phi^1, \phi^2, \phi^3, \phi^4 \in \Phi_n\right\}$$

has pseudo-dimension at most $O(n \log n)$.

Proof. By Definition F.18, we need to bound the VC dimension of the function class $\mathcal{F}_n^+ \subseteq (\mathbb{R}^n \times \{0,1\} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R})$

$$\mathcal{F}_n^+ := \left\{(x, a, x', \xi) \mapsto \mathbb{1}[f(\phi^{1:2}(x), \phi^{3:4}(x'), a) > \xi] \mid f : \{0,1\}^5 \to [0,1], \phi^1, \phi^2, \phi^3, \phi^4 \in \Phi_n\right\}.$$

Fix any set $\mathcal{D} = (x_i, a_i, x_i', \xi_i)_{i=1}^m$. We would like to upper bound the number of attainable vectors

$$(\mathbb{1}[f(\phi^{1:2}(x_1), \phi^{3:4}(x_1'), a_1) > \xi_1], \dots, \mathbb{1}[f(\phi^{1:2}(x_m), \phi^{3:4}(x_m'), a_m) > \xi_m]) \in \{0,1\}^m$$

as f and $\phi^1, \dots, \phi^4$ vary. First, note that the numbers $\xi_1, \dots, \xi_m$ partition $\mathbb{R}$ into $m+1$ intervals, and two functions f, f' such that $f(b)$ and $f'(b)$ lie in the same interval for all $b \in \{0,1\}^5$ induce the same vector (for any fixed $\phi^1, \dots, \phi^4$). Thus, we can restrict focus to $(m+1)^{32} = \mathrm{poly}(m)$ choices of f.

Fix one such f. By the Milnor-Warren bound Milnor (1964), the set $\{(\phi(x_1), \dots, \phi(x_m)) : \phi \in \Phi_n\}$ has size at most $m^{O(n)}$. Thus, as $\phi^1, \dots, \phi^4 \in \Phi_n$ vary, the number of attainable vectors is bounded by $m^{O(n)}$. Summing over the $\mathrm{poly}(m)$ choices of f, we find that the total number of attainable vectors as all parameters vary is still $m^{O(n)}$, so $\mathcal{D}$ cannot be shattered by $\mathcal{F}_n^+$ unless $m \le O(n \log n)$. Thus, the pseudo-dimension of $\mathcal{F}_n$ is at most $O(n \log n)$. ∎

Proof of Proposition F.17. We apply Theorem F.19 with $\mathcal{S} := \mathcal{A} := \{0,1\}$, $\mathcal{X} := \mathbb{R}^n$, $H := (\log n)^{\log\log n}$, and $\Phi := \Phi_n$. By Lemma F.20, the pseudo-dimension of the resulting set $\mathcal{F}$ defined in Eq. (47) is at most $O(n \log n)$. Invoke the algorithm $\texttt{Alg}^M$ guaranteed by Theorem F.19 with parameters $\epsilon := \frac{1}{512}$ and $\delta := 1/2$. For any $M \in \mathcal{M}_n$, since M is a generalized Φ_n-decodable block MDP, we get that with probability at least $1/2$, the sets $\Psi_{1:H}$ produced by $\texttt{Alg}^M$ satisfy Eq. (48). Let $\mathcal{X}_{\mathsf{good}}$ be the set of $x \in \mathcal{X}$ satisfying

$$\max_{\pi \in \Pi} d_H^{M,\pi}(x) \ge \epsilon \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{P}_H^M(x \mid s, a).$$

Suppose that M has decoding functions $\phi_1^\star, \dots, \phi_H^\star$. For any $s \in \mathcal{S}$, we have

$$\begin{aligned}
\max_{\pi \in \Psi} d_H^{M,\pi}(s) &\ge \mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Psi_H)} \left[d_H^{M,\pi}(s)\right] \\
&\ge \sum_{x \in \mathcal{X}_{\mathsf{good}} : \phi_H^\star(x) = s} \mathop{\mathbb{E}}_{\pi \sim \mathrm{Unif}(\Psi_H)} \left[d_H^{M,\pi}(x)\right] \\
&\ge \frac{1}{64} \sum_{x \in \mathcal{X}_{\mathsf{good}} : \phi_H^\star(x) = s} \max_{\pi \in \Pi} d_H^{M,\pi}(x)
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{64} \max_{\pi \in \Pi} \sum_{x \in \mathcal{X}_{\mathsf{good}} : \phi^\star_H(x) = s} d_H^{M,\pi}(x) \\
&\geq \frac{1}{64} \max_{\pi \in \Pi} \left(d_H^{M,\pi}(s) - \sum_{x \in \mathcal{X} \setminus \mathcal{X}_{\mathsf{good}} : \phi^\star_H(x) = s} d_H^{M,\pi}(x) \right) \\
&\geq \frac{1}{64} \max_{\pi \in \Pi} \left(d_H^{M,\pi}(s) - \epsilon \sum_{x \in \mathcal{X}} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{P}_H^M(x|s,a) \right) \\
&= \frac{1}{64} \max_{\pi \in \Pi} \left(d_H^{M,\pi}(s) - \epsilon |\mathcal{S}||\mathcal{A}| \right) \\
&= \frac{1}{64} \max_{\pi \in \Pi} \left(d_H^{M,\pi}(s) - \frac{1}{8} \right)
\end{aligned}
$$

by choice of ϵ. Finally, the sample complexity of the algorithm is $\mathrm{poly}(n)$ by Theorem F.19, the fact that the pseudo-dimension is at most $O(n \log n)$, the choice of parameters $\epsilon, \delta = \Omega(1)$, and the fact that $H, |\mathcal{S}|, |\mathcal{A}| \leq n$. ■

Algorithm 7 $\texttt{PSDP}(k, \texttt{Reg}, R, \Psi_{1:k}, \Gamma, N)$: Policy Search by Dynamic Programming (variant of Bagnell et al. (2003); see also Mhammedi et al. (2023b); Golowich et al. (2024b))

input: Step $k \in [H]$; regression oracle $\texttt{Reg}$; reward function $R : \mathcal{X} \to [0,1]$; policy covers $\Psi_1, \dots, \Psi_k$; backup policy cover Γ; number of samples $N \in \mathbb{N}$.
for $h = k, \dots, 1$ **do**
  for $a \in \mathcal{A}$ **do**
    $\mathcal{D}_{h,a} \leftarrow \emptyset$.
    for N times **do**
      Sample policy $\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h) + \mathrm{Unif}(\Gamma))$.
      Sample trajectory $(x_1, a_1, \dots, x_k, a_k, x_{k+1}) \sim \pi \circ_h a \circ_{h+1} \widehat{\pi}^{h+1:k}$.
      Sample $r_{k+1} \sim \mathrm{Ber}(R(x_{k+1}))$.
      Update dataset: $\mathcal{D}_{h,a} \leftarrow \mathcal{D}_{h,a} \cup \{(x_h, r_{k+1})\}$.
    Solve regression:

$$\widehat{Q}_h(\cdot, a) \leftarrow \texttt{Reg}(\mathcal{D}_{h,a}).$$

  Define $\widehat{\pi}_h : \mathcal{X} \to \mathcal{A}$ by

$$\widehat{\pi}_h(x) := \arg\max_{a \in \mathcal{A}} \widehat{Q}_h(x, a),$$

  and write $\widehat{\pi}^{h:k} = (\widehat{\pi}_h, \dots, \widehat{\pi}_k)$.
return: Policy $\widehat{\pi}^{1:k} \in \Pi$.

