# Rich-Observation Reinforcement Learning with Continuous Latent Dynamics

Yuda Song [1] Lili Wu [2] Dylan J. Foster [2] Akshay Krishnamurthy [2]

## Abstract

Sample-efficiency and reliability remain major bottlenecks toward wide adoption of reinforcement learning algorithms in continuous settings with high-dimensional perceptual inputs. Toward addressing these challenges, we introduce a new theoretical framework, RichCLD ("Rich-Observation RL with Continuous Latent Dynamics"), where the agent performs control based on high-dimensional observations, but the environment is governed by low-dimensional latent states and Lipschitz continuous dynamics. Our main contribution is an algorithm for RichCLD that is provably statistically and computationally efficient. The core of our algorithm is a new representation learning objective; we show that prior representation learning schemes tailored to discrete dynamics do not naturally extend to the continuous setting. Our objective is amenable to practical implementation, and empirically, it compares favorably to prior schemes in a standard evaluation protocol. We further provide several insights into the statistical complexity of the RichCLD framework, in particular proving that certain notions of Lipschitzness that admit sample-efficient learning in the absence of rich observations are insufficient in the rich-observation setting.

## 1. Introduction

It is becoming increasingly common to deploy algorithms for reinforcement learning and control in systems where the underlying ("latent") dynamics are nonlinear, continuous, and low-dimensional, yet the agent perceives the environment through high-dimensional ("rich") observations such as images from a camera (Wahlström et al., 2015; Levine et al., 2016; Kumar et al., 2021; Nair et al., 2023; Baker et al., 2022; Brohan et al., 2022). These domains demand that agents (i) efficiently explore in the face of complex nonlinearities, and (ii) learn continuous representations that respect the structure of the latent dynamics, ideally in tandem with exploration. In spite of extensive empirical investigation into modeling and algorithm design (Laskin et al., 2020; Yarats et al., 2021a; Hafner et al., 2023), sample-efficiency and reliability remain major challenges (Dean et al., 2020), and our understanding of fundamental algorithmic principles for representation learning and exploration is still in its infancy.

Toward understanding algorithmic principles and fundamental limits for reinforcement learning and control with high-dimensional observations, a recent line of theoretical research adopts the framework of *rich-observation reinforcement learning* (c.f., Du et al., 2019; Misra et al., 2020; Mhammedi et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b). This line of work has led to new provably efficient algorithms, but, to date, has focused on systems with discrete ("tabular") or linear latent dynamics, which is unsuitable for most real-world control applications.

**The RichCLD framework.** We initiate the study of theoretically-sound algorithms for reinforcement learning with continuous latent dynamics by introducing a new framework, the RichCLD ("Rich-Observation RL with Continuous Latent Dynamics") framework. In the RichCLD framework, the agent performs control based on high-dimensional observations (e.g., images from a camera), but the underlying state obeys Lipschitz continuous dynamics. Lipschitz continuous dynamics have been studied extensively in the absence of rich observations (Kakade et al., 2003; Shah & Xie, 2018; Henaff, 2019; Ni et al., 2019; Song & Sun, 2019; Cao & Krishnamurthy, 2020; Sinclair et al., 2023) and are versatile enough to capture applications ranging from robotic control (Tedrake, 2023) to online resource allocation (Sinclair et al., 2023). Our framework addresses these applications in a more realistic setting where the agent perceives the system through high-dimensional feedback.

The central challenge in rich-observation RL—the RichCLD framework included—is that representation learning and exploration must be interleaved. Agents need to learn a good representation to guide exploration, but doing so requires gathering data from throughout the space, and is difficult unless the agent already knows how to explore. Since exploration under Lipschitz dynamics in the absence of rich observations is relatively well-understood, the main

[1]Carnegie Mellon University [2]Microsoft Research. Correspondence to: Yuda Song <yudas@cs.cmu.edu>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

algorithmic challenge can be summarized as:

*How can we design representation learning schemes that gracefully compose with exploration in the presence of continuous latent dynamics?*

In this paper, we develop representation learning schemes that address this challenge, and combine them with principled exploration, yielding provable algorithms for reinforcement learning in the RichCLD framework.

**Contributions.** We develop new statistical, algorithmic, and empirical results in the RichCLD framework.

(C1) **Statistical complexity.** We establish that the RichCLD framework is statistically tractable by analyzing a variant of a general-purpose algorithm for reinforcement learning, GOLF (Jin et al., 2021; Xie et al., 2023; Amortila et al., 2024). The sample complexity has a nonparametric flavor, scaling exponentially with the latent state dimension, as expected. We also show that weaker notions of continuity are insufficient for statistically tractable learning, which separates the rich-observation setting from the classical one.

(C2) **Representation learning.** We provide a new representation learning scheme, Bellman Consistent Representation Learning (BCRL.C), which provably learns a representation that enables downstream exploration and reward maximization in the RichCLD framework. We complement this result by showing that standard representation learning procedures considered in prior work do not readily adapt to handle Lipschitz continuous dynamics as-is (Misra et al., 2020; Lamb et al., 2023; Mhammedi et al., 2023b).

(C3) **End-to-end algorithm.** By interleaving BCRL.C with exploration, we obtain CRIEE, a provably (statistically and computationally) efficient algorithm for learning in the RichCLD framework.

(C4) **Practical implementation and empirical evaluation.** We derive a practical variant of BCRL.C and provide a qualitative and quantitative evaluation in visual navigation environments (Koul et al., 2023) and the visual D4RL benchmark (Lu et al., 2023).

Together, we believe these results constitute a useful starting point for further theoretical investigation into continuous RL with rich observations. We discuss avenues for future work, including improved sample complexity, weakened continuity assumptions, and adaptivity, in Section 6.

**Paper organization.** Section 2 introduces the RichCLD framework and our sample complexity desiderata. Section 3 addresses (C1) while Section 4 addresses (C2) and (C3). Experiments (C4) are presented in Section 5, and we conclude with a discussion in Section 6. Proofs and additional details, including further related work, are deferred to the appendix.

## 2. Problem Setting

**Markov decision processes.** We consider an episodic finite-horizon Markov decision process (MDP) with horizon $H \in \mathbb{N}$. An MDP $M := (\mathcal{X}, \mathcal{A}, H, P, R)$ consists of state space $\mathcal{X}$, action space $\mathcal{A}$, transition distribution $P = \{P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})\}_{h=1}^H$, and reward function[1] $R = \{R_h : \mathcal{X} \times \mathcal{A} \to [0,1]\}_{h=1}^H$. Executing a nonstationary policy $\pi = (\pi_1, \ldots, \pi_H)$, where each $\pi_h \in (\mathcal{X} \to \Delta(\mathcal{A}))$, for an episode induces a trajectory $\tau = (x_1, a_1, r_1, \ldots, x_H, a_H, r_H)$ via the process $x_h \sim P_h(x_{h-1}, a_{h-1})$, $a_h \sim \pi_h(x_h)$ and $r_h = R_h(x_h, a_h)$ for all $h \in [H]$, without loss generality, we assume that there is a fixed initial state $x_1$. We let $\mathbb{P}^\pi$ and $\mathbb{E}^\pi$ denote the law and expectation under this process, and we define the occupancy measures $d_h^\pi(x) = \mathbb{P}^\pi(x_h = x)$ and $d_h^\pi(x,a) = \mathbb{P}^\pi(x_h = x, a_h = a)$. Following convention, we assume $\sum_{h=1}^H r_h \le 1$ almost surely. We define $J(\pi) = \mathbb{E}^\pi\left[\sum_{h=1}^H r_h\right]$ as the expected reward under the policy $\pi$ and let $\pi^\star \in \arg\max_{\pi \in \Pi} J(\pi)$ be the optimal policy that satisfies Bellman's equations, where $\Pi$ is the set of all randomized non-stationary policies.

**The RichCLD model.** A RichCLD model is an MDP with particularly structured dynamics and rewards, corresponding to a rich-observation MDP with a continuous latent state space.

A rich-observation MDP (Krishnamurthy et al., 2016; Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b) is an MDP with a *latent state space* $\mathcal{S}$ and decoders $\{\phi_h^\star\}_{h=1}^H : \mathcal{X} \to \mathcal{S}$ such that: (1) the reward function depends only on the latent state $s_h := \phi_h^\star(x_h)$, i.e., $R_h(x_h, a_h) = R_h^{\text{latent}}(\phi_h^\star(x_h), a_h)$, and (2) the transition dynamics operate on the latent state in the sense that, for each $x_h, a_h$ the dynamics evolve as $s_{h+1} \sim P_h^{\text{latent}}(\phi_h^\star(x_h), a_h)$ and $x_{h+1} \sim E_{h+1}(s_{h+1})$. Here $P_h^{\text{latent}} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the *latent dynamics*, $R_h^{\text{latent}} : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the *latent reward* (we often omit the superscript latent from these objects), and $E_h : \mathcal{S} \to \Delta(\mathcal{X})$ is an *emission distribution*. $\mathrm{supp} E_h(\cdot \mid s_h) \cap \mathrm{supp} E_h(\cdot \mid s_h') = \emptyset$, $\forall s_h \neq s_h' \in \mathcal{S}$.

In a RichCLD model, we further posit that (a) the latent state and action spaces are continuous and (b) the latent dynamics are Lipschitz continuous w.r.t. the latent states and actions. Concretely, we assume that the latent state space is a metric space $(\mathcal{S}, D_\mathcal{S})$ with *covering dimension* $\dim_\mathcal{S} \in \mathbb{R}_+$. That is, for any $\eta > 0$, there exists a set of *covering states* $\mathcal{S}_\eta$ with size $S_\eta := |\mathcal{S}_\eta| \le (2/\eta)^{\dim_\mathcal{S}}$ such that

$$\forall s \in \mathcal{S},\ \exists s_\eta \in \mathcal{S}_\eta\ :\ D_\mathcal{S}(s, s_\eta) \le \eta/2.$$

Analogously, we assume that the action space $\mathcal{A}$ is a metric space $(\mathcal{A}, D_\mathcal{A})$ with covering dimension $\dim_\mathcal{A}$ and covering

[1] For simplicity, we assume that the reward is known.

set $\mathcal{A}_\eta$ with size $A_\eta$. We use $\mathcal{S}_\eta$ and $\mathcal{A}_\eta$ to refer to *fixed* but arbitrary coverings. We define a joint metric over state-action pairs via $D((s,a),(s',a')) = D_{\mathcal{S}}(s,s') + D_{\mathcal{A}}(a,a')$, and abbreviate $\dim_{\mathcal{SA}} := \dim_{\mathcal{S}} + \dim_{\mathcal{A}}$.

To enable sample-efficient learning guarantees and take advantage of the metric structure for $\mathcal{S}$ and $\mathcal{A}$, we make a continuity assumption on the latent dynamics and posit that they are Lipschitz continuous in *total variation distance*.

**Assumption 2.1** (Lipschitz dynamics). *For every $h \in [H]$, for all $s, s' \in \mathcal{S}$ and $a, a' \in \mathcal{A}$,*[2]

$$\|P_h(\cdot \mid s,a) - P_h(\cdot \mid s',a')\|_{\mathsf{TV}} \le D((s,a),(s',a')),$$
$$|R_h(s,a) - R_h(s',a')| \le D((s,a),(s',a')).$$

This assumption asserts that nearby states and actions lead to similar transitions and rewards, but otherwise allows the dynamics to be arbitrarily nonlinear. Thus it captures many control-theoretic settings, as described in the next example.

**Example 2.1.** *Consider a system with $\mathcal{S} = \mathbb{R}^{\dim_{\mathcal{S}}}, \mathcal{A} = \mathbb{R}^{\dim_{\mathcal{A}}}$, and where transitions and rewards follow the law*

$$s_{h+1} = f(s_h, a_h) + \omega_h, \quad \text{and} \quad r_h = g(s_h, a_h),$$

*for $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and $\omega_h \sim \mathcal{N}(0, I_{\dim_{\mathcal{S}}})$. Then Assumption 2.1 holds whenever $f$ and $g$ are Lipschitz, i.e., $\|f(s,a) - f(s',a')\|_2 \le D((s,a),(s',a'))$ and $|g(s,a) - g(s',a')| \le D((s,a),(s',a'))$.* ◁

Lipschitz dynamics and several weaker assumptions have been studied in the absence of rich observations in prior work (Kakade et al., 2003; Shah & Xie, 2018; Ni et al., 2019; Song & Sun, 2019; Cao & Krishnamurthy, 2020; Sinclair et al., 2023). We show later that some of these weaker assumptions, which are sufficient in the classical setting, do not enable sample efficient learning with rich observations. However, we leave a deeper understanding of more refined continuity assumptions to future work.

**Function approximation and learning objective.** We consider online reinforcement learning in the function approximation setting. Here, the algorithm interacts with an unknown RichCLD MDP in episodes, where in the $t^{\text{th}}$ episode, the algorithm selects policy $\pi^t$ and collects trajectory $\tau^t$ by executing $\pi^t$ in the MDP. The goal of the algorithm is to identify an $\varepsilon$-optimal policy $\widehat{\pi}$ such that $J(\pi^\star) - J(\widehat{\pi}) \le \varepsilon$ with probability at least $1 - \delta$.

To facilitate achieving this learning goal in a sample-efficient manner, we assume access to a *decoder class* $\Phi \subset (\mathcal{X} \to \mathcal{S})$ containing the true decoders $\phi^\star_h$.

**Assumption 2.2** (Decoder realizability). *We have $\phi^\star_h \in \Phi$ for all $h \in [H]$.*

Given this, we say an algorithm is sample efficient if it learns an $\varepsilon$-optimal policy in $\mathrm{poly}(H, \log(|\Phi|/\delta), \varepsilon^{-\mathrm{poly}(\dim_{\mathcal{SA}})})$ episodes.[3] Crucially, there is no dependence on $|\mathcal{X}|$. We note that this type of guarantee generalizes existing results for (a) Block MDPs, which have finite $\mathcal{S}$ and $\mathcal{A}$ with the identity metric, and (b) Lipschitz MDPs in the absence of rich observations, where a "nonparametric" sample complexity of $\varepsilon^{-(\dim_{\mathcal{SA}}+2)}$ is optimal (Sinclair et al., 2023).

**Additional notation.** For a pair of policies $\pi, \pi' \in \Pi$, we define $\pi \circ_t \pi'$ as the policy that acts according to $\pi$ for the first $t-1$ steps and $\pi'$ for the remaining steps $t, \ldots, H$. We use the shorthand $x_h \sim \pi$ to indicate that $x_h$ is drawn from the law $\mathbb{P}^\pi$, and use the shorthand $(x_h, a_h) \sim \pi$ analogously. The notation $\widetilde{\mathcal{O}}(\cdot)$ indicates that a bound holds up to factors polylogarithmic in parameters appearing in the expression.

## 3. Statistical Complexity for the RichCLD Framework

At first glance, it may not be apparent to the reader whether the RichCLD framework is even tractable. Hence, in this section, we perform a preliminary investigation into statistical complexity, deferring the development of computationally efficient algorithms to Section 4. We present two results: (1) we show that the RichCLD framework is indeed tractable, and the sample complexity scaling as $O(\varepsilon^{-\mathrm{poly}(\dim_{\mathcal{SA}})})$ is achievable (Theorem 3.1), and (2) we prove a statistical separation between the framework and its non-rich-observation counterpart, the Lipschitz MDP, showing that weaker notions of Lipschitzness that lead to sample-efficient learning in the latter are intractable in the former (Theorem 3.2). **Upper bound for the RichCLD framework.** We provide an upper bound for the RichCLD model by instantiating a variant of the computationally inefficient GOLF algorithm of Jin et al. (2021); Xie et al. (2023). Prior analyses of this algorithm consider general value function approximation and obtain sample complexity scaling with the structural parameters for the MDP such as the *Bellman-Eluder dimension* (Jin et al., 2021) or *coverability* (Xie et al., 2023). Roughly speaking, these structural parameters measure the number of distributions in the MDP (induced by policies) that one must visit before one can extrapolate to any other distribution. Both quantities are known to be small in several MDP classes of interest; notably, for tabular Block MDPs, both scale only with the number of latent states $|\mathcal{S}|$ and number of actions $|\mathcal{A}|$, and

[2] For probability measures $\mathbb{P}$ and $\mathbb{Q}$ over a measurable space $(\mathcal{X}, \mathscr{E})$, total variation distance is defined via $\|\mathbb{P} - \mathbb{Q}\|_{\mathsf{TV}} = \sup_{E \in \mathscr{E}} |\mathbb{P}(E) - \mathbb{Q}(E)| = \frac{1}{2}\int |d\mathbb{P} - d\mathbb{Q}|$.

[3] Following the convention in the rich-observation literature, we assume for simplicity that $|\Phi| < \infty$ and provide sample complexity bounds that scale with $\log|\Phi|$, but it is trivial to extend our results to other notions of statistical complexity for $\Phi$.

are independent of the size of the observation space.

Unfortunately, due to the continuity of the latent state space in the RichCLD model, Bellman-Eluder dimension (as well as other complexity measures (Jiang et al., 2017; Du et al., 2021)) and coverability can both be unbounded, leading to vacuous guarantees from prior analyses. To address this, we introduce a notion of *approximate coverability* (Definition E.3), which extends coverability to allow for a certain form of misspecification. A second challenge arises because the natural value function class to use in GOLF (Lipschitz functions composed with decoders) is infinitely large and must be discretized to admit uniform convergence. Discretization introduces an approximation error, which has an unfavorable interaction with the misspecification of the MDP (i.e., the approximation parameter in approximate coverability), and results in a slow convergence rate even with careful treatment of these error terms. We address this by employing the recent disagreement-based regression (DBR) technique of Amortila et al. (2024) that avoids the interaction between these error terms arising from misspecification. To conclude, we show that RichCLD framework satisfies approximate coverability, and by carefully trading off misspecification with distribution shift, we can show that the RichCLD model is indeed learnable.

**Theorem 3.1** (PAC upper bound for RichCLD framework; informal)**.** *Suppose Assumptions 2.1 and 2.2 hold. For any $\delta \in (0,1)$ and $\varepsilon \in (0,1)$, with probability at least $1-\delta$, GOLF.DBR (Algorithm 5) outputs a policy $\widehat{\pi}$ satisfying $J(\pi^\star) - J(\widehat{\pi}) \le \varepsilon$ with sample complexity*

$$\mathcal{O}\left(\frac{H^{(2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+3)}\log(TH|\Phi|/\delta\varepsilon)}{\varepsilon^{(2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2)}}\right).$$

See Appendix E for a formal statement and proof. Regarding the sample complexity, the exponent on $\varepsilon$ scales with $2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2$, which is worse than the exponent $\dim_{\mathcal{SA}}+2$ in the minimax rate for Lipschitz MDPs without rich observations (Sinclair et al., 2023). There are two primary sources for this in our analysis. First, we incur a quadratic dependence on the effective state space size; specialized to tabular Block MDPs, this takes $O(|\mathcal{S}|^2)$ instead of $O(|\mathcal{S}|)$. The second arises from sampling uniformly over the action covering set $\mathcal{A}_\eta$ to estimate Bellman errors; this yields the additional $\dim_{\mathcal{A}}$ term. Both of these can be avoided in Lipschitz MDPs but manifest in all existing analyses for rich-observation settings (the former in tabular Block MDPs (Jiang et al., 2017; Misra et al., 2020; Jin et al., 2021; Zhang et al., 2022; Mhammedi et al., 2023b) and the latter in Theorem 4 of Jiang et al. (2017)). However, it remains open to determine if either of these dependencies are necessary in rich-observation settings.

**Lower bound under weaker continuity.** Assumption 2.1 places a rather stronger Lipschitz continuity assumption on the latent dynamics. In the absence of rich observations, prior work obtains sample-efficient algorithms under weaker conditions. Specifically, the best existing results for model-free methods (Sinclair et al., 2019; Song & Sun, 2019; Cao & Krishnamurthy, 2020) assume only that $Q^\star$ and $V^\star$ (the optimal value functions) are Lipschitz continuous with respect to the metric $D(\cdot,\cdot)$, while the best results for model-based methods (Sinclair et al., 2023) measure continuity via the 1-Wasserstein distance rather than via total variation.[4] It is therefore natural to ask whether these weaker conditions enable tractable learning in the RichCLD framework.

Our next result shows that the weakest of these assumptions— $Q^\star$ and $V^\star$ Lipschitzness—is not sufficient for sample-efficient learning with rich observations. Formally, the assumption is that for all $h, s, s', a, a'$:

$$\begin{aligned} |Q_h^\star(s,a) - Q_h^\star(s',a')| &\le D((s,a),(s',a')) \quad \text{and} \\ |V_h^\star(s) - V_h^\star(s')| &\le D_{\mathcal{S}}(s,s'). \end{aligned} \tag{1}$$

**Theorem 3.2** (Lipschitz $Q^\star/V^\star$ lower bound; informal)**.** *For rich-observation MDPs satisfying the $Q^\star/V^\star$-Lipschitz latent dynamics assumption (1), any algorithm requires $\widetilde{\Omega}(\min\{|\Phi|, |\mathcal{X}|^{1/2}, 2^{\Omega(H)}\})$ episodes to learn an $\varepsilon$-optimal policy in the worst case for an absolute constant $\varepsilon > 0$, even when $|\mathcal{A}| = 2$ and $|\mathcal{S}_\eta| = \mathcal{O}(H)$ for all $\eta \ge 0$.*

See Appendix E.4 for a formal statement and proof. The result shows that sample complexity scaling with $\log|\Phi|$ and independent of $|\mathcal{X}|$—the gold standard for rich-observation RL—is not possible under $Q^\star/V^\star$-Lipschitz latent dynamics. The mechanism at play is that $Q^\star/V^\star$-Lipschitzness does not ensure Bellman completeness—a function approximation condition crucial to the analysis of GOLF—while TV-Lipschitzness does. Interestingly, the lower bound does not apply to the intermediate assumption of latent Wasserstein Lipschitzness, which also does not ensure completeness; understanding the statistical complexity of the latter setting is an important open problem.

## 4. Efficient Algorithms for the RichCLD Framework

In this section, we turn our focus to algorithm development, and present efficient algorithms for learning in the RichCLD framework. As highlighted in the introduction, the central challenge for rich-observation RL in the presence of continuous dynamics is to develop a representation learning approach that (i) is provably sample efficient, (ii) captures the dynamics structure of the latent state space, and (iii)

[4] For probability measures $\mathbb{P}$ and $\mathbb{Q}$ the 1-Wasserstein distance is defined as $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{W}} := \sup_{f\in\mathcal{F}}\{\int f d\mathbb{P} - f d\mathbb{Q}\}$, where $\mathcal{F}$ is the set of all 1-Lipschitz functions. Total variation distance upper bounds 1-Wasserstein distance, but the converse is not true.

is computationally tractable. We address this challenge in Section 4.1 by providing a new representation learning objective, BCRL.C, then build on this development in Section 4.2 to provide a new algorithm for online exploration in the RichCLD framework.

**Remark 4.1.** BCRL was introduced by Modi et al. (2024) with no name and was further studied by Zhang et al. (2022); Mhammedi et al. (2023a), who both use the name RepLearn. We introduce the name BCRL for "Bellman Consistent Representation Learning" for this procedure, and call our variation BCRL.C, as it is designed for continuous dynamics.

### 4.1. Representation Learning with Continuous Latent Dynamics: BCRL.C

A principle shared by many prior works on representation learning for RL is that one should capture the information necessary to represent *Bellman backups* of functions of interest (e.g., value functions (Zhang et al., 2022; Mhammedi et al., 2023a; Modi et al., 2024)). For the RichCLD framework, we show that the Bellman backup of *any* bounded function is Lipschitz with respect to the true latent state $\phi^\star(x)$. Building on these prior works, we aim to learn a representation such that Bellman backups of functions of interest can be approximated by Lipschitz functions of the learned latent state; stated equivalently: *we learn a representation that respects the Lipschitz structure of the latent dynamics.*

Concretely, fix time step $h$ and suppose we have a dataset of $(x_h, a_h, x_{h+1})$ tuples in which $(x_h, a_h)$ are drawn from a data distribution $\rho_h \in \Delta(\mathcal{X} \times \mathcal{A})$ and $x_{h+1} \sim P_h(x_h, a_h)$. Let $\mathrm{Lip} \subset \mathcal{S} \times \mathcal{A} \to [0, L]$ denote the set of $L$-bounded functions that are 1-Lipschitz with respect to the metric $D$. We aim to learn a decoder $\phi_h$ that minimizes the following population-level objective:

$$\max_{f \in \mathcal{F}} \min_{g \in \mathrm{Lip}} \mathbb{E}_{\rho_h}\left[\Big(g(\phi_h(x_h), a_h) - \mathcal{P}_h[f](x_h, a_h)\Big)^2\right] \quad (2)$$

where $\mathcal{F} : \mathcal{X} \to [0, L]$ is a given class of *discriminators* whose Bellman backups we would like to approximate and $\mathcal{P}_h[\cdot]$ is the (reward-free) Bellman backup operator, defined via $\mathcal{P}_h[f](x_h, a_h) := \mathbb{E}[f(x_{h+1}) \mid x_h, a_h]$. The novel twist over prior work (Modi et al., 2024) is that by constraining the "prediction head" $g \in \mathrm{Lip}$ (which is composed with the representation $\phi_h$), we ensure that any decoder $\phi$ with low objective value can approximate Bellman backups via Lipschitz functions of the learned latent state. This constrains the learned latent space to respect the continuity structure of the true latent dynamics. We leave $\mathcal{F}$ and $\rho_h$ as free parameters here but will instantiate them concretely in Section 4.2.

**Algorithm and guarantee.** The main challenge in minimizing the population objective Eq. (2) from samples is that it involves the conditional expectation $\mathbb{E}[f(x_{h+1}) \mid x_h, a_h]$,

**Algorithm 1** BCRL.C: Bellman Consistent Representation Learning with Continuous Latent Dynamics

**input:** Layer $h \in [H]$, dataset $\mathcal{D}_h$ of $(x_h, a_h, x_{h+1})$ tuples, decoder class $\Phi$, discriminator class $\mathcal{F}$.

Define loss function $\ell_{\mathcal{D}_h}(\phi, g, f) :=$
$$\widehat{\mathbb{E}}_{(x_h, a_h, x_{h+1}) \sim \mathcal{D}_h}\left[(g(\phi(x_h, a_h)) - f(x_{h+1}))^2\right].$$

Solve the *min-max-min* optimization problem:
$$\phi_h \leftarrow \arg\min_{\phi_h \in \Phi} \max_{f \in \mathcal{F}} \left\{ \min_{g \in \mathrm{Lip}} \ell_{\mathcal{D}_h}(\phi_h, g, f) - \min_{\widetilde{\phi}_h \in \Phi, \widetilde{g} \in \mathrm{Lip}} \ell_{\mathcal{D}}(\widetilde{\phi}_h, \widetilde{g}, f) \right\}.$$

**return** $\phi_h$.

which leads to the well-known "double sampling" bias. Following Modi et al. (2024), we introduce a nested inner optimization problem to de-bias. With this bias correction, we obtain our main representation learning algorithm: for each $h \in [H]$, given dataset $\mathcal{D}_h$ of $(x_h, a_h, x_{h+1})$ tuples, we solve the following min-max-min problem:

$$\phi_h \leftarrow \arg\min_{\phi_h \in \Phi} \max_{f \in \mathcal{F}} \left\{ \min_{g \in \mathrm{Lip}} \widehat{\ell}(\phi_h, g, f) - \widehat{\mathsf{opt}}(f) \right\} \quad (3)$$

where[5]

$$\widehat{\ell}_{\mathcal{D}_h}(\phi, g, f) := \widehat{\mathbb{E}}_{\mathcal{D}_h}\left[(g(\phi(x_h), a_h) - f(x_{h+1}))^2\right],$$
$$\text{and} \quad \widehat{\mathsf{opt}}(f) := \min_{\widetilde{\phi}_h \in \Phi, \widetilde{g} \in \mathrm{Lip}} \widehat{\ell}(\widetilde{\phi}_h, \widetilde{g}, f).$$

We call this algorithm BCRL.C, and provide the full pseudocode in Algorithm 1. The guarantee for BCRL.C is stated in terms of a new concept called a *pseudobackup* operator, which we define as

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathcal{V}} : f \mapsto \arg\min_{v \in \mathcal{V}} \widehat{\mathbb{E}}_{\mathcal{D}_h}\left[\big(v(x_h, a_h) - f(x_{h+1})\big)^2\right], \quad (4)$$

given dataset $\mathcal{D}_h$ and function class $\mathcal{V} \subset (\mathcal{X} \times \mathcal{A}) \to [0, L]$. Informally, this represents the best approximation to $\mathcal{P}_h[f]$ over the class $\mathcal{V}$. For the remainder of the paper, we choose $\mathcal{V} := \mathrm{Lip} \circ \phi_h$ for some $\phi_h \in \Phi$, and we condense the notation to $\widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h}$ to make the dependence on the decoder explicit. We now state the main guarantee for BCRL.C.

**Theorem 4.1** (Guarantee of BCRL.C)**.** *Suppose Assumptions 2.1 and 2.2 hold. Fix $h \in [H]$ and $\delta \in (0, 1)$, and define $\mathcal{F}_{h+1} = \mathrm{Lip} \circ \Phi : \mathcal{X} \to [0, L]$. Let $\mathcal{D}_h$ be a dataset*

[5] $\widehat{\mathbb{E}}[\cdot]$ denotes sample average: $\widehat{\mathbb{E}}_{(x) \sim \mathcal{D}}[f] := \frac{1}{|\mathcal{D}|} \sum_{(x) \in \mathcal{D}} f(x)$.

*of $N$ i.i.d. tuples $(x_h, a_h, x_{h+1})$ sampled as $(x_h, a_h) \sim \rho_h$ and $x_{h+1} \sim P_h(x_h, a_h)$ where $\rho_h \in \Delta(\mathcal{X} \times \mathcal{A})$. Then with probability at least $1-\delta$, the decoder $\phi_h$ produced by BCRL.C ensures that for all $f \in \mathcal{F}_{h+1}$*

$$\mathbb{E}_{\rho_h}\left[\left(\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f](x,a) - \mathcal{P}_h[f](x,a)\right)^2\right] \leq \varepsilon_{\mathrm{rep}}(N,\delta),$$

*where*

$$\varepsilon_{\mathrm{rep}}(N,\delta) := \widetilde{\mathcal{O}}\left(\frac{L^{2\dim_{\mathcal{SA}}}\log(|\Phi|/\delta)}{N^{\frac{1}{\dim_{\mathcal{SA}}+1}}}\right).$$

As an immediate consequence, in Appendix B, we show how the theorem can be applied to offline reinforcement learning, yielding guarantees for approximate dynamic programming (ADP) algorithms (e.g., fitted Q iteration) in the learned latent space. As should be expected, the sample complexity bound has a nonparametric flavor, scaling exponentially with the latent dimensionality, but crucially it does not depend on the size of the observation space.

Regarding computational complexity, the optimization problem in Eq. (3) is amenable to gradient based techniques, but may be difficult to solve in practice. For the simpler tabular Block MDP setting, prior work (Zhang et al., 2022; Mhammedi et al., 2023a; Modi et al., 2024) has developed an iterative counterpart to Algorithm 1, which obtains a slightly worse statistical guarantee but is much more computationally viable. In Appendix C, we give a continuous extension of this iterative scheme (see pseudocode in Algorithm 4), which we use in our experiments in Section 5.

**Other representation learning objectives.** Within the line of theoretical research on representation learning for RL, a number of other schemes have been developed and analyzed, with a focus on tabular Block MDPs. The most notable objectives are contrastive learning (Misra et al., 2020) and multistep inverse kinematics (Lamb et al., 2023; Mhammedi et al., 2023b). Here, we briefly highlight that these schemes do not extend as-is to accommodate continuous latent dynamics in the RichCLD framework, providing further motivation behind our objective (2). For simplicity, we discuss these objectives at the population level.

The contrastive learning objective of Misra et al. (2020) involves distinguishing transition tuples that can be generated by the dynamics from those that cannot. The procedure samples pairs of tuples $(x_h^1, a_h^1, x_{h+1}^1)$ and $(x_h^2, a_h^2, x_{h+1}^2)$ i.i.d. where $x_h, a_h \sim \rho_h$ and $x_{h+1} \sim P_h(x_h, a_h)$ for a data collection distribution $\rho_h \in \Delta(\mathcal{X} \times \mathcal{A})$. Then, it assigns the "real transition" $(x_h^1, a_h^1, x_{h+1}^1)$ a positive label and the "fake transition" $(x_h^1, a_h^1, x_{h+1}^2)$ a negative label. Finally, a classifier of the form $f(\phi_h(x_h), a_h, \phi_{h+1}(x_{h+1}))$ is trained to minimize classification error on this dataset. One can show that the optimal population-level classifier takes the form

$$\frac{P_h(\phi^\star_{h+1}(x_{h+1}) \mid \phi^\star_h(x_h), a_h)}{P_h(\phi^\star_{h+1}(x_{h+1}) \mid \phi^\star_h(x_h), a_h) + \widetilde{\rho}_h(\phi^\star_{h+1}(x_{h+1}))}, \quad (5)$$

where $\widetilde{\rho}_h(\cdot)$ is the marginal distribution over $x_{h+1}$ under the data collection process. The key property of this optimal classifier—used by Misra et al. (2020)—is that it only depends on the latent states, and does not directly depend on the observations themselves. This is also true in the RichCLD framework; unfortunately, even though the latent dynamics are Lipschitz, the optimal classifier in Eq. (5) is not guaranteed to be a Lipschitz function of the latent state. This is problematic, as we need to construct a low-complexity function class that contains the predictor in Eq. (5) in order to provide sample-efficient learning guarantees.