Appendix G. Supporting Technical Results

This section contains supporting technical results. Section G.1 gives a self-contained presentation of the PSDP algorithm, while Section G.2 gives miscellaneous regression reductions used throughout our main results.

G.1. Policy Search by Dynamic Programming (PSDP)

The following lemma provides an analysis of *Policy Search by Dynamic Programming (PSDP)* (Bagnell et al., 2003)—specifically, an implementation where the Q-functions are fit using one-context regression (Algorithm 7). This shows that (approximate) policy optimization with a given reward function is efficiently reducible to one-context regression, and is a key element in both Theorem 3.1 and Theorem 4.1, as a subroutine in PCE (Algorithm 1) and PCR (Algorithm 5) respectively. While the below statement is technically novel, since it abstracts generalization arguments into the regression oracle, at a technical level the analysis is entirely standard, see e.g. Mhammedi et al. (2023b); Golowich et al. (2024b).

Value functions. For a Block MDP M, policy π, and collection $\mathbf{r} = (\mathbf{r}_h)_{h=1}^H$ of reward functions $\mathbf{r}_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we define value functions $Q_h^{M,\pi,\mathbf{r}}(x,a) := \mathbb{E}^{M,\pi}[\sum_{k=h}^H \mathbf{r}_k(x_k, a_k) \mid x_h = x, a_h = a]$ and $V_h^{M,\pi,\mathbf{r}}(x) := \mathbb{E}^{M,\pi}[\sum_{k=h}^H \mathbf{r}_k(x_k, a_k) \mid x_h = x]$.

Lemma G.1 (PSDP analysis) *Fix $\alpha, \epsilon, \delta \in (0,1)$, $k \in [H]$, and $N \in \mathbb{N}$. Let $\Phi_{1:k}$ be α-truncated policy covers for M at steps $1, \dots, k$ (Definition B.2), let $\texttt{Reg}$ be an oracle that solves N_{reg}-efficient one-context regression over Φ, and let $R : \overline{\mathcal{X}} \to [0,1]$ be a function with $R(\mathfrak{t}) = 0$. Let $\Gamma \subset \Pi$ be a finite set of policies. If $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, then $\widehat{\pi} \leftarrow \texttt{PSDP}(k, \texttt{Reg}, R, \Psi_{1:k}, \Gamma, N)$ satisfies*

$$\mathbb{E}^{M,\widehat{\pi}}[R(x_{k+1})] \geq \max_{\pi \in \Pi} \mathbb{E}^{\overline{M}(\Gamma),\pi}[R(x_{k+1})] - \frac{4H\sqrt{|\mathcal{A}|\epsilon}}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}$$

with probability at least $1 - H|\mathcal{A}|\delta$.

Proof. Define reward function $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_H)$ by

$$\mathbf{r}_h(x,a) = \begin{cases} R(x) & \text{if } h = k+1 \\ 0 & \text{otherwise} \end{cases}.$$

Let $\pi^\star \in \arg\max_{\pi \in \Pi} \mathbb{E}^{M,\pi}[R(x_{k+1})]$. Applying Lemma B.12 with reward function R and policies $\widehat{\pi}$ and $\pi^\star$ gives

$$\begin{aligned} &\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[R(x_{k+1})] - \mathbb{E}^{M,\widehat{\pi}}[R(x_{k+1})] \\ &\leq \sum_{h=1}^{k} \mathbb{E}^{\overline{M}(\Gamma),\pi^\star}\left[Q_h^{M,\widehat{\pi},\mathbf{r}}(x_h, a_h) - V_h^{M,\widehat{\pi},\mathbf{r}}(x_h)\right]. \end{aligned}$$

Fix $h \in [k]$. By definition (Line 11) we have $\widehat{\pi}_h(x) \in \arg\max_{a \in \mathcal{A}} \widehat{Q}_h(x,a)$. Let us define $\Delta_h : \overline{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ by

$$\Delta_h(x) := \begin{cases} \max_{a \in \mathcal{A}} |Q_h^{M,\widehat{\pi},\mathbf{r}}(x,a) - \widehat{Q}_h(x,a)| & \text{if } x \in \mathcal{X} \\ 0 & \text{if } x = \mathfrak{t} \end{cases}.$$

Then for any $x \in \mathcal{X}$, we have

$$\begin{aligned} &Q_h^{M,\widehat{\pi},\mathbf{r}}(x, \pi_h^\star(x)) - V_h^{M,\widehat{\pi},\mathbf{r}}(x) \\ &= Q_h^{M,\widehat{\pi},\mathbf{r}}(x, \pi_h^\star(x)) - Q_h^{M,\widehat{\pi},\mathbf{r}}(x, \widehat{\pi}_h(x)) \\ &\leq \widehat{Q}_h(x, \pi_h^\star(x)) - \widehat{Q}_h(x, \widehat{\pi}_h(x)) + 2\Delta_h(x) \\ &\leq 2\Delta_h(x) \end{aligned}$$

where the first inequality uses the definition of $\Delta_h(x)$ and the second inequality uses the fact that $\widehat{\pi}_h(x) \in \arg\max_{a \in \mathcal{A}} \widehat{Q}_h(x,a)$. Note that the above inequality also holds for $x = \mathfrak{t}$, since $Q_h^{M,\pi,\mathbf{r}}(\mathfrak{t}, a) = V_h^{M,\pi,\mathbf{r}}(\mathfrak{t}) = 0$ for any $\pi \in \Pi$, $a \in \mathcal{A}$. It follows that

$$\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[R(x_{k+1})] - \mathbb{E}^{M,\widehat{\pi}}[R(x_{k+1})] \leq \sum_{h=1}^{k-1} \mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[2\Delta_h(x_h)], \tag{49}$$

and it only remains to upper bound each term $\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[\Delta_h(x_h)]$. Once more, fix $h \in [k]$ and $a \in \mathcal{A}$. The dataset $\mathcal{D}_{h,a}$ consists of N independent and identically distributed samples (x,r) with

$$\mathbb{E}[r|x] = \mathbb{E}^{M,a\circ_h\widehat{\pi}^{h+1:k}}[R(x_{k+1})|x_h = x] = \mathbb{E}^{M,a\circ_h\widehat{\pi}^{h+1:k}}[R(x_{k+1})|s_h = \phi^\star(x)] = Q_h^{M,\widehat{\pi},\mathbf{r}}(x,a).$$

In particular, by the penultimate equality, $\mathbb{E}[r|x]$ only depends on $\phi^\star(x)$, so by the guarantee on `Reg` (Definition 2.2) and the assumption that $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, it holds with probability at least $1-\delta$ that

$$\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}(\widehat{Q}_h(x_h, a) - Q_h^{M,\widehat{\pi},\mathbf{r}}(x_h, a))^2 \leq \epsilon.$$

Let $\mathcal{E}_{h,a}$ be the event that this inequality holds. Under $\bigcap_{a\in\mathcal{A}} \mathcal{E}_{h,a}$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}[\Delta_h(x_h)^2] &\leq \sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}(\widehat{Q}_h(x_h, a) - Q_h^{M,\widehat{\pi},\mathbf{r}}(x_h, a))^2 \\ &\leq |\mathcal{A}|\epsilon,\end{aligned}$$

which yields, via Jensen's inequality, that

$$\mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}[\Delta_h(x_h)] \leq \sqrt{|\mathcal{A}|\epsilon}.$$

It follows that under event $\bigcap_{a\in\mathcal{A}} \mathcal{E}_{h,a}$,

$$\begin{aligned}\mathbb{E}^{\overline{M}(\Gamma),\pi^\star}[\Delta_h(x_h)] &\leq \frac{2}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})} \mathop{\mathbb{E}}_{\pi \sim \frac{1}{2}(\mathrm{Unif}(\Psi_h)+\mathrm{Unif}(\Gamma))} \mathbb{E}^{M,\pi}[\Delta_h(x_h)] \\ &\leq \frac{2\sqrt{|\mathcal{A}|\epsilon}}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}.\end{aligned}$$

where the first inequality uses Item 3 of Lemma B.9 together with the assumption that Ψ_h is an α-truncated cover (Definition B.2) and the fact that $\Delta_h(\mathfrak{t}) = 0$ and $\Delta_h(x) \geq 0$ for all $x \in \mathcal{X}$. Substituting into Eq. (49), we conclude that, under the event $\bigcap_{h=1}^k \bigcap_{a\in\mathcal{A}} \mathcal{E}_{h,a}$ (which occurs with probability at least $1 - H|\mathcal{A}|\delta$),

$$\mathbb{E}^{\overline{M},\pi^\star}[R(x_{k+1})] - \mathbb{E}^{M,\widehat{\pi}}[R(x_{k+1})] \leq \frac{4H\sqrt{|\mathcal{A}|\epsilon}}{\min(\alpha\sigma_{\mathsf{trunc}}, \sigma_{\mathsf{bkup}})}$$

as claimed. ■

G.2. Miscellaneous Reductions

In Section G.2.1, we prove that one-context regression is necessary for reward-free episodic RL, modifying a proof of Golowich et al. (2024a); this reduction is one piece in the proof of Theorem 3.2. In Section G.2.2, we introduce noiseless one-context regression and prove that it is necessary for reward-free RL in the reset access model, which complements our result from Section 4. In Section G.2.3, we show that one-context regression is a special case of two-context regression, which is needed for our episodic RL algorithm `PCE` (see Section 3). In Section G.2.4 and Section G.2.5, we show that one-context regression and two-context regression for concept class Φ^{aug} reduce to one-context regression and two-context regression for Φ; the latter is one piece in the proof of Theorem 3.2.

In these reductions, fix a concept class $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$ and recall the definition of sets $\mathcal{X}^{\mathsf{aug}}, \mathcal{S}^{\mathsf{aug}}$ and augmented concept class $\Phi^{\mathsf{aug}} \subseteq (\mathcal{X}^{\mathsf{aug}} \to \mathcal{S}^{\mathsf{aug}})$ from Definition D.1.

Algorithm 8 $\texttt{OneRed}(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: Reduction from one-context regression to reward-free episodic RL

input: Oracle $\mathcal{O}$ for reward-free episodic RL; samples $(x^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.
Set $\varepsilon_A := \sqrt{\epsilon/4}$ and $i = 1$. Initialize $\mathcal{O}$ with tolerance $\epsilon/4$, failure probability $\delta/2$, horizon $H := 2$, and action set $\mathcal{A} := \{0, \varepsilon_A, \dots, 1 - \varepsilon_A\}$. Simulate $\mathcal{O}$ as follows:
repeat
    When $\mathcal{O}$ queries for a new episode, pass $x^{(i)}$.
    When $\mathcal{O}$ plays an action $a_1 \in \mathcal{A}$, pass observation 0 with probability $(a - y^{(i)})^2$. Otherwise, pass observation 1. In either case, set $i \leftarrow i + 1$.
until $\mathcal{O}$ returns policy cover Ψ
$m \leftarrow 16\epsilon^{-2} \log(4|\Psi|/\delta)$.
for $\pi \in \Psi$ **do**
    Compute $\widehat{E}(\pi) := \frac{1}{m} \sum_{i=n-m+1}^n (\pi(x^{(i)}) - y^{(i)})^2$.
return: $\widehat{\pi} := \arg\min_{\pi \in \Psi} \widehat{E}(\pi)$.

G.2.1. THE OneRed REDUCTION

In recent work, Golowich et al. (2024a) showed that one-context regression is necessary for reward-directed episodic RL. Here we adapt their argument to reward-free RL; the modification is straightforward under regularity (using the "extra" states $\{0, 1\}$).

Concretely, the following theorem shows that $\texttt{OneRed}$ (Algorithm 8) reduces one-context regression for concept class Φ to reward-free episodic RL for concept class Φ^{aug}. We use this reduction as a component of our reduction from *two*-context regression to reward-free RL (see Appendix D).

Proposition G.2 (Modification of (Golowich et al., 2024a, Proposition B.2)) *Suppose that $\mathcal{O}$ is a $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free episodic RL oracle for Φ^{aug}. Then $\texttt{OneRed}(\mathcal{O}, \cdot)$ is a N_{reg}-efficient one-context regression algorithm for Φ with*

$$N_{\mathsf{reg}}(\epsilon, \delta) = N_{\mathsf{RL}}\left(\frac{\epsilon}{4}, \frac{\delta}{2}, 2, \sqrt{\frac{4}{\epsilon}}\right) + \frac{16 \log(4K_{\mathsf{RL}}\left(\frac{\epsilon}{4}, \frac{\delta}{2}, 2, \sqrt{\frac{4}{\epsilon}}\right)/\delta)}{\epsilon^2}.$$