The multistep inverse kinematics approach (Lamb et al., 2023; Mhammedi et al., 2023b) involves predicting the action $a_h \sim \mathrm{unif}(\mathcal{A})$ at time $h$ from the observation $x_h$ and a future observation $x_t$ for fixed $t > h$ under a roll-out policy $\pi$. When the action space is finite, one can show that the optimal population-level objective takes the form

$$(x_h, a_h, x_t) \mapsto \frac{\mathbb{P}^\pi(\phi^\star(x_t) \mid \phi^\star(x_h), a_h)}{\sum_{a \in \mathcal{A}} \mathbb{P}^\pi(\phi^\star(x_t) \mid \phi^\star(x_h), a)}. \quad (6)$$

Similar to contrastive learning, the key property of this objective is that it depends on the observation only through the corresponding latent state, a central property used by Mhammedi et al. (2023b). However, analogously to contrastive learning, this property alone is not sufficient for sample-efficient learning in the RichCLD, because we need to construct a low-complexity function class to express the optimal predictor (27) in this latent space. Unfortunately, the optimal predictor for multistep inverse kinematics objective may not be a Lipschitz function of the latent state, even when the transition dynamics themselves are Lipschitz.

**Proposition 4.1** (Informal)**.** *In the RichCLD framework, the optimal population-level regression function for contrastive learning and multistep inverse kinematics may not be Lipschitz with respect to the metric on latent state-action space.*

See Appendix F.4 for a formal statement and proof.

### 4.2. Efficient Exploration with Continuous Latent Dynamics: CRIEE

Equipped with BCRL.C, we can turn our attention to online reinforcement learning. Our high-level approach, also used in prior work (Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b), is to interleave representation learning and exploration: we iteratively learn a new decoder based on data the algorithm has gathered, then use this decoder within an exploration scheme to acquire new information. In what follows, we describe the algorithm in detail and present theoretical guarantees.

**Algorithm 2** CRIEE: Continuous Representation Learning with Interleaved Explore-Exploit

1: **input:** Decoders $\Phi$, iterations $T$, discretization scale $\eta$, parameters $\{\lambda^t\}_{t\in[T]}$, $\{\widehat{\alpha}^t\}_{t\in[T]}$.
2: Initialize datasets $\mathcal{D}^0_{1,h} \leftarrow \emptyset$, $\mathcal{D}^0_{2,h} \leftarrow \emptyset$ for all $h \in [H]$.
3: Initialize $\pi^0 = \{\pi_h\}_{h=1}^H$ arbitrarily.
4: **for** $t = 1, \dots, T$ **do**
   `// Gather data.`
5: For $h \in [H]$, gather tuples:

$$(x_h, a_h, x_{h+1}) \sim \pi^{t-1} \circ_h \pi^{\mathsf{unif}}_\eta,$$
$$(x_h, a_h, x_{h+1}) \sim \pi^{t-1} \circ_{h-1} \pi^{\mathsf{unif}}_\eta.$$

Add the first to $\mathcal{D}^t_{1,h}$ and the second to $\mathcal{D}^t_{2,h}$. Let $\mathcal{D}^t_h \leftarrow \mathcal{D}^t_{1,h} \cup \mathcal{D}^t_{2,h}$.

   `// Learn representation.`
6: Call Alg. 1 (or Alg. 4) with $\mathcal{F}_{h+1}$ in Eq. (24) :

$$\phi^t_h \leftarrow \mathsf{BCRL.C}_h(\mathcal{D}^t_h, \Phi, \mathcal{F}_{h+1}),\ \forall h \in [H].$$

7: Define exploration bonus (cf. Eq. (7)):

$$\widehat{b}^t_h(x,a) := \min\left\{\widehat{\alpha}^t \sqrt{\frac{1}{N_{\eta,\phi^t_h}(x,a,\mathcal{D}^t_{1,h}) + \lambda^t}}, 2\right\}.$$

   `// Learn policy.`
8: Call Alg. 3 with estimated decoder and bonuses:

$$\pi^t \leftarrow \mathsf{OptDP}\big(\{\mathcal{Q}_h\}_{h=1}^H, \{\mathcal{D}_h\}_{h=1}^H, \{\widehat{b}^t_h\}_{h=1}^H, \eta\big),$$
$$\text{with } \mathcal{Q}_h := \{(x,a) \mapsto w^\top \mathsf{disc}_\eta[\phi^t_h](x,a) : \|w\|_\infty \le 2\}.$$

9: **return:** $\widehat{\pi} := \frac{1}{T}\sum_{t=1}^T \pi^t$.

**Notation for discretization.** Our exploration scheme is based on *discretization* of the learned latent state space, which requires additional notation. Recall that for any $\eta > 0$, we have covering sets $\mathcal{S}_\eta$ and $\mathcal{A}_\eta$ for the latent state and action spaces, respectively, and define $\pi^{\mathsf{unif}}_\eta(x) = \mathsf{unif}(\mathcal{A}_\eta)$ as the policy that uniformly explores over the cover. Given a pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we define $\mathsf{disc}_\eta(s,a) \in \mathcal{S}_\eta \times \mathcal{A}_\eta$ as any covering element for $(s,a)$ in $\mathcal{S}_\eta \times \mathcal{A}_\eta$ such that $D((s,a), \mathsf{disc}_\eta(s,a)) \le \eta$.[6] Next, we define $\mathsf{ball}_\eta(s,a) := \{(\tilde{s},\tilde{a}) \in \mathcal{S} \times \mathcal{A} : \mathsf{disc}_\eta(s,a) = \mathsf{disc}_\eta(\tilde{s},\tilde{a})\}$ as the "ball" of $(s,a)$ pairs that map to the same covering element. Finally, we define $\mathcal{B}_\eta := \{\mathsf{ball}_\eta(s_\eta, a_\eta) : s_\eta, a_\eta \in \mathcal{S}_\eta \times \mathcal{A}_\eta\}$; note that the definitions ensure that $\mathcal{B}_\eta$ is a partition of $\mathcal{S} \times \mathcal{A}$.

It is also useful to lift these definitions to observation space. In particular, for any decoder $\phi \in \Phi$, we use the notation $\mathsf{disc}_\eta[\phi](x,a) := \mathsf{disc}_\eta(\phi(x),a)$ and $\mathsf{ball}_\eta[\phi](x,a) :=$

[6] If $(s,a)$ is covered by more than one element in $\mathcal{S}_\eta \times \mathcal{A}_\eta$, we break ties in an arbitrary but consistent fashion.

**Algorithm 3** OptDP: Optimistic Dynamic Programming

1: **input:** Function class $\{\mathcal{Q}_h\}_{h=1}^H$, dataset $\{\mathcal{D}_h\}_{h=1}^H$, exploration bonus $\{\widehat{b}_h\}_{h=1}^H$, discretization scale $\eta > 0$.
2: Set $V_{H+1}(x) = 0$, $\forall x \in \mathcal{X}$.
3: **for** $h = H, \dots, 1$ **do**
   `// Recursively update value functions via pseudobackups (cf. Eq. (4)).`
4: Learn value functions with pseudobackups:

$$q_h := \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathcal{Q}_h}[V_{h+1}],$$
$$Q_h(x,a) := R_h(x,a) + \widehat{b}_h(x,a) + q_h(x,a),$$
$$V_h(x) := \max_{a \in \mathcal{A}_\eta} Q_h(x,a),$$
$$\pi_h(x) := \operatorname*{arg\,max}_{a \in \mathcal{A}_\eta} Q_h(x,a).$$

5: **return:** $\{\pi_h\}_{h=1}^H$.

$\{(\tilde{x},\tilde{a}) \in \mathcal{X} \times \mathcal{A} : \mathsf{disc}_\eta[\phi](x,a) = \mathsf{disc}_\eta[\phi](\tilde{x},\tilde{a})\}$.

**Algorithm and guarantee.** Our main algorithm, which we call CRIEE (Continuous Representation Learning with Interleaved Explore-Exploit), is displayed in Algorithm 2. Our algorithm builds upon the work of Zhang et al. (2022) in the context of tabular Block MDPs, and interleaves representation learning, via BCRL.C, with an exploration scheme based on optimistic dynamic programming.

The algorithm proceeds for $T$ iterations. For each iteration $t$, after collecting data in Line 5, we learn decoders $\{\phi^t_h\}_{h=1}^H$ by applying BCRL.C (Algorithm 1) to all of the data gathered so far, with a particular choice of discriminator class defined in Eq. (24) of Appendix F.3. The decoders are then used to define (i) an exploration bonus (Line 7) and (ii) a function class $\mathcal{Q}$, with which we perform an optimistic form of approximate dynamic programming (Line 8).

In more detail, the exploration bonus is a count-based bonus over balls *in the learned latent space*. Formally, for a decoder $\phi$ and dataset $\mathcal{D}$, we define

$$N_{\eta,\phi}(x,a,\mathcal{D}) := \sum_{\tilde{x},\tilde{a} \in \mathcal{D}} \mathbb{1}\big\{(\tilde{x},\tilde{a}) \in \mathsf{ball}_\eta[\phi](x,a)\big\}, \quad (7)$$

which counts the number of times we have visited each ball in the learned latent space. We define the exploration bonus $\widehat{b}^t_h(x,a)$ in Line 7 to incentivize the agent to visit balls with low counts, and then invoke optimistic dynamic programming (OptDP; Algorithm 3). OptDP is a standard approach, but we apply it with a function class derived from discretizing the learned latent state—specifically $\mathcal{Q} = \big\{(x,a) \mapsto w^\top \mathsf{disc}_\eta[\phi^t_h](x,a)\big\}$, where $\|w\|_\infty \le 2$—for fitting Bellman backups, which results in a policy that we deploy in the next iteration. The main guarantee for CRIEE is as follows.

**Theorem 4.2** (PAC guarantee for CRIEE; informal)**.** *Suppose Assumptions 2.1 and 2.2 hold. For any $\varepsilon, \delta \in (0, 1)$, with an appropriate choice of parameters, CRIEE outputs a policy $\widehat{\pi}$ such that $J(\pi^\star) - J(\widehat{\pi}) \leq \varepsilon$ with probability at least $1 - \delta$, and with sample complexity*

$$\mathcal{O}\left(\frac{H^{\mathcal{O}(\dim_{\mathcal{SA}}^2)} \log(TH|\Phi|/\delta\varepsilon)}{\varepsilon^{\mathcal{O}(\dim_{\mathcal{SA}}^2)}}\right).$$

CRIEE achieves a similar statistical guarantee to the one in Theorem 3.1, however the constants in the exponents on $\varepsilon$ and $H$ are somewhat worse. On the other hand, CRIEE is much more computationally viable than GOLF, because BCRL.C—the main computational bottleneck—admits a practical implementation.

**Analysis and technical challenges.** The main technical challenge in CRIEE—beyond the design and analysis of BCRL.C—arises from the continuity of the learned representation, which precludes us from analyzing CRIEE using the "implicit" latent model approach of Zhang et al. (2022). Instead, we rely on the pseudobackup operators defined in Eq. (4), but a key challenge is that these operators do not inherit standard properties of Bellman backups. In particular, the pseudobackup $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}$ is not guaranteed to be a linear operator (i.e., $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f+g] \neq \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f] + \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[g]$), and it is not guaranteed to be "monotone" (i.e. $f \leq g$ pointwise does not imply $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f] \leq \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[g]$ pointwise). Without these properties, we cannot naively appeal to standard techniques used in the analysis of optimistic algorithms; for example, we cannot apply the simulation lemma, performance difference lemma (Kakade, 2003), or telescoping decomposition (Jiang et al., 2017) for regret.

We resolve this issue algorithmically, by discretizing the learned representation in the optimistic DP phase of the algorithm (via the definition of the bonuses in Line 7 and the class $\mathcal{Q}$ in Line 8). This allows us to define *linear pseudobackups*, and, by a careful analysis, we can show that these operators satisfy linearity and monotonicity, addressing the above issue. This solution introduces discretization error between the continuous representation obtained from BCRL.C and the one used for planning, but by carefully tracking these errors we obtain Theorem 4.2.

## 5. Experiments

We now present proof-of-concept empirical validations for our algorithms in Section 4, with a focus on the representation learning component BCRL.C. We consider a maze environment (Koul et al., 2023) and a locomotion benchmark (Lu et al., 2023), both with visual (rich) observations. We study the following questions: (1) Do the representations from BCRL.C respect the structure (e.g., Lipschitzness) of the latent dynamics? (2) Are they useful for downstream reward optimization? and (3) Is the Lipschitz constraint for the inner minimization in Eq. (3) essential? We prioritize studying representation learning in isolation (as opposed to studying it in tandem with exploration) because in practice, the BCRL.C subroutine can be composed with a variety of RL methods (as we show below), so we expect it may be more broadly useful than the full CRIEE algorithm.

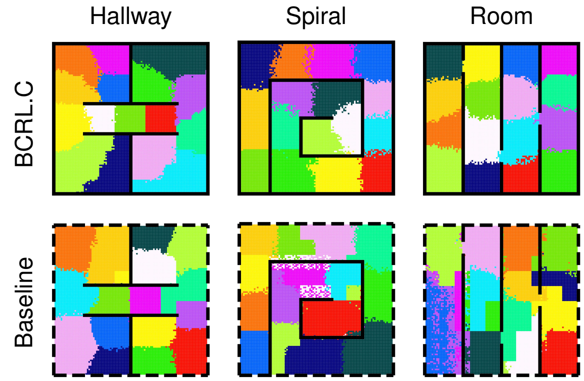


*Figure 1.* Results for BCRL.C on maze environments: $K$-means clusterings ($K = 16$) visualize the learned latent space. Baseline does not enforce Lipschitzness of the $g$ functions in Eq. (3). BCRL.C learns to respect walls and boundaries; Baseline does not.

**Implementation.** We implement the iterative version of BCRL.C mentioned in Section 4.1; full pseudocode is displayed in Algorithm 4 in Appendix C. At a high level, the algorithm learns a representation (decoder) by growing a set of discriminators over multiple iterations, where in each iteration we (i) find a decoder that can approximate the Bellman backups of the current discriminator set, (ii) find a new discriminator that witnesses large error for the current decoder (if possible), and (iii) add this to the discriminator set; this scheme approximates the idealized "min-max-min" objective in Eq. (3). We use deep neural networks to parameterize the decoders $\phi \in \Phi$, the discriminators $f \in \mathcal{F}$, and the prediction heads $g \in \text{Lip}$; architecture details are given in Appendix C. We ensure Lipschitzness of the prediction heads using spectral normalization (Miyato et al., 2018), which rescales all weight matrices in the neural network to have spectral norm 1 after each update. We train one decoder for all time steps since our experimental domains admit stationary dynamics. See Appendix C for hyperparameter settings and additional details.

### 5.1. Maze Environment

We study questions (1) and (3) above in a set of three synthetic maze environments (Koul et al., 2023), visualized in Figure 1. In each maze, we control a point mass in a two-dimensional continuous space ($\dim_{\mathcal{S}} = 2$); the three mazes correspond to different layouts. Actions correspond to continuous displacement in the environment ($\dim_{\mathcal{A}} = 2$), but are perturbed with noise through the dynamics. The observation is a $100 \times 100$ pixel image of the maze with pixel value 1 in the agent's position, and pixel value 0 in all other coordinates. The latent dynamics are Lipschitz in the sense of Assumption 2.1, so the environment falls into the RichCLD framework.

We perform a *qualitative* evaluation here, analogous to Koul et al. (2023). For each maze environment, we train the decoder via BCRL.C with 500k samples that are collected from a random policy. We visualize the learned representation by performing $K$-means clustering ($K = 16$) in the learned latent space: we uniformly sample 10k points in the environment and display their $(x, y)$ position, using color to represent the cluster assignment in the learned latent space. The results are visualized in the top row of Figure 1, where we find that the learned decoder correctly captures the local dynamics of the environment, respecting the boundaries and the walls. To investigate the role of the Lipschitz constraint (question (3)), the bottom row of Figure 1 visualizes the $K$-means clusters obtained from a decoder trained via the same protocol, but without spectral normalization ($g \notin \mathrm{Lip}$) so that the prediction heads are not constrained to be Lipschitz. We find that without the Lipschitz constraint, this learned decoder does not respect the boundaries and walls and is much worse at capturing the local dynamics.

### 5.2. Locomotion Benchmark

For a quantitative evaluation, and to address question (2) above, we consider two MuJoCo environments from the visual D4RL benchmark (Lu et al., 2023). We focus on two environments, walker and cheetah. We use an evaluation protocol from Islam et al. (2022): given offline data, we train decoders using each representation learning method, then train an agent that takes the learned latent state as input via offline reinforcement learning (specifically TD3-BC (Fujimoto & Gu, 2021)), and measure the reward obtained by the agent. The only deviation from the setup of Islam et al. (2022) is that we remove the exogenous noise from the observations, because filtering exogenous noise is not the main focus of our representation learning algorithm. The results of this evaluation are visualized in Figure 2, along with another recently proposed method for learning continuous representations (Koul et al., 2023), and (ii) a randomly initialized decoder with the same architecture as that of the other methods. Learning curves are obtained by, for each $t$, taking a checkpoint of the decoder after $t$ training epochs, running TD3-BC for 1k epochs using this decoder, and recording the reward obtained by the final policy. We pick PCLaSt as a representative baseline because it outperforms other representation learning methods (such as contrastive learning (Misra et al., 2020) and multistep inverse kinematics (Lamb et al., 2023)) in the visual D4RL benchmark with exogenous noise (Koul et al., 2023). In both environments, the representations obtained by BCRL.C are competitive with those of PCLaSt and significantly better than the randomly initialized decoder, suggesting that indeed they are useful for downstream tasks.

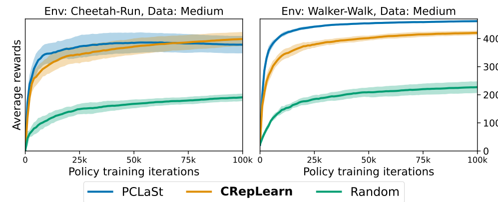


*Figure 2.* Results on visual D4RL: learning curves of TD3-BC using pre-trained (frozen) decoders from each representation learning method. Random denotes a randomly initialized decoder.

## 6. Discussion

Our work lays the foundation for investigation into rich observation RL with continuous latent states—both in terms of algorithm design and statistical complexity—and raises several open questions and directions for future research.

- **Statistical complexity.** Although our guarantees have a similar nonparametric form to the optimal bounds for Lipschitz MDPs (Ni et al., 2019; Song & Sun, 2019) , it remains to sharply characterize the minimax sample complexity for the RichCLD framework. It would also be interesting to understand whether it is possible to adapt to the intrinsic dimension of the latent space to obtain improved guarantees in benign instances (Cao & Krishnamurthy, 2020; Sinclair et al., 2023).
- **The role of continuity.** Our work shows that TV-Lipschitz continuity suffices for learning in the RichCLD framework, and that the $Q^\star$-Lipschitz assumption considered in prior work (Ni et al., 2019; Song & Sun, 2019; Cao & Krishnamurthy, 2020; Sinclair et al., 2023) is insufficient. How does the optimal sample complexity in the RichCLD framework change under other notions of Lipschitzness (e.g., Wasserstein) or continuity (e.g., Hölder). Are there more general principles under which assumptions and guarantees for Lipschitz MDPs transfer to the RichCLD framework?
- **Exogenous noise.** Developing representation learning and exploration schemes that filter exogenous noise and other irrelevant information is an important practical problem, with some recent progress in the context of discrete latent dynamics (Efroni et al., 2021; Lamb et al., 2023; Islam et al., 2022; Koul et al., 2023). Can we lift these techniques to the continuous setting?
- **General principles for representation learning.** Our work shows that existing representation learning schemes such as contrastive learning (Misra et al., 2020) and inverse dynamics (Pathak et al., 2017; Badia et al., 2020; Baker et al., 2022; Bharadhwaj et al., 2022) or multi-step inverse dynamics (Lamb et al., 2023; Mhammedi et al., 2023b) do not succeed as-is with continuous latent dynamics. Can we adapt these techniques to handle continuous dynamics, and are there natural assumptions under which they succeed?

## Impact Statement

This paper presents work whose goal is to advance the theoretical foundations of reinforcement learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## References


Agarwal, A., Henaff, M., Kakade, S., and Sun, W. PC-PG: Policy cover directed exploration for provable policy gradient learning. *Advances in Neural Information Processing Systems*, 2020a.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Advances in Neural Information Processing Systems*, 2020b.

Amortila, P., Cao, T., and Krishnamurthy, A. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. *arXiv:2401.12216*, 2024.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., and Blundell, C. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, 2020.

Baker, B., Akkaya, I., Zhokov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 2022.

Banijamali, E., Shu, R., Bui, H., and Ghodsi, A. Robust locally-linear controllable embedding. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Bharadhwaj, H., Babaeizadeh, M., Erhan, D., and Levine, S. Information prioritization through empowerment in visual model-based RL. In *International Conference on Learning Representations*, 2022.

Boffi, N. M., Tu, S., and Slotine, J.-J. E. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, 2021.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.-H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambo, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-1: Robotics transformer for real-world control at scale. *arXiv:2212.06817*, 2022.

Cao, T. and Krishnamurthy, A. Provably adaptive reinforcement learning in metric spaces. *Advances in Neural Information Processing Systems*, 2020.

Dean, S. and Recht, B. Certainty equivalent perception-based control. In *Learning for Dynamics and Control*, 2021.

Dean, S., Matni, N., Recht, B., and Ye, V. Robust guarantees for perception-based control. *Learning for Dynamics and Control*, 2020.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, 2021.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in RL. *International Conference on Machine Learning*, 2021.

Efroni, Y., Misra, D., Krishnamurthy, A., Agarwal, A., and Langford, J. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2021.

Farebrother, J., Greaves, J., Agarwal, R., Lan, C. L., Goroshin, R., Castro, P. S., and Bellemare, M. G. Proto-value networks: Scaling representation learning with auxiliary tasks. In *The Eleventh International Conference on Learning Representations*, 2023.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv:2112.13487*, 2021.

Foster, D. J., Golowich, N., and Han, Y. Tight guarantees for interactive decision making with the decision-estimation coefficient. *Conference on Learning Theory*, 2023a.

Foster, D. J., Golowich, N., Qian, J., Rakhlin, A., and Sekhari, A. Model-free reinforcement learning with the decision-estimation coefficient. In *Advances in Neural Information Processing Systems*, 2023b.

Freedman, D. A. On tail probabilities for martingales. *The Annals of Probability*, 1975.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M. G. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2019.

Guo, Z., Thakoor, S., Pîslar, M., Avila Pires, B., Altché, F., Tallec, C., Saade, A., Calandriello, D., Grill, J.-B., Tang, Y., Valko, M., Munos, R., Azar, M. G., and Piot, B. BYOL-explore: Exploration by bootstrapped prediction. *Advances in Neural Information Processing Systems*, 2022.

Ha, D. and Schmidhuber, J. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 2018.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv:2301.04104*, 2023.

Henaff, M. Explicit explore-exploit algorithms in continuous state spaces. *Advances in Neural Information Processing Systems*, 2019.

Islam, R., Tomar, M., Lamb, A., Efroni, Y., Zang, H., Didolkar, A., Misra, D., Li, X., van Seijen, H., Combes, R. T. d., and Langford, J. Agent-controller representations: Principled offline rl with rich exogenous information. *arXiv:2211.00164*, 2022.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of RL problems and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 2021.

Jonschkowski, R. and Brock, O. Learning state representations with robotic priors. *Autonomous Robots*, 2015.

Kakade, S., Kearns, M. J., and Langford, J. Exploration in metric state spaces. In *International Conference on Machine Learning*, 2003.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 2020.

Kakade, S. M. *On the sample complexity of reinforcement learning*. University of London, University College London, 2003.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *arXiv:1312.1277*, 2013.

Koul, A., Sujit, S., Chen, S., Evans, B., Wu, L., Xu, B., Chari, R., Islam, R., Seraj, R., Efroni, Y., Molu, L., Dud'ik, M., Langford, J., and Lamb, A. PcLast: Discovering plannable continuous latent states. *arXiv:2311.03534*, 2023.

Krishnamurthy, A., Agarwal, A., and Langford, J. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2016.

Kumar, A., Fu, Z., Pathak, D., and Malik, J. Rma: Rapid motor adaptation for legged robots. *arXiv:2107.04034*, 2021.

Lamb, A., Islam, R., Efroni, Y., Didolkar, A., Misra, D., Foster, D., Molu, L., Chari, R., Krishnamurthy, A., and Langford, J. Guaranteed discovery of controllable latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, 2020.

Levine, N., Chow, Y., Shu, R., Li, A., Ghavamzadeh, M., and Bui, H. Prediction, consistency, curvature: Representation learning for locally-linear control. *International Conference on Learning Representations*, 2020.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 2016.

Lu, C., Ball, P. J., Rudner, T. G. J., Parker-Holder, J., Osborne, M. A., and Teh, Y. W. Challenges and opportunities in offline reinforcement learning from visual observations. *Transactions on Machine Learning Research*, 2023.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 2022.

Mhammedi, Z., Foster, D. J., Simchowitz, M., Misra, D., Sun, W., Krishnamurthy, A., Rakhlin, A., and Langford, J. Learning the linear quadratic regulator from nonlinear observations. *Advances in Neural Information Processing Systems*, 2020.

Mhammedi, Z., Block, A., Foster, D. J., and Rakhlin, A. Efficient model-free exploration in low-rank MDPs. *Advances in Neural Information Processing Systems*, 2023a.

Mhammedi, Z., Foster, D. J., and Rakhlin, A. Representation learning with multi-step inverse kinematics: An efficient and optimal approach to rich-observation rl. In *International Conference on Machine Learning*, 2023b.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, 2020.

Misra, D., Liu, Q., Jin, C., and Langford, J. Provable rich observation reinforcement learning with combinatorial latent states. In *International Conference on Learning Representations*, 2021.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 2024.

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., and Gupta, A. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, 2023.

Ni, C., Yang, L. F., and Wang, M. Learning to control in metric space with optimal regret. In *Allerton Conference on Communication, Control, and Computing*, 2019.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv:1608.02732*, 2016.

Pari, J., Shafiullah, N. M., Arunachalam, S. P., and Pinto, L. The surprising effectiveness of representation learning for visual imitation. *arXiv:2112.01511*, 2021.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

Pfrommer, D., Zhang, T., Tu, S., and Matni, N. Tasil: Taylor series imitation learning. *Advances in Neural Information Processing Systems*, 2022.

Pfrommer, D., Simchowitz, M., Westenbroek, T., Matni, N., and Tu, S. The power of learned locally linear models for nonlinear policy optimization. In *International Conference on Machine Learning*, 2023.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, 2013.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., and Silver, D. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020.

Shah, D. and Xie, Q. Q-learning with nearest neighbors. *Advances in Neural Information Processing Systems*, 2018.

Shu, R., Nguyen, T., Chow, Y., Pham, T., Than, K., Ghavamzadeh, M., Ermon, S., and Bui, H. H. Predictive coding for locally-linear control. *International Conference on Machine Learning*, 2020.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, 2019.

Sinclair, S., Wang, T., Jain, G., Banerjee, S., and Yu, C. Adaptive discretization for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 2020.

Sinclair, S. R., Banerjee, S., and Yu, C. L. Adaptive discretization for episodic reinforcement learning in metric spaces. *ACM Conference on Measurement and Analysis of Computing Systems*, 2019.

Sinclair, S. R., Banerjee, S., and Yu, C. L. Adaptive discretization in online reinforcement learning. *Operations Research*, 2023.

Slivkins, A. Contextual bandits with similarity information. In *Conference on Learning Theory*, 2011.

Slotine, J.-J. E. and Li, W. *Applied nonlinear control*. Prentice Hall, 1991.

Song, Y. and Sun, W. PC-MLP: Model-based reinforcement learning with policy cover guided exploration. In *International Conference on Machine Learning*, 2021.

Song, Z. and Sun, W. Efficient model-free reinforcement learning in metric spaces. *arXiv:1905.00475*, 2019.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, 2019.

Tedrake, R. *Underactuated Robotics*. 2023.

Tian, Y., Zhang, K., Tedrake, R., and Sra, S. Toward understanding state representation learning in MuZero: A case study in linear quadratic gaussian control. In *IEEE Conference on Decision and Control (CDC)*, 2023.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 2021.

Tu, S., Robey, A., Zhang, T., and Matni, N. On the sample complexity of stability constrained imitation learning. In *Learning for Dynamics and Control Conference*, 2022.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline RL in low-rank mdps. In *International Conference on Learning Representations*, 2022.

Wagenmaker, A., Shi, G., and Jamieson, K. G. Optimal exploration for model-based rl in nonlinear systems. *Advances in Neural Information Processing Systems*, 2024.

Wahlström, N., Schön, T. B., and Deisenroth, M. P. From pixels to torques: Policy learning with deep dynamical models. *arXiv:1502.02251*, 2015.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Watter, M., Springenberg, J., Boedecker, J., and Riedmiller, M. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems*, 2015.

Xiao, T., Radosavovic, I., Darrell, T., and Malik, J. Masked visual pre-training for motor control. *arXiv:2203.06173*, 2022.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *International Conference on Learning Representations*, 2023.

Yarats, D., Fergus, R., and Kostrikov, I. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021a.

Yarats, D., Zhang, A., Kostrikov, I., Amos, B., Pineau, J., and Fergus, R. Improving sample efficiency in model-free reinforcement learning from images. In *AAAI Conference on Artificial Intelligence*, 2021b.

Zhang, X., Song, Y., Uehara, M., Wang, M., Agarwal, A., and Sun, W. Efficient reinforcement learning in block MDPs: A model-free representation learning approach. In *International Conference on Machine Learning*, 2022.

# Contents of Appendix



## A. Additional Related Work

**Reinforcement learning in metric spaces.** There is a large body of literature on reinforcement learning in metric spaces (also known as the Lipschitz MDP framework), either with generative model access (Kakade et al., 2003; Shah & Xie, 2018; Henaff, 2019) or in the fully online reinforcement learning framework (Sinclair et al., 2019; Ni et al., 2019; Song & Sun, 2019; Sinclair et al., 2020; Cao & Krishnamurthy, 2020; Sinclair et al., 2023). All of these works assume that the state is observed directly, and thus do not address the representation learning problem. Compared to our results, these works often operate under weaker Lipschitz continuity assumptions; as shown in Section 3, these assumptions do not readily lift to the rich-observation setting. Additionally, many of these works establish adaptive (also known as instance-dependent) guarantees that are analogous to gap-dependent bounds in the tabular setting (Simchowitz & Jamieson, 2019) and that generalize prior zooming guarantees for Lipschitz bandits (Slivkins, 2011; Kleinberg et al., 2013). Obtaining similar adaptive guarantees for RichCLD is an interesting direction for future work.

**Reinforcement learning with rich observations.** Reinforcement learning with rich observations has received extensive investigation in recent years, mostly focused on the *Block MDP* framework in which the latent state space is tabular/finite (Krishnamurthy et al., 2016; Du et al., 2019; Misra et al., 2020; Zhang et al., 2022; Mhammedi et al., 2023b), as well as the closely related *Low-Rank MDP* framework (Agarwal et al., 2020b; Zhang et al., 2022; Uehara et al., 2022; Mhammedi et al., 2023a; Modi et al., 2024). Notable works that go beyond tabular latent spaces include Dean et al. (2020); Mhammedi et al. (2020); Dean & Recht (2021), which considers continuous *linear* dynamics, and Misra et al. (2021), which considers factored (but discrete) latent dynamics. To the best of our knowledge, there are no works from this line of research that address continuous latent states with *nonlinear* dynamics.

**General complexity measures for reinforcement learning.** Another line of research provides general complexity measures that enable sample-efficient reinforcement learning, including Bellman rank (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Jin et al., 2021), eluder dimension (Russo & Van Roy, 2013), coverability (Xie et al., 2023), and the Decision-Estimation Coefficient (DEC) (Foster et al., 2021; 2023a;b). Naively applying Bellman rank and variants (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Jin et al., 2021) to the RichCLD framework is problematic due to issues around misspecification. In particular, while the RichCLD framework can be shown to admit low *approximate* Bellman rank via discretization, the sample complexity guarantees for approximate Bellman rank in these works are too weak to give non-trivial guarantees (in detail, the misspecification error is typically scaled by the Bellman rank, which is problematic when the rank depends on the misspecification error itself). Our analysis of GOLF in Section 3 can be viewed as an approximate version of the coverability-based analysis in Xie et al. (2023), but it delicately exploits the specific structure of the RichCLD framework. We expect that it is possible to bound the Decision-Estimation Coefficient (Foster et al., 2021; 2023a;b) for the RichCLD framework, but deriving efficient algorithms using this framework is non-trivial.

**Empirical research on continuous control with high-dimensional observations.** There is a large body of empirical research that addresses continuous control from high-dimensional observations via representation learning (Wahlström et al., 2015; Watter et al., 2015; Banijamali et al., 2018; Ha & Schmidhuber, 2018; Hafner et al., 2020; 2019; Gelada et al., 2019; Levine et al., 2020; Laskin et al., 2020; Shu et al., 2020; Schrittwieser et al., 2020; Yarats et al., 2021b; Guo et al., 2022; Hafner et al., 2023; Farebrother et al., 2023; Koul et al., 2023), often through the reconstruction of observations or prediction of future latent states, though recent work considers more principled objectives (Koul et al., 2023). There is a similarly large body of work from the robotics community developing representation learning methods for use in robotics applications (Jonschkowski & Brock, 2015; Pari et al., 2021; Kumar et al., 2021; Nair et al., 2023; Xiao et al., 2022). These works do not provide provable guarantees or systematically address the exploration problem.

**Continuous control.** Control of continuous, nonlinear systems is a classical topic in control theory (e.g., Slotine & Li (1991)). With a few exceptions (Mania et al., 2022; Kakade et al., 2020; Song & Sun, 2021; Boffi et al., 2021; Pfrommer et al., 2022; Tu et al., 2022; Pfrommer et al., 2023; Tian et al., 2023; Wagenmaker et al., 2024), this literature does not address the exploration problem when the underlying system is unknown, nor does it consider the issue of sample complexity.

## B. Offline RL Results with Representation Learning

In this section, we will show how to leverage the representation learning result Theorem 4.1 under an exploratory distribution in the offline RL setting. In the offline RL setting, the learner only has access to a fixed dataset, instead of the online interaction with the environment. We assume we have one dataset $\mathcal{D}_h$ for each time step $h \in [H]$, and each dataset consist of $N$ i.i.d. samples $\{(x_h^i, a_h^i, x_{h+1}^i)\}_{i=1}^N$ generated by $x_h^i, a_h^i \sim \rho_h, x_{h+1}^i \sim P_h(x_h^i, a_h^i)$. We call the distribution $\rho_h \in \Delta(\mathcal{X} \times \mathcal{A})$ the offline distribution.

### B.1. A General Result for Offline RL with BCRL.C

Suppose for each $h \in [H]$, we have offline dataset $\mathcal{D}_h$ and function classes $\{\mathcal{V}_h\}_{h=1}^{H+1} \subset (\mathcal{X} \to \mathbb{R})$ that contain the optimal state-value functions $V_h^\star \in \mathcal{V}_h$. Recall that $\Phi_h : \mathcal{X} \to \mathcal{S}$ is the decoder class, and we define $\Theta_h : \mathcal{S} \to \mathbb{R}$ as the prediction head class (generalizing the set of Lipschitz functions that we used in the RichCLD framework), such that $Q_h^\star \in \mathcal{F}_h := \Theta_h \circ \Phi_h$ for all $h$. We use these as inputs to the representation learning algorithm, to obtain a decoders $\{\phi_h\}_{h=1}^H$ that satisfy

$$\max_{v \in \mathcal{V}_{h+1}} \mathbb{E}_{x,a \sim \rho_h} \left[ \left( \widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h}[v](x, a) - \mathcal{P}_h[v](x, a) \right)^2 \right] = \varepsilon_{\text{rep}}(N, \delta), \tag{8}$$

where

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h}[v] = \min_{f \in \mathcal{F}_h} \widehat{\mathbb{E}}_{x,a,x' \sim \mathcal{D}_h} \left[ (f(x, a) - v(x'))^2 \right],$$

and $\varepsilon_{\text{rep}}(N, \delta)$ is the accuracy guarantee, which we will leave abstract in this section. (In the context of the RichCLD framework, $\varepsilon_{\text{rep}}(N, \delta)$ is controlled by Theorem 4.1). Note that for $\varepsilon_{\text{rep}}(N, \delta)$ to decay to 0 as $N \to \infty$, one requires a form of completeness, in the sense that $\mathcal{P}_h[v] \in \mathcal{F}_h$ for all $v \in \mathcal{V}_{h+1}$.

Then let us consider the following approximate dynamic programming (ADP) algorithm. The goal is to construct $f_h \in \mathcal{F}_h$ to estimate $Q_h^\star$, the optimal value function. Let $f_{H+1}(x, a) = 0,\ \forall x, a \in \mathcal{X} \times \mathcal{A}$, and $\pi_{H+1}(x)$ be an arbitrary action. For each $h$, we define $f_h^\pi(x) = f_h(x, \pi_h(x))$. Then for each $h = H, \dots, 1$, for each $x, a \in \mathcal{X} \times \mathcal{A}$, let

$$f_h(x, a) = R_h(x, a) + \widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h}[f_{h+1}^\pi](x, a), \quad \pi_h(x) = \arg\max_{a \in \mathcal{A}} f_h(x, a). \tag{9}$$

We show that, with an exploratory offline distribution, running this ADP algorithm with the representation learning result in Eq. (8) will lead to a policy that is close to the optimal policy. In this abstract setting, we assume that the offline distribution is exploratory in the sense that it enables transferring of Bellman error:

**Definition B.1.** *The Bellman transfer coefficient of a distribution $\rho$ with respect to value function class $\mathcal{F}$ is*

$$C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) := \max_{f\in\mathcal{F}} \max_{g\in\mathcal{F}} \frac{\sum_{h=1}^{H} \mathbb{E}_{x,a\sim d_h^{\pi_g}}[|f_h(x,a) - \mathcal{T}_h f_{h+1}(x,a)|]}{\sum_{h=1}^{H}\sqrt{\mathbb{E}_{x,a\sim\rho_h}\left[(f_h(x,a) - \mathcal{T}_h f_{h+1}(x,a))^2\right]}}.$$

Whenever $C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F})$ is small, we can show that the ADP algorithm in Eq. (9) will return a policy that is close to the optimal policy.

**Proposition B.1.** *If for all $h \in [H] : V_h^\star \in \mathcal{V}_h, Q_h^\star \in \mathcal{F}_h$, and Eq. (8) holds, then the policy $\pi$ returned by Eq. (9) satisfies*

$$J(\pi^\star) - J(\pi) \le H \cdot C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) \cdot \sqrt{\varepsilon_{\mathrm{rep}}(N,\delta)}.$$

Note that conditioned on the representation learning result, Proposition B.1 does not require any additional assumptions about the dynamics of the MDPs such as Assumption 2.1. In other words, the Bellman error coverage assumption is a general coverage notion that can guarantee optimality as long as the representation learning guarantee holds.

### B.2. Results for Offline RL in the RichCLD Framework

We now turn to the RichCLD framework. The first result follows immediately from instantiating the above general result to the RichCLD framework: let $\mathcal{F} = \mathrm{Lip}\circ\Phi$, we have the following result:

**Corollary B.1.** Suppose Assumption 2.1 holds. Let $\mathcal{F} = \mathrm{Lip}\circ\Phi$, $\mathcal{V} = \{\max_a f(x,a) : f \in \mathcal{F}\}$, and assume Eq. (8) holds for $\mathcal{F}$, then the policy $\pi$ returned by Eq. (9) satisfies

$$J(\pi^\star) - J(\pi) \le H \cdot C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) \cdot \sqrt{\varepsilon_{\mathrm{rep}}(N,\delta)},$$

where $\varepsilon_{\mathrm{rep}}(N,\delta)$ is given in Theorem 4.1.

Corollary B.1 follows from Proposition B.1 and Theorem 4.1, along with the fact that, under Assumption 2.1, we have the required realizability conditions. This is verified in Lemma E.2.

We obtain a second result by considering another coverage notion specific to the RichCLD framework. The intuition is that, the Bellman transfer coefficient is defined with respect to the observations, but in the RichCLD framework it is more natural to measure coverage in the latent state. Toward a suitable definition, let us consider several options. The first is the $\ell_\infty$ notion of coverage $\sup_{s\in\mathcal{S},h\in[H]} \frac{d_h^\pi(s)}{\rho_h(s)}$; this can be unbounded since the latent state space $\mathcal{S}$ is continuous. The second is the Bellman transfer coefficient in the latent space, i.e., Definition B.1 with $s = \phi^\star(x)$ instead of $x$; this yields exactly the same guarantee as Proposition B.1. The third candidate, is $\ell_\infty$ coverage over a *discretization* of the latent state space; as we will see, this yields a different guarantee than Proposition B.1.

Recall that for any $\eta > 0$, we can construct an $\eta$-cover $\mathcal{S}_\eta$ of the latent space $\mathcal{S}_\eta$ such that for any $s \in \mathcal{S}$, there exists $s' \in \mathcal{S}_\eta$ such that $D_{\mathcal{S}}(s,s') \le \eta/2$. As in Section 4.2 we can define covering elements as follow: given a pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define $\mathsf{disc}_\eta(s,a) \in \mathcal{S}_\eta \times \mathcal{A}_\eta$ as any covering element for $(s,a)$ in $\mathcal{S}_\eta\times\mathcal{A}_\eta$ such that $D((s,a),\mathsf{disc}_\eta(s,a)) \le \eta$.[7] Next, we define $\mathsf{ball}_\eta(s,a) := \{(\tilde{s},\tilde{a}) \in \mathcal{S}\times\mathcal{A} : \mathsf{disc}_\eta(s,a) = \mathsf{disc}_\eta(\tilde{s},\tilde{a})\}$ as the "ball" of $(s,a)$ pairs that map to the same covering element. Finally, we define $\mathcal{B}_\eta := \{\mathsf{ball}_\eta(s_\eta,a_\eta) : s_\eta, a_\eta \in \mathcal{S}_\eta\times\mathcal{A}_\eta\}$; note that the definitions ensure that $\mathcal{B}_\eta$ is a partition of $\mathcal{S}\times\mathcal{A}$. Then for any distribution $\rho$ over $\mathcal{S}\times\mathcal{A}$, given any $\mathsf{ball}_\eta \in \mathcal{B}_\eta$, we define $\rho(\mathsf{ball}_\eta)$ as the probability mass of $\rho$ on $\mathsf{ball}_\eta$ (since there are finitely many balls).

Given these definitions, we can define the $\ell_\infty$ coverage over the discretization.

**Definition B.2.** *Given any $\eta > 0$, the $\ell_\infty$-discretized concentrability of the offline distribution $\rho$ with respect to value function class $\mathcal{F}$ is defined as*

$$C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta) := \max_{f\in\mathcal{F}} \max_{\mathsf{ball}_\eta\in\mathcal{B}_\eta, h\in[H]} \frac{d_h^{\pi_f}(\mathsf{ball}_\eta)}{\rho(\mathsf{ball}_\eta)}.$$

[7] If $(s,a)$ is covered by more than one element in $\mathcal{S}_\eta\times\mathcal{A}_\eta$, we break ties in an arbitrary but consistent fashion.

To build intuition, observe that if $\eta$ is small, e.g., if we take $\eta \to 0$, then we recover $\ell_\infty$ coverage over the continuous latent space. Thus $C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta)$ will likely be unbounded as $\eta \to 0$. On the other hand, if we take $\eta$ to be large, then the coverage will small, e.g, if $\eta = 1$, then the coverage will trivially be 1 (assuming the diameter of the latent state-action space is 1).

As we will see later, we will also require the offline distribution to provide coverage over the action space, which we formalize via the following assumption:

**Assumption B.1** (Action coverage)**.** *We assume that for all $h \in [H]$, $\eta > 0$, we have:*

$$\sup_{x\in\mathcal{X},a\in\mathcal{A},f\in\mathcal{F}} \frac{\pi_{h;f}(a \mid x)}{\rho_h(\mathsf{ball}_\eta(a) \mid x)} \leq A_{\mathsf{cov};\eta},$$

*where* $\mathsf{ball}_\eta(a) := \{\tilde{a} : \mathsf{disc}_\eta(a) = \mathsf{disc}_\eta(\tilde{a})\}$*, and* $\mathsf{disc}_\eta(a)$ *denotes the covering element of $a$ in $\mathcal{A}_\eta$.*

Note that control on conditional probabilities of the actions does not guarantee (tight) control on the discretized concentrability coefficient and vice versa. Indeed, the offline data distribution $\rho$ is induced by a policy that takes all actions uniformly from the covering set $A_\eta$, then $A_{\mathsf{cov};\eta} \leq |\mathcal{A}_\eta|$, but the discretized concentrability coefficient can be larger then $|\mathcal{A}_\eta|^H$ in general.

**Proposition B.2.** *Under Assumption 2.1, if Eq. (8) holds, then the policy $\pi$ returned by Eq. (9) satisfies*

$$J(\pi^\star) - J(\pi) \leq \inf_{\eta>0}\left\{HA_{\mathsf{cov};\eta}\cdot\left(C^{\textit{ball}}_{\mathrm{conc}}(\mathcal{F};\eta)\right)\sqrt{\varepsilon_{\mathrm{rep}}(N,\delta)} + 2H\eta\right\}.$$

The bound reveals the tradeoff in the discretization level $\eta$; the approximation error term decreases as $\eta \to 0$, but both coverage terms increase.

### B.3. Proofs of the Offline Results

In this section we provide the proofs of the results earlier in this section. We require additional notations for the proofs, whose introduction is deferred to Appendix D for a more streamlined presentation.

**Proof of Proposition B.1.** By the standard decomposition, we have

$$\begin{aligned}
&J(\pi^\star) - J(\pi)\\
&= \mathbb{E}[V_1^\star(x_1) - f_1^\pi(x_1)] + \mathbb{E}[f_1^\pi(x) - V_1^\pi(x)]\\
&\leq \mathbb{E}\Big[V_1^\star(x_1) - f_1^{\pi^\star}(x_1)\Big] + \mathbb{E}[f_1^\pi(x) - V_1^\pi(x)]\\
&\leq \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^{\pi^\star}}[-(f_h(x,a) - \mathcal{T}_h f_h(x,a))] + \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^\pi}[f_h(x,a) - \mathcal{T}_h f_{h+1}(x,a)]\\
&\leq \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^{\pi^\star}}[|f_h(x,a) - \mathcal{T}_h f_h(x,a)|] + \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^\pi}[|f_h(x,a) - \mathcal{T}_h f_{h+1}(x,a)|]\\
&= \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^{\pi^\star}}\Big[-\Big(\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\Big)\Big] + \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^\pi}\Big[\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\Big]\\
&\leq \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^{\pi^\star}}\Big[\Big|\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\Big|\Big] + \sum_{h=1}^H \mathbb{E}_{x,a\sim d_h^\pi}\Big[\Big|\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\Big|\Big].
\end{aligned}$$

then we can show that let $\pi^\star$ be the optimal policy, and let $\pi$ be the policy returned by our ADP procedure, then by

Definition B.1,

$$
\begin{aligned}
&J(\pi^\star) - J(\pi) \\
\leq & C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) \sum_{h=1}^{H} \sqrt{\mathbb{E}_{x,a\sim\rho_h}\left[(f_h(x,a) - \mathcal{T}_h f_{h+1}(x,a))^2\right]} + C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) \sum_{h=1}^{H} \sqrt{\mathbb{E}_{x,a\sim\rho_h}\left[(f_h(x,a) - \mathcal{T}_h f_{h+1}(x,a))^2\right]} \\
= & 2C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) \sum_{h=1}^{H} \sqrt{\mathbb{E}_{x,a\sim d_h^{\pi^\star}}\left[\left(\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\right)^2\right]} \\
= & 2C^{\mathsf{be}}_{\mathrm{conc}}(\mathcal{F}) H \sqrt{\varepsilon_{\mathrm{rep}}(N,\delta)}.
\end{aligned}
$$

The last line is because $f_h$ is Lipschitz and thus the representation learning guarantee holds for all $f_h$ defined using pseudobackups in the ADP procedure. □

**Proof of Proposition B.2.** Again by the same decomposition as the previous proof, we have

$$
\begin{aligned}
&J(\pi^\star) - J(\pi) \\
\leq & \sum_{h=1}^{H} \mathbb{E}_{x,a\sim d_h^{\pi^\star}}\left[\left|\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\right|\right] + \sum_{h=1}^{H} \mathbb{E}_{x,a\sim d_h^{\pi}}\left[\left|\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\right|\right] \\
= & \sum_{h=1}^{H} \int d_h^{\pi^\star}(x,a)\delta_h(x,a)\mathrm{d}\nu(x,a) + \sum_{h=1}^{H} \int d_h^{\pi}(x,a)\delta_h(x,a)\mathrm{d}\nu(x,a),
\end{aligned}
$$

where we use the shorthand $\delta_h(x,a) := \left|\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f_{h+1}](x,a) - \mathcal{P}_h[f_{h+1}](x,a)\right|$.

Next, for any $\pi$, we can rewrite

$$
\begin{aligned}
&\sum_{h=1}^{H} \int d_h^{\pi}(x,a)\delta_h(x,a)\mathrm{d}\nu(x,a) \\
= & \sum_{h=2}^{H} \int d_{h-1}^{\pi}(x,a)\left\{\int P_h(\widetilde{x} \mid x,a)\pi_h(\widetilde{a} \mid \widetilde{x})\delta_h(\widetilde{x},\widetilde{a})\mathrm{d}\nu(\widetilde{x},\widetilde{a})\right\}\mathrm{d}\nu(x,a) + \int d_1^{\pi}(x,a)\delta_1(x,a)\mathrm{d}\nu(x,a) \\
\leq & A_{\mathsf{cov};\eta} \sum_{h=2}^{H} \int d_{h-1}^{\pi}(x,a)\left\{\int P_h(\widetilde{x} \mid x,a)\rho_h(\widetilde{a} \mid \widetilde{x})\delta_h(\widetilde{x},\widetilde{a})\mathrm{d}\nu(\widetilde{x},\widetilde{a})\right\}\mathrm{d}\nu(x,a) \\
& + A_{\mathsf{cov};\eta} \int \rho_1(x,a)\delta_1(x,a)\mathrm{d}\nu(x,a) + H\eta,
\end{aligned}
$$

where the last line is due to Assumption B.1 and the fact that $\delta_h$ is Lipschitz with respect to action (but not necessarily Lipschitz in observation). Then let us denote another shorthand $\widetilde{\delta}_{h-1}(x,a) := \int P_h(\widetilde{x} \mid x,a)\rho_h(\widetilde{a} \mid x)\delta_h(\widetilde{x},\widetilde{a})\mathrm{d}\nu(\widetilde{x},\widetilde{a})$, and due to the Lipschitz assumption on the latent dynamics (Assumption 2.1), we can observe that $\tilde{\delta}$ is Lipschitz in both latent state ($\phi^\star(x)$) and action.

Now for any distribution $p$ over $\mathcal{X} \times \mathcal{A}$, we denote the probability of visiting a state action pair $x,a$ conditioned on visiting a ball $b$: $p(x,a \mid b)$ such that $p(x,a) = p(b) \cdot p(x,a \mid b)$, where $p(b)$ is the probability over a ball that we defined above.

Now for any $\pi$ and $h \geq 2$ and $\eta > 0$, we have

$$\begin{aligned}\int d^{\pi}_{h-1}(x,a)\widetilde{\delta}_{h-1}(x,a)\mathrm{d}\nu(x,a) &= \sum_{\mathsf{ball}_\eta \in \mathcal{B}_\eta} d^{\pi}_{h-1}(\mathsf{ball}_\eta) \int d^{\pi}_{h-1}(x,a \mid \mathsf{ball}_\eta)\widetilde{\delta}_{h-1}(x,a)\mathrm{d}\nu(x,a) \\ &\leq \sum_{\mathsf{ball}_\eta \in \mathcal{B}_\eta} d^{\pi}_{h-1}(\mathsf{ball}_\eta) \int \rho_{h-1}(x,a \mid \mathsf{ball}_\eta)\widetilde{\delta}_{h-1}(x,a)\mathrm{d}\nu(x,a) + \eta \\ &\leq C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta) \sum_{\mathsf{ball}_\eta \in \mathcal{B}_\eta} \rho_{h-1}(\mathsf{ball}_\eta) \int \rho_{h-1}(x,a \mid \mathsf{ball}_\eta)\widetilde{\delta}_{h-1}(x,a)\mathrm{d}\nu(x,a) + \eta \\ &= C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta) \int \rho_{h-1}(x,a)\widetilde{\delta}_{h-1}(x,a)\mathrm{d}\nu(x,a) + \eta \\ &= C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta) \int \rho_{h}(x,a)\delta_{h}(x,a)\mathrm{d}\nu(x,a) + \eta,\end{aligned}$$

where the second line is due to $\widetilde{\delta}$ differs at most $\delta$ within observations from the same $\delta$-ball, and the conditional probability marginalizes to 1. Line 3 is due to the definition of the ball coverage and importance weighting (note that $\widetilde{\delta}$ and $\delta$ are always non-negative). The last line is due to the construction of $\widetilde{\delta}$.

Then putting everything together we have:

$$\begin{aligned}&J(\pi^\star) - J(\pi) \\ &\leq \sum_{h=1}^{H} \int d^{\pi^\star}_{h}(x,a)\delta_h(x,a)\mathrm{d}\nu(x,a) + \sum_{h=1}^{H} \int d^{\pi}_{h}(x,a)\delta_h(x,a)\mathrm{d}\nu(x,a) \\ &\leq 2A_{\mathsf{cov};\eta} \cdot C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta) \sum_{h=1}^{H} \int \rho_h(x,a)\delta_h(x,a)\mathrm{d}\nu(x,a) + 2H\eta \\ &= 2HA_{\mathsf{cov};\eta} \cdot C^{\mathsf{ball}}_{\mathrm{conc}}(\mathcal{F};\eta)\sqrt{\varepsilon_{\mathrm{rep}}(N,\delta)} + 2H\eta.\end{aligned}$$

□

# C. Additional Experimental Results and Details

## C.1. Pseudocode of Iter-BCRL.C

**Algorithm 4** Iter-BCRL.C

**require** Dataset $\mathcal{D}_h = \{x, a, x'\}$, decoder class $\Phi$, discriminator class $\mathcal{F}_{h+1}$, latent Lipschitz class Lip, number of iterations $T$, termination threshold $\beta$.

1: Randomly initialize $\phi_h^1 \in \Phi$.
2: Denote loss function

$$\ell_{\mathcal{D}_h}(\phi, g, f) = \mathbb{E}_{(x_h, a_h, x_{h+1}) \sim \mathcal{D}_h}\left[(g(\phi(x_h, a_h)) - f(x_{h+1}))^2\right].$$

3: **for** $t = 1, \dots, T$ **do**
4: Select the adversarial discriminator for the current decoder:

$$f^t = \arg\max_{f \in \mathcal{F}_{h+1}} \left[ \min_{g \in \mathrm{Lip}} \ell_{\mathcal{D}_h}(\phi_h^t, g, f) - \min_{\widetilde{\phi}_h \in \Phi, \widetilde{g} \in \mathrm{Lip}} \ell_{\mathcal{D}_h}(\widetilde{\phi}, \widetilde{g}, f) \right]. \tag{10}$$

5: **if** $f^t$ induces a loss smaller than $\beta$ **then**
6: **return** $\phi^t$.
7: Update the decoder to minimize Bellman error for all discriminators:

$$\phi_h^{t+1} = \arg\min_{\phi_h \in \Phi} \left[ \min_{\{g^i \in \mathrm{Lip}\}_{i=1}^t} \sum_{i=1}^t \ell_{\mathcal{D}_h}(\phi_h, g^i, f^i) \right]. \tag{11}$$

## C.2. Maze Environments

**Environments.** Here we give more details about the maze environment. Each maze is a two-dimensional point-mass control problem, and the latent state is the $x, y$ coordinates of the point mass. Specifically, we consider $\mathcal{S} = [0,1] \times [0,1]$ and $\dim_{\mathcal{S}} = 2$. The observation space is a $100 \times 100$ pixel image of the maze, and the value of the pixel is 1 if the point mass is in the corresponding position, and 0 otherwise. The action space is $\mathcal{A} = [-0.2, 0.2] \times [-0.2, 0.2]$, and $\dim_{\mathcal{A}} = 2$, which corresponds to the intended displacement of the point mass. The latent dynamics proceed as: given state $s$ and action $a$, we first sample a uniform action noise from $\xi \sim \mathsf{Unif}([0, 0.01] \times [0, 0.01])$, and then proceed to $s' = w(s + a + \xi)$, where $w$ is the projection of the point mass according to the wall dynamics: for example, the point mass cannot move through walls. Note that the latent dynamics and the emission distribution are identical across all time steps. The three mazes in our experiments share the same $\mathcal{S}, \mathcal{A}, \mathcal{X}$, emission distribution and action noise, and the only difference is the configuration of the walls.

Note that in the maze environment, the horizon does not affect the representation learning algorithm due to the stationary dynamics, and we do not perform policy optimization in this environment.

For the collection of offline data, we run a random policy (i.e., $a \sim \mathsf{Unif}([-0.2, 0.2] \times [-0.2, 0.2])$) for 500k steps in each maze. Specifically, every 2000 steps, we reset the point mass to a random position in the maze and proceed rolling out the random policy for 2000 steps. The collected data is then used to train the representation learning algorithm. Note that, for these environments, this offline data-generating process is exploratory, as can be seen from the example in Figure 3.

**BCRL.C implementation.** We start with a concrete description of the way in which we implement the iterative version of BCRL.C given in Algorithm 4: in each iteration $i$, we first fix the previous decoder $\phi^{i-1}$, and then we solve for the discriminator $f^i$ that maximize the prediction error of $\phi^{i-1}$ (with the double sampling correction, i.e., solve Eq. (10)). We solve the minimax problem in Eq. (10) by alternating between performing several gradient descent steps on $g$ and $\tilde{g}$, and one gradient ascent step on $f$ and $\widetilde{\phi}$ (This is a form of two-time scale gradient-descent-ascent).

After we solve for $f^i$, to solve Eq. (11), we fix the discriminators from all iterations $f^1, \dots, f^i$, and we find the decoder $\phi^i$ that minimize the backup error for all $f^1, \dots, f^i$ at the same time (with a different Lipschitz prediction head for each $f^i$). In the following hyperparameter table, we denote the number of gradient descent steps on $\phi$ as "decoder steps", and the

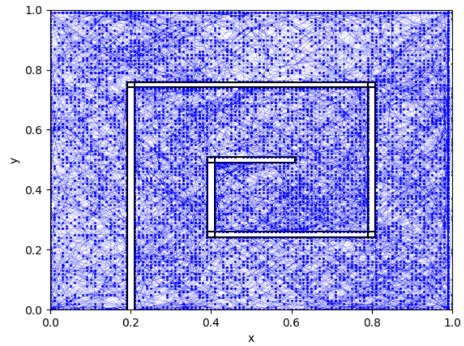


*Figure 3.* Visualization of the coverage of the offline dataset in the latent space for the Spiral Maze environment. Each dot denotes one sample and each line denotes the transition between the states. Here we subsample 20000 samples out of the 500k samples collected in the maze environment.

number of gradient ascent steps on $f$ and $\tilde{\phi}$ as "discriminator steps", and gradient descent steps on $g$ and $\tilde{g}$ as "prediction head steps".

**Baseline.** For the baseline displayed in Figure 1, we consider a variant of BCRL.C without the Lipschitz constraint. That is, for the BCRL.C implementation above, we use spectral normalization for each weight matrix in the prediction head, but in the baseline we do not add spectral normalization (the baseline is otherwise identical).

**Architecture.** In maze, we use MLP-Mixer (Tolstikhin et al., 2021) to parameterize the decoder, we use a two-layer network with spectral normalization (Miyato et al., 2018) to parameterize the Lipschitz prediction head (Lip), and the discriminator is the composition of the decoder and the Lipschitz prediction head. More hyperparameters are given in Table 1.

*Table 1.* Hyperparameters for Maze

| | BCRL.C Value | Baseline Value | Values considered |
|---|---|---|---|
| Offline sample size | 500000 | 500000 | 500000 |
| Decoder steps | 200 | 200 | 200 |
| Discriminator steps | 50 | 50 | 50 |
| Prediction head steps | 10 | 10 | 10 |
| Minibatch size | 128 | 128 | 128 |
| Latent representation dimension | 32 | 128 | 32;128 |
| MLP hidden layer size | 128 | 128 | 128 |
| MLP-mixer number of layers | 2 | 2 | 2 |
| MLP-mixer number of channels | 32 | 32 | 32 |
| MLP-mixer patch size | 10 | 10 | 10 |
| MLP-mixer hidden layer size | 32 | 32 | 32 |

### C.3. Locomotion Environments

**Environments.** We consider two MuJoCo environments, walker-walk and cheetah-run. We train representations on the environments using the visual D4RL benchmark datasets (Lu et al., 2023).[8]

The action space is 6 dimensional where each dimension has a real value ranging from -1 to 1. Each frame consists of 3 channels resulting in the size $3 \times 84 \times 84$, and each observation $x_h$ in a batch consists of 3 sequential frames resulting in the size $9 \times 84 \times 84$ where each value is the pixel value range from 0 to 255, as shown in Figure 4 for visual illustration.

**Baselines.** For baselines, we consider (i), the PCLaSt method (Koul et al., 2023), which was shown to have strong performances in the D4RL benchmark experiments among various baselines, and (ii) a randomly initialized decoder. PCLaSt

[8]The datasets that we use can be downloaded in data source: cheetah run medium and data source: walker walk medium.

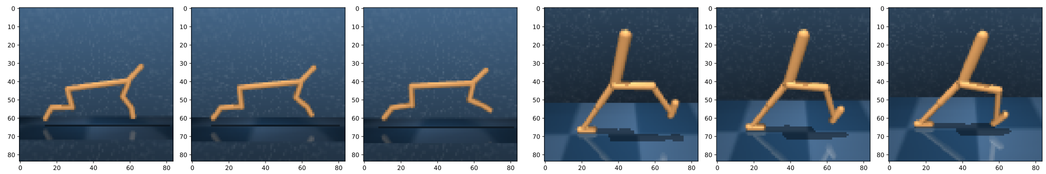

*Figure 4.* Left: 3-frames Cheetah Run; Right: 3-frames Walker Walk.

learns the decoder by combining a multi-step inverse kinematics (predicting the current action from the current and future state, the same as the MuSIK objective in (27)) which helps filter out exogenous noises and a temporal contrastive objective on top of the learned representation from the multi-step inverse kinematics objective. The goal of the temporal contrastive objective is to cluster the states that are temporally close to each other together, and push the states that are temporally far away from each other apart, in the $\ell_2$ metric. Here the temporal relationship is determined by the dynamics of the environment. For (ii) the random initialized decoder, we use the same encoder architecture as that of the other methods, but it is randomly initialized without further training, i.e., we directly run the offline RL algorithm with the randomly initialized decoder frozen through the entire training process.

**Architecture.** We parameterize BCRL.C with an architecture similar to the one used for the Maze environments, except that we use a four-layer convolutional neural network rather than MLP-Mixer to parameterize the decoder. The configuration of the four-layer convolutional neural network is presented in Table 2.

**Evaluation protocol.** We first learn the decoders via the three representation learning methods, and then we use an offline RL method, TD3-BC (Fujimoto & Gu, 2021) with frozen decoders to evaluate performance. Given a frozen decoder, we run TD3-BC for 100k iterations. Every 1k iterations, we evaluate the learned policy in the environment by running 10 episodes online and recording the average episodic rewards over the 10 episodes. Finally we plot the average rewards over the iterations. We plot the mean and the shaded area denotes 2 standard errors across 5 replicates as shown in Figure 2. In each replicate we re-train the representation and the TD3-BC agent, using different random seeds.

For computational efficiency and GPU memory usage consideration, we bound the set of discriminators to have size $M = 125$, and discard one at random when the set reaches this size.

We use the same parameter configuration as in Koul et al. (2023) with two differences. The first concerns data augmentation, where we add the same random noise to both $x_h$ and $x_{h+1}$ while Koul et al. (2023) add different random noise to all the observations. The second is that we train the PCLaSt decoder for 100k iterations but train the BCRL.C decoder for 1450 iterations. However, note that for BCRL.C, each training iteration involves iteratively solving the min-max-min problem, so the number of iterations is not directly comparable. The BCRL.C-related hyperparameters are presented in Table 2.