Proof. Let $\epsilon, \delta > 0$, $\mathcal{D} \in \Delta(\mathcal{X})$, and $f : \mathcal{S} \to \{0, 1\}$. Let $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$. Let $(x^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $x^{(i)} \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi(x^{(i)}))$ for some $\phi \in \Phi$. We analyze the execution of $\texttt{OneRed}(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$. We know that $n - m \geq N_{\mathsf{RL}}(\epsilon/4, \delta/2, 2, \sqrt{4/\epsilon})$, since $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$. Fix any episode i of interaction with the oracle $\mathcal{O}$. Observe that conditioned on the initial observation $x^{(i)}$ and action a_1,

$$\begin{aligned}
\mathbb{E}[(a_1 - y^{(i)})^2 \mid x^{(i)}, a_1] &= (a_1)^2 - 2a_1 \mathbb{E}[y^{(i)} \mid x^{(i)}] + \mathbb{E}[(y^{(i)})^2 \mid x^{(i)}] \\
&= (a_1)^2 + (1 - 2a_1) f(\phi(x^{(i)}))
\end{aligned}$$

where the final equality uses that $y^{(i)} \in \{0, 1\}$. Thus, $\texttt{OneRed}$ simulates $\mathcal{O}$ on a Φ^{aug}-decodable block MDP M with horizon 2, observation space $\mathcal{X}^{\mathsf{aug}}$, latent state space $\mathcal{S}^{\mathsf{aug}}$, initial observation distribution $\mathcal{D}$, and transition distribution defined by

$$\mathbb{P}_2(0 \mid x_1, a_1) := a_1^2 + (1 - 2a_1) f(\phi(x_1)),$$

$$\mathbb{P}_2(1 \mid x_1, a_1) := 1 - \mathbb{P}_2(0 \mid x_1, a_1).$$

By Definition 2.1, the output of $\mathcal{O}$ is a set of policies Ψ of size at most $K_{\mathsf{RL}}(\epsilon/4, \delta/2, 2, \sqrt{4/\epsilon})$, such that with probability at least $1 - \delta/2$, there is some $\pi^\star \in \Psi$ such that

$$d_2^{M,\pi^\star}(1) \geq \max_{\pi \in \Pi} d_2^{M,\pi}(1) - \frac{\epsilon}{4}.$$

Condition on this event, and observe that for any $\pi \in \Pi$,

$$\begin{aligned} d_2^{M,\pi}(0) &= \mathbb{E}^{M,\pi}[a_1^2 + (1 - 2a_1) f(\phi(x_1))] \\ &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[(\pi(x) - f(\phi(x)))^2 + f(\phi(x)) - f(\phi(x))^2] \\ &= \mathop{\mathbb{E}}_{x,y}[(\pi(x) - y)^2] + Z \end{aligned}$$

where $Z := \mathbb{E}_{x,y}[f(\phi(x)) - f(\phi(x))^2 - (f(\phi(x)) - y)^2]$, and the expectations are over a fresh sample (x, y) from the same distribution as $(x^{(i)}, y^{(i)})$. But now by Hoeffding's inequality, the union bound, and choice of $m := 16\epsilon^{-2} \log(4|\Psi|/\delta)$, we have with probability at least $1 - \delta/2$ that for all $\pi \in \Psi$,

$$\left| \widehat{E}(\pi) - \mathop{\mathbb{E}}_{x,y}[(\pi(x) - y)^2] \right| \leq \epsilon/4,$$

where $\widehat{E}(\pi)$ is the empirical loss for π computed in Algorithm 8 of `OneRed`. In this event, we get that

$$\begin{aligned} \mathop{\mathbb{E}}_{x,y}[(\widehat{\pi}(x) - y)^2] &\leq \frac{\epsilon}{2} + \mathop{\mathbb{E}}_{x,y}[(\pi^\star(x) - y)^2] \\ &= \frac{\epsilon}{2} + 1 - d_2^{M,\pi^\star}(1) - Z \\ &\leq \frac{\epsilon}{2} + 1 - \max_{\pi \in \Pi} d_2^{M,\pi}(1) - Z \\ &\leq \frac{3\epsilon}{4} + \min_{\pi \in \Pi} \mathop{\mathbb{E}}_{x,y}[(\pi(x) - y)^2]. \end{aligned}$$

It follows that

$$\mathop{\mathbb{E}}_{x}[(\widehat{\pi}(x) - f(\phi(x)))^2] \leq \frac{3\epsilon}{4} + \min_{\pi \in \Pi} \mathop{\mathbb{E}}_{x}[(\pi(x) - f(\phi(x)))^2] \leq \epsilon$$

since the policy $\pi(x) = \sqrt{4/\epsilon} \lfloor f(\phi(x) \cdot \sqrt{4/\epsilon} \rfloor$ has squared error at most $\epsilon/4$. ■

G.2.2. THE `NoiselessOneRed` REDUCTION

In this section, we adapt the reduction from Section G.2.1 to the reset access model. However, this requires weakening the regression problem to be *noiseless*:

Algorithm 9 `NoiselessOneRed`$(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: Reduction from noiseless one-context regression to RL with resets

input: Oracle $\mathcal{O}$ for reward-free RL with resets; samples $(x^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.
Set $\varepsilon_A := \sqrt{\epsilon/4}$ and $i = 1$. Initialize $\mathcal{O}$ with tolerance $\epsilon/4$, failure probability $\delta/2$, horizon $H := 2$, and action set $\mathcal{A} := \{0, \varepsilon_A, \dots, 1 - \varepsilon_A\}$. Simulate $\mathcal{O}$ as follows:
repeat
    When $\mathcal{O}$ queries the first sampling oracle, pass $x^{(i)}$ and set $i \leftarrow i + 1$.
    When $\mathcal{O}$ queries the second sampling oracle with inputs $x_1 \in \mathcal{X}^{\mathsf{aug}}$ and $a_1 \in \mathcal{A}$, identify any $j < i$ with $x_1 = x^{(j)}$. With probability $(a - y^{(j)})^2$, pass observation 0. Otherwise, pass observation 1.
until $\mathcal{O}$ returns policy cover Ψ
$m \leftarrow 16\epsilon^{-2}\log(4|\Psi|/\delta)$.
for $\pi \in \Psi$ **do**
    Compute $\widehat{E}(\pi) := \frac{1}{m}\sum_{i=n-m+1}^n (\pi(x^{(i)}) - y^{(i)})^2$.
return: $\widehat{\pi} := \arg\min_{\pi \in \Psi} \widehat{E}(\pi)$.