*Table 2.* Hyperparameters for Locomotion Environments

| | Value |
|---|---|
| The number of the offline frames | 100000 |
| Decoder steps | 200 |
| Discriminator steps | 50 |
| Prediction head steps | 10 |
| Batch size | 256 |
| Latent representation dimension | 512 |
| Number of the decoder pretraining steps | 1450 |
| Convolutional Layer 1: | |
| · Input channels | 9x84x84 |
| · Output channels | 32 |
| · Kernel size | 3x3 |
| · Stride | 2 |
| · Activation function | ReLU |
| Convolutional Layer 2–4: | |
| · Input and output channels both | 32 |
| · Kernel size | 3x3 |
| · Stride | 1 |
| · Activation function | ReLU |

# D. Technical Background

## D.1. Measure-theoretic Notation

For all the proofs in the appendix, we will adopt the following measure-theoretic notation. We assume that there exists a $\sigma$-finite measure $\nu_{\mathcal{X}}$ for the state space $\mathcal{X}$ such that for every $h \in [H], x \in \mathcal{X}, a \in \mathcal{A}$, we have $P_h(x' \mid x, a)$ as a probability density function with respect to $\nu_{\mathcal{X}}$, i.e.,

$$\forall \mathcal{X}' \subset \mathcal{X}: \;\; \mathbb{P}(\mathcal{X}' \mid x_h = x, a_h = a) = \int_{\mathcal{X}'} P_h(x' \mid x, a) \mathrm{d}\nu_{\mathcal{X}}(x').$$

Similarly, we assume there is a $\sigma$-finite measure $\nu_{\mathcal{A}}$ for the action space $\mathcal{A}$, such that for any policy $\pi$ and $x \in \mathcal{X}$, $\pi(x)$ is a probability density function with respect to $\nu_{\mathcal{A}}$, i.e.,

$$\forall \pi \in \Pi, x \in \mathcal{X}, \mathcal{A}' \subset \mathcal{A}: \;\; \pi(\mathcal{A}' \mid x) = \int_{\mathcal{A}'} \pi(a' \mid x) \mathrm{d}\nu_{\mathcal{A}}(a').$$

We denote the joint measure of the state-action space as $\nu(x, a) = \nu_{\mathcal{X}}(x) \times \nu_{\mathcal{A}}(a)$.

We typically use the same notation for a probability distribution and density function; the object we refer to will be clear from context. For example, recall that $d_h^\pi \in \Delta(\mathcal{X} \times \mathcal{A})$ is the occupancy of policy $\pi$ at time $h$. Then given a function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we can define the expectation of $f$ under $d_h^\pi$ as

$$\mathbb{E}_{x,a \sim d_h^\pi}[f(x, a)] = \int_{\mathcal{X} \times \mathcal{A}} f(x, a) d_h^\pi(x, a) \mathrm{d}\nu(x, a),$$

where it is clear that $d_h^\pi$ on the left hand side denotes the probability distribution, and $d_h^\pi$ on the right hand side denotes the probability density function.

Unless otherwise specified, all integrals are over the entire domain of the variable being integrated.

## D.2. Auxiliary Lemmas

**Lemma D.1** (Freedman's inequality (Freedman, 1975))**.** *Let $X_1, \dots, X_n \in \mathbb{R}$ be a martingale difference sequence with $|X_i| \le L$ almost surely for all $i$. Then for all $\varepsilon > 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^n X_i \ge \varepsilon\right] \le \exp\left(-\frac{\varepsilon^2}{\sum_{i=1}^n \mathbb{E}[X_i^2 \mid X_{1:i-1}] + L\varepsilon/3}\right).$$

**Lemma D.2** (Lemma G.2 of Agarwal et al. (2020a))**.** *For $t = 1, \dots, T$, consider a sequence of symmetric matrices $M_t = M_{t-1} + G_t \in \mathbb{R}^{d \times d}$, with $M_0 := \lambda_0 I$ for some $\lambda_0 > 0$, $G_t \succeq 0$, and $\|G_t\|_{\mathsf{op}} \le 1$. Then we have that*

$$2 \ln \det(M_T) - 2 \ln \det(M_0) \ge \sum_{t=1}^T \mathrm{tr}(G_t M_{t-1}^{-1}).$$

**Lemma D.3** (Potential function lemma (Lemma 20 of Zhang et al. (2022)))**.** *Consider the sequence of matrices defined in Lemma D.2. If $\mathrm{tr}(G_t) \le B^2$ for all $t \in [T]$, then*

$$2 \ln \det(M_T) - 2 \ln \det(M_0) \le d \ln\left(1 + \frac{TB^2}{d\lambda_0}\right).$$

# E. Proofs from Section 3

**Algorithm 5** GOLF.DBR (variant of Amortila et al. (2024))

1: **input:** Function class $\mathcal{F}$, estimation policy $\psi$, confidence radius $\beta > 0$, discretization scale $\eta > 0$, filter level $\kappa$.
2: Initialize $\mathcal{F}^1 \leftarrow \mathcal{F}$, $\mathcal{D}^1 \leftarrow \emptyset$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: $f^t \leftarrow \arg\max_{f \in \mathcal{F}^t} f_1(x_1, \pi_{1;f}(x_1))$, where $\pi_f := \left\{\pi_{h;f}(x_h) = \arg\max_{a_h \in \mathcal{A}} f_h(x_h, a_h)\right\}_{h=1}^H$.
5: $\pi^t \leftarrow \left\{\pi_h(x_h) = \arg\max_{a_h \in \mathcal{A}_\eta} f_h^t(x_h, a_h)\right\}_{h=1}^H$.
6: For each $h \in [H]$, collect $(x_h^t, a_h^t, r_h^t) \sim \pi^t \circ_h \psi$ and update $\mathcal{D}_h^{t+1} \leftarrow \mathcal{D}_h^t \cup \left\{x_h^t, a_h^t, r_h^t, x_{h+1}^t\right\}$.
7: Compute version space:

$$\mathcal{F}^{t+1} \leftarrow \left\{ f \in \mathcal{F}^t \mid \max_{g_h \in \mathcal{F}_h} \ell_h^t(f_h, f_{h+1}, g_h) - \ell_h^t(g_h, f_{h+1}, f_h) \leq \beta, \ \forall h \in [H] \right\},$$

where $\ell_h^t(f_h, f_{h+1}, f'_h) :=$

$$\sum_{(x_h, a_h, r_h, x_{h+1}) \in \mathcal{D}_h^{t+1}} W^\kappa_{f_h, f'_h}(x_h) \left( f_h(x_h, a_h) - r_h - \max_{a_{h+1} \in \mathcal{A}} f_{h+1}(x_{h+1}, a_{h+1}) \right)^2,$$

and $W^\kappa_{f_h, f'_h}(x_h) := \mathbb{1}\{|f_h(x_h) - f'_h(x_h)| \geq \kappa\}$.

8: **return:** $\widehat{\pi} = \frac{1}{T} \sum_{t=1}^T \pi^t$.

In this section we discuss the statistical results for the RichCLD framework. We begin with a description of the GOLF.DBR algorithm and the formal upper bound result statement in Appendix E.1. Then we discuss the discrepancy in sample complexity between Theorem 3.1 and the optimal sample complexity for the classical Lipschitz MDP setting, without rich observations. In Appendix E.2, we provide the proof of Theorem 3.1, begining by introducing the key concept of approximate coverability (Definition E.3), a general tool for capturing the complexity of settings with misspecification. We prove that Algorithm 5, with an appropriate choice of parameters, has low sample complexity in any MDP with bounded approximate coverability. Finally, we prove that RichCLD framework indeed has bounded approximate coverability and satisfies all the other conditions for the general analysis, allowing us to derive Theorem 3.1 as a corollary.

## E.1. Formal Version of Theorem 3.1 and Discussion

Our upper bound result is achieved by a variant of the GOLF algorithm, GOLF.DBR (Amortila et al., 2024). Pseudocode is displayed in Algorithm 5. GOLF is a version space algorithm based on the principle of optimistic in the face of uncertainty: in each iteration $t$, for each $h \in [H]$, we first select the most optimistic value function $f^t$ from the version space $\mathcal{F}^t$. Then we collect data from the policy $\pi^t$, the greedy policy with respect to $f^t$ (restricted to taking actions in the discretized set $\mathcal{A}_\eta$): for each $h \in [H]$, we first follow $\pi^t$ for $h - 1$ steps, and then we take an action with the estimation policy $\psi$ (similar to Du et al. (2021), discussed in the sequel), and collect the transition $(x_h, a_h, r_h, x_{h+1})$. Finally, we update the version space $\mathcal{F}^{t+1}$ by removing all the value functions that are inconsistent with the collected data.

Note that the first difference between our variant and the original GOLF algorithm is that we use an exploration policy $\psi$ to take a single action in the data collection process. We take $\psi$ to be uniform over $\mathcal{A}_\eta$, an $\eta$-cover of the action space, and discuss the need for this in detail subsequently. The second difference is that all deployed policies choose actions from the discretized action set $\mathcal{A}_\eta$. The third difference is in how we update the version space, the elimination condition is based on the "disagreement-based regression" loss introduced in Amortila et al. (2024), which as we show later, carefully controls misspecification terms that arise in the RichCLD framework.

We start by stating the formal result of our upper bound.

**Theorem 3.1′** (PAC upper bound for RichCLD; formal version of Theorem 3.1)**.** *Suppose Assumptions 2.1 and 2.2 hold. For any $\delta \in (0, 1)$ and $\varepsilon \in (0, 1)$, with an appropriate choice of parameters $\beta$, $\eta$, function class $\mathcal{F}$ and estimation policy $\psi$,* *GOLF.DBR (Algorithm 5), outputs a policy $\widehat{\pi}$ satisfying $J(\pi^\star) - J(\widehat{\pi}) \le \varepsilon$ with probability at least $1-\delta$, using at most*

$$\mathcal{O}\left(\frac{H^{\widetilde{d}+1}\log(H|\Phi|/\delta\varepsilon)}{\varepsilon^{\widetilde{d}}}\right),$$

*samples, where $\widetilde{d} := 2\dim_{\mathcal{S}\mathcal{A}} + \dim_{\mathcal{A}} + 2$.*

We prove this result in Appendix E.2. We first discuss the sample complexity guarantee.

**Discussion.** The dependence on $\dim_{\mathcal{S}\mathcal{A}}$ and $\dim_{\mathcal{A}}$ in Theorem 3.1 is significantly worse than what can be achieved for Lipschitz MDPs (Sinclair et al., 2019; Song & Sun, 2019; Cao & Krishnamurthy, 2020). Here, we briefly explain these differences through the lens of pseudo-regret, defined via

$$\mathsf{Reg}(T) := \sum_{t=1}^{T} J(\pi^\star) - J(\pi^t).$$

Pseudoregret is a central quantity in the analysis of Algorithm 5, as the policy $\widehat{\pi}$ returned has

$$J(\pi^\star) - J(\widehat{\pi}) = \frac{1}{T}\mathsf{Reg}(T); \tag{12}$$

it is similar to standard regret, except it is defined with respect to the policy sequence $\pi^1, \dots, \pi^T$ instead of the data collection policies. In what follows, we highlight three factors that result in differences between the RichCLD framework and Lipschitz MDPs.

*Dependence on state covering number.* In the Lipschitz MDP setting, previous works (Sinclair et al., 2019; Song & Sun, 2019; Cao & Krishnamurthy, 2020) have shown the optimal regret bound (ignoring dependence on the horizon $H$ and terms that are not essential to this discussion) is

$$\mathsf{Reg}(T) \le \widetilde{\mathcal{O}}\left(\inf_{\eta>0}\left\{\sqrt{\left(\frac{1}{\eta}\right)^{\dim_{\mathcal{S}\mathcal{A}}} T} + T\eta\right\}\right) = \widetilde{\mathcal{O}}\left(T^{1-\frac{1}{\dim_{\mathcal{S}\mathcal{A}}+2}}\right). \tag{13}$$

Above, $\eta$ corresponds to discretization level, and the optimal choice is $\eta = T^{-\frac{1}{\dim_{\mathcal{S}\mathcal{A}}+2}}$. Through Eq. (12), this also yields the optimal PAC sample complexity $\widetilde{\mathcal{O}}(\varepsilon^{-(\dim_{\mathcal{S}\mathcal{A}}+2)})$. Note that Eq. (13) generalizes the minimax optimal rate $\widetilde{\mathcal{O}}(\sqrt{SAT})$ (Azar et al., 2017) for tabular MDPs with $S$ states and $A$ actions, with the dependence on $SA$ replaced by the covering number $(1/\eta)^{\dim_{\mathcal{S}\mathcal{A}}}$.

Now let us consider the rich-observation setting, and draw an analogy to tabular rich-observation MDPs. Even though the minimax regret for tabular MDPs with $S$ states scales with $\widetilde{\mathcal{O}}(\sqrt{ST})$, to the best of our knowledge the best existing upper bound for tabular rich-observation MDPs (i.e., Block MDPs) is $\widetilde{\mathcal{O}}(\sqrt{S^2A^2T})$, which is obtained through GOLF. In particular, the two bounds of Jin et al. (2021); Xie et al. (2023) both give $\mathsf{Reg}(T) \le \widetilde{\mathcal{O}}\big(\sqrt{(SAT\beta)}\big)$, which has an $SA$ factor arising from distribution shift (either Bellman-eluder dimension or coverability). An additional $SA$ factor arises from the confidence radius $\beta = \mathcal{O}(\log|\mathcal{F}|) = \widetilde{\mathcal{O}}(SA)$ (ignoring dependence on $\log|\Phi|$), which gives the $\widetilde{\mathcal{O}}(\sqrt{S^2A^2T})$ rate.

Observing that Lipschitz MDPs generalize the standard tabular setting with $(1/\eta)^{\dim_{\mathcal{S}\mathcal{A}}}$ replacing $SA$, and extending to the rich-observation setting, existing Block MDP results would suggest the bound

$$\mathsf{Reg}(T) \le \widetilde{\mathcal{O}}\left(\inf_{\eta>0}\left\{\sqrt{\left(\frac{1}{\eta}\right)^{2\dim_{\mathcal{S}\mathcal{A}}} T} + T\eta\right\}\right),$$

for the RichCLD framework. This already degrades the PAC bound (when compared with Lipschitz MDPs) to $\widetilde{\mathcal{O}}(\varepsilon^{-(2\dim_{\mathcal{S}\mathcal{A}}+2)})$, with optimal tuning. Thus, it does not seem to be possible to match the Lipschitz MDP rate (Eq. (13)) in the RichCLD setting without first closing the $(SA)^2$ vs. $SA$ gap for tabular Block MDPs.

*Dependence on action covering number.* Another unique property of rich-observation RL that further degrades the rate is the dependence on the number of actions (or action space covering number). Again, consider tabular Block MDPs. Recall that we sample a single action from the uniform policy within each trajectory during the data collection process. This form of randomization appears in most existing algorithms for rich-observation RL (e.g., (Agarwal et al., 2020b)), and is used in *every* algorithm we are aware of for Low-Rank MDPs. The only exception we are aware of is GOLF, where the coverability-based analysis in (Xie et al., 2023) avoids uniform sampling by leveraging the structure of the emission process. This yields a regret bound of $\widetilde{\mathcal{O}}(\sqrt{S^2A^2T})$, but if random actions are taken (as we do), one obtains

$$\mathsf{Reg}(T) \le \widetilde{\mathcal{O}}(\sqrt{S^2A^3T}),$$

where the extra $\sqrt{A}$ factor arises from a standard importance weighting argument.

For the RichCLD, it is unclear if one can avoid this additional factor in the same vein as (Xie et al., 2023).

As a result, we incur regret

$$\mathsf{Reg}(T) \le \widetilde{\mathcal{O}}\left(\inf_{\eta>0}\left\{\sqrt{\left(\frac{1}{\eta}\right)^{2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}}T} + T\eta\right\}\right),$$

where taking $\eta = T^{-\frac{1}{2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2}}$ gives us the PAC bound of $\widetilde{\mathcal{O}}(\varepsilon^{-(2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2)})$.

*Dependence on misspecification.* Finally, let us remark on the role of misspecification and highlight the improvement obtained from disagreement-based regression (Amortila et al., 2024), which is incorporated into Theorem 3.1. In Theorem 3.1, we take $\mathcal{F} = \mathrm{Lip} \circ \Phi$. Obtaining uniform convergence with this function class requires a discretization argument, and if we discretize at level $\gamma > 0$, we can set $\beta = (1/\gamma)^{\dim_{\mathcal{SA}}} + T\gamma^2$. Unfortunately, as identified by Amortila et al. (2024), the standard GOLF algorithm manifests *misspecification amplifciation*, an undesirable interaction between the distribution shift parameter (approximate coverability) and the misspecification of the function class. Misspecification amplification implies that we cannot set $\gamma = \eta$, as doing so results in a trivial regret of $\widetilde{\mathcal{O}}\Big(T(1/\eta)^{\dim_{\mathcal{SA}}} + T\eta\Big) = \Omega(T)$. Instead, we must choose $\gamma \neq \eta$ and this gives:

$$\mathsf{Reg}(T) \le \widetilde{\mathcal{O}}\left(\sqrt{\left(\frac{1}{\eta}\right)^{\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}}\left(\frac{1}{\gamma}\right)^{\dim_{\mathcal{SA}}}T + T^2\gamma^2} + T\eta\right).$$

Tuning $\gamma = T^{-\frac{1}{(\dim_{\mathcal{SA}}+2)}}$, $\eta = T^{-\frac{2}{((\dim_{\mathcal{A}}+\dim_{\mathcal{SA}}+2)(\dim_{\mathcal{SA}}+2))}}$, we obtain a $\widetilde{\mathcal{O}}(\varepsilon^{-\dim^2_{\mathcal{SA}}})$ dependence.

However, this degradation from $\dim_{\mathcal{SA}}$ to $\dim^2_{\mathcal{SA}}$ due to misspecification amplification is not fundamental. Indeed, Amortila et al. (2024) introduce "disagreement-based regression" precisely to avoid this phenomenon, and using it here results in the guarantee in Theorem 3.1′.

### E.2. Proof of Theorem 3.1′

In this section we prove Theorem 3.1′. The basic idea behind our proof is to discretize the state and action space, then carefully proceed with an analysis similar to the finite state/action setting. To this end, we introduce *approximate coverage*, a new complexity measure that builds on the coverability framework from Xie et al. (2023), but handles the misspecification error induced by discretizing in the nonparametric setting. Similarly, we introduce the approximate version of the completeness assumption, which is required by previous analysis (Jin et al., 2021; Xie et al., 2023). We first introduce the definition of approximate coverability, then show that any MDP with bounded approximate coverability is PAC-learnable via running Algorithm 5 with an approximately complete function class. Finally, we show that the RichCLD framework admits bounded approximate coverability.

In the remaining part of the appendix, when using a decoder $\phi : \mathcal{X} \to \mathcal{S}$ sometimes we write $\phi(x, a) := (\phi(x), a)$ to avoid superfluous parenthesis.

#### E.2.1. Approximate Coverability

We introduce *approximate coverability* as a structural property that generalizes the coverability property from Xie et al. (2023) to exploit the structure of a class of MDPs. Approximate coverability involves two important concepts: the *one-step-back operator* and *approximate occupancy measures*. These are introduced below.

The one-step-back operator is defined as follows:

**Definition E.1** (One-step-back operator)**.** *Fix $h \in [H]$. Given function class $\mathcal{G} : \mathcal{X} \times \mathcal{A} \to [0, L]$ and policy class $\Pi : \mathcal{X} \to \Delta(\mathcal{A})$, an operator $\mathrm{osb}_{h,\psi} : \mathcal{G} \times \Pi \to ((\mathcal{X} \times \mathcal{A}) \to [0, L])$ is called a* one-step-back operator *for an estimation policy $\psi$ and time step $h$ with coefficient $A$ if it satisfies the following property:*

$$\forall g \in \mathcal{G}, \pi \in \Pi : \quad \mathbb{E}^{\pi}\left[\mathsf{conv}^{+}(\mathrm{osb}_{h,\psi}[g, \pi](x_{h-1}, a_{h-1}))\right] \leq \mathbb{E}^{\pi \circ_h \psi}\left[A \cdot \mathsf{conv}^{+}(g_h(x_h, a_h))\right].$$

*where we use the notation $\mathsf{conv}^{+}(\cdot)$ to indicate that the inequality holds for all non-negative convex functions.*

In the following, we often instantiate $\psi$ with $\pi_{\eta}^{\mathsf{unif}}$, and $A = A_{\eta}$.

Definition E.1 is an abstract definition. When a one-step back operator exists, we can compare the expected value of any function from timestep $h$ to a related quantity at time step $h - 1$. Unsurprisingly, the construction of such an operator will rely on the dynamics of the MDP. This one-step-back operator can be seen as a generalization of a central technique used in prior works (Agarwal et al., 2020b; Jin et al., 2021; Uehara et al., 2022).

Next, we introduce the notion of approximate occupancy measure.

**Definition E.2** (Approximate occupancy)**.** *Fix $h \in [H]$, function class $\mathcal{G}$, policy class $\Pi$ and let $\mathrm{osb}_{h,\psi}$ be a one step back operator with estimation policy $\psi$ and coefficient $A$. We say that $\Xi_{\eta} \subset \Delta(\mathcal{X} \times \mathcal{A})$ is an $\eta$-approximate occupancy class for $\Pi$ if for all $\pi$ there exists $\xi_{\eta} \in \Xi_{\eta}$ such that for all $g \in \mathcal{G}$ with $\widetilde{g}_{\pi} = \mathrm{osb}_{h,\psi}[g, \pi]$:*

- ***Average distribution shift:** $d_h^{\pi} \to \xi_{\eta}$.*

$$\mathbb{E}^{\pi}[g(x_h, a_h)] \leq \mathbb{E}^{\xi_{\eta}}\left[\widetilde{g}_{\pi}(x_h, a_h)\right] + \eta,$$

  *where $\mathbb{E}^{\xi_{\eta}}[\cdot] := \mathbb{E}_{(x,a)\sim\xi_{\eta}}[\cdot]$.*

- ***Pointwise distribution shift:** $\xi_{\eta} \to d_h^{\pi}$. There exists a coupling*[9] *$\sigma \in \Gamma(\xi_{\eta}, d_h^{\pi})$ such that for all $x, a \in \mathcal{X} \times \mathcal{A}$ we have*

$$\xi_{\eta}(x, a)\widetilde{g}_{\pi}^{2}(x, a) \leq \int \sigma(x, a, \widetilde{x}, \widetilde{a})\widetilde{g}_{\pi}^{2}(\widetilde{x}, \widetilde{a})\mathrm{d}\nu(\widetilde{x}, \widetilde{a}) + \xi_{\eta}(x, a)\eta^{2}.$$

Approximate occupancy measures formalize discretization over the continuous state-action space, allowing for a controlled discretization error. Introducing approximate occupancy measures allows us to appropriately control distribution shift, via the following notion of approximate coverability.

**Definition E.3** (Approximate coverability)**.** *The approximate coverability of an MDP with respect to $\Xi_{\eta}$ that satisfies Definition E.2 (with any $\mathrm{osb}$, $\mathcal{G}$, $\Pi$, $\psi$, and $A$) is defined as*

$$C_{\mathrm{cov}}(\Xi_{\eta}) := \inf_{\mu_1,\ldots,\mu_H \in \Delta(\mathcal{X}\times\mathcal{A})} \sup_{\xi_{\eta}\in\Xi_{\eta}, h\in[H]} \left\| \frac{\xi_{\eta;h}}{\mu_h} \right\|_{\infty}.$$

Above, the function class $\mathcal{G}, \Pi$ as well as the exploration policy $\psi$ and coefficient $A$ are implicit. We instantiate these objects in the sequel. The essential difference between approximate coverability and coverability (Xie et al., 2023) is that in our definition, the supremum is taken over approximate occupancy measures (Definition E.2), instead of occupancy measures. This is a natural generalization of the coverability property, as it allows us to handle misspecification error induced by discretization in the nonparametric setting.

### E.2.2. Analysis under Approximate Coverability

We now give a general analysis of the GOLF.DBR algorithm under approximate coverability. A basic assumption that is required by all previous analyses is the realizability and completeness of the value function class $\mathcal{F}$. Below we state a more general misspecified version of this assumption. As we show later, with the class of Lipschitz function, the RichCLD framework satisfies these properties.

First, for any function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we define the reward-based Bellman backup operator $\mathcal{T}_h : \mathcal{F}_{h+1} \to \mathcal{F}_h$, where $\mathcal{T}_h[f_{h+1}](x, a) = R_h(x, a) + \mathbb{E}\left[\max_{a_{h+1}} f_{h+1}(x_{h+1}, a_{h+1}) \mid x_h = x, a_h = a\right]$.

[9] Given two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ and $(\Omega_2, \mathcal{F}_2, \mathbb{Q})$, a coupling of $\mathbb{P}$ and $\mathbb{Q}$, $\sigma \in \Gamma(\mathbb{P}, \mathbb{Q})$, is a joint probability measure over $(\Omega_1 \times \Omega_2)$ such that the marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$. I.e., $\sigma(A, \Omega_2) = \mathbb{P}(A),\ \forall A \in \mathcal{F}_1$, and $\sigma(\Omega_1, B) = \mathbb{Q}(B),\ \forall B \in \mathcal{F}_2$.

**Assumption E.1** (Approximate realizability and completeness)**.** *We assume the function class $\mathcal{F}_\gamma$ (given as an argument to GOLF.DBR) is $\gamma$ misspecified in the $\ell_\infty$ norm: For all $h \in [H]$, we have $f^\star_{\infty;h} \in \mathcal{F}_{\gamma;h}$ such that $\left\| f^\star_{\infty;h} - Q^\star_h \right\|_\infty \leq \gamma$. Also, we have for any $f_{h+1} \in \mathcal{F}_{\gamma;h+1}$, there exists $f_h \in \mathcal{F}_{\gamma;h+1}$ such that $\|f_h - \mathcal{T}_h[f_{h+1}]\|_\infty \leq \gamma$. Furthermore, we assume there exists a finite subset of actions $\mathcal{A}_\gamma$ such that for all $h \in [H]$, for any $x \in \mathcal{X}$ and $f_h \in \mathcal{F}_h$, $\max_{a \in \mathcal{A}} f_h(x,a) - \max_{a \in \mathcal{A}_\gamma} f_h(x,a) \leq \gamma$.*

The next results gives a sample complexity bound for GOLF.DBR for any MDP that satisfies Definition E.3 and any function that satisfies Assumption E.1.

**Lemma E.1.** *Consider any MDP that satisfies Definition E.3 with approximate occupancy $\Xi_\eta$, and parameter $\eta \in (0,1)$, $\psi$ and $A$ for classes $\mathcal{F}_\gamma, \Pi = \{\pi_f : f \in \mathcal{F}_\gamma\}$. Then running GOLF.DBR (Algorithm 5) with function $\mathcal{F}_\eta$ that satisfies Assumption E.1 and $\beta = c\log(TH|\mathcal{F}_\eta|/\delta)$, outputs a sequence of policies $\pi^1, \ldots, \pi^T$, such that with probability at least $1-\delta$,*

$$\mathsf{Reg}(T) := \sum_{t=1}^{T} J(\pi^\star) - J(\pi^t) \leq \mathcal{O}\left(H\sqrt{AC_{\mathrm{cov}}(\Xi_\eta)T\beta\log(T)} + TH\eta\right),$$

*where $C_{\mathrm{cov}}(\Xi_\eta)$ is the approximate coverability parameter defined in Definition E.3.*

**Proof of Lemma E.1.** For each $f^t$ induced by the algorithm, for each $h \in [H]$, let $\mathrm{apx}[f^t_{h+1}] \in \mathcal{F}_{\eta,h}$ be the $\ell_\infty$ approximation of $\mathcal{T}_h[f^t_{h+1}]$ such that $\|\mathrm{apx}[f^t_{h+1}] - \mathcal{T}_h[f^t_{h+1}]\|_\infty \leq \eta$. Note that this is always possible due to Assumption E.1. Then we define

$$\delta^t_h := f^t_h - \mathrm{apx}[f^t_{h+1}] \quad \text{and} \quad \delta'^t_h := \mathrm{apx}[f^t_{h+1}] - \mathcal{T}_h[f^t_{h+1}].$$

Under completeness assumption (Assumption E.1), we have

$$\begin{aligned}
\mathsf{Reg}(T) &= \sum_{t=1}^{T} J(\pi^\star) - J(\pi^t) \\
&\leq \sum_{t=1}^{T} \left| f^t_1(x_1, \pi^t_1(x_1)) - J(\pi^t) \right| + 2T\eta && \text{(Almost optimism (Lemma E.6) with } \gamma = \eta) \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \left| \mathbb{E}^{\pi^t_h}[f_h - \mathcal{T}_h[f_{h+1}]] \right| + 3T\eta && \text{(Lemma E.4)} \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \left| \mathbb{E}^{\pi^t_h}\left[\delta^t_h + \delta'^t_h\right] \right| + 3T\eta \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \left| \mathbb{E}^{\pi^t_h}\left[\delta^t_h\right] \right| + 4T\eta \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^{\pi^t_h}\left[ \left| \mathbb{1}\{\delta^t_h \geq 3\eta\}\delta^t_h \right| \right] + 7T\eta.
\end{aligned}$$

Now define $\mathcal{G}_h = |\mathcal{F}_{\eta;h} - \mathcal{T}_h[\mathcal{F}_{\eta;h+1}]|$, we see first that $\mathcal{G}_h$ is bounded by 2, and second $|\mathbb{1}\{\delta^t_h \geq 3\eta\}\delta^t_h| \in \mathcal{G}_h$, then by Definition E.1, let $\widetilde{\delta}^t_{h-1} = \mathrm{osb}_h(|\mathbb{1}\{\delta^t_h \geq 3\eta\}\delta^t_h|, \pi^t_h)$ be defined with respect to the test policy $\psi$, we have

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^{\pi^t_h}\left[ \left| \mathbb{1}\{\delta^t_h \geq 3\eta\}\delta^t_h \right| \right] \leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\right] + TH\eta,$$

where $\xi^t_{\eta;h-1}$ is the distribution over $(\mathcal{X} \times \mathcal{A})$ from the family of distributions $\Xi_\eta$ defined in Definition E.2 such that the average distribution shift property holds for $\pi^t_h$. We can now proceed by adapting the coverability analysis in Xie et al. (2023). We first define

$$\bar{\xi}^t_{\eta;h}(x,a) := \sum_{i=1}^{t-1} \xi^i_{\eta;h}(x,a), \quad \text{and} \quad \mu^\star_h = \operatorname*{arg\,min}_{\mu_h \in \Delta(\mathcal{X}\times\mathcal{A})} \sup_{\xi_\eta \in \Xi_\eta} \left\| \frac{\xi_{\eta;h}}{\mu_h} \right\|_\infty, \tag{14}$$

and define the "burn-in" time

$$\tau_h(x,a) = \min\{t \mid \bar{\xi}^t_{\eta;h}(x,a) \geq C_{\mathrm{cov}}(\Xi_\eta) \cdot \mu^\star_h(x,a)\}.$$

Similar to the definition of $\bar{\xi}^t_{\eta;h}$, we also define

$$\bar{d}^{\pi^t}_h(x,a) = \sum_{i=1}^{t-1} d^{\pi^i}_h(x,a).$$

Finally we will use the following shorthand to denote the data collection distribution in Algorithm 5: at iteration $t$, recall for each $h \in H$, the data collection policy is $\pi^t \circ_h \psi$, and denote the data collection distribution as $d^t_h(x,a) = d^{\pi^t \circ_h \psi}_h(x,a)$, then we denote the sum of historical data collection distribution as

$$\bar{d}^t_h(x,a) = \sum_{i=1}^{t-1} d^i_h(x,a). \tag{15}$$

We can then decompose the regret for each layer $h$ as

$$\sum_{t=1}^T \mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\right] = \underbrace{\sum_{t=1}^T \mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\mathbb{1}\{t < \tau_{h-1}\}\right]}_{\text{burn-in term}} + \underbrace{\sum_{t=1}^T \mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\mathbb{1}\{t \geq \tau_{h-1}\}\right]}_{\text{stable term}}.$$

For the burn-in term, we have by the construction that $|\widetilde{\delta}^t_{h-1}| \leq 1$, then

$$\sum_{t=1}^T \mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\mathbb{1}\{t < \tau_{h-1}\}\right] \leq \int \sum_{t<\tau_{h-1}(x,a)} \xi^t_{\eta;h-1}(x,a)\mathrm{d}\nu(x,a) = \int \bar{\xi}^{\tau_{h-1}(x,a)}_{\eta;h-1}(x,a)\mathrm{d}\nu(x,a).$$