Definition G.3 (Noiseless one-context regression) *Let $N_{\mathsf{reg}} : (0, 1/2)^2 \to \mathbb{N}$ be a function. An algorithm `Alg` is an N_{reg}-efficient noiseless one-context regression algorithm for Φ if the following holds. Fix $\epsilon, \delta \in (0, 1/2)$, $n \in \mathbb{N}$, and $\phi \in \Phi$. Let $\mathcal{D} \in \Delta(\mathcal{X})$ be a distribution, and let $f : \mathcal{S} \to \{0, 1\}$. Let $(x^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $x^{(i)} \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi(x^{(i)}))$. If $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$, then with probability at least $1 - \delta$, the output of `Alg`$((x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$ is a circuit $\mathcal{R} : \mathcal{X} \to [0, 1]$ satisfying*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} (\mathcal{R}(x) - f(\phi(x)))^2 \leq \epsilon.$$

With this definition, the following theorem shows that `NoiselessOneRed` (Algorithm 9) reduces noiseless one-context regression for concept class Φ to reward-free RL for concept class Φ^{aug} in the reset model. By combining with Proposition G.6, this implies that there is a reduction to reward-free RL for concept class Φ itself, so long as Φ is regular. We leave it as an open problem whether the reduction can be strengthened to work with noisy one-context regression.

Proposition G.4 *Suppose that $\mathcal{O}$ is a $(N_{\mathsf{RL}}, K_{\mathsf{RL}})$-efficient reward-free reset RL algorithm for Φ^{aug}. Then `NoiselessOneRed`$(\mathcal{O}, \cdot)$ is a N_{reg}-efficient noiseless one-context regression algorithm for Φ with*

$$N_{\mathsf{reg}}(\epsilon, \delta) = N_{\mathsf{RL}}\left(\frac{\epsilon}{4}, \frac{\delta}{2}, 2, \sqrt{\frac{4}{\epsilon}}\right) + \frac{16\log(4K_{\mathsf{RL}}\left(\frac{\epsilon}{4}, \frac{\delta}{2}, 2, \sqrt{\frac{4}{\epsilon}}\right)/\delta)}{\epsilon^2}.$$

Proof. Let $\epsilon, \delta > 0$, $\mathcal{D} \in \Delta(\mathcal{X})$, and $f : \mathcal{S} \to \{0, 1\}$. Let $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$. Let $(x^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $x^{(i)} \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi(x^{(i)}))$ for some $\phi \in \Phi$. We analyze the execution of `NoiselessOneRed`$(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$. We know that $n - m \geq N_{\mathsf{RL}}(\epsilon/4, \delta/2, 2, \sqrt{4/\epsilon})$, since $n \geq N_{\mathsf{reg}}(\epsilon, \delta)$. Now the first sampling oracle provides independent

Algorithm 10 $\texttt{OneTwo}(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: One-context regression to two-context regression reduction

1: **input:** Two-context regression oracle $\mathcal{O}$; samples $(x^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.
2: Pick arbitrary $\overline{x} \in \mathcal{X}$.
3: Compute

$$\widetilde{\mathcal{R}} \leftarrow \mathcal{O}((x_1^{(i)}, \overline{x}, y^{(i)})_{i=1}^n, \epsilon, \delta).$$

4: **return:** $\mathcal{R}$ defined by $\mathcal{R}(x) := \widetilde{\mathcal{R}}(x, \overline{x})$.

samples from $\mathcal{D}$. For the second sampling oracle, since $\mathcal{O}$ can only query $x_1 \in \mathcal{X}^{\mathsf{aug}}$ which it has previously seen, it must be that there exists $j < i$ with $x_1 = x^{(j)}$. Moreover, since the range of f is in $\{0, 1\}$, we have deterministically that $y^{(j)} = f(\phi(x^{(j)}))$. Conditioned on the queries x_1 and a_1, the output of the second sampling oracle is therefore independent of all prior queries, and the probability of observing 0 is

$$\begin{aligned}\mathbb{E}[(a_1 - y^{(j)})^2 \mid x^{(j)}, a_1] &= (a_1)^2 - 2a_1 \mathbb{E}[y^{(j)} \mid x^{(j)}] + \mathbb{E}[(y^{(j)})^2 \mid x^{(j)}] \\ &= (a_1)^2 + (1 - 2a_1) f(\phi(x_1))\end{aligned}$$

where the final equality uses that $y^{(i)} \in \{0, 1\}$. Thus, $\texttt{NoiselessOneRed}$ simulates $\mathcal{O}$ on a Φ^{aug}-decodable block MDP M with horizon 2, observation space $\mathcal{X}^{\mathsf{aug}}$, latent state space $\mathcal{S}^{\mathsf{aug}}$, initial observation distribution $\mathcal{D}$, and transition distribution defined by

$$\mathbb{P}_2(0 \mid x_1, a_1) := a_1^2 + (1 - 2a_1) f(\phi(x_1)),$$

$$\mathbb{P}_2(1 \mid x_1, a_1) := 1 - \mathbb{P}_2(0 \mid x_1, a_1).$$

The remainder of the proof is identical to that of Proposition G.2. ■

G.2.3. THE $\texttt{OneTwo}$ REDUCTION

The following proposition shows that $\texttt{OneTwo}$ (Algorithm 10) is an efficient reduction from one-context regression to two-context regression.

Proposition G.5 *Fix sets $\mathcal{X}, \mathcal{S}$ and $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$. Suppose that $\mathcal{O}$ is an N_{reg}-efficient two-context regression oracle for Φ. Then $\texttt{OneTwo}(\mathcal{O}, \cdot)$ is an N_{reg}-efficient one-context regression oracle for Φ.*

Proof. Let $(x^{(i)}, y^{(i)})_{i=1}^n$ be i.i.d. samples with $x^{(i)} \sim \mathcal{D}$, $y^{(i)} \in \{0, 1\}$, and $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi(x^{(i)}))$ for some $\phi \in \Phi$ and $f : \mathcal{S} \to [0, 1]$. Then for any fixed $\overline{x} \in \mathcal{X}$, the samples $(x_1^{(i)}, \overline{x}, y^{(i)})_{i=1}^n$ are i.i.d., the distribution of $(x_1^{(i)}, \overline{x})$ is ϕ-realizable (since $\overline{x}$ is independent of $x_1^{(i)}$), and $\mathbb{E}[y^{(i)} \mid x_1^{(i)}, \overline{x}] = g(\phi(x^{(i)}), \phi(\overline{x}))$ where $g(s_1, s_2) := f(s_1)$. By Definition 2.4, so long as $N \geq N_{\mathsf{reg}}(\epsilon, \delta)$, it holds with probability at least $1 - \delta$ that the predictor $\widetilde{\mathcal{R}}$ computed in Line 3 satisfies

$$\mathop{\mathbb{E}}_{x_1 \sim \mathcal{D}} (\widetilde{\mathcal{R}}(x_1, \overline{x}) - g(\phi(x_1), \phi(\overline{x})))^2 \leq \epsilon.$$

Algorithm 11 `OneAug`$(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: One-context regression for Φ^{aug}

1: **input:** One-context regression oracle $\mathcal{O}$ for Φ; samples $(x^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.
2: Let $S := \{i \in [n] : x^{(i)} \in \mathcal{X}\}$.
3: Compute

$$\mathcal{R}_{\mathcal{X}} \leftarrow \mathcal{O}((x^{(i)}, y^{(i)})_{i \in S}, \epsilon/6, \delta/6).$$