Then by the definition of accumulated distribution (Eq. (14)), we have

$$\begin{aligned}\int \bar{\xi}^{\tau_{h-1}(x,a)}_{\eta;h-1}(x,a)\mathrm{d}\nu(x,a) &= \int \bar{\xi}^{\tau_{h-1}(x,a)-1}_{\eta;h-1}(x,a)\mathrm{d}\nu(x,a) + \int \xi^{\tau_{h-1}(x,a)-1}_{\eta;h-1}(x,a)\mathrm{d}\nu(x,a)\\ &\leq 2C_{\mathrm{cov}}(\Xi_\eta)\sum_{x,a}\mu^\star_h(x,a) = 2C_{\mathrm{cov}}(\Xi_\eta),\end{aligned}$$

where the last inequality is due to the definition of $\tau_h(x,a)$ and $C_{\mathrm{cov}}(\Xi_\eta)$. Thus we proved that the burn-in term is bounded by $2C_{\mathrm{cov}}(\Xi)$. For the stable term, recall the definition of $\sigma_{\eta;h}$ the coupling between $\xi_{\eta;h}$ and $d^\pi_h$ that satisfies the pointwise distribution shift property, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\mathbb{1}\{t \geq \tau_{h-1}\}\right] \\
&= \sum_{t=1}^{T}\int \xi^t_{\eta;h-1}(x,a)\left(\frac{\bar{\xi}^t_{\eta;h-1}(x,a)}{\bar{\xi}^t_{\eta;h-1}(x,a)}\right)^{\frac{1}{2}}\widetilde{\delta}^t_{h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\mathrm{d}\nu(x,a) \\
&= \sum_{t=1}^{T}\int \sqrt{\frac{\left(\xi^t_{\eta;h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\right)^2}{\bar{\xi}^t_{\eta;h-1}(x,a)}}\sqrt{\bar{\xi}^t_{\eta;h-1}(x,a)\widetilde{\delta}^{t^2}_{f;\pi;h-1}(x,a)}\mathrm{d}\nu(x,a) \\
&\leq \sum_{t=1}^{T}\int \sqrt{\frac{\left(\xi^t_{\eta;h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\right)^2}{\bar{\xi}^t_{\eta;h-1}(x,a)}}\sqrt{2\int \bar{\sigma}^t_{h-1}(\widetilde{x},\widetilde{a},x,a)\widetilde{\delta}^{t^2}_{h-1}(\widetilde{x},\widetilde{a}) + 2\xi^t_{\eta;h-1}(x,a)\eta^2\mathrm{d}\nu(\widetilde{x},\widetilde{a})\mathrm{d}\nu(x,a)} \\
&\hspace{20em}\text{(Point-wise distribution shift (Definition E.2))} \\
&\leq \sum_{t=1}^{T}\int \sqrt{\frac{\left(\xi^t_{\eta;h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\right)^2}{\bar{\xi}^t_{\eta;h-1}(x,a)}}\sqrt{2\int \bar{\sigma}^t_{h-1}(\widetilde{x},\widetilde{a},x,a)\widetilde{\delta}^{t^2}_{h-1}(\widetilde{x},\widetilde{a})\mathrm{d}\nu(\widetilde{x},\widetilde{a})\mathrm{d}\nu(x,a) + 2T\eta} \quad \text{(Normalization)} \\
&\leq \sqrt{\sum_{t=1}^{T}\int \frac{\left(\xi^t_{\eta;h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\right)^2}{\bar{\xi}^t_{\eta;h-1}(x,a)}\mathrm{d}\nu(x,a)}\sqrt{\sum_{t=1}^{T}\int 2\bar{\sigma}^t_{h-1}(\widetilde{x},\widetilde{a},x,a)\widetilde{\delta}^{t^2}_{h-1}(\widetilde{x},\widetilde{a})\mathrm{d}\nu(\widetilde{x},\widetilde{a})\mathrm{d}\nu(x,a) + 2T\eta} \\
&\hspace{28em}\text{(Cauchy-Schwarz)} \\
&= \sqrt{\sum_{t=1}^{T}\int \frac{\left(\xi^t_{\eta;h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\right)^2}{\bar{\xi}^t_{\eta;h-1}(x,a)}\mathrm{d}\nu(x,a)}\sqrt{\sum_{t=1}^{T}\int 2\bar{d}^{\pi^t}_{h-1}(x,a)\widetilde{\delta}^{t^2}_{f;\pi;h-1}(x,a)\mathrm{d}\nu(x,a) + 2T\eta}. \\
&\hspace{24em}\text{(Marginalization of } \sigma_{\eta,h}\text{)}
\end{aligned}
$$

Now define

$$
\mathrm{err}^t_h =:= \mathbb{1}\left\{\left|f^t_h - \mathrm{apx}(f^t_{h+1})\right| \geq 3\gamma\right\}\cdot\left\{\left(f^t_h - \mathcal{T}_h f^t_{h+1}\right)^2 - \left(\mathrm{apx}[f^t_{h+1}] - \mathcal{T}_h\left[f^t_{h+1}\right]\right)^2\right\},
$$

the term inside the second square root can be bounded by

$$
\begin{aligned}
\sum_{t=1}^{T}\int 2\bar{d}^{\pi^t}_{h-1}(x,a)\widetilde{\delta}^{t^2}_{h-1}(x,a)\mathrm{d}\nu(x,a) &\leq \sum_{t=1}^{T}\int 2\bar{d}^t_h(x,a)A\left\{\mathbb{1}^2\left\{\delta^t(h) \geq 3\eta\right\}\cdot\delta^{t^2}_h(x,a)\right\}\mathrm{d}\nu(x,a) \\
&\leq 6A\sum_{t=1}^{T}\int \bar{d}^t_h(x,a)\mathrm{err}^t_h(x,a)\mathrm{d}\nu(x,a),
\end{aligned}
$$

where the first line is by the property of the one-step-back operator (Definition E.1). Plugging it back into the above we have:

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}^{\xi^t_{\eta;h-1}}\left[\widetilde{\delta}^t_{h-1}\mathbb{1}\{t \geq \tau_{h-1}\}\right] \leq \\
&\underbrace{\sqrt{\sum_{t=1}^{T}\int \frac{\left(\xi^t_{\eta;h-1}(x,a)\mathbb{1}\{t \geq \tau_{h-1}\}\right)^2}{\bar{\xi}^t_{\eta;h-1}(x,a)}\mathrm{d}\nu(x,a)}}_{\text{A: Distribution shift}}\underbrace{\sqrt{\sum_{t=1}^{T}6A\int \bar{d}^t_h(x,a)\mathrm{err}^t_h(x,a)\mathrm{d}\nu(x,a)}}_{\text{B: In-sample error}}+2T\eta.
\end{aligned}
$$

The distribution shift term is bounded by $O(\sqrt{C_{\mathrm{cov}}(\Xi)\log(T)})$, which follows from an identical argument to Theorem 1 of Xie et al. (2023), and the in-sample error term is bounded by $O(\sqrt{\beta T})$ by invoking Lemma E.5 with our choice of $\beta$. Thus, we have that

$$\mathsf{Reg}(T) \leq O\Big(H\sqrt{AC_{\mathrm{cov}}(\Xi_\eta)\beta T\log(T)} + TH\eta\Big).$$

□

### E.2.3. Instantiating the Regret Bound for RichCLD

Having proven Lemma E.1, it remains to show that we can construct a function class that satisfies the approximate realizability and completeness assumption (Assumption E.1) for the RichCLD framework and RichCLD framework with Lipschitz dynamics (Assumption 2.1) has bounded approximate coverability.

Before we prove the realizability and completeness result, we first introduce the notation for the covering set of a function class. Given a function class $\mathcal{F}$, we use $\mathcal{N}_\infty(\mathcal{F};\gamma)$ to denote the $\gamma$-covering set of the function class $\mathcal{F}$, that is, for any $f\in\mathcal{F}$, there exist $f'\in\mathcal{N}_\infty(\mathcal{F};\gamma)$ such that $\|f-f'\|_\infty \leq \gamma$. We denote the size of the covering set $N_\infty(\mathcal{F};\gamma) := |\mathcal{N}_\infty(\mathcal{F};\gamma)|$.

**Lemma E.2** (Realizability and completeness)**.** *Consider the RichCLD framework and suppose Assumptions 2.1 and 2.2 hold. Let* $\mathrm{Lip}_2 : \mathcal{S}\times\mathcal{A}\to[0,1]$ *be the set of all* $[0,1]$*-bounded 2-Lipschitz functions. Define* $\mathcal{F} := \mathrm{Lip}_2\circ\Phi$*, and then* $\mathcal{N}_\infty(\mathcal{F},\gamma)$ *satisfies the approximate realizability and completeness properties in Assumption E.1.*

**Proof of Lemma E.2.** All the results follows if we prove $\mathcal{F} = \mathrm{Lip}_2\circ\Phi$ satisfies exact realizability and completeness. The realizability result is just a combination of Assumption 2.2 and a standard fact from the non-parametric RL literature that Lipschitzness in dynamics implies Lipschitzness in $Q^\star$ (see, e.g., Lemma 2.4 of Sinclair et al. (2019)). For the completeness results, we recall that

$$\begin{aligned}\mathcal{T}_h[f_{h+1}](x,a) &= R_h(x,a) + \mathbb{E}\Big[\max_{a_{h+1}} f_{h+1}(x_{h+1},a_{h+1}) \mid x_h = x, a_h = a\Big]\\ &= R_h(\phi_h^\star(x),a) + \mathbb{E}\Big[\max_{a_{h+1}} f_{h+1}(x_{h+1},a_{h+1}) \mid s_h = \phi_h^\star(x), a_h = a\Big].\end{aligned}$$

This implies that we just require Lipschitzness of the function

$$(s,a)\mapsto \mathcal{T}_h[f_{h+1}](s,a) := R_h(s,a) + \mathbb{E}\Big[\max_{a_{h+1}} f_{h+1}(x_{h+1},a_{h+1}) \mid s_h = s, a_h = a\Big].$$

By Assumption 2.1, $R_h$ is 1-Lipschitz in $(s,a)$, and since $f_{h+1}$ is bounded by 1, $(s,a)\mapsto\mathbb{E}[\max_{a_{h+1}} f_{h+1}(x_{h+1},a_{h+1}) \mid s_h = s, a_h = a]$ is 1-Lipschitz, so that $\mathcal{T}_h[f_{h+1}](s,a)$ is 2-Lipschitz. By Assumption 2.2 and our construction of $\mathcal{F}$, this proves the completeness.

Finally, by realizing any function in $\mathcal{F}$ is Lipschitz with respect to the action, then for all $h\in[H]$, for any $x\in\mathcal{X}$ and $f_h\in\mathcal{F}_{\gamma,h}$, we have $\max_{a\in\mathcal{A}} f_h(x,a) - \max_{a'\in\mathcal{A}_\gamma} f_h(x,a')\leq\gamma$, where $\mathcal{A}_\gamma$ is a $\gamma$-covering set of the action space. □

Next we will prove that the RichCLD framework indeed has a bounded approximate coverability.

**Lemma E.3.** *For the RichCLD framework, there exists one-step-back functions* $\{\mathrm{osb}_h\}_{h=1}^H$ *that satisfies Definition E.1, and there exists a distribution family* $\Xi_\eta$ *which satisfies the conditions of Definition E.2 and* $\Xi_\eta$ *has finite approximate coverability:* $C_{\mathrm{cov}}(\Xi_\eta)\leq d_\eta := \left(\frac{1}{\eta}\right)^{\dim_{\mathcal{SA}}}$.

**Proof of Lemma E.3.** The proof consists of three major parts: we first construct the mapping $\mathrm{osb}_h$ that satisfies Definition E.1, and then the construction of the family of distributions $\Xi_\eta$ that satisfies all the properties in Definition E.2, and finally we show with the constructed $\Xi_\eta$, $C_{\mathrm{cov}}(\Xi_\eta)$ is bounded for any $\eta$. **Construction of** $\mathrm{osb}$**.** For all $h\in[H]$, take $\mathcal{G}_h : \mathcal{X}\times\mathcal{A}\to[0,L]$ to be any positive bounded function class and for any $g_h\in\mathcal{G}$, we have

$$\begin{aligned}\mathbb{E}^\pi[g_h] &\leq \int d_h^\pi(x_h,a_h)g_h(x_h,a_h)\mathrm{d}\nu(x_h,a_h)\\ &= \int d_{h-1}^\pi(x_{h-1},a_{h-1})\underbrace{\int P_{h-1}(x_h\mid x_{h-1},a_{h-1})\pi_h(a_h\mid x_h)g_h(x_h,a_h)\mathrm{d}\nu(x_h,a_h)}_{=:\widetilde{g}_{\pi;h-1}(x_{h-1},a_{h-1})}\mathrm{d}\nu(x_{h-1},a_{h-1}),\end{aligned}$$

thus we have the construction that

$$\mathrm{osb}_h(g_h, \pi_h) = \int P_{h-1}(x_h \mid x_{h-1}, a_{h-1}) \pi_h(a_h \mid x_h) g_h(x_h, a_h) \mathrm{d}\nu(x_h, a_h).$$

By construction, we can easily verify that this choice satisfies

$$\mathbb{E}^{\pi}\left[\mathsf{conv}^{+}(\widetilde{g}_{\pi;h-1})\right] \le \mathbb{E}^{\pi \circ_h \pi_\eta^{\mathsf{unif}}}\left[A_\eta \cdot \mathsf{conv}^{+}(g_h)\right]$$

by Jensen's inequality and importance weighting. We can also this inequality can be extended to any policy $\psi$ such that $\sup_{x,a} \frac{1}{\psi(a|x)} = A < \infty$, and we have

$$\mathbb{E}^{\pi}\left[\mathsf{conv}^{+}(\widetilde{g}_{\pi;h-1})\right] \le \mathbb{E}^{\pi \circ_h \psi}\left[A \cdot \mathsf{conv}^{+}(g_h)\right].$$

**Construction of $\Xi_\eta$.** Now let us consider any fixed $\eta > 0$, $f \in \mathcal{F}$ and $\pi \in \Pi$. We now observe that our construction for $\widetilde{g}$ that satisfies the following property: if the ground-truth latents for two observations $x$ and $x'$ are close to each other, i.e., $D_{\mathcal{S}}(\phi_h^\star(x), \phi_h^\star(x')) \le \eta$, then by to the Lipschitz continuity of $P_h$ with respect to $\mathcal{S}$, we have $|\widetilde{g}_{\pi;h}(x,a) - \widetilde{g}_{\pi;h}(x',a)| \le \eta$ for any $f \in \mathcal{F}$, $\pi \in \Pi$. Note that this property is independent of the original function class $\mathcal{F}$ or $\mathcal{G}$, but instead it is a property inherited from the dynamics of the MDPs from the RichCLD framework. This property hints at a natural construction of $\xi$: instead of focusing on each observation, there should be a distribution that "groups" the support (probability of being visited by $\pi$) where the latent states are close to each other without changing the value of the expectation of $\widetilde{g}_{\pi;h}$ under $d^\pi$ by too much. Recall the definition of $\mathsf{ball}_\eta[\phi_h^\star](x,a)$,

$$\mathsf{ball}_\eta[\phi_h^\star](x,a) := \{\widetilde{x}, \widetilde{a} \in \mathcal{X} \times \mathcal{A} \mid \mathsf{disc}_\eta[\phi_h^\star](\widetilde{x}, \widetilde{a}) = \mathsf{disc}_\eta[\phi_h^\star](x,a)\},$$

which is the set of observations whose corresponding latent states are $\eta$-close to each other in the latent space. Given a policy $\pi$ and an observation-space ball $\mathsf{ball}_\eta[\phi_h^\star](x,a)$, we can define $d_h^\pi(\mathsf{ball}_\eta[\phi_h^\star](x,a))$ be the probability of $\pi$ visiting the ball $\mathsf{ball}_\eta[\phi_h^\star](x,a)$ at timestep $h$, and define $\upsilon(x, a \mid \mathsf{ball}_\eta[\phi_h^\star](x,a))$ be the uniform density over the $\mathsf{ball}_\eta[\phi_h^\star](x,a)$. Then we can define the following distribution $\xi_{\eta,h}$ and coupling $\sigma_{\eta,h}$:

$$\begin{aligned}
&\sigma_{\eta,h}(x,a,\widetilde{x},\widetilde{a}) := \mathbb{1}\{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi_h^\star](x,a)\}\upsilon(x,a \mid \mathsf{ball}_\eta[\phi_h^\star](x,a)) d_h^\pi(\widetilde{x},\widetilde{a}) \\
&\xi_{\eta,h}(x,a) := \int \sigma_{\eta,h}(x,a,\widetilde{x},\widetilde{a}) \mathrm{d}\nu(\widetilde{x},\widetilde{a}) = \upsilon(x,a \mid \mathsf{ball}_\eta[\phi_h^\star](x,a)) d_h^\pi(\mathsf{ball}_\eta[\phi_h^\star](x,a)).
\end{aligned}$$

That is, $\xi$ assigns equal density to all the observations within each observation ball. We can also verify that $d_h^\pi(\widetilde{x},\widetilde{a}) = \int \sigma_{\eta,h}(x,a,\widetilde{x},\widetilde{a}) \mathrm{d}\nu(x,a)$ as required. This concludes the construction of $\xi$.

*Average distribution shift:* $d_h^\pi \to \xi_{\eta,h}$. From the construction above, it follows that

$$\begin{aligned}
&\int d_h^\pi(x,a)\widetilde{g}_{\pi;h}(x,a)\mathrm{d}\nu(x,a) \\
&= \int \upsilon(x,a \mid \mathsf{ball}_\eta[\phi_h^\star](x,a)) \int_{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi_h^\star](x,a)} d_h^\pi(\widetilde{x},\widetilde{a})\widetilde{g}_{\pi;h}(\widetilde{x},\widetilde{a})\mathrm{d}\nu(\widetilde{x},\widetilde{a})\mathrm{d}\nu(x,a) \\
&\le \int \upsilon(x,a \mid \mathsf{ball}_\eta[\phi_h^\star](x,a)) \int_{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi_h^\star](x,a)} d_h^\pi(\widetilde{x},\widetilde{a})\widetilde{g}_{\pi;h}(x,a)\mathrm{d}\nu(\widetilde{x},\widetilde{a})\mathrm{d}\nu(x,a) + \eta \\
&= \int \xi_{\eta,h}(x,a)\widetilde{g}_{\pi;h}(x,a)\mathrm{d}\nu(x,a) + \eta.
\end{aligned}$$

*Pointwise distribution shift:* $\xi_{\eta,h} \to d_h^\pi$. To prove the second distribution shift property, we have for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned}
&\xi_{\eta,h}(x,a) g^2_{\pi;h}(x,a) \\
&= \upsilon(x,a \mid \mathsf{ball}_\eta[\phi^\star_h](x,a)) \int_{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi^\star_h](x,a)} d_h^\pi(\widetilde{x},\widetilde{a}) \widetilde{g}^2_{\pi;h}(x,a) \mathrm{d}\nu(\widetilde{x},\widetilde{a}) \\
&\leq \upsilon(x,a \mid \mathsf{ball}_\eta[\phi^\star_h](x,a)) \int_{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi^\star_h](x,a)} d_h^\pi(\widetilde{x},\widetilde{a}) 2\Big( \big(\widetilde{g}_{\pi;h}(\widetilde{x},\widetilde{a}) - \widetilde{g}_{\pi;h}(x,a)\big)^2 + \widetilde{g}^2_{\pi;h}(\widetilde{x},\widetilde{a}) \Big) \mathrm{d}\nu(\widetilde{x},\widetilde{a}) \quad \text{(AM-GM)} \\
&\leq 2\upsilon(x,a \mid \mathsf{ball}_\eta[\phi^\star_h](x,a)) \int_{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi^\star_h](x,a)} d_h^\pi(\widetilde{x},\widetilde{a}) \widetilde{g}^2_{\pi;h}(\widetilde{x},\widetilde{a}) \mathrm{d}\nu(\widetilde{x},\widetilde{a}) + 2\xi_{\eta,h}(x,a)\eta^2 \\
&= 2 \int \sigma_{\eta,h}(\widetilde{x},\widetilde{a},x,a) \widetilde{g}^2_{\pi;h}(\widetilde{x},\widetilde{a}) \mathrm{d}\nu(\widetilde{x},\widetilde{a}) + 2\xi_{\eta,h}(x,a)\eta^2.
\end{aligned}$$

This finishes the construction of $\Xi$. **Bounding $C_{\mathrm{cov}}(\Xi)$.** Recall the definition of $C_{\mathrm{cov}}(\Xi)$:

$$C_{\mathrm{cov}}(\Xi_\eta) := \inf_{\mu_1,\ldots,\mu_H \in \Delta(\mathcal{X}\times\mathcal{A})} \sup_{\xi_\eta \in \Xi_\eta, h \in [H]} \left\| \frac{\xi_{\eta;h}}{\mu_h} \right\|_\infty.$$

Our goal is to construct $\mu_h$ for each $h \in [H]$ such that $C_{\mathrm{cov}}(\Xi_\eta)$ is bounded. We first make the following observation: by the construction of $\xi$, we can see that it is the same as the distribution of a tabular Block MDP where the states are the collections of states in each ball on the level of discretization $\eta$. Follow this intuition, recall the notation $\mathcal{B}_\eta := \{\mathsf{ball}_\eta(s_\eta, a_\eta) : s_\eta, a_\eta \in \mathcal{S}_\eta \times \mathcal{A}_\eta\}$, which defines the set of all $\eta$-balls in the latent space. Now fix $h \in [H]$, for each ball $b_\eta \in \mathcal{B}_\eta$, we define the policy that maximizes its probability of visiting this ball:

$$\pi_{b_\eta} = \arg\max_{\pi \in \Pi} \mathbb{P}^\pi(b_\eta).$$

Then we can construct the following distribution $\mu_h$:

$$\mu_h(x,a) = \frac{1}{S_\eta A_\eta} \sum_{b_\eta \in \mathcal{B}_\eta} \upsilon(x \mid b_\eta) d_h^{\pi_{b_\eta}}(b_\eta),$$

that is, the distribution $\mu_h$ takes the uniform distribution over the grid actions and for each observation, it takes the average of the "ball-reaching" policy's occupancy measure over each ball (similar to how we design $\xi$), and take the uniform distribution over all the "ball-reaching" policy's occupancy measure. Now given any $x \in \mathcal{X}$, and policy $\pi$ and its corresponding $\xi_{\eta,h}$, we define the approximate occupancy measure on observation:

$$\xi_{\eta,h}(x,a) = \upsilon(x \mid \mathsf{ball}_\eta[\phi^\star_h](x)) d_h^\pi(\mathsf{ball}_\eta[\phi^\star_h](x)) \pi_h(a \mid x),$$

i.e., the average occupancy measure of the ball $\mathsf{ball}_\eta[\phi^\star_h](x)$ under policy $\pi$. Then denote $b_\eta = \mathsf{ball}_\eta[\phi^\star_h](x)$, we have that

$$\frac{\xi_{\eta,h}(x,a)}{\mu_h(x,a)} \leq S_\eta A_\eta \cdot \frac{\upsilon(x \mid b_\eta) d_h^\pi(b_\eta)}{\upsilon(x \mid b_\eta) d_h^{\pi_{b_\eta}}(b_\eta)} = S_\eta A_\eta \cdot \frac{d_h^\pi(b_\eta)}{d_h^{\pi_{b_\eta}}(b_\eta)} \leq S_\eta A_\eta,$$

where the last inequality follows the definition of $\pi_{b_\eta}$ such that $\pi_{b_\eta} = \arg\max_{\pi \in \Pi} \mathbb{P}^\pi(s_h \in b_\eta)$. And the proof is completed by the fact that $S_\eta A_\eta \leq \left(\frac{1}{\eta}\right)^{\dim_{\mathcal{SA}}}$. □

With Lemma E.1 and Lemma E.3, we are ready to prove Theorem 3.1′.

**Proof of Theorem 3.1′.** By Lemma E.1, we have that GOLF.DBR satisfies that

$$\mathsf{Reg}(T) \leq \mathcal{O}\bigg( H\sqrt{C_{\mathrm{cov}}(\Xi_\eta) A_\eta \beta T \log(T)} + TH\eta \bigg),$$

as long as we take

$$\beta = c \log(N_\infty(\mathcal{F},\eta) TH/\delta),$$

where we choose function class $\mathcal{N}_\infty(\mathcal{F},\eta)$ where $\mathcal{F} = \mathrm{Lip} \circ \Phi$ such that approximate completeness holds (Lemma E.2), and thus the event in Lemma E.5 holds with probability at least $1-\delta$ for all $t \in [T]$, and choose estimation policy $\psi$ to the be uniform policy over the covering action space $\mathcal{A}_\eta$.

Plugging in this choice for $\beta$ and the bound on $\Xi_\eta$ from Lemma E.3, we have with probability as least $1-\delta$,

$$\mathsf{Reg}(T) \le O\left(H\sqrt{(1/\eta)^{\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}} T \log(THN_\infty(\mathcal{F},\eta)/\delta)} + TH\eta\right).$$

Then to calculate the covering number of $\mathcal{F}$, note that the covering number of Lip is $\left(\frac{1}{\eta}\right)^{\frac{1}{\eta}^{\dim_{\mathcal{SA}}}}$ (Wainwright, 2019), and $\Phi$ is finite, we have

$$\mathsf{Reg}(T) \le O\left(H\sqrt{(1/\eta)^{2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}} T \log(TH|\Phi|/\eta\delta)} + TH\eta\right).$$

To conclude, we choose $\eta = T^{-\frac{1}{2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2}}$, which gives

$$\mathsf{Reg}(T) \le \mathcal{O}\left(HT^{\frac{2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+1}{2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2}}\sqrt{\log(TH|\Phi|/\eta\delta)}\right).$$

Finally, we convert this to a PAC bound. We have that with number of total samples

$$H \cdot T = \mathcal{O}\left(H^{(2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+3)} \log(TH|\Phi|/\delta\varepsilon)/\varepsilon^{(2\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}+2)}\right)$$

sufficiently large, the policy $\widehat{\pi}$ satisfies

$$J(\pi^\star) - J(\widehat{\pi}) \le \varepsilon.$$

□

### E.3. Auxiliary Lemmas

**Lemma E.4.** *Given any value function* $\{f_h\}_{h=1}^H$ *that satisfies the last condition in Assumption E.1, and let* $\pi_f := \left\{\pi_{f;h}(x_h) = \arg\max_{a_h \in \mathcal{A}_\eta} f_h(x_h,a_h)\right\}_{h=1}^H$*, we have*

$$\mathbb{E}[f_1(x_1,\pi_{f;1}(x_1)) - J(\pi_f)] \le \sum_{h=1}^H \mathbb{E}_{x_h,a_h\sim d_h^{\pi_f}}[|f_h(x_h,a_h) - \mathcal{T}_h[f_{h+1}](x_h,a_h)|] + H\eta.$$

**Proof of Lemma E.4.** The proof is mostly the same as the standard results such as Lemma 1 of Jiang et al. (2017), and the only difference is that we need to handle the misspecification due to the continuous action space. By the standard decomposition, we have

$$\begin{aligned}
&\mathbb{E}[f_1(x_1,\pi_{f;1}(x_1)) - J(\pi_f)] \\
=&\mathbb{E}_{a_1\sim\pi_{f;1}(x_1)}\left[f_1(x_1,a_1) - \left(r_1(x_1,a_1) + \mathbb{E}_{x_2\sim P_1(x_1,a_1)}\left[V_2^{\pi_f}(x_2)\right]\right)\right] \\
=&\mathbb{E}_{a_1\sim\pi_{f;1}(x_1)}\Big[f_1(x_1,a_1) - \left(r_1(x_1,a_1) + \mathbb{E}_{x_2\sim P_1(x_1,a_1)}[f_2(x_2,\pi_{f;2}(x_2))]\right) + \\
&\qquad\qquad \mathbb{E}_{x_2\sim P_1(x_1,a_1)}\left[f_2(x_2,\pi_{f;2}(x_2)) - V_2^{\pi_f}(x_2)\right]\Big].
\end{aligned}$$

Note that the second term is the recursion term, and for the first term, we have

$$\begin{aligned}
&\mathbb{E}_{a_1\sim\pi_{f;1}(x_1)}\left[f_1(x_1,a_1)-\left(r_1(x_1,a_1)+\mathbb{E}_{x_2\sim P_1(x_1,a_1)}[f_2(x_2,\pi_{f;2}(x_2))]\right)\right]\\
=&\mathbb{E}_{a_1\sim\pi_{f;1}(x_1)}\left[\left|f_1(x_1,a_1)-\left(r_1(x_1,a_1)+\mathbb{E}_{x_2\sim P_1(x_1,a_1)}\left[\max_{a_2\in\mathcal{A}_\eta} f_2(x_2,a_2)\right]\right)\right|\right]\\
\leq&\mathbb{E}_{a_1\sim\pi_{f;1}(x_1)}\left[\left|f_1(x_1,a_1)-\left(r_1(x_1,a_1)+\mathbb{E}_{x_2\sim P_1(x_1,a_1)}\left[\max_{a_2\in\mathcal{A}} f_2(x_2,a_2)\right]\right)\right|\right]+\eta \\
&\text{(last property in Assumption E.1)}\\
=&\mathbb{E}_{a_1\sim\pi_{f;1}(x_1)}[|f_1(x_1,a_1)-\mathcal{T}_1[f_2(x_1,a_1)]|]+\eta.
\end{aligned}$$

Then by applying the recursion on the second term we complete the proof. □

**Lemma E.5** (Theorem 2.1 and Eq. (13) of Amortila et al. (2024))**.** *Suppose Assumption E.1 holds for $\mathcal{F}_\eta$. For any $f\in\mathcal{F}_\eta$, denote $\mathrm{apx}[f_{h+1}]\in\mathcal{N}_\infty(\mathcal{F},\gamma)$ as the $\ell_\infty$-approximation of $\mathcal{T}_h[f_{h+1}]$. Then for each $t\in[T]$, denote*

$$\begin{aligned}
&\delta^t_{f;h}:=f^t_h-\mathcal{T}_h\left[f^t_{h+1}\right],\quad \textit{and}\\
&\mathrm{err}^t_{f;h}:=\mathbb{1}\left\{\left|f^t_h-\mathrm{apx}(f^t_{h+1})\right|\geq 3\gamma\right\}\cdot\left\{\left(f^t_h-\mathcal{T}_hf^t_{h+1}\right)^2-\left(\mathrm{apx}[f^t_{h+1}]-\mathcal{T}_h\left[f^t_{h+1}\right]\right)^2\right\}.
\end{aligned}$$

*Then running Algorithm 5 with $\mathcal{F}_\gamma$ and*

$$\beta=c\log(TH|\mathcal{F}_\eta|/\delta),$$

*with probability at least $1-\delta$, we have that for all $t\in[T],h\in[H]$,*

$$f^\star_\infty\in\mathcal{F}^t\quad\textit{and}\quad\sum_{i<t}\mathbb{E}_{x,a\sim d^i_h}\,\mathrm{err}^t_{f^t;h}(x,a)\leq\mathcal{O}(\beta),$$

*where $d^i_h$ is the data collection distribution of Algorithm 5 defined in Eq. (15), and the $f^\star_\infty$ is the $\ell_\infty$ approximation of $Q^\star$, i.e., $\|f^\star_\infty-Q^\star\|_\infty=\gamma$.*

**Lemma E.6** (Almost optimism)**.** *Suppose Lemma E.5 holds. Then for each iteration $t\in[T]$, we have*

$$f^t_1(x_1,\pi^t_1(x_1))-J(\pi^\star)\geq-(\eta+\gamma).$$

**Proof of Lemma E.6.** Conditioned on the event that Lemma E.5 holds, by the construction of $f^t$, we have

$$\max_{a\in\mathcal{A}}f^t_1(x_1,a)\geq f^\star_{\infty;1}(x_1,\pi^\star(x_1))\geq J(\pi^\star)-\gamma.$$

By the the last property of Assumption E.1, we have

$$\max_{a\in\mathcal{A}}f^t_1(x_1,a)-f^t_1(x_1,\pi^t_1(x_1))=\max_{a\in\mathcal{A}}f^t_1(x_1,a)-\max_{a\in\mathcal{A}_\eta}f^t_1(x_1,a)\leq\eta.$$

Then put everything together we complete the proof. □

### E.4. Proof of Theorem 3.2 (Lower Bound for $Q^\star/V^\star$-Lipschitz Latent Dynamics)

The following result gives the formal version of Theorem 3.2.

**Theorem 3.2′** (Lipschitz $Q^\star/V^\star$ lower bound; formal version of Theorem 3.2)**.** *For every $N\in\mathbb{N}$, there exists a decoder class $\Phi$ with $|\Phi|=N$ and a family of rich observation MDPs $\mathcal{M}$ with (i) $H\leq\mathcal{O}(\log(N))$, (ii) $|\mathcal{S}_\eta|\leq 4H\leq\mathcal{O}(\log(N))$ for all $\eta\geq 0$, (iii) $|\mathcal{A}|=2$, (iv) $|\mathcal{X}|\leq N^2$, (v) $Q^\star/V^\star$-Lipschitz latent dynamics, and (vi) decoder realizability, such that any algorithm requires $\Omega(N/\log(N))$ episodes to learn an c-optimal policy for a worst-case MDP in $\mathcal{M}$, where $c>0$ is an absolute constant.*

**Proof of Theorem 3.2′.** Let $N \geq 4$ be given, and assume without loss of generality that it is a power of 2. We first construct the class of rich-observation MDPs, then verify the Lipschitz assumption and prove a sample complexity lower bound.

**Latent MDP.** Our construction has a single "known" latent MDP $M^{\text{latent}}$; that is, the only uncertainty in the family of rich-observation MDPs we construct arises from the emission processes. Set $H = \log_2(N) + 1$ and $\mathcal{A} = \{0, 1\}$. We define the state space and latent transition dynamics as follows.