4: For $b \in \{0, 1\}$, let $S_b := \{i \in [n] : x^{(i)} = b\}$ and compute

$$\mathcal{R}_b := \frac{1}{|S_b|} \sum_{i \in S_b} y^{(i)}.$$

5: **return:** $\mathcal{R} : \mathcal{X}^{\mathsf{aug}} \to [0, 1]$ defined by

$$\mathcal{R}(x) := \begin{cases} \mathcal{R}_{\mathcal{X}}(x) & \text{if } x \in \mathcal{X} \\ \mathcal{R}_0 & \text{if } x = 0 \\ \mathcal{R}_1 & \text{if } x = 1 \end{cases}.$$

In this event, by definition of g, the output $\mathcal{R}(\cdot) := \widetilde{\mathcal{R}}(\cdot, \overline{x})$ of `OneTwo` satisfies

$$\mathop{\mathbb{E}}_{x_1 \sim \mathcal{D}} (\mathcal{R}(x_1) - f(\phi(x_1)))^2 \leq \epsilon$$

as required for one-context regression (Definition 2.2). ■

G.2.4. THE `OneAug` REDUCTION

The following proposition shows that `OneAug` (Algorithm 11) is an efficient reduction from one-context reduction for concept class Φ^{aug} to the same problem for concept class Φ. The basic idea is that the states $\{0, 1\}$ are fully observed, so they can be regressed separately via mean estimation.

Proposition G.6 *There is a constant $C_{G.6} > 0$ so that the following holds. Fix sets $\mathcal{X}, \mathcal{S}$ and $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$. Suppose that $\mathcal{O}$ is an N_{reg}-efficient one-context regression oracle for Φ. Then* `OneAug`$(\mathcal{O}, \cdot)$ *is an N'_{reg}-efficient one-context regression oracle for Φ^{aug} with*

$$N'_{\mathsf{reg}}(\epsilon, \delta) = C_{G.6} \left(\epsilon^{-1} N_{\mathsf{reg}}(\epsilon/6, \delta/6) + \epsilon^{-2} \log(12/\delta) \right).$$

Proof. Fix $\epsilon, \delta \in (0, 1)$, $\mathcal{D} \in \Delta(\mathcal{X}^{\mathsf{aug}})$, $f : \mathcal{S}^{\mathsf{aug}} \to [0, 1]$, and $\phi^{\mathsf{aug}} \in \Phi^{\mathsf{aug}}$. By definition there is $\phi \in \Phi$ with $\phi^{\mathsf{aug}} = \mathrm{aug}(\phi)$. We invoke Lemma G.7 with the following parameters. Set $\mathcal{Z} := \mathcal{X}^{\mathsf{aug}} \times \{0, 1\}$, and define $\mathcal{Z}_0 := \{0\} \times \{0, 1\}$, $\mathcal{Z}_1 = \{1\} \times \{0, 1\}$, and $\mathcal{Z}_2 := \mathcal{X} \times \{0, 1\}$. Let $\mu \in \Delta(\mathcal{Z})$ be the distribution of (x, y) where $x \sim \mathcal{D}$ and $y \in \{0, 1\}$ with $\mathbb{E}[y \mid x] = f(\phi^{\mathsf{aug}}(x))$.

Let μ_0, μ_1, μ_2 be the conditional distributions associated with $\mathcal{Z}_0, \mathcal{Z}_1, \mathcal{Z}_2$. Finally, define h_0, h_1, h_2 by

$$h_0((x^{(i)}, y^{(i)})_{i=1}^m)(x, y) := \left(\frac{1}{m} \sum_{i=1}^m y^{(i)} - f(0) \right)^2,$$

$$h_1((x^{(i)}, y^{(i)})_{i=1}^m)(x, y) := \left(\frac{1}{m} \sum_{i=1}^m y^{(i)} - f(1) \right)^2,$$

$$h_2((x^{(i)}, y^{(i)})_{i=1}^m)(x, y) := \left(\mathcal{O}((x^{(i)}, y^{(i)})_{i=1}^m, \epsilon/6, \delta/6)(x) - f(\phi^{\mathsf{aug}}(x)) \right)^2.$$

By Hoeffding's inequality, if $(x^{(i)}, y^{(i)})_{i=1}^m$ are i.i.d. samples from μ_0 and $m \geq 6\epsilon^{-1} \log(12/\delta)$, then since $\mathbb{E}[y^{(i)}] = f(\phi^{\mathsf{aug}}(0)) = f(0)$, it holds with probability at least $1 - \delta/6$ that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mu_0} [h_0((x^{(i)}, y^{(i)})_{i=1}^m)(x, y)] \leq \epsilon/6.$$

The same argument holds for h_1. Finally, if $(x^{(i)}, y^{(i)})_{i=1}^m$ are i.i.d. samples from μ_2 and $m \geq N_{\mathsf{reg}}(\epsilon/6, \delta/6)$, then by the assumption on $\mathcal{O}$ and the fact that $\mathbb{E}[y^{(i)} \mid x^{(i)}] = f(\phi^{\mathsf{aug}}(x^{(i)})) = f(\phi(x^{(i)}))$ since $x^{(i)} \in \mathcal{X}$, it holds that with probability at least $1 - \delta/6$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mu_2} [h_2((x^{(i)}, y^{(i)})_{i=1}^m)(x, y)] \leq \epsilon/6.$$

We conclude from Lemma G.7 that if $n \geq N'_{\mathsf{reg}}(\epsilon, \delta) = C_{G.6} \left(\epsilon^{-1} N_{\mathsf{reg}}(\epsilon/6, \delta/6) + \epsilon^{-2} \log(12/\delta) \right)$ and $C_{G.6}$ is a sufficiently large constant, then with probability at least $1 - \delta$ over i.i.d. samples $(x^{(i)}, y^{(i)})_{i=1}^n$ from μ, the function H defined in Lemma G.7 satisfies $\mathbb{E}_{(x,y)\sim\mu}[H(x, y)] \leq \epsilon$. But we can write

$$H(x, y) = (\mathcal{R}(x) - f(\phi^{\mathsf{aug}}(x)))^2$$

where

$$\mathcal{R}(x) = \begin{cases} \frac{1}{\#\{i : x^{(i)} = 0\}} \sum_{i:x^{(i)}=0} y^{(i)} & \text{if } x = 0 \\ \frac{1}{\#\{i : x^{(i)} = 1\}} \sum_{i:x^{(i)}=1} y^{(i)} & \text{if } x = 1 \\ \mathcal{O}((x^{(i)}, y^{(i)})_{i:x^{(i)} \in \mathcal{X}}, \epsilon/6, \delta/6)(x) & \text{if } x \in \mathcal{X} \end{cases}.$$

This is precisely the predictor computed by $\texttt{OneAug}(\mathcal{O}, (x^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$, so we have shown that $\texttt{OneAug}(\mathcal{O}, \cdot)$ is an N'_{reg}-efficient one-context regression algorithm for Φ^{aug}. ■