- The state space can be partitioned as $\mathcal{S} = \mathcal{S}^1, \ldots, \mathcal{S}^N$.
- Each block $\mathcal{S}^i$ corresponds to a standard depth-$H$ binary tree MDP with deterministic dynamics (e.g., Osband & Van Roy (2016); Domingues et al. (2021)). There is a single "root" node at layer $h = 1$, which we denote by $s^i_{\mathsf{root}}$, and $N$ "leaf" nodes at layer $H$, which we denote by $\left\{s^{i,j}_{\mathsf{leaf}}\right\}_{j\in[N]}$. For each $h = 1, \ldots, H-1$, choosing action 0 leads to the left successor of the current state deterministically, and choosing action 1 leads to the right sucessor; this process continues until we reach a leaf node at layer $H$.
- The initial state distribution is $P^{\text{latent}}_1(\emptyset) = \mathsf{unif}(s^1_{\mathsf{root}}, \ldots, s^N_{\mathsf{root}})$.
- There are no rewards for layers $1, \ldots, H-1$. For layer $H$, the reward is

$$R_H(s^{i,j}_{\mathsf{leaf}}, \cdot) = \mathbb{1}\{j = i\}. \tag{16}$$

Informally, this construction can summarized as follows. At layer 1, we draw the index of one of $N$ binary trees uniformly at random, and initialize into the root of the tree. From here, we receive a reward of 1 if we successfully navigate to the leaf node whose index agrees with the index of the tree itself, and receive a reward of 0 otherwise.

Note that the total number of latent states in this construction is $|\mathcal{S}| = N \cdot |\mathcal{S}_1| = N(2N-1)$; we will now choose a metric for which the covering number is significantly smaller.

**Verifying the Lipschitz assumption.** Let $Q^\star$ denote the optimal Q-function for the MDP $M^{\text{latent}}$. Since $M^{\text{latent}}$ is known/fixed in our construction, we can choose the state metric as

$$D_{\mathcal{S}}(s, s') = \mathbb{1}\{Q^\star_h(s, \cdot) \neq Q^\star_h(s', \cdot)\},$$

for states $s$ and $s'$ that are both reachable at layer $h \in [H]$, and set $D_{\mathcal{S}}(s, s') = 1$ if $s$ and $s'$ are reachable at different layers. We also choose $D_{\mathcal{A}}(a, a') = \mathbb{1}\{a \neq a'\}$. With this construction, it is immediate that for all $\eta \geq 0$, the latent space covering number with respect to $D_{\mathcal{S}}$ is bounded as $|\mathcal{S}_\eta| \leq 4H = 4(\log_2(N) + 1)$. Indeed, our reward construction ensures that $Q^\star$ is $0/1$-valued, so there are only four possible profiles ($(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$) that $Q^\star_h(s, \cdot)$ can take on. We form the cover by merging all states in a given layer whose profiles agree.

**Observation space and decoder class.** Let us introduce some additional notation. For each block $\mathcal{S}^i$, let $\mathcal{S}^i_h := \{s^{i,j}_h\}_{j\in[2^{h-1}]}$ denote the states in block $i$ that are reachable at layer $h$, so that $\mathcal{S}^i_1 = \left\{s^i_{\mathsf{root}}\right\}$ and $\mathcal{S}^i_H = \{s^{i,j}_{\mathsf{leaf}}\}_{j\in[N]}$. We define $\mathcal{X} = \mathcal{S}$ so that $|\mathcal{X}| \leq 4N^2$, and consider a class of emission processes corresponding to deterministic maps. Let

$$\Psi = \{\psi_i\}_{i\in[N]}$$

denote the set of cyclic permutations on $N$ elements, excluding the identity permutation. That is, each $\psi \in \Psi$ takes the form $\psi_i : k \mapsto k + i \mod N$ for $i \in \{1, \ldots, N\}$. For each $\psi \in \Psi$, we consider the emission process

$$E^\psi_h(\cdot \mid s^{i,j}_h) = \mathbb{1}_{s^{\psi(i),j}_h}.$$

That is, $E^\psi$ shifts the index of the binary tree containing $s^{i,j}_h$ according to $\psi$. We consider the class of rich-observation MDPs given by

$$\mathcal{M} := \left\{M^i := E^{\psi_i} \circ M^{\text{latent}} \mid \psi_i \in \Psi\right\}, \tag{17}$$

where $M^\psi := E^\psi \circ M^{\text{latent}}$ denotes the rich-observation MDP obtained by equipping $M^{\text{latent}}$ with the emission process $E^\psi$. It is clear that this class of rich-observation MDPs satisfies the decodability assumption for the class

$$\Phi = \left\{s^{i,j}_h \mapsto s^{\psi^{-1}(i),j}_h \mid \psi \in \Psi\right\},$$

which has $|\Phi| = N$.

**Sample complexity lower bound.** To lower bound, the sample complexity, we prove a lower bound on the constrained PAC Decision-Estimation Coefficient (DEC) of Foster et al. (2023a). For an arbitrary MDP $\overline{M}$ (defined over the space $\mathcal{X}$) and $\varepsilon \in [0, 2^{1/2}]$, define[10]

$$\mathsf{dec}_{\varepsilon}(\mathcal{M}, \overline{M}) = \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[J^M(\pi_M) - J^M(\pi)] \mid \mathbb{E}_{\pi \sim q}\left[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \leq \varepsilon^2 \right\},$$

where $M(\pi)$ denotes the law over trajectories $(x_1, a_1, r_1), \ldots, (x_H, a_H, r_H)$ induced by executing the policy $\pi$ in the MDP $M$, $J^M(\pi)$ denotes the expected reward for policy $\pi$ under $M$, and $\pi_M$ denotes the optimal policy for $M$. We further define

$$\mathsf{dec}_{\varepsilon}(\mathcal{M}, \overline{M}) = \sup_{\overline{M}} \mathsf{dec}_{\varepsilon}(\mathcal{M}, \overline{M}),$$

where the supremum ranges over all MDPs defined over $\mathcal{X}$ and $\mathcal{A}$. We now appeal to the following lemma.

**Lemma E.7.** *For all $\varepsilon^2 \geq 4/N$, we have that* $\sup_{\overline{M}} \mathsf{dec}_{\varepsilon}(\mathcal{M}, \overline{M}) \geq \frac{1}{2}$.

In light of Lemma E.7, it follows from Theorem 2.1 in Foster et al. (2023a)[11] that any PAC RL algorithm that uses $T$ episodes of interaction for $T \log(T) \leq c \cdot N$ must have $\mathbb{E}[J^M(\pi_M) - J^M(\widehat{\pi})] \geq c'$ for a worst-case MDP in $\mathcal{M}$, where $c, c' > 0$ are absolute constants. □

**Proof of Lemma E.7.** Define $\overline{M}^{\text{latent}}$ as the latent-space MDP that has identical dynamics to $M^{\text{latent}}$ but, has zero reward in every state, and define $\overline{M} := \text{id} \circ \overline{M}^{\text{latent}}$ as the rich-observation MDP obtained by composing $\overline{M}^{\text{latent}}$ with the identity emission process that sets $x_h = s_h$. Observe that $\overline{M}$ and $M_i$, induce identical dynamics in observation space if rewards are ignored: For all policies $\pi$,

$$\mathbb{P}^{\overline{M},\pi}[(x_1, a_1), \ldots, (x_H, a_H) = \cdot] = \mathbb{P}^{M^i,\pi}[(x_1, a_1), \ldots, (x_H, a_H) = \cdot]. \tag{18}$$

This is because, e.g., under $M^i$ the probability that $x_1 = j$ is $1/N$ and this is true for all choices of the cyclic permutation. It follows that for each $i$, for all policies $\pi$, we have

$$\begin{aligned}
D^2_{\mathsf{H}}\big(M^i(\pi), \overline{M}(\pi)\big) &= D^2_{\mathsf{H}}\Big((E^{\psi_i} \circ M^{\text{latent}})(\pi), (\text{id} \circ \overline{M}^{\text{latent}})(\pi)\Big) \\
&= \sum_{j=1}^{N} \mathbb{P}^{\overline{M},\pi}\left[x_H = s^{\psi_i(j),j}_{\mathsf{leaf}}\right] \cdot D^2_{\mathsf{H}}(\mathbb{1}_1, \mathbb{1}_0) \\
&= 2\sum_{j=1}^{N} \mathbb{P}^{\overline{M},\pi}\left[x_H = s^{\psi_i(j),j}_{\mathsf{leaf}}\right] \\
&= \frac{2}{N}\sum_{j=1}^{N} \mathbb{P}^{\overline{M},\pi}\left[x_H = s^{\psi_i(j),j}_{\mathsf{leaf}} \mid x_1 = s^{\psi_i(j)}_{\mathsf{root}}\right], \qquad (19) \\
&= \frac{2}{N}\sum_{j=1}^{N} \mathbb{P}^{\overline{M},\pi}\left[x_H = s^{j,\psi_i^{-1}(j)}_{\mathsf{leaf}} \mid x_1 = s^{j}_{\mathsf{root}}\right]. \qquad (20)
\end{aligned}$$

The second identity is the conditioning rule for Hellinger distance, which states that:

$$D^2_{\mathsf{H}}\big(\mathbb{P}_{Y|X} \circ \mathbb{P}_X, \mathbb{Q}_{Y|X} \circ \mathbb{P}_X\big) = \mathbb{E}_{X \sim \mathbb{P}_X}\left[D^2_{\mathsf{H}}\big(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\big)\right].$$

We instantiate this with $X$ as the trajectory, i.e., $(x_1, a_1), \ldots, (x_H, a_H)$, and $Y$ as the reward, and use the fact that the law of the trajectory is the same for $M^i(\pi)$ and $\overline{M}(\pi)$. For the rest of the calculation, we use that the learner receives identical

[10] For measures $\mathbb{P}$ and $\mathbb{Q}$, we define squared Hellinger distance by $D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}})^2$.

[11] Theorem 2.1 in Foster et al. (2023a) is stated with respect to $\sup_{\overline{M} \in \text{conv}(\mathcal{M})} \mathsf{dec}_{\varepsilon}(\mathcal{M}, \overline{M})$, but the actual proof (Section 2.2) gives a stronger result that scales with $\sup_{\overline{M}} \mathsf{dec}_{\varepsilon}(\mathcal{M}, \overline{M})$.

feedback in the MDPs $M^i$ and $\overline{M}$ unless they reach the observation $x_H = s_{\mathsf{leaf}}^{\psi_i(j),j}$ for some $j$ (corresponding to latent state $s_{\mathsf{leaf}}^{j,j}$ in $M^i$), in which case they receiver reward 1 in $M^i$ but reward 0 in $\overline{M}$. We now claim that for any $q \in \Delta(\Pi)$, there exists a set of at least $N/2$ indices $\mathcal{I}_q \subset [N]$ such that

$$\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M^i(\pi), \overline{M}(\pi)\big)\big] \leq \frac{4}{N} \tag{21}$$

for all $i \in \mathcal{I}_q$. To see this, note that by Eq. (20), we have

$$\begin{aligned}\mathbb{E}_{i\sim\mathsf{unif}([N])}\,\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M^i(\pi), \overline{M}(\pi)\big)\big] &\leq \mathbb{E}_{\pi\sim q}\left[\frac{2}{N}\sum_{j=1}^N \frac{1}{N}\sum_{i=1}^N \mathbb{P}^{\overline{M},\pi}\Big[x_H = s_{\mathsf{leaf}}^{j,\psi_i^{-1}(j)} \mid x_1 = s_{\mathsf{root}}^j\Big]\right] \\ &\leq \mathbb{E}_{\pi\sim q}\left[\frac{2}{N}\sum_{j=1}^N \frac{1}{N}\right] = \frac{2}{N},\end{aligned}$$

where the second inequality uses that $\sum_{i=1}^N \mathbb{P}^{\overline{M},\pi}\Big[x_H = s_{\mathsf{leaf}}^{j,\psi_i^{-1}(j)} \mid x_1 = s_{\mathsf{root}}^j\Big] \leq 1$, as the events in the sum are mutually exclusive (and the event we condition on does not depend on $i$). We conclude by Markov's inequality that $\mathbb{P}_{i\sim\mathsf{unif}([N])}\big[\mathbb{E}_{\pi\sim q}\big[D^2_{\mathsf{H}}\big(M^i(\pi), \overline{M}(\pi)\big)\big] \geq 4/N\big] \leq 1/2$, giving $\mathcal{I}_q \geq N/2$.

From Eq. (21), we conclude that for all $\varepsilon^2 \geq 4/N$,

$$\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \inf_{q\in\Delta(\Pi)}\inf_{p\in\Delta(\Pi)}\sup_{i\in\mathcal{I}_q}\Big\{\mathbb{E}_{\pi\sim p}\Big[J^{M^i}(\pi_{M^i}) - J^{M^i}(\pi)\Big]\Big\}.$$

To lower bound this quantity, observe that for any index $i$ and any policy $\pi$, we have

$$\begin{aligned}J^{M^i}(\pi_{M^i}) - J^{M^i}(\pi) &= \frac{1}{N}\sum_{j=1}^N \mathbb{P}^{M^i,\pi}\Big[x_H \neq s_{\mathsf{leaf}}^{\psi_i(j),j} \mid x_1 = s_{\mathsf{root}}^{\psi_i(j)}\Big] \\ &= 1 - \frac{1}{N}\sum_{j=1}^N \mathbb{P}^{M^i,\pi}\Big[x_H = s_{\mathsf{leaf}}^{\psi_i(j),j} \mid x_1 = s_{\mathsf{root}}^{\psi_i(j)}\Big] \\ &= 1 - \frac{1}{N}\sum_{j=1}^N \mathbb{P}^{\overline{M},\pi}\Big[x_H = s_{\mathsf{leaf}}^{\psi_i(j),j} \mid x_1 = s_{\mathsf{root}}^{\psi_i(j)}\Big] \\ &= 1 - \frac{1}{N}\sum_{j=1}^N \mathbb{P}^{\overline{M},\pi}\Big[x_H = s_{\mathsf{leaf}}^{j,\psi_i^{-1}(j)} \mid x_1 = s_{\mathsf{root}}^{j}\Big],\end{aligned}$$

where the third inequality uses Eq. (18). We conclude that for any distribution $p, q \in \Delta(\Pi)$,

$$\begin{aligned}&\sup_{i\in\mathcal{I}_q}\Big\{\mathbb{E}_{\pi\sim p}\Big[J^{M^i}(\pi_{M^i}) - J^{M^i}(\pi)\Big]\Big\} \\ &\geq \mathbb{E}_{i\sim\mathsf{unif}(\mathcal{I}_q)}\Big\{\mathbb{E}_{\pi\sim p}\Big[J^{M^i}(\pi_{M^i}) - J^{M^i}(\pi)\Big]\Big\} \\ &\geq 1 - \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{i\sim\mathsf{unif}(\mathcal{I}_q)}\,\mathbb{P}^{\overline{M},\pi}\Big[x_H = s_{\mathsf{leaf}}^{j,\psi_i^{-1}(j)} \mid x_1 = s_{\mathsf{root}}^{j}\Big] \\ &= 1 - \frac{1}{N}\sum_{j=1}^N \frac{1}{|\mathcal{I}_q|}\sum_{i\in\mathcal{I}_q}\mathbb{P}^{\overline{M},\pi}\Big[x_H = s_{\mathsf{leaf}}^{j,\psi_i^{-1}(j)} \mid x_1 = s_{\mathsf{root}}^{j}\Big] \geq 1 - \frac{1}{|\mathcal{I}_q|} \geq \frac{1}{2}\end{aligned}$$

as long as $N \geq 4$, where the second-to-last inequality uses that for all $j$, the events $\big\{x_H = s_{\mathsf{leaf}}^{j,\psi_i^{-1}(j)} \mid x_1 = s_{\mathsf{root}}^{j}\big\}$ are disjoint for all $i$. Since this lower bound holds uniformly for all $q, p \in \Delta(\Pi)$, we conclude that

$$\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \frac{1}{2}.$$

□

# F. Proofs from Section 4

The structure of this section is as follows. We first provide the proof of the representation learning result (Theorem 4.1) in Appendix F.1. Then in Appendix F.2, we prove the formal version of the guarantee of CRIEE (Theorem 4.2′), and in Appendix F.3 we present the proof for Theorem 4.2′. Finally in Appendix F.4, we prove Proposition 4.1, the negative result for the prior approaches in RichCLD.

## F.1. Proof of Theorem 4.1

The proof of Theorem 4.1 follows standard concentration arguments, combined with structural properties of the RichCLD framework that follow from Assumptions 2.1 and 2.2. We first prove some technical lemmas, as well as a more general version of Theorem 4.1 (Appendix F.1.1), then prove Theorem 4.1 as a consequence (Appendix F.1.2). Throughout this section of the appendix, when using a decoder $\phi : \mathcal{X} \to \mathcal{S}$, we abbreviate $\phi(x, a) := (\phi(x), a)$ to keep notation compact.

### F.1.1. Technical Tools

In this section, we prove a more general version of Theorem 4.1, Theorem F.1, which supports a setting in which the dataset is collected in an online fashion; this result will be used later to prove our online RL guarantees for Algorithm 2. Beyond allowing for an online dataset, we also prove a result that holds for an arbitrary discriminator class $\mathcal{F}$; Theorem 4.1 will follow by instantiating the discriminator class as the composition of the decoder class and a class of Lipschitz functions.

We recall notation for covering numbers for a function class. Denote the $\ell_\infty$ norm for a function $f : \mathcal{X} \to \mathbb{R}$ as $\|f\|_\infty = \sup_{x\in\mathcal{X}}|f(x)|$. For any function class $\mathcal{F}$, we denote a $\gamma$-cover in the $\ell_\infty$ norm by $\mathcal{N}_\infty(\mathcal{F}, \gamma)$; i.e., for any $f \in \mathcal{F}$, there exists $\tilde{f} \in \mathcal{N}_\infty(\mathcal{F}, \gamma)$ such that $\|f - \tilde{f}\|_\infty \le \gamma$. We define $N_\infty(\mathcal{F}, \gamma) = |\mathcal{N}_\infty(\mathcal{F}, \gamma)|$.

**Theorem F.1** (Guarantee of BCRL.C). *Let $h \in [H]$ be fixed, and consider a dataset $\mathcal{D}_h := \{(x_h^i, a_h^i, x_{h+1}^i)\}_{i=1}^t$, where $(x_h^i, a_h^i, x_{h+1}^i) \sim \bar{\rho}_h^t \circ P_h$, and $\bar{\rho}_h^t = \frac{1}{t}\sum_{i=1}^t \rho_h^i$, where $\rho_h^i \in \Delta(\mathcal{X} \times \mathcal{A})$ is an arbitrary state-action distribution that may depend upon the randomness of rounds $1 \dots i-1$. Suppose Assumptions 2.1 and 2.2 hold, and let $\phi_h^t$ be the output of Algorithm 1 with inputs $\mathcal{D}_h$, $\mathcal{F} : \mathcal{X} \to [0, L]$, and $\mathrm{Lip} : \mathcal{S} \to [0, L]$. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$, for all $\gamma \in (0, 1)$,*

$$\max_{f\in\mathcal{F}_{h+1}} \mathbb{E}_{x,a\sim\bar{\rho}_h^t}\left[\left(\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h^t}[f](x, a) - \mathcal{P}_h[f](x, a)\right)^2\right] \le \varepsilon_{\mathrm{rep}}(t, \delta, \gamma),$$

*where*

$$\varepsilon_{\mathrm{rep}}(t, \delta, \gamma) = \frac{88L^2(2L)^{\dim_{\mathcal{SA}}}\left(\frac{1}{\gamma}\right)^{\dim_{\mathcal{SA}}}\log(2L|\Phi|\cdot N_\infty(\mathcal{F}, \gamma)/(\delta\gamma))}{t} + 24L^2\gamma^2.$$

Before proving Theorem F.1, we state a basic concentration lemma, proven in the sequel. Recall that for any $h \in [H]$, given dataset $\mathcal{D}_h$, the least squares loss between a decoder $\phi$, lipschitz function $g$ and discriminator $f$ is defined as

$$\ell_{\mathcal{D}_h}(\phi, g, f) := \widehat{\mathbb{E}}_{(x_h,a_h,x_{h+1})\sim\mathcal{D}_h}\left[(g(\phi(x_h, a_h)) - f(x_{h+1}))^2\right].$$

Similarly, the population least square loss between a decoder $\phi$, lipschitz function $g$ and discriminator $f$ under the distribution $\bar{\rho}_h^t$ is defined as

$$\ell_{\bar{\rho}_h^t}(\phi, g, f) := \mathbb{E}_{(x_h,a_h)\sim\bar{\rho}_h^t}\left[(g(\phi(x_h, a_h)) - \mathbb{E}[f \mid x_h, a_h])^2\right].$$

**Lemma F.1** (Concentration of the BCRL.C loss). *Let $h \in [H]$ be fixed, and consider a dataset $\mathcal{D}_h := \{(x_h^i, a_h^i, x_{h+1}^i)\}_{i=1}^t$, where $(x_h^i, a_h^i, x_{h+1}^i) \sim \bar{\rho}_h^t \circ P_h$, and $\bar{\rho}_h^t = \frac{1}{t}\sum_{i=1}^t \rho_h^i$, where $\rho_h^i \in \Delta(\mathcal{X} \times \mathcal{A})$ is an arbitrary state-action distribution that may depend upon the randomness of rounds $1 \dots i-1$. Suppose Assumptions 2.1 and 2.2 hold. If all $f \in \mathcal{F}_{h+1}$ have $f(x) \in [0, L]$, then for all $\delta \in (0, 1)$, with probability $1 - \delta$, we have for all $\phi \in \Phi$, $f \in \mathcal{F}_{h+1}$, $g \in \mathrm{Lip}$, and $\gamma \in (0, 1)$,*

$$\begin{aligned}&\left|\left(\ell_{\bar{\rho}_h^t}(\phi, g, f) - \ell_{\bar{\rho}_h^t}(\phi^\star, g_f^\star, f)\right) - (\ell_{\mathcal{D}_h}(\phi, g, f) - \ell_{\mathcal{D}_h}(\phi^\star, g^\star, f))\right| \\ &\le \frac{1}{2}\left|\left(\ell_{\bar{\rho}_h^t}(\phi, g, f) - \ell_{\bar{\rho}_h^t}(\phi^\star, g_f^\star, f)\right)\right| + \frac{22L^2(2L/\gamma)^{\dim_{\mathcal{SA}}}\log(2L|\Phi|\cdot N_\infty(\mathcal{F}, \gamma)/(\delta\gamma))}{t} + 6L^2\gamma^2,\end{aligned}$$

*where $g_f^\star := \arg\min_{g\in\mathrm{Lip}} \ell_{\bar{\rho}_h^t}(\phi^\star, g, f)$.*

Before proving this result, we use it to prove Theorem F.1.

**Proof of Theorem F.1.** Let us condition on the event from Lemma F.1. Going forward, we fix $\gamma$ and $h$, and omit the dependence on $h$ and $t$ for compactness. Define as shorthand

$$\varepsilon := \varepsilon_{\text{rep}}(t,\delta) = \frac{22L^2(2L)^{\dim_{\mathcal{SA}}} d_\gamma \log(2L|\Phi| \cdot N_\infty(\mathcal{F},\gamma)/(\delta\gamma))}{t} + 6L^2\gamma^2.$$

For all $f \in \mathcal{F}_h$, denote the solution of the inner optimization from Eq. (3) by $\widetilde{\phi}_f, \widetilde{g}_f$, and let $g^\star_f$ be the predictor of backup of $f$ using ground truth decoder $\phi^\star$, i.e., $g^\star_f := \arg\min_{g \in \text{Lip}} \ell_{\bar{\rho}^t_h}(\phi^\star, g, f)$. Conditioned the event in Lemma F.1, we have

$$\begin{aligned}
\ell_{\mathcal{D}_h}(\phi^\star, g^\star_f, f) - \ell_{\mathcal{D}_h}(\widetilde{\phi}_f, \widetilde{g}_f, f) &\le \frac{1}{2}\ell_{\bar{\rho}^t_h}(\phi^\star, g^\star_f, f) - \frac{1}{2}\ell_{\bar{\rho}^t}(\widetilde{\phi}_f, \widetilde{g}_f, f) + \varepsilon && \text{(Lemma F.1)}\\
&\le \varepsilon. && (\ell_{\bar{\rho}^t_h}(\phi^\star, g^\star_f, f) = 0.)
\end{aligned}$$

Then, letting $g_f$ be the predictor of pseudobackup of $f$ using learned decoder $\phi$, i.e., $g_f := \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi}[f]$, we have

$$\begin{aligned}
\ell_{\mathcal{D}_h}(\phi, g_f, f) - \ell_{\mathcal{D}_h}(\widetilde{\phi}_f, \widetilde{g}_f, f) &= \ell_{\mathcal{D}_h}(\phi, g_f, f) - \ell_{\mathcal{D}_h}(\phi^\star, g^\star_f, f) + \ell_{\mathcal{D}_h}(\phi^\star, g^\star_f, f) - \ell_{\mathcal{D}_h}(\widetilde{\phi}_f, \widetilde{g}_f, f)\\
&\ge \ell_{\mathcal{D}_h}(\phi, g_f, f) - \ell_{\mathcal{D}_h}(\phi^\star, g^\star_f, f) && \text{(Construction of } \widetilde{\phi}_f, \widetilde{g}_f)\\
&\ge \frac{1}{2}\Big(\ell_{\bar{\rho}^t_h}(\phi, g_f, f) - \ell_{\bar{\rho}^t_h}(\phi^\star, g^\star_f, f)\Big) - \varepsilon. && \text{(Lemma F.1)}
\end{aligned}$$

Finally, combining the results above gives

$$\begin{aligned}
\ell_{\bar{\rho}^t_h}(\phi, g_f, f) - \ell_{\bar{\rho}^t_h}(\phi^\star, g^\star_f, f) &\le 2\Big(\ell_{\mathcal{D}_h}(\phi, g_f, f) - \ell_{\mathcal{D}_h}(\widetilde{\phi}_f, \widetilde{g}_f, f)\Big) + 2\varepsilon\\
&\le 2 \max_{k \in \mathcal{F}_{h+1}} \Big(\ell_{\mathcal{D}_h}(\phi, g_k, k) - \ell_{\mathcal{D}_h}(\widetilde{\phi}_f, \widetilde{g}_k, k)\Big) + 2\varepsilon\\
&\le 2 \max_{k \in \mathcal{F}_{h+1}} \Big(\ell_{\mathcal{D}_h}(\phi^\star, g^\star_k, k) - \ell_{\mathcal{D}_h}(\widetilde{\phi}_f, \widetilde{g}_k, k)\Big) + 2\varepsilon\\
&\le 4\varepsilon,
\end{aligned}$$

completing the proof. □

Finally, we provide the proof of Lemma F.1.

**Proof of Lemma F.1.** First let us use the shorthand $g^\star = g^\star_f$. At any round $t$, denote $\mathscr{F}^{t-1}$ as the filtration for the random variable

$$Y^t := \big(g(\phi(x^t, a^t)) - f(x'^t)\big)^2 - \big(g^\star(\phi^\star(x^t, a^t)) - f(x'^t)\big)^2.$$

Then by Lemma D.1, we have that at round $t$, for all $\delta' \in (0,1)$, with probability $1 - \delta'$,

$$\Big|\Big(\ell_{\bar{\rho}^t_h}(\phi, g, f) - \ell_{\bar{\rho}^t_h}(\phi^\star, g^\star, f)\Big) - \big(\ell_{\mathcal{D}_h}(\phi, g, f) - \ell_{\mathcal{D}_h}(\phi^\star, g^\star, f)\big)\Big| \le \sqrt{\frac{2\mathbb{V}(Y^t \mid \mathscr{F}^{t-1}) \log(2/\delta')}{t}} + \frac{16L^2 \log(2/\delta')}{3t}, \tag{22}$$

where $\mathbb{V}\big(Y^t \mid \mathscr{F}^{t-1}\big) := \mathbb{E}\Big[\big(Y^t - \mathbb{E}\big[Y^t \mid \mathscr{F}^{t-1}\big]\big)^2 \mid \mathscr{F}^{t-1}\Big] = \mathbb{E}\Big[\big(Y^t\big)^2 \mid \mathscr{F}^{t-1}\Big]$ denotes the conditional variance of $Y^t$ given $\mathscr{F}^{t-1}$. To bound $\mathbb{V}[Y^t \mid \mathscr{F}^{t-1}]$, we start with:

$$\begin{aligned}
\mathbb{E}\big[Y^t \mid \mathscr{F}^{t-1}\big] &= \mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}\Big[\big(g(\phi(x^t, a^t)) - f(x'^t)\big)^2 - \big(g^\star(\phi^\star(x^t, a^t)) - f(x'^t)\big)^2\Big]\\
&= \mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}\big[\big(g(\phi(x^t, a^t)) + g^\star(\phi^\star(x^t, a^t)) - 2f(x'^t)\big)\big(g(\phi(x^t, a^t)) - g^\star(\phi^\star(x^t, a^t))\big)\big]\\
&= \mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}\Big[\big(g(\phi(x^t, a^t)) - g^\star(\phi^\star(x^t, a^t))\big)^2\Big].\\
&\qquad (\mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}[(g^\star(\phi^\star(x^t, a^t)) - f(x'))(g(\phi(x^t, a^t)) - g^\star(\phi^\star(x^t, a^t)))] = 0.)
\end{aligned}$$

Then we can show that

$$\begin{aligned}
\mathbb{V}[Y^t \mid \mathscr{F}^{t-1}] &= \mathbb{E}\Big[\big(Y^t\big)^2 \mid \mathscr{F}^{t-1}\Big] \\
&= \mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}\Big[\Big(\big(g(\phi(x^t,a^t)) - f(x'^t)\big)^2 - \big(g^\star(\phi^\star(x^t,a^t)) - f(x'^t)\big)^2\Big)^2\Big] \\
&= \mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}\Big[\big(g(\phi(x^t,a^t)) + g^\star(\phi^\star(x^t,a^t)) - 2f(x')\big)^2\big(g(\phi(x^t,a^t)) - g^\star(\phi^\star(x^t,a^t))\big)^2\Big] \\
&\le 16L^2\,\mathbb{E}_{\rho^t|\mathscr{F}^{t-1}}\Big[\big(g(\phi(x^t,a^t)) - g^\star(\phi^\star(x^t,a^t))\big)^2\Big] \\
&= 16L^2\,\mathbb{E}[Y^t \mid \mathscr{F}^{t-1}].
\end{aligned}$$

Plugging back into Eq. (22), we have, with probability $1-\delta'$,

$$\Big|\Big(\ell_{\bar\rho_h^t}(\phi,g,f) - \ell_{\bar\rho_h^t}(\phi^\star,g^\star,f)\Big) - \big(\ell_{\mathcal{D}_h}(\phi,g,f) - \ell_{\mathcal{D}_h}(\phi^\star,g^\star,f)\big)\Big| \le \sqrt{\frac{32L^2\,\mathbb{E}[Y^t]\log(2/\delta')}{t}} + \frac{16L^2\log(2/\delta')}{3t}.$$

Now, let $\widetilde{\mathcal{G}}$ denote any $\gamma$-cover for the Lip function w.r.t., the $\ell_\infty$ norm. Since Lip is bounded by $L$, we have that $\big|\widetilde{\mathcal{G}}\big| = \left(\frac{1}{\gamma}\right)^{\left(\frac{2L}{\gamma}\right)^{\dim_{\mathcal{SA}}}}$ (Wainwright, 2019). We now take a union bound over $\Phi \times \widetilde{\mathcal{G}} \times \widetilde{\mathcal{F}}$, where $\widetilde{\mathcal{F}}$ is an $\ell_\infty$-cover of the discriminator class $\mathcal{F}$, denoted by $\mathcal{N}_\infty(\mathcal{F},\gamma)$. Then by union bound, with probability $1-|\Phi|\big|\widetilde{\mathcal{G}}\big|\big|\widetilde{\mathcal{F}}\big|\delta'$, we have that for all $g \in \mathrm{Lip}$ and $f \in \mathcal{F}$, let $\widetilde{g} \in \widetilde{\mathcal{G}}$ such that $\|g-\widetilde{g}\|_\infty \le \gamma$, $\widetilde{f}$ is defined similarly, and $\widetilde{Y}^t := \Big(\widetilde{g}(\phi(x^t,a^t)) - \widetilde{f}(x'^t)\Big)^2 - (g^\star(\phi^\star(x^t,a^t)) - f(x'^t))^2$ we have

$$\begin{aligned}
&\Big|\Big(\ell_{\bar\rho_h^t}(\phi,g,f) - \ell_{\bar\rho_h^t}(\phi^\star,g^\star,f)\Big) - \big(\ell_{\mathcal{D}_h}(\phi,g,f) - \ell_{\mathcal{D}_h}(\phi^\star,g^\star,f)\big)\Big| \\
\le&\Big|\Big(\ell_{\bar\rho_h^t}(\phi,\widetilde{g},\widetilde{f}) - \ell_{\bar\rho_h^t}(\phi^\star,g^\star,f)\Big) - \Big(\ell_{\mathcal{D}_h}(\phi,\widetilde{g},\widetilde{f}) - \ell_{\mathcal{D}_h}(\phi^\star,g^\star,f)\Big)\Big| + 4L^2\gamma^2 \\
\le&\sqrt{\frac{32L^2\,\mathbb{E}[\widetilde{Y}^t]\log(2/\delta')}{t}} + \frac{16L^2\log(2/\delta')}{3t} + 4L^2\gamma^2 \\
\le&\frac{1}{2}\,\mathbb{E}[\widetilde{Y}^t] + \frac{16L^2\log(2/\delta')}{t} + \frac{16L^2\log(2/\delta')}{3t} + 4L^2\gamma^2 \qquad \text{(AM-GM)} \\
\le&\frac{1}{2}\,\mathbb{E}[Y^t] + \frac{22L^2\log(2/\delta')}{t} + 6L^2\gamma^2 \\
=&\frac{1}{2}\Big(\ell_{\bar\rho_h^t}(\phi,g,f) - \ell_{\bar\rho_h^t}(\phi^\star,g^\star,f)\Big) + \frac{22L^2\log(2/\delta')}{t} + 6L^2\gamma^2.
\end{aligned}$$

Finally, we apply the result above with $\delta' = \frac{\delta}{|\Phi||\widetilde{\mathcal{G}}||\widetilde{\mathcal{F}}|}$ to complete the proof. □

#### F.1.2. PROOF OF THEOREM 4.1

**Proof of Theorem 4.1.** By applying Theorem F.1 with the discriminator class $\mathcal{F}_{h+1} = \mathrm{Lip} \circ \Phi : \mathcal{X} \to [0, L]$, we have that with probability at least $1-\delta$,

$$\max_{f \in \mathcal{F}_{h+1}} \mathbb{E}_{x,a\sim\bar\rho_h^t}\Big[\Big(\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h^t}[f](x,a) - \mathcal{P}_h[f](x,a)\Big)^2\Big] \le \varepsilon_{\mathrm{rep}}(t,\delta),$$

where

$$\varepsilon_{\mathrm{rep}}(t,\delta) = \frac{88L^2(2L)^{\dim_{\mathcal{SA}}}\,d_\gamma\log(2L|\Phi|\cdot N_\infty(\mathcal{F}_{h+1},\gamma)/(\delta\gamma))}{t} + 24L^2\gamma^2.$$

By a similar calculation to Lemma F.14, we have that the covering number of $\mathcal{F}_{h+1}$ is

$$N_\infty(\mathcal{F}_{h+1},\gamma) \le \left(\frac{1}{\gamma}\right)^{\left(\frac{2L}{\gamma}\right)^{\dim_{\mathcal{SA}}}} \cdot |\Phi|,$$

so that

$$\varepsilon_{\mathrm{rep}}(t,\delta)=\widetilde{\mathcal{O}}\left(\frac{L^{2\dim_{\mathcal{SA}}}\left(\frac{1}{\gamma}\right)^{2\dim_{\mathcal{SA}}}\log\left(2L|\Phi|\cdot\left(\frac{1}{\gamma}\right)/(\delta\gamma)\right)}{t}+L^2\gamma^2\right).$$

Taking $\gamma=t^{\frac{1}{2\dim_{\mathcal{SA}}+2}}$ completes the proof. □

### F.2. Formal Version of Theorem 4.2

The following theorem, proved in Appendix F.3, is the formal version of Theorem 4.2.