The preceding proof used the following convenient technical lemma about composing statistical predictors on different subsets of a space:

Lemma G.7 *Let $\mathcal{Z}$ be a set, and let $\mathcal{Z}_1 \sqcup \cdots \sqcup \mathcal{Z}_k$ be a partition of $\mathcal{Z}$. Let $\mu \in \Delta(\mathcal{Z})$ be a distribution. For each $i \in [k]$ let μ_i be the distribution of $z \sim \mu$ conditioned on $z \in \mathcal{Z}_i$, and let $h_i : (\mathcal{Z}_i)^\star \to (\mathcal{Z}_i \to [0, 1])$ be a function with the following property: given at least m i.i.d. samples $(z_i^{(j)})_j$ from μ_i, it holds with probability at least $1 - \delta$ that $\mathbb{E}_{z\sim\mu_i}[h_i((z_i^{(j)})_j)(z)] \leq \epsilon$.*

Let $(z^{(j)})_{j=1}^n$ be n i.i.d. samples from μ, and for each $i \in [k]$ let $S_i = \{j : z^{(j)} \in \mathcal{Z}_i\}$. Define $H : \mathcal{Z} \to [0, 1]$ by

$$H(z) := h_i((z^{(j)})_{j\in S_i})(z) \text{ for } z \in \mathcal{Z}_i.$$

Algorithm 12 TwoAug$(\mathcal{O}, (x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$: Two-context regression for Φ^{aug}

1: **input:** Two-context regression oracle $\mathcal{O}$ for Φ; samples $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$; tolerances ϵ, δ.

2: Fix $\overline{x} \in \mathcal{X}$.

3: For all $i \in [n]$, define $\widetilde{x}_1^{(i)} = x_1^{(i)}$ if $x_1^{(i)} \in \mathcal{X}$, and otherwise $\overline{x}$. Similarly define $\widetilde{x}_2^{(i)}$.

4: For all pairs $B, B' \in \{\{0\}, \{1\}, \mathcal{X}\}$, define

$$\mathcal{R}_{B,B'} \leftarrow \mathcal{O}((\widetilde{x}_1^{(i)}, \widetilde{x}_2^{(i)}, y^{(i)})_{i:x_1^{(i)} \in B, x_2^{(i)} \in B'}, \epsilon/18, \delta/18).$$

5: **return:** the predictor $\mathcal{R} : \mathcal{X}^{\mathsf{aug}} \times \mathcal{X}^{\mathsf{aug}} \to [0,1]$ defined as follows. Given x_1, x_2, define $\widetilde{x}_1 = x_1$ if $x_1 \in \mathcal{X}$ and $\widetilde{x}_1 = \overline{x}$ otherwise; similarly define $\widetilde{x}_2$. Then output $\mathcal{R}_{B,B'}(\widetilde{x}_1, \widetilde{x}_2)$ for the unique B, B' with $x_1 \in B$ and $x_2 \in B'$.

If $n \geq 2m/\epsilon + 8\log(1/\delta)/\epsilon$, then with probability at least $1 - 2k\delta$, it holds that

$$\mathop{\mathbb{E}}_{z \sim \mu}[H(z)] \leq 2k\epsilon.$$

Proof. Let $\mathcal{I}_{\mathsf{good}} = \{i \in [k] : \mu(\mathcal{Z}_i) \geq \epsilon\}$. Let $\mathcal{E}$ be the event that $|S_i| \geq m$ for all $i \in \mathcal{I}_{\mathsf{good}}$. For each such i, we have by a Chernoff bound and choice of n that,

$$\Pr[|S_i| < m] \leq \Pr\left[|S_i| < \frac{1}{2}\mu(\mathcal{Z}_i)n\right] \leq e^{-\mu(\mathcal{Z}_i)n/8} \leq \delta.$$

Therefore $\Pr[\mathcal{E}] \geq 1 - \delta k$. Condition on $S_1, \dots, S_k$ and suppose that $\mathcal{E}$ holds. For each $i \in \mathcal{I}_{\mathsf{good}}$, the tuple $(z^{(j)})_{j \in S_i}$ consists of at least m i.i.d. samples from μ_i. Therefore by the lemma assumption, with probability at least $1 - \delta k$, we have for all $i \in \mathcal{I}_{\mathsf{good}}$ that

$$\mathop{\mathbb{E}}_{z \sim \mu_i}[h_i((z^{(j)})_{j \in S_i})(z)] \leq \epsilon.$$

Condition additionally on this event. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{z \sim \mu}[H(z)] &= \sum_{i=1}^k \mu(\mathcal{Z}_i) \mathop{\mathbb{E}}_{z \sim \mu_i}[h_i((z^{(j)})_{j \in S_i}(z)] \\
&\leq \sum_{i \in \mathcal{I}_{\mathsf{good}}} \mu(\mathcal{Z}_i)\epsilon + \sum_{i \in [k] \setminus \mathcal{I}_{\mathsf{good}}} \mu(\mathcal{Z}_i) \\
&\leq (k+1)\epsilon
\end{aligned}$$

which suffices for the claimed bound, and holds in an event with probability at least $1 - 2\delta k$. ■

G.2.5. THE TwoAug REDUCTION

The following proposition shows that TwoAug (Algorithm 12) is an efficient reduction from two-context reduction for concept class Φ^{aug} to the same problem for concept class Φ. Similar to OneAug, the reduction decomposes the regression problem into several parts, though now some of the parts are effectively one-context regression problems.

Proposition G.8 *There is a constant $C_{G.8} > 0$ so that the following holds. Fix sets $\mathcal{X}, \mathcal{S}$ and $\Phi \subseteq (\mathcal{X} \to \mathcal{S})$. Suppose that $\mathcal{O}$ is an N_{reg}-efficient two-context regression oracle for Φ. Then $\mathsf{TwoAug}(\mathcal{O}, \cdot)$ is an N'_{reg}-efficient two-context regression oracle for Φ^{aug} with*

$$N'_{\mathsf{reg}}(\epsilon, \delta) = C_{G.8}\left(\epsilon^{-1} N_{\mathsf{reg}}(\epsilon/18, \delta/18) + \epsilon^{-2}\log(36/\delta)\right).$$

Proof. Fix $\epsilon, \delta \in (0,1)$, $\mathcal{D} \in \Delta(\mathcal{X}^{\mathsf{aug}} \times \mathcal{X}^{\mathsf{aug}})$, $f : \mathcal{S}^{\mathsf{aug}} \times \mathcal{S}^{\mathsf{aug}} \to [0,1]$, and $\phi^{\mathsf{aug}} \in \Phi^{\mathsf{aug}}$. Suppose that $\mathcal{D}$ is ϕ^{aug}-realizable (Definition 2.3). By definition of Φ^{aug}, there is $\phi \in \Phi$ with $\phi^{\mathsf{aug}} = \mathrm{aug}(\phi)$. We invoke Lemma G.7 with the following parameters. Set $\mathcal{Z} := \mathcal{X}^{\mathsf{aug}} \times \mathcal{X}^{\mathsf{aug}} \times \{0,1\}$. Let $\mu \in \Delta(\mathcal{Z})$ be the distribution of (x_1, x_2, y) where $(x_1, x_2) \sim \mathcal{D}$ and $y \in \{0,1\}$ with $\mathbb{E}[y \mid x_1, x_2] = f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))$. For each $B, B' \in \{\{0\}, \{1\}, \mathcal{X}\}$ define $\mathcal{Z}_{B,B'} := \mathcal{B} \times \mathcal{B}' \times \{0,1\}$, and let $\mu_{B,B'}$ be the associated conditional distribution. Fix $\overline{x} \in \mathcal{X}$, and define $h_{B,B'}$ by

$$\begin{aligned} &h_{B,B'}((x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^m)(x_1, x_2, y) \\ &:= \begin{cases} \left(\mathcal{O}((x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(x_1, x_2) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))\right)^2 & \text{if } B, B' = \mathcal{X} \\ \left(\mathcal{O}((x_1^{(i)}, \overline{x}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(x_1, \overline{x}) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))\right)^2 & \text{if } B = \mathcal{X}, B' \neq \mathcal{X} \\ \left(\mathcal{O}((\overline{x}, x_2^{(i)}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(\overline{x}, x_2) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))\right)^2 & \text{if } B \neq \mathcal{X}, B' = \mathcal{X} \\ \left(\mathcal{O}((\overline{x}, \overline{x}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(\overline{x}, \overline{x}) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))\right)^2 & \text{if } B, B' \neq \mathcal{X} \end{cases}. \end{aligned}$$

Fix B, B'. Let $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^m$ be i.i.d. samples from $\mu_{B,B'}$ and suppose $m \geq N_{\mathsf{reg}}(\epsilon/18, \delta/18)$. If $B, B' = \mathcal{X}$ then $\mathbb{E}[y^{(i)} \mid x_1^{(i)}, x_2^{(i)}] = f(\phi^{\mathsf{aug}}(x_1^{(i)}), \phi^{\mathsf{aug}}(x_2^{(i)})) = f(\phi(x_1^{(i)}), \phi(x_2^{(i)}))$. Moreover, the marginal distribution of $(x_1^{(i)}, x_2^{(i)})$ is ϕ-realizable since it can be expressed as the conditional distribution of $\mathcal{D}$ under the event that $\phi^{\mathsf{aug}}(x_1^{(i)}), \phi^{\mathsf{aug}}(x_2^{(i)}) \in \mathcal{S}$. Thus, by the assumption on $\mathcal{O}$, it holds with probability at least $1 - \delta/18$ that

$$\mathop{\mathbb{E}}_{(x_1, x_2, y) \sim \mu_{B,B'}}\left[\left(\mathcal{O}((x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(x_1, x_2) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))\right)^2\right] \leq \epsilon/18.$$

If $B = \mathcal{X}, B' = \{0\}$ then $\mathbb{E}[y^{(i)} \mid x_1^{(i)}, \overline{x}] = \mathbb{E}[y^{(i)} \mid x_1^{(i)}] = f(\phi(x_1^{(i)}), 0)$ since $x_2^{(i)} = \phi^{\mathsf{aug}}(x_2^{(i)}) = 0$ is fixed under $\mu_{B,B'}$. Moreover, the marginal distribution of $(x_1^{(i)}, \overline{x})$ is ϕ-realizable since $\overline{x}$ is fixed and hence independent of $x_1^{(i)}$). Thus, by the assumption on $\mathcal{O}$, it holds with probability at least $1 - \delta/18$ that

$$\begin{aligned} &\mathop{\mathbb{E}}_{(x_1, x_2, y) \sim \mu_{B,B'}}\left[\left(\mathcal{O}((x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(x_1, x_2) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2))\right)^2\right] \\ &= \mathop{\mathbb{E}}_{(x,y) \sim \mu_{B,B'}}\left[\left(\mathcal{O}((x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^m, \epsilon/18, \delta/18)(x_1, x_2) - f(\phi(x_1), 0)\right)^2\right] \\ &\leq \epsilon/18. \end{aligned}$$

The remaining cases follow by analogous arguments. Thus, we can apply Lemma G.7. If $n \geq N'_{\mathsf{reg}}(\epsilon, \delta) = C_{G.8}(\epsilon^{-1} N_{\mathsf{reg}}(\epsilon/18, \delta/18) + \epsilon^{-1}\log(36/\delta))$, where $C_{G.8}$ is a sufficiently large constant, then with probability at least $1 - \delta$ over i.i.d. samples $(x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n$ from μ, the function H defined in Lemma G.7 satisfies $\mathbb{E}_{(x_1, x_2, y) \sim \mu}[H(x_1, x_2, y)] \leq \epsilon$. But we can write

$$H(x_1, x_2, y) = (\mathcal{R}(x) - f(\phi^{\mathsf{aug}}(x_1), \phi^{\mathsf{aug}}(x_2)))^2$$

where

$$\mathcal{R}(x_1, x_2) = \begin{cases} \mathcal{O}((x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i:x_1^{(i)}, x_2^{(i)} \in \mathcal{X}}, \epsilon/18, \delta/18)(x_1, x_2) & \text{if } x_1, x_2 \in \mathcal{X} \\ \mathcal{O}((x_1^{(i)}, \overline{x}, y^{(i)})_{i:x_1^{(i)} \in \mathcal{X}, x_2^{(i)} = x_2}, \epsilon/18, \delta/18)(x_1, \overline{x}) & \text{if } x_1 \in \mathcal{X}, x_2 \notin \mathcal{X} \\ \mathcal{O}((\overline{x}, x_2^{(i)}, y^{(i)})_{i:x_1^{(i)} = x_1, x_2^{(i)} \in \mathcal{X}}, \epsilon/18, \delta/18)(\overline{x}, x_2) & \text{if } x_1 \notin \mathcal{X}, x_2 \in \mathcal{X} \\ \mathcal{O}((\overline{x}, \overline{x}, y^{(i)})_{i:x_1^{(i)} = x_1, x_2^{(i)} = x_2}, \epsilon/18, \delta/18)(\overline{x}, \overline{x}) & \text{if } x_1, x_2 \notin \mathcal{X} \end{cases}.$$

This is exactly the predictor computed by $\mathsf{TwoAug}(\mathcal{O}, (x_1^{(i)}, x_2^{(i)}, y^{(i)})_{i=1}^n, \epsilon, \delta)$, so we have shown that $\mathsf{TwoAug}(\mathcal{O}, \cdot)$ is an N'_{reg}-efficient two-context regression algorithm for Φ^{aug}. ■