**Theorem 4.2′** (Guarantee for CRIEE; formal version of Theorem 4.2)**.** *With probability at least* $1-\delta$*, setting parameters*

$$\lambda^t=\Theta\left(t^{\frac{\dim_{\mathcal{SA}}}{\bar{d}+2}}\log\left(\frac{t|\Phi|}{\delta}\right)\right),\quad \widehat{\alpha}^t=\Theta\left(t^{\frac{\bar{d}}{\tilde{d}}}\log\left(\frac{t|\Phi|}{\delta}\right)\right),$$

*let* $\widehat{\pi}$ *be the output of the CRIEE, we have*

$$J(\pi^\star)-J(\widehat{\pi})\le\varepsilon,$$

*with a total number of samples at most*

$$H\cdot T=\mathcal{O}\left(\frac{H^{2\widetilde{d}+3}\log(TH|\Phi|/\delta\varepsilon)}{\varepsilon^{\widetilde{d}+1}}\right),$$

*where* $\widetilde{d}=3\dim_{\mathcal{SA}}^2+4\dim_{\mathcal{SA}}\dim_{\mathcal{A}}+5\dim_{\mathcal{SA}}+4\dim_{\mathcal{A}}+1$*,* $\bar{d}=1.5\dim_{\mathcal{SA}}^2+2\dim_{\mathcal{SA}}\dim_{\mathcal{A}}+\dim_{\mathcal{SA}}+\dim_{\mathcal{A}}$.

### F.3. Proof of Theorem 4.2′

We begin by introducing some preliminary notation and giving an overview of the proof structure.

#### F.3.1. Preliminaries and Proof Organization

**Additional notation for pseudobackups.**

Recall that in Eq. (4), given a function class $\mathcal{V}\subset(\mathcal{X}\times\mathcal{A})\to[0,L]$, we defined the pseudobackup operator as

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{V}}:f\mapsto\arg\min_{v\in\mathcal{V}}\widehat{\mathbb{E}}_{\mathcal{D}_h}\left[\left(v(x_h,a_h)-f(x_{h+1})\right)^2\right],$$

and we specialized to $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}$ where $\mathcal{V}:=\mathrm{Lip}\circ\phi_h$ for some $\phi_h\in\Phi$. We call this pseudobackup operator the *continuous pseudobackup* as we will distinguish it from the other pseudobackup operators involving discretization.

Before we describe the other pseudobackup operators, we first introduce notation for the discretized version of the decoder. Given a decoder $\phi_h$, we let $\mathsf{disc}_\eta[\phi](x,a):=\mathsf{disc}_\eta[\phi(x,a)]$, and one should interpret $\mathsf{disc}_\eta[\phi_h](\cdot)$ as a one-hot vector in $\mathbb{R}^{d_\eta}$ (recall that $d_\eta=\left(\frac{1}{\eta}\right)^{\dim_{\mathcal{SA}}}$). With this we can define the function class

$$\mathcal{W}=\left\{w^\top\mathsf{disc}_\eta[\phi_h]\mid w\in\mathbb{R}^{d_\eta},\|w\|_\infty\le L\right\},$$

the linear functions over the discretized decoder. We call the pseudobackup operator induced by this class as *linear pseudobackup*, denoted as $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}$. Following the definition, the linear pseudobackup is given by

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}[f]=w_f^\top\mathsf{disc}_\eta[\phi_h],\quad\text{where}$$
$$w_f=\arg\min_{w\in\mathbb{R}^{d_\eta},\|w\|_\infty\le L}\mathbb{E}_{x,a,x'\sim\mathcal{D}_h}(w^\top\mathsf{disc}_\eta[\phi_h](x,a)-f(x'))^2,$$

Now since $\mathsf{disc}_\eta[\phi_h](\cdot)$ is a one-hot vector, and the target function $f$ is bounded, we obtain a closed-form solution for the linear pseudobackup:

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}[f](x,a) = \frac{\sum_{(x,a,x')\in\mathcal{D}_h} \mathbb{1}\{x,a \in \mathsf{ball}_\eta[\phi_h](x,a)\} f(x')}{\sum_{(x,a)\in\mathcal{D}_h} \mathbb{1}\{x,a \in \mathsf{ball}_\eta[\phi_h](x,a)\}},$$

with the convention that $\frac{0}{0} = 0$. Note that the linear pseudobackup $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}[\cdot]$ is equivalent to the value iteration backup operator in Algorithm 2.

Finally, we will define the *discretized pseudobackup* operator $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathsf{disc}_\eta[\phi_h]}[f]$. The intuition is that the discretized pseudobackup bridges between the continuous pseudobackup and the linear pseudobackup, which is using a linear function over the discretized decoder to predict the value of the continuous pseudobackup. Formally, we will deviate from the definition Eq. (4) and define

$$\begin{aligned}
&\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathsf{disc}_\eta[\phi_h]}[f] = w^\top \mathsf{disc}_\eta[\phi_h], \quad \text{where} \\
&w = \underset{w\in\mathbb{R}^{d_\eta}, \|w\|_\infty \le L}{\arg\min} \quad \mathbb{E}_{x,a\sim\mathcal{D}_h}(w^\top \mathsf{disc}_\eta[\phi_h](x,a) - \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f](x,a))^2.
\end{aligned}$$

and similarly as above, for each $x, a$, we have

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathsf{disc}_\eta[\phi_h]}[f](x,a) := \frac{\sum_{(\widetilde{x},\widetilde{a})\in\mathcal{D}_h} \mathbb{1}\{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi_h](x,a)\} \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}[f](\widetilde{x},\widetilde{a})}{\sum_{(\widetilde{x},\widetilde{a})\in\mathcal{D}} \mathbb{1}\{\widetilde{x},\widetilde{a} \in \mathsf{ball}_\eta[\phi_h](x,a)\}}, \tag{23}$$

and now we can see that the discretized pseudobackup is serving as a discretized version of the continuous pseudobackup, capturing the value of the continuous pseudobackup averaged over the ball defined by the same decoder that defines the pseudobackups.

Finally, we will often nest $h$ applications of the pseudobackup operator, with actions selected according to a given policy $\pi$. To do this, we use the notation

$$\left[\widetilde{\mathcal{P}}^{\pi}_{\rho,\mathcal{H}}\right]^{\otimes h}[g] := f_1(x,\pi(x)), \text{ where } f_h := \widetilde{\mathcal{P}}_{\rho_h,\mathcal{H}_h}[g], \quad \text{and} \quad f_i = \widetilde{\mathcal{P}}_{\rho_i,\mathcal{H}_i}[f_{i+1}(\cdot,\pi(\cdot))], \ \ \forall i \in [h-1].$$

Similarly, we will define nested reward-free Bellman backup

$$[\mathcal{P}]^{\otimes h}[g] := f_1(x,\pi(x)), \text{ where } f_h := \mathcal{P}_h[g], \quad \text{and} \quad f_i = \mathcal{P}_i[f_{i+1}(\cdot,\pi(\cdot))], \ \ \forall i \in [h-1].$$

**Construction of the discriminator class.** We now specify the discriminator class $\mathcal{F}_h$ used in CRIEE. Recalling that $\eta > 0$ is the discretization parameter for CRIEE, we define the discriminator class to be the union of the following two function classes:

$$\mathcal{F}^1_{\eta,h+1} = \left\{ f(x) : \mathbb{E}_{a\sim\pi^{\mathsf{unif}}_\eta}\left[w^\top \mathsf{disc}_\eta[\phi](x,a) - g\big(\widetilde{\phi}(x,a)\big)\right] \mid \phi,\widetilde{\phi} \in \Phi_{h+1}, g \in \mathrm{Lip}, w \in \mathbb{R}^{d_\eta}, \|w\|_\infty \le 1 \right\}, \tag{24}$$

$$\mathcal{F}^2_{\eta,h+1} =$$

$$\left\{ f(x) : \max_a \left( \frac{R_{h+1}(x,a) + \min\{w^\top \mathsf{disc}_\eta[\phi](x,a), 2\}}{2H+1} + \widetilde{w}^\top \mathsf{disc}_\eta[\phi](x,a) \right) \mid \phi \in \Phi_{h+1}, w, \widetilde{w} \in \mathbb{R}^{d_\eta}, \|w\|_\infty \le c, \|\widetilde{w}\|_\infty \le 2 \right\}.$$

We then set $\mathcal{F}_{h+1} = \mathcal{F}^1_{\eta,h+1} \cup \mathcal{F}^2_{\eta,h+1}$, leaving the dependence on $\eta$ implicit.

Let us give some brief intuition behind the construction of the discriminator class. The first class $\mathcal{F}^1_{\eta,h+1}$ is the class of functions that can represent the following: given a state $x_{h+1}$, what is the expected error between the linear pseudobackup $\widetilde{\mathcal{P}}_{\mathcal{D},\mathcal{W}}$ (the first term in the expectation) and the Bellman backup $\mathcal{P}$ (the second term in the expectation), under an action $a_{h+1}$ sampled uniformly via $\pi^{\mathsf{unif}}_\eta$? We will show later how to use this construction to transfer the error under the linear pseudobackup to the continuous pseudobackup $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi_h}$. The second discriminator class $\mathcal{F}^2_{\eta,h+1}$, is more direct, and aims to represent the optimistic value functions induced by Algorithm 3 (weighted by $\frac{1}{2H+1}$).

### F.3.2. Proof Organization

We organize the proof of Theorem 4.2′ into the following modules:

- In Appendix F.3.3, we show how to decompose the pseudoregret for Algorithm 2 into an error term defined as a difference between the linear pseudobackups defined in the prequel and the true Bellman backups—conditioned on establishing optimism, a property we will return to later. As in prior work (Uehara et al., 2022; Zhang et al., 2022), the challenge in proceeding from here is that we need to control on-policy error, yet we are only guaranteed control over error under the data collection distribution (**Challenge 1**). In addition, we have a mismatch between the linear pseudobackup and the continuous pseudobackup; the latter of which is close to the true Bellman backup with respect to the data collection distributions (**Challenge 2**).
- In Appendix F.3.4, we address **Challenge 2** by showing show that the linear pseudobackup is close to the continuous pseudobackup *under the data collection distribution*.
- In Appendix F.3.5 we address **Challenge 1** by adapting the "one-step-back" trick used in Agarwal et al. (2020b); Uehara et al. (2022); Zhang et al. (2022), but with a careful treatment to account for misspecification arising from discretization.
- Next, in Appendix F.3.6, after combining the results above with an in-distribution guarantee for the continuous pseudobackup from Theorem 4.1, we establish optimism.
- Finally, in Appendix F.3.7, we combine the results above to prove Theorem 4.2′.

Supporting technical lemmas are deferred to Appendix F.3.8.

### F.3.3. Regret Decomposition by Simulation Lemma

We start with a simulation lemma for the pseudobackups.

**Lemma F.2** (Simulation Lemma for pseudobackups)**.** *Fixed a pseudobackup $\widetilde{\mathcal{P}}_h := \widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{V}}$ defined in Eq. (4) for any dataset $\mathcal{D}_h$, $\mathcal{V} \subset (\mathcal{X} \times \mathcal{A}) \to [0, L]$, and reward function $\widetilde{R}_h$, for any policy $\pi$, let $f_h$ be its estimated Q functions through the pseudobackup $\widetilde{\mathcal{P}}_h$, i.e., for all $h \in [H]$, $f_h = \widetilde{R}_h + \widetilde{\mathcal{P}}^\pi_h[f_{h+1}]$, and $f_{H+1} = 0$. Denote the value function induced by $f$ and $\pi$ as $f^\pi_h(x) = f_h(x, \pi(x))$. Let $Q^\pi$ and $V^\pi$ be the true Q function and value function induced by $\pi$ and reward function $R$. Then for any $x_1 \in \mathcal{X}$,*

$$f^\pi_1(x_1) - V^\pi_1(x_1) = \sum_{h=1}^{H} \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} [\delta_h](x_1, \pi_1(x_1)),$$

*where*

$$\delta_h = \widetilde{R}_h - R_h + (\widetilde{\mathcal{P}}^\pi_h - \mathcal{P}^\pi_h)[Q^\pi_{h+1}] + (\widetilde{\mathcal{P}}^\pi_h - \mathcal{P}^\pi_h)[f_{h+1}] + (\widetilde{\mathcal{P}}^\pi_h - \mathcal{P}^\pi_h)[(f_{h+1} - Q^\pi_{h+1})] + (\mathcal{P}^\pi_h - \widetilde{\mathcal{P}}^\pi_h)[Q^\pi_{h+1}].$$

*When the pseudobackup operator is linear, i.e., $\widetilde{\mathcal{P}}[f + f'] = \widetilde{\mathcal{P}}[f] + \widetilde{\mathcal{P}}[f']$, we have*

$$f^\pi_1(x_1) - V^\pi_1(x_1) = \sum_{h=1}^{H} \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} [\widetilde{R}_h - R_h + (\widetilde{\mathcal{P}}^\pi_h - \mathcal{P}^\pi_h)[Q^\pi_{h+1}]](x_1, \pi(x_1)).$$

**Proof of Lemma F.2.** By construction, we have

$$\begin{aligned}
&f^\pi_1(x_1) - V^\pi_1(x_1) \\
&= \widetilde{R}_1(x_1, \pi(x_1)) + \widetilde{\mathcal{P}}_1[f^\pi_2](x_1, \pi(x_1)) - R_1(x_1, \pi_1(x_1)) - \mathcal{P}_1[V^\pi_2](x_1, \pi(x_1)) \\
&= (\widetilde{R}_1 - R_1)(x_1, \pi(x_1)) + \widetilde{\mathcal{P}}_1[f^\pi_2](x_1, \pi(x_1)) - \widetilde{\mathcal{P}}_1[V^\pi_2](x_1, \pi(x_1)) + \widetilde{\mathcal{P}}_1[V^\pi_2](x_1, \pi(x_1)) - \mathcal{P}_1[V^\pi_2](x_1, \pi(x_1)) \\
&= (\widetilde{R}_1 - R_1)(x_1, \pi(x_1)) + (\widetilde{\mathcal{P}}_1 - \mathcal{P}_1)[V^\pi_2](x_1, \pi(x_1)) + \widetilde{\mathcal{P}}_1[f^\pi_2](x_1, \pi(x_1)) - \widetilde{\mathcal{P}}_1[V^\pi_2](x_1, \pi(x_1)) \\
&= (\widetilde{R}_1 - R_1)(x_1, \pi(x_1)) + (\widetilde{\mathcal{P}}_1 - \mathcal{P}_1)[V^\pi_2](x_1, \pi(x_1)) + (\widetilde{\mathcal{P}}_1 - \mathcal{P}_1)[f^\pi_2](x_1, \pi(x_1)) + (\widetilde{\mathcal{P}}_1 - \mathcal{P}_1)[(f^\pi_2 - V^\pi_2)](x_1, \pi(x_1)) + \\
&\quad (\mathcal{P}_1 - \widetilde{\mathcal{P}}_1)[V^\pi_2](x_1, \pi(x_1)) + \underbrace{\widetilde{\mathcal{P}}_1[(f^\pi_2 - V^\pi_2)](x_1, \pi(x_1))}_{\text{recursion}}.
\end{aligned}$$

Then let

$$\delta_h := \widetilde{R}_h - R_h + (\widetilde{\mathcal{P}}_h - \mathcal{P}_h)[V^\pi_{h+1}] + (\widetilde{\mathcal{P}}_h - \mathcal{P}_h)[f^\pi_{h+1}] + (\widetilde{\mathcal{P}}_h - \mathcal{P}_h)[(f^\pi_{h+1} - V^\pi_{h+1})] + (\mathcal{P}^\pi_h - \widetilde{\mathcal{P}}^\pi_h)[V^\pi_{h+1}],$$

we get

$$f^\pi - V^\pi = \sum_{h=1}^{H} \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} [\delta_h](x_1, \pi(x_1)).$$

Note that when $\widetilde{\mathcal{P}}$ is linear, i.e., $\widetilde{\mathcal{P}}[f_1 + f_2] = \widetilde{\mathcal{P}}[f_1] + \widetilde{\mathcal{P}}[f_2]$, we have

$$(\widetilde{\mathcal{P}}_h - \mathcal{P}_h)[f^\pi_{h+1}] + (\widetilde{\mathcal{P}}_h - \mathcal{P}_h)[(f^\pi_{h+1} - V^\pi_{h+1})] + (\mathcal{P}^\pi_h - \widetilde{\mathcal{P}}^\pi_h)[V^\pi_{h+1}] = 0,$$

and thus

$$f^\pi - V^\pi = \sum_{h=1}^{H} \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} [\widetilde{R}_h - R_h + (\widetilde{\mathcal{P}}_h - \mathcal{P}_h)[V^\pi_{h+1}]](x_1, \pi(x_1)),$$

finally, note that $\mathcal{P}_h[V^\pi_{h+1}] = \mathcal{P}^\pi_h[Q^\pi_{h+1}]$ and thus we complete the proof. □

**Lemma F.3** (Simulation Lemma for Bellman backups)**.** *Under the similar setup as Lemma F.2, we have*

$$f^\pi_1(x_1) - V^\pi_1(x_1) = \sum_{h=1}^{H} \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} [\widetilde{R}_h - R_h + (\widetilde{\mathcal{P}}^\pi_h - \mathcal{P}^\pi_h)[f_{h+1}]](x_1, \pi(x_1)).$$

The above result is the classic simulation lemma rewritten in the reward-free Bellman backup notation so we omit the proof here (e.g., see (Sun et al., 2019)).

### F.3.4. Closeness between Pseudobackups

In this section we solve one major difficulty of our analysis: to show that the following two pseudobackups are close to each other: 1) $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi^t_h}$, which we have the representation learning guarantee, and 2) $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}$, which we use for planning and exploration. First we can see that for any $f \in \mathcal{F}_{h+1}$, due to the fact that the predictor class used to define $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi^t_h}$ is 1-Lipschitz, we have

$$\left\|\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathsf{disc}_\eta[\phi^t_h]}[f] - \widetilde{\mathcal{P}}_{\mathcal{D}_h,\phi^t_h}[f]\right\|_\infty \leq \eta.$$

Then the remaining part is to show that $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathsf{disc}_\eta[\phi^t_h]}$ and $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}$ are close. For this part, we only need to show that $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathsf{disc}_\eta[\phi^t_h]}$ and $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}$ are close under the training distribution. We first introduce some notations that we use in this section: for any function $f \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$, and $P$ which is a probability measure over $\mathcal{X}$, we denote $\|f\|^2_{L_2(P)} = \mathbb{E}_{x\sim P}\left[f^2(x)\right]$. With this we are ready to state the closeness result:

**Lemma F.4** (Closeness between pseudobackups)**.** *For any round $t$, for all $h \in [H]$ and $f \in \mathcal{F}_{h+1}$, let $\rho^t_h$ be the data generating distribution of $\mathcal{D}^t_h$, we have*

$$\|\widetilde{\mathcal{P}}_{\mathcal{D}^t_h,\mathcal{W}}[f] - \widetilde{\mathcal{P}}_{\mathcal{D}^t_h,\mathsf{disc}_\eta(\phi^t_h)}[f]\|_{L_2(\rho^t_h)} \leq \varepsilon_{\mathrm{rep}}(t) + 2\varepsilon_{\mathrm{hist}}(t).$$

*And thus*

$$\|\widetilde{\mathcal{P}}_{\mathcal{D}^t_h,\mathcal{W}}[f] - \widetilde{\mathcal{P}}_{\mathcal{D}^t_h,\phi^t_h}[f]\|_{L_2(\rho^t_h)} \leq \varepsilon_{\mathrm{rep}}(t) + 2\varepsilon_{\mathrm{hist}}(t) + 2\eta.$$

**Proof of Lemma F.4.** The proof is based on the following observation that both the linear pseudobackup and the discretized pseudobackup are histograms. Given a distribution $\rho$ over $\mathcal{X}$, let $\mathcal{D}^n$ denote a dataset with $n$ samples drawn i.i.d. from $\rho$. Given any function $f \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$, the population histogram with bins $\mathcal{B} = \{b_i\}_{i=1}^B$ of $\Gamma(f)$ is defined as

$$\Gamma(f)(x) = \sum_{i=1}^{B} \mathbb{1}\{x \in b_i\} \int_{b_i} f(x')\rho(x'),$$

and the empirical histogram of $\Gamma_n(f)$ is defined as

$$\Gamma_n(f)(x) = \sum_{i=1}^{B} \mathbb{1}\{x \in b_i\} \frac{1}{n_i} \sum_{j=1}^{n} \mathbb{1}\{x_j \in b_i\} f(x_j).$$

where $n_i$ is the number of data in bin $b_i$.

Then we observe that, for any $t, h$, denote the histogram according to the discretized feature $\mathsf{disc}_\eta(\phi_h^t)$ and the distribution $\rho_h^t$ as $\Gamma_{h,t}$, and the empirical histogram according to $\mathcal{D}_h^t$ as $\widehat{\Gamma}_{h,t}$, then we have the following identities: for any $f \in \mathcal{F}_h$,

$$\widetilde{\mathcal{P}}_{\mathcal{D},\mathcal{W}}[f](x,a) = \widehat{\Gamma}_{h,t}(\mathbb{E}[f(x_{h+1}) \mid x_h = x, a_h = a]) \quad \text{and} \quad \widetilde{\mathcal{P}}_{\mathcal{D},\mathsf{disc}_\eta[\phi]}[f](x,a) = \widehat{\Gamma}_{h,t}\Big(\widetilde{\mathcal{P}}_{\mathcal{D},\phi_h^t}[f](x,a)\Big). \tag{25}$$

In Lemma F.7, we prove that, if the target functions of two empirical histograms (with the same set of bins) are close to each other, then the two empirical histograms are close to each other in distribution as well. In our case, the target functions $\mathcal{P}[f]$ and $\widetilde{\mathcal{P}}_{\mathcal{D},\phi_h^t}[f]$ are indeed close to each other in distribution (by Theorem 4.1), then plugging in the guarantee of Theorem 4.1 into Lemma F.7 we complete the proof. □

Now we state and prove the result that histograms approximately preserve closeness of the target functions:

**Lemma F.5** (Empirical histograms approximately preserve closeness). *Using notations from Lemma F.4, suppose we have functions $f, g \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$ are close: $\|f - g\|_{L_2(\rho)} \le \varepsilon$. Then we have*

$$\|\Gamma_n(f) - \Gamma_n(g)\|_{L_2(\rho)} \le \varepsilon + 2\varepsilon_{\text{hist}}(n).$$

**Proof of Lemma F.5.** Since $\|\cdot\|_{L_2(P)}$ is a metric, by triangle inequality we have

$$\|\Gamma_n(f) - \Gamma_n(g)\|_{L_2(\rho)} \le \|\Gamma_n(f) - \Gamma(f)\|_{L_2(\rho)} + \|\Gamma(f) - \Gamma(g)\|_{L_2(\rho)} + \|\Gamma_n(g) - \Gamma(g)\|_{L_2(\rho)}.$$

Now the first and third terms are the differences between an empirical histogram and the population one with the same target, and we prove the difference is small by Lemma F.7 with standard concentration result. For the second term, since histogram is a convex projection, then by Lemma F.6 we have that convex projection preserves closeness, and by the assumption that $f, g$ are close we complete the proof. □

**Lemma F.6** (Convex projections preserve closeness). *Let $f, g \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$ be two functions, and $P$ be a probability measure over $\mathcal{X}$. Suppose that $f$ and $g$ are close:*

$$\|f - g\|_{L_2(P)}^2 = \mathbb{E}_{x \sim P}\big[(f(x) - g(x))^2\big] \le \varepsilon.$$

*Then suppose we have a convex projection $\Gamma : \mathcal{F} \to \mathcal{H}$, where $\mathcal{H}$ is a convex function class, i.e.,*

$$\Gamma(f) = \arg\min_{h \in \mathcal{H}} \|h - f\|_{L_2(P)},$$

*then we have*

$$\|\Gamma(f) - \Gamma(g)\|_{L_2(P)}^2 \le \|f - g\|_{L_2(P)}^2 \le \varepsilon.$$

**Proof of Lemma F.6.** Let us define the notation of inner product in the function space under $P$:

$$\langle f, g \rangle_P = \mathbb{E}_{x \sim P}[f(x) \cdot g(x)].$$

Then by the convexity of $\mathcal{H}$, we have

$$\langle \Gamma(f) - \Gamma(g), \Gamma(f) - f \rangle_P \le 0 \quad \text{and} \quad \langle \Gamma(g) - \Gamma(f), \Gamma(g) - g \rangle_P \le 0.$$

Expanding the inner product and summing them together and rearranging the terms, we have

$$\|\Gamma(f) - \Gamma(g)\|_{L_2(P)}^2 \le \langle \Gamma(g) - \Gamma(f), f - g \rangle_P.$$

Finally by AM-GM we complete the proof. □

**Lemma F.7** (Concentration of histograms)**.** *Let $\rho$ be a distribution over $\mathcal{X}$, where $\rho = \frac{1}{n}\sum_{i=1}^{n}\rho^i$, and each $\rho^i$ may depend on the randomness in previous rounds. Let $\mathcal{D}^n$ be a dataset with $n$ samples drawn i.i.d. from $\rho$. Then for any $f \in \mathcal{F} : \mathcal{X} \to \mathbb{R}, \|f\|_\infty \leq L$, for histogram $\Gamma$ with $B$ bins, with probability at least $1-\delta$,*

$$\|\Gamma(f) - \Gamma_n(f)\|^2_{L_2(\rho)} \leq \varepsilon_{\text{hist}}(n) = \mathcal{O}\left(\frac{BL^2\log(|\mathcal{F}|/\delta)}{n}\right).$$

**Proof of Lemma F.7.** Let us fix $f \in \mathcal{F}$, and define $\rho_{\mathcal{B}}$ as the distribution of each bin under $\rho$. Then we have

$$\begin{aligned}\|\Gamma(f) - \Gamma_n(f)\|^2_{L_2(\rho)} &= \mathbb{E}_{x\sim\rho}\left[(\Gamma(f)(x) - \Gamma_n(f)(x))^2\right] \\ &= \mathbb{E}_{b\sim\rho_{\mathcal{B}}}\left[\mathbb{E}\left[(\Gamma(f)(x) - \Gamma_n(f)(x))^2 \mid b\right]\right] \\ &\leq \mathbb{E}_{b\sim\rho_{\mathcal{B}}}\frac{L^2\log(1/\delta)}{n(b)} \\ &\leq \frac{BL^2\log(1/\delta)}{n},\end{aligned}$$

where the first inequality is by standard concentration argument from observing that conditioned on bin $b$, the empirical histogram $\Gamma_n(f)$ converges to expected histogram $\Gamma(f)$ in bin $b$. Note that $b$ here is a random variable, and $n(b)$ denotes the number of data from bin $b$ in the dataset $\mathcal{D}^n$. Finally taking a union bound over $\mathcal{F}$ we complete the proof. □

### F.3.5. Error Transfer by the One-step-back Trick

One difficulty we mentioned in the main text is that, in general, the pseudobackup operators do not preserve the order of functions that they take on. Without this property, it is hard to bound the representation error under the induced policies by transferring to the representation error under the data collection distribution, in order to prove optimism. Luckily, we can show that the linear pseudobackup operator preserves the order of functions, but in general it is not clear if the continuous pseudobackup preserves the order of functions. Combined with the result in the last section that the linear pseudobackup is close to the continuous pseudobackup in-distribution, we can use the one-step-back trick to leverage the representation learning results. We show that the linear pseudobackup preserves the order of functions in Lemma F.10, which gives the following distribution shift result. We instantiate $\tilde{\mathcal{P}}$ to be the linear pseudobackup, but the result holds for any pseudobackup that is piecewise constant with respect to $\mathsf{disc}_\eta[\phi_h^t]$ and monotone:

**Lemma F.8** (One-step-back for linear pseudobackup)**.** *Let $\tilde{P} := \widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}$. Conditioned on the event that Theorem 4.1 holds for testing distribution $\rho_h$ with error $\varepsilon_{\text{rep}}$, and Lemma F.4 holds with error $\varepsilon_{\text{hist}}$. Then for any set of functions $\{f_h\}_{h=1}^H$ where $f_h \in (\mathcal{X}\times\mathcal{A} \to [-L, L])$, and $f_h(\cdot, \pi_\eta^{\mathsf{unif}}) \in \mathcal{F}_h$ for all $h \in [H]$, for any policy $\pi$,*

$$\sum_{h=1}^{H}\left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)}[f_h](x_1, \pi(x_1)) \leq$$

$$\sum_{h=2}^{H}\left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)}\min\left\{\sqrt{\frac{1}{t\bar{\rho}_h^t[\mathsf{ball}_\eta[\phi_h^t](\cdot)] + \lambda^t}}\sqrt{2tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\gamma}_h^t}\left[f_{h+1}^2(x,a)\right] + \zeta(t)}, L\right\} + \sqrt{A\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[f_1^2(x,a)]} + 4H\eta,$$

*where*

$$\zeta(t) = 4\varepsilon_{\text{hist}}(t) + 4A_\eta^2\varepsilon_{\text{rep}}(t) + 18\lambda^t L^2 d_\eta.$$

**Proof of Lemma F.8.** For $h = 1$, we have:

$$\begin{aligned}g_1(x_1, \pi(x_1)) &= \mathbb{E}_{a\sim\pi(x_1)}[g_1(x_1, a)] \\ &\leq \sqrt{\max_{a\in\mathcal{A}_\eta}\frac{\pi(a \mid x_1)}{\pi_\eta^{\mathsf{unif}}(a \mid x_1)}\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[g_1^2(x,a)]} \qquad \text{(Jensen)} \\ &\leq \sqrt{A_\eta\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[g_1^2(x,a)]}.\end{aligned}$$

For $h = 2, \ldots, H$, we have

$$
\begin{aligned}
&\left[\widetilde{\mathcal{P}}\right]^{\otimes h}[f_{h+1}] \\
=&\left[\widetilde{\mathcal{P}}\right]^{\otimes h-1}\left[\widetilde{\mathcal{P}}_h^{\pi}[f_{h+1}]\right] \\
=&\left[\widetilde{\mathcal{P}}\right]^{\otimes h-1}\left[\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h]} \mathbb{1}\{\phi_h(\cdot) \in b_\eta\}\widetilde{\mathcal{P}}_h^{\pi}[f_{h+1}](x_\eta, a_\eta)\right] \\
\leq&\left[\widetilde{\mathcal{P}}\right]^{\otimes h-1}\left[\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h]} \min\left\{\frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}}\sqrt{(t\bar{\rho}_h^t(b_\eta) + \lambda^t)\Big(\widetilde{\mathcal{P}}_h^{\pi}[f_{h+1}](x_\eta, a_\eta)\Big)^2}, L\right\}\right],
\end{aligned}
$$

where each $x_\eta, a_\eta \in {\phi_h^t}^{-1}(s_\eta, a_\eta)$, where $s_\eta, a_\eta$ is the covering point in the latent space for the ball $b_\eta$, and the second equality is due to the fact that $\widetilde{\mathcal{P}}_h$ is piecewise constant with respect to $\mathsf{disc}_\eta[\phi_h^t]$.

Focusing on the function inside the pseudobackup, we have the following pointwise inequality:

$$
\begin{aligned}
&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}}\sqrt{(t\bar{\rho}_h^t(b_\eta) + \lambda^t)\Big(\widetilde{\mathcal{P}}_h^{\pi}[f_{h+1}](x_\eta, a_\eta)\Big)^2} \\
\leq&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}}\sqrt{\int_{b_\eta}(t\bar{\rho}_h^t(x,a) + \lambda^t)\Big(\widetilde{\mathcal{P}}_h^{\pi}[f_{h+1}](x,a)\Big)^2 \mathrm{d}\nu(x,a)} \qquad \text{(piecewise constant)} \\
\leq&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}}\sqrt{\int_{b_\eta}(t\bar{\rho}_h^t(x,a) + \lambda^t)\Big(\Big(\widetilde{\mathcal{P}}_h^{\pi} - \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h^t]}^{\pi}\Big)[f_{h+1}](x,a) + \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta}^{\pi}[f_{h+1}](x,a)\Big)^2 \mathrm{d}\nu(x,a)} \\
\leq&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \cdot \\
&\qquad\sqrt{\int_{b_\eta}(t\bar{\rho}_h^t(x,a) + \lambda^t)\Big(\Big(\widetilde{\mathcal{P}}_h^{\pi} - \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h^t]}^{\pi}\Big)[f_{h+1}](x,a) + \widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h^t}^{\pi}[f_{h+1}](x,a) + 2\eta\Big)^2 \mathrm{d}\nu(x,a)} \\
\leq&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \\
&\qquad\sqrt{2\int_{b_\eta}(t\bar{\rho}_h^t(x,a) + \lambda^t)\Big(\Big(\widetilde{\mathcal{P}}_h^{\pi} - \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h^t]}^{\pi}\Big)[f_{h+1}](x,a)\Big)^2 + \Big(\widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h^t}^{\pi}[f_{h+1}](x,a)\Big)^2 \mathrm{d}\nu(x,a)} + 4\eta + \lambda^t L d_\eta \\
\leq&\sqrt{\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot) \in b_\eta\}}{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \cdot \\
&\sqrt{2\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]}\int_{b_\eta}(t\bar{\rho}_h^t(x,a) + \lambda^t)\Big(\Big(\widetilde{\mathcal{P}}_h^{\pi} - \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h^t]}^{\pi}\Big)[f_{h+1}](x,a)\Big)^2 + \Big(\widetilde{\mathcal{P}}_{\mathcal{D}_h, \phi_h^t}^{\pi}[f_{h+1}](x,a)\Big)^2 \mathrm{d}\nu(x,a)} \\
&\qquad + 4\eta + \lambda^t L d_\eta.
\end{aligned}
$$

Then we focus on the terms inside the second square root:

$$\begin{aligned}
&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^t]} \int_{b_\eta} (t\bar{\rho}_h^t(x,a) + \lambda^t)\left(\left(\widetilde{\mathcal{P}}_h^\pi - \widetilde{\mathcal{P}}^\pi_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h^t]}\right)[f_{h+1}](x,a) + \widetilde{\mathcal{P}}^\pi_{\mathcal{D}_h,\phi_h^t}[f_{h+1}](x,a)\right)^2 \mathrm{d}\nu(x,a) \\
\leq & t\,\mathbb{E}_{x,a\sim\bar{\rho}_h^t}\left[\left(\left(\widetilde{\mathcal{P}}_h^\pi - \widetilde{\mathcal{P}}^\pi_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h^t]}\right)[f_{h+1}](x,a)\right)^2 + \left(\widetilde{\mathcal{P}}^\pi_{\mathcal{D}_h,\phi_h^t}[f_{h+1}](x,a)\right)^2\right] + 9L^2\lambda^t d_\eta \\
\leq & t\varepsilon_{\mathrm{rep}}(t) + 2t\varepsilon_{\mathrm{hist}}(t) + t\,\mathbb{E}_{x,a\sim\bar{\rho}_h^t}\left[\left(\widetilde{\mathcal{P}}^\pi_{\mathcal{D}_h,\phi_h^t}[f_{h+1}](x,a)\right)^2\right] + 9L^2\lambda^t d_\eta && \text{(Lemma F.4)} \\
\leq & t\varepsilon_{\mathrm{rep}}(t) + 2t\varepsilon_{\mathrm{hist}}(t) + tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\rho}_h^t}\left[\left(\widetilde{\mathcal{P}}^{\pi_\eta^{\mathsf{unif}}}_{\mathcal{D}_h,\phi_h^t}[f_{h+1}](x,a)\right)^2\right] + 9L^2\lambda^t d_\eta && \text{(Importance sampling)} \\
\leq & t\varepsilon_{\mathrm{rep}}(t) + 2t\varepsilon_{\mathrm{hist}}(t) + tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\rho}_h^t}\left[\left[\mathbb{E}_h^{\pi_\eta^{\mathsf{unif}}}[f_{h+1}](x,a)\right]^2\right] + tA_\eta^2\varepsilon_{\mathrm{rep}}(t) + 9L^2\lambda^t d_\eta && \text{(Theorem 4.1)} \\
\leq & t\varepsilon_{\mathrm{rep}}(t) + 2t\varepsilon_{\mathrm{hist}}(t) + tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\gamma}_h^t}\left[f_{h+1}^2(x,a)\right] + tA_\eta^2\varepsilon_{\mathrm{rep}}(t) + 9L^2\lambda^t d_\eta. && \text{(Jensen)}
\end{aligned}$$

Since every inequality above holds in the point-wise way, then by Lemma F.10, putting everything together we have for each $h \geq 2$,

$$\begin{aligned}
&\left[\widetilde{\mathcal{P}}\right]^{\otimes h}[f_{h+1}] \\
\leq & \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} \sqrt{\sum_{b_\eta\in\mathcal{B}_\eta[\phi_h^t]} \frac{\mathbb{1}\{\phi_h(\cdot)\in b_\eta\}}{t\bar{\rho}_h^t(b_\eta)+\lambda^t}} \sqrt{2tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\gamma}_h^t}\left[f_{h+1}^2(x,a)\right] + 4tA_\eta^2\varepsilon_{\mathrm{rep}}(t) + 4t\varepsilon_{\mathrm{hist}}(t) + 18L^2\lambda^t d_\eta} + 4\eta \\
\leq & \left[\widetilde{\mathcal{P}}\right]^{\otimes(h-1)} \sqrt{\frac{1}{t\bar{\rho}_h^t[\mathsf{ball}_\eta[\phi_h^t](\cdot)] + \lambda^t}} \sqrt{2tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\gamma}_h^t}\left[f_{h+1}^2(x,a)\right] + \zeta(t)} + 4\eta,
\end{aligned}$$

where

$$\zeta(t) := 4tA_\eta^2\varepsilon_{\mathrm{rep}}(t) + 4t\varepsilon_{\mathrm{hist}}(t) + 18L^2\lambda^t d_\eta, \tag{26}$$

finally, summing over $h \in [H]$ we complete the proof. □

After the distribution shift result for linear pseudobackup, we next state a similar result for the reward-free Bellman backup:

**Lemma F.9** (One-step-back for Bellman backup)**.** *For any set of functions $\{f_h\}_{h=1}^H$ where $f_h \in (\mathcal{X}\times\mathcal{A}\to[-L,L])$, and any policy $\pi$, we have*

$$\begin{aligned}
&\sum_{h=1}^H [\mathcal{P}]^{\otimes(h-1)}[f_h](x_1,\pi(x_1)) \leq \\
&\sum_{h=2}^H [\mathcal{P}]^{\otimes(h-1)} \sqrt{\frac{1}{t\bar{\rho}_h^t[\mathsf{ball}_\eta[\phi_h^\star](\cdot)] + \lambda^t}} \sqrt{tA_\eta^2\,\mathbb{E}_{x,a\sim\bar{\gamma}_h^t}\left[f_{h+1}^2(x,a)\right] + \lambda^t L^2 d_\eta} + \sqrt{A\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[f_1^2(x,a)]} + 3H\eta.
\end{aligned}$$

**Proof of Lemma F.9.** The proof is mostly similar to the proof of the previous one-step-back lemma. To start, for $h = 1$, by Jensen's inequality, we have:

$$f_1(x_1,\pi(x_1)) = \mathbb{E}_{a\sim\pi(x_1)}[f_1(x_1,a)] \leq \sqrt{\max_{a\in\mathcal{A}_\eta}\frac{\pi(a\mid x_1)}{\pi_\eta^{\mathsf{unif}}(a\mid x_1)}\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[f_1^2(x,a)]} \leq \sqrt{A_\eta\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[f_1^2(x,a)]}.$$

Then for $h \geq 2$, we have

$$\begin{aligned}
&[\mathcal{P}]^{\otimes h}[f_{h+1}] \\
=&[\mathcal{P}]^{\otimes h-1}[\mathcal{P}_h[f_{h+1}]] \\
\leq&[\mathcal{P}]^{\otimes h-1}\left[\mathcal{P}_{\mathsf{disc}_\eta[\phi_h^\star]}[f_{h+1}]\right] + \eta \\
=&[\mathcal{P}]^{\otimes h-1}\left[\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{(t\bar{\rho}_h^t(b_\eta) + \lambda^t)\left[\mathcal{P}_{\mathsf{disc}_\eta[\phi_h^\star]}[f_{h+1}](x_\eta, a_\eta)\right]^2}\right] + \eta,
\end{aligned}$$

Once again focusing on the terms inside the expectation, we have the following pointwise inequality:

$$\begin{aligned}
&\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{(t\bar{\rho}_h^t(b_\eta) + \lambda^t)\left[\mathcal{P}_{\mathsf{disc}_\eta[\phi_h^\star]}[f_{h+1}](x_\eta, a_\eta)\right]^2} \\
\leq& \sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{\int_{b_\eta} (t\bar{\rho}_h^t(x,a) + \lambda^t)\left[\mathcal{P}_{\mathsf{disc}_\eta[\phi_h^\star]}[f_{h+1}](x, a)\right]^2 \mathrm{d}\nu(x,a)} && \text{(piecewise constant)} \\
\leq& \sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{\int_{b_\eta} (t\bar{\rho}_h^t(x,a) + \lambda^t)[\mathcal{P}_h[f_{h+1}](x, a) + \eta]^2 \mathrm{d}\nu(x,a)} \\
\leq& \sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{\sqrt{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{\int_{b_\eta} 2(t\bar{\rho}_h^t(x,a) + \lambda^t)[\mathcal{P}_h[f_{h+1}](x, a)]^2 \mathrm{d}\nu(x,a) + 2\eta} \\
\leq& \sqrt{\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{2\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \int_{b_\eta} (t\bar{\rho}_h^t(x,a) + \lambda^t)[\mathcal{P}_h[f_{h+1}](x, a)]^2 \mathrm{d}\nu(x,a) + 2\eta} \\
\leq& \sqrt{\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{2\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \int_{b_\eta} t\bar{\rho}_h^t(x,a)[\mathcal{P}_h[f_{h+1}](x, a)]^2 \mathrm{d}\nu(x,a) + L^2\lambda^t d_\eta + 2\eta} \\
=& \sqrt{\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{2t\,\mathbb{E}_{\bar{\rho}_h^t}[\mathcal{P}_h[f_{h+1}](x, a)]^2 \mathrm{d}\nu(x,a) + L^2\lambda^t d_\eta + 2\eta} \\
\leq& \sqrt{\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{2tA_\eta^2\,\mathbb{E}_{\bar{\rho}_h^t}\left[\left[\mathbb{E}_h^{\pi_\eta^{\mathsf{unif}}}[f_{h+1}](x, a)\right]^2\right] + L^2\lambda^t d_\eta + 2\eta} && \text{(Importance sampling)} \\
\leq& \sqrt{\sum_{b_\eta \in \mathcal{B}_\eta[\phi_h^\star]} \frac{\mathbb{1}\{\phi_h^\star(\cdot) \in b_\eta\}}{t\bar{\rho}_h^t(b_\eta) + \lambda^t}} \sqrt{2tA_\eta^2\,\mathbb{E}_{\bar{\gamma}_h^t}\left[f_{h+1}^2(x, a)\right] + L^2\lambda^t d_\eta + 2\eta}. && \text{(Jensen)}
\end{aligned}$$

Finally putting everything together we complete the proof. □

**Lemma F.10** (Monotonicity of linear pseudobackup). *Let $f, f' \in \mathcal{F}_{h+1} : \mathcal{X} \to [0, L]$ be two functions such that $f(x) \leq f'(x)$ for all $x \in \mathcal{X}$. Then we have $\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}[f](x) \leq \widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}[f'](x)$.*

**Proof of Lemma F.10.** Recall the definition of the linear pseudobackup, fix the decoder $\phi$, the discretized decoder $\mathsf{disc}_\eta[\phi](x, a)$, is a $d_\eta := \left(\frac{1}{\eta}\right)^{\dim_{\mathcal{SA}}}$ dimensional one-hot vector, for any $x, a \in \mathcal{X} \times \mathcal{A}$. Then if we set the $\ell_\infty$ constraint on $\mathcal{W}$ to be $\mathcal{W} = \{w \in \mathbb{R}_\eta^d \mid \|w\|_\infty \leq L\}$, we have that

$$\widetilde{\mathcal{P}}_{\mathcal{D}_h,\mathcal{W}}[f](x, a) := \frac{\sum_{(\widetilde{x},\widetilde{a},x') \in \mathcal{D}_h} \mathbb{1}\{\widetilde{x}, \widetilde{a} \in \mathsf{ball}_\eta[\phi_h](x, a)\} f(x')}{\sum_{(\widetilde{x},\widetilde{a}) \in \mathcal{D}} \mathbb{1}\{\widetilde{x}, \widetilde{a} \in \mathsf{ball}_\eta[\phi_h](x, a)\}},$$

and by the non-negativity of indicator functions we complete the proof. □

F.3.6. PROVING OPTIMISM

**Lemma F.11** (Almost optimism)**.** *For any round $t$, let*

$$\widehat{\alpha}^t = \sqrt{2tA_\eta^2 \varepsilon_{\text{rep}}(t) + \zeta(t)}/c \quad \text{and} \quad \lambda_T = \Theta(d_\eta \ln(T|\Phi|/\delta)),$$

*where $\zeta(t)$ is defined in Eq. (26), and $c$ is a constant from Lemma F.13. Then for any policy $\pi^t$ derived by the linear pseudobackup operator, let $f^{\pi^t}$ denotes the value function induced from the pseudobackup operator in Algorithm 3, then with probability at least $1-\delta$, we have*

$$f_1^{\pi^t}(x_1) - V_1^{\pi^t}(x_1) \geq -\left(\sqrt{A_\eta \varepsilon_{\text{rep}}(t)} + 3H\eta\right).$$

**Proof of Lemma F.11.** By Lemma F.2, we have

$$f^{\pi^t}(x_1) - V^{\pi^t}(x_1) = \sum_{h=1}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi^t]}\right)^{\otimes(h-1)} \left[\widehat{b}_h^t + (\widetilde{\mathcal{P}}_{\mathcal{D}_h^t, \mathsf{disc}_\eta[\phi_h^t]} - \mathcal{P}_h)[V_{h+1}^\pi]\right](x_1, \pi(x_1)).$$

In the following we focus on iteration $t$, so we will drop the superscript $t$ for notational simplicity. Let $g_h := (\mathcal{P}_h - \widetilde{\mathcal{P}}_{\mathcal{D}_h, \mathsf{disc}_\eta[\phi_h]})[V_{h+1}^\pi]$, note that $g_h \in \mathcal{F}_h^1$ (by construction of $\mathcal{F}_h^1$). We first check that Lemma F.10 holds since $\|g_h\|_\infty \leq 2$. Then by Lemma F.8:

$$\begin{aligned}
&\sum_{h=1}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}, \mathsf{disc}_\eta[\phi]}\right)^{\otimes(h-1)} [g_h] \\
&\leq \sum_{h=2}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}, \mathsf{disc}_\eta[\phi]}\right)^{\otimes(h-1)} \min\left\{\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](\cdot)) + \lambda^t}} \sqrt{2tA_\eta^2\, \mathbb{E}_{x,a\sim\bar{\gamma}_h^t}[g_h^2(x,a)] + \zeta(t)}, 2\right\} + \sqrt{A\, \mathbb{E}_{x,a\sim\bar{\rho}_1^t}[g_1^2(x,a)]} + 3H\eta \\
&\leq \sum_{h=2}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}, \mathsf{disc}_\eta[\phi]}\right)^{\otimes(h-1)} \min\left\{\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](\cdot)) + \lambda^t}} \underbrace{\sqrt{2tA_\eta^2 \varepsilon_{\text{rep}}(t) + \zeta(t)}}_{\alpha^t}, 2\right\} + \sqrt{A\varepsilon_{\text{rep}}(t)} + 3H\eta.
\end{aligned}$$

By the construction of bonus we have

$$\begin{aligned}
\sum_{h=1}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}^t, \mathsf{disc}_\eta[\phi^t]}\right)^{\otimes(h-1)} \left[\widehat{b}_h^t\right] &= \sum_{h=1}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}^t, \mathsf{disc}_\eta[\phi^t]}\right)^{\otimes(h-1)} \left[\min\left\{\widehat{\alpha}^t \cdot \sqrt{\frac{1}{N_{\eta,\phi_h^t}(\cdot, \mathcal{D}_{1,h}^t) + \lambda^t}}, 2\right\}\right] \\
&\geq \sum_{h=1}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}^t, \mathsf{disc}_\eta[\phi^t]}\right)^{\otimes(h-1)} \left[\min\left\{c\widehat{\alpha}^t \cdot \sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](\cdot)) + \lambda^t}}, 2\right\}\right].
\end{aligned}$$

(Concentration of the bonus, Lemma F.13)

Putting everything together we have

$$\begin{aligned}
f^{\pi^t}(x_1) - V^{\pi^t}(x_1) \geq \sum_{h=1}^{H} \left(\widetilde{\mathcal{P}}_{\mathcal{D}^t, \mathsf{disc}_\eta[\phi^t]}\right)^{\otimes(h-1)} &\left[\min\left\{c\widehat{\alpha}^t \cdot \sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](\cdot)) + \lambda^t}}, 2\right\} - \right. \\
&\left. \left(\min\left\{\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](\cdot)) + \lambda^t}}\alpha^t, 2\right\} + \sqrt{A\varepsilon_{\text{rep}}(t)} + 3H\eta\right)\right],
\end{aligned}$$

and finally by construction of $\widehat{\alpha}^t$ we complete the proof. □

### F.3.7. Proving the Regret

**Theorem F.2** (Pseudo regret for pseudobackups)**.** *With probability at least* $1-\delta$*, setting parameters*

$$\lambda^t = \Theta\left(t^{\frac{\dim_{\mathcal{SA}}}{\tilde{d}+2}} \log\left(\frac{t|\Phi|}{\delta}\right)\right), \quad \widehat{\alpha}^t = \Theta\left(t^{\frac{\bar{d}}{\tilde{d}}} \log\left(\frac{t|\Phi|}{\delta}\right)\right),$$

*let* $\widehat{\pi}$ *be the output of the* CRIEE*, we have*

$$\mathsf{Reg}(T) \le O\left(H^2 T^{\frac{\tilde{d}}{\tilde{d}+2}} \sqrt{\log(T|\Phi|/\delta)}\right),$$

*where* $\widetilde{d} = 6\dim_{\mathcal{SA}}^2 + 8\dim_{\mathcal{SA}}\dim_{\mathcal{A}} + 10\dim_{\mathcal{SA}} + 8\dim_{\mathcal{A}} + 2$.

**Proof of Theorem F.2.** By the standard decomposition (Jiang et al., 2017) we have

$$\sum_{t=1}^T J(\pi^\star) - J(\pi^t) = \sum_{t=1}^T \left(V^{\pi^\star}(x_1) - f^{t;\pi^\star}(x_1)\right) + \left(f^{t;\pi^\star}(x_1) - V^{\pi^t}(x_1)\right).$$

By Lemma F.11, we have

$$\sum_{t=1}^T \left(V^{\pi^\star} - V^{\pi^t}\right) \le \sum_{t=1}^T \left(f^{t;\pi^\star} - V^{\pi^t}\right) + \left(\sqrt{A\varepsilon_{\mathrm{rep}}(t)} + 3H\eta\right) \le \sum_{t=1}^T \left(f^{\pi^t} - V^{\pi^t}\right) + \left(\sqrt{A\varepsilon_{\mathrm{rep}}(t)} + 3H\eta\right).$$

By Lemma F.3, we have

$$\sum_{t=1}^T \left(f^{\pi^t}(x_1) - V^{\pi^t}(x_1)\right) \le \sum_{t=1}^T \sum_{h=1}^H (\mathcal{P})^{\otimes(h-1)}[\underbrace{\widehat{b}_h^t}_{A} + \underbrace{(\widetilde{\mathcal{P}}_{\mathcal{D}_h^t, \mathsf{disc}_\eta[\phi_h^t]} - \mathcal{P}_h)[f_{h+1}^{\pi^t}]}_{B}](x_1, \pi^t(x_1)).$$

To bound $A$, note that by construction $\left\|\widehat{b}_h^t\right\|_\infty \le 2$, i.e., $L=2$, and $b_h^t \in \mathcal{F}_h$, then by Lemma F.9, we have

$$\begin{aligned}
&\sum_{h=1}^H (\mathcal{P})^{\otimes(h-1)}[\widehat{b}_h^t] \\
&\le \sum_{h=2}^H (\mathcal{P})^{\otimes(h-1)} \sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](\cdot)) + \lambda^t}} \sqrt{tA_\eta^2\, \mathbb{E}_{x,a\sim\bar{\gamma}_h^t}\left[\left(\widehat{b}_h^t\right)^2(x,a)\right] + \lambda^t 4 d_\eta} + \sqrt{tA\, \mathbb{E}_{x,a\sim\bar{\rho}_1^t}\left[\left(\widehat{b}_1^t\right)^2(x,a)\right]} + 3H\eta.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}_{x,a\sim\rho_h}\left[\left(\widehat{b}_h^t\right)^2(x,a)\right] &\le c^2 \left(\widehat{\alpha}^t\right)^2 \mathbb{E}_{x,a\sim\rho_h}\left[\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^t](x,a)) + \lambda^t}\right] && \text{(Lemma F.13)} \\
&= \left(\alpha^t\right)^2 \mathbb{E}_{x,a\sim\rho_h}\left[\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^t](x,a)) + \lambda^t}\right] \\
&\le \left(\alpha^t\right)^2 \frac{d_\eta}{t}.
\end{aligned}$$

To bound term $B$, first let us denote the shorthand:

$$\delta_h := \frac{1}{2H+1} (\widetilde{\mathcal{P}}_{\mathcal{D}_h^t, \mathsf{disc}_\eta[\phi_h^t]} - \mathcal{P}_h)[f_{h+1}^\pi],$$

since again $\left\|\widehat{b}_h^t\right\|_\infty \leq 2$, we have $\left\|f_{h+1}^\pi\right\|_\infty \leq 2H+1$, and thus $\|\delta_h\|_\infty \leq 2$. Then by Lemma F.8, term $B$ can be bounded as the following:

$$B \leq (2H+1)\sum_{h=2}^{H}(\mathcal{P})^{\otimes(h-1)}\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](\cdot))+\lambda^t}}\sqrt{tA_\eta^2\,\mathbb{E}_{x,a\sim\gamma_h^t}[\delta_h^2(x,a)]+\lambda^t 4d_\eta}+$$
$$(2H+1)\sqrt{tA\,\mathbb{E}_{x,a\sim\bar{\rho}_1^t}[\delta_1^2(x,a)]}+3H\eta$$
$$\leq (2H+1)\sum_{h=2}^{H}(\mathcal{P})^{\otimes(h-1)}\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](\cdot))+\lambda^t}}\sqrt{tA_\eta^2\varepsilon_{\text{rep}}(t)+\lambda^t 4d_\eta}+(2H+1)\sqrt{tA\varepsilon_{\text{rep}}(t)}+3H\eta,$$

where the second line is because $\frac{1}{2H+1}f_{h+1}^\pi \in \mathcal{F}_{h+1}$ and thus we invoke Theorem 4.1. Finally we bound the potential term in front of term $A$ and $B$, by Lemma F.12, we have

$$\sum_{t=1}^{T}(\mathcal{P})^{\otimes(h-1)}\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](\cdot))+\lambda^t}}(x_1,\pi^t(x_1))$$
$$\leq\sqrt{T\sum_{t=1}^{T}(\mathcal{P})^{\otimes(h-1)}\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](\cdot))+\lambda^t}(x_1,\pi^t(x_1))} \qquad \text{(Cauchy Schwarz)}$$
$$\leq\sqrt{d_\eta T\ln\left(1+\frac{T}{\lambda_1 d_\eta}\right)}.$$

Now recall the construction of $\alpha^t$:

$$\alpha^t=\sqrt{tA_\eta^2\varepsilon_{\text{rep}}(t)+\zeta(t)}=O\left(\sqrt{tA_\eta^2\varepsilon_{\text{rep}}(t)+\lambda^t d_\eta}\right). \qquad (\varepsilon_{\text{hist}}(t)\leq\varepsilon_{\text{rep}}(t))$$

By Theorem 4.1 and note that $L=2$, we have

$$\varepsilon_{\text{rep}}(t)=\frac{352(4^{\dim_{\mathcal{SA}}})d_\gamma\log(4|\Phi|\cdot\mathcal{N}_\infty(\mathcal{F},\gamma)/(\delta\gamma))}{t}+96\gamma^2,$$

then by Lemma F.14, we have

$$\varepsilon_{\text{rep}}(t)=O\left(\frac{d_\eta^2 d_\gamma^2\log(t|\Phi|/(\gamma\delta))}{t}+\gamma^2\right),$$

combining everything together and taking leading terms we have:

$$\sum_{t=1}^{T}\left(V^{\pi^\star}-V^{\pi^t}\right)\leq O\left(H^2A_\eta^2\sqrt{d_\eta^3 d_\gamma^2 T\log(Td_\eta|\Phi|/(\gamma\delta))+T^2 d_\eta\gamma^2}+TH\eta\right).$$

Now we can take $\gamma=T^{-\frac{1}{2d+2}}$, and $d=\dim_{\mathcal{SA}}$, and we get

$$\sum_{t=1}^{T}\left(V^{\pi^\star}-V^{\pi^t}\right)\leq O\left(H^2A_\eta^2\sqrt{d_\eta^3 T^{\frac{2d+1}{d+1}}\log(Td_\eta|\Phi|/\delta)}+TH\eta\right).$$

Finally taking (note that $A_\eta=(1/\eta)^{\dim_{\mathcal{A}}}$)

$$\eta=T^{-\frac{2}{(3\dim_{\mathcal{SA}}+4\dim_{\mathcal{A}}+2)(2\dim_{\mathcal{SA}}+2)}}$$

we get:

$$\sum_{t=1}^{T}\left(V^{\pi^\star}-V^{\pi^t}\right)\leq O\left(H^2 T^{\frac{\widetilde{d}}{\widetilde{d}+2}}\sqrt{\log(T|\Phi|/\delta)}\right),$$

where $\widetilde{d}=6\dim_{\mathcal{SA}}^2+8\dim_{\mathcal{SA}}\dim_{\mathcal{A}}+10\dim_{\mathcal{SA}}+8\dim_{\mathcal{A}}+2$. □

Finally, to prove Theorem 4.2′, we divide the right-hand-side by $T$, bound by $\varepsilon$ and solve for $T$.

F.3.8. SUPPORTING LEMMAS

**Lemma F.12** (Concentration of potential)**.** *We have that*

$$\sum_{t=1}^{T}(\mathcal{P})^{\otimes(h-1)}\left[\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](\cdot))+\lambda_t}\right](x_1,\pi(x_1))\le d_\eta\ln\left(1+\frac{T}{\lambda^1\eta^{d_\eta}}\right),$$

*where* $d_\eta=\left(\frac{1}{\eta}\right)^{\dim_{\mathcal{SA}}}$.

**Proof of Lemma F.12.** First let us define the vector $\phi^\star_{\eta;h}$ as a $d_\eta$-dimensional one-hot vector, where the $i$-th entry is 1 if $\phi^\star_h(x,a)\in\mathsf{ball}^i_\eta[\phi^\star_h]$, and 0 otherwise. Then we have the following identity:

$$\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](x,a))+\lambda_t}=\phi^\star_{\eta;h}(x,a)\Sigma^{-1}_{t;\phi^\star_{\eta;h};\rho_h}\phi^\star_{\eta;h}(x,a),$$

where

$$\Sigma_{t;\phi^\star_{\eta;h};\rho_h}=\sum_{\tau=1}^{t}\mathbb{E}_{x,a\sim\rho_h^\tau}\,\phi^\star_{\eta;h}(x,a)\big(\phi^\star_{\eta;h}(x,a)\big)^\top+\lambda_t I.$$

Hence we establish the relationship between the potential function that we are interested in and linear models. Then by Lemma D.2 and Lemma D.3, we have that:

$$\sum_{t=1}^{T}\mathbb{E}_{x,a\sim\rho_h^t}\left[\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h^\star](x,a))+\lambda_t}\right]\le 2\ln\det\left(\Sigma_{T;\phi^\star_{\eta;h};\rho_h}\right)-2\ln\log(\lambda I)\le d_\eta\ln\left(1+\frac{T}{\lambda_1 d_\eta}\right),$$

where we use the fact $B=1$ because the vectors are one-hot. □

Finally we can also leverage the connection to linear models to prove the concentration of bonus:

**Lemma F.13** (Concentration of bonus; Lemma 22 of Zhang et al. (2022))**.** *Set* $\lambda_T=\Theta(d_\eta\ln(T|\Phi|/\delta))$*, then with probability at least* $1-\delta$*, we have for all* $t\in[T]$ *and* $\phi\in\Phi$*:*

$$c_1\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](x,a))+\lambda_t}}\le\sqrt{\frac{1}{N_{\eta,\phi}(x,a,\mathcal{D}^t_{1,h})+\lambda_t}}\le c_2\sqrt{\frac{1}{t\bar{\rho}_h^t(\mathsf{ball}_\eta[\phi_h](x,a))+\lambda_t}}.$$

**Lemma F.14** (Covering number of the discriminator class)**.** *For any* $h\in[H]$*, define*

$$\mathcal{F}^1_{\eta,h}=\left\{f(x):\mathbb{E}_{a\sim\pi^{\mathsf{unif}}_\eta}\left[w^\top\mathsf{disc}_\eta[\phi](x,a)-g\left(\widetilde{\phi}(x,a)\right)\right]\mid\phi,\widetilde{\phi}\in\Phi_{h+1},g\in\mathrm{Lip},w\in\mathbb{R}^{d_\eta},\|w\|_\infty\le 1\right\},$$

$$\mathcal{F}^2_{\eta,h}=\left\{f(x):\max_a\left(\frac{R_{h+1}(x,a)+\min\{w^\top\mathsf{disc}_\eta[\phi](x,a),2\}}{2H+1}+\widetilde{w}^\top\mathsf{disc}_\eta[\phi](x,a)\right)\right.$$
$$\left.\mid\phi\in\Phi_{h+1},w,\widetilde{w}\in\mathbb{R}^{d_\eta},\|w\|_\infty\le c,\|\widetilde{w}\|_\infty\le 2\right\}.$$

*Then* $\mathcal{F}^1_{\eta,h}$ *and* $\mathcal{F}^2_{\eta,h}$ *have* $\gamma$*-covering number*

$$N_\infty(\mathcal{F}^1_{\eta,h},\gamma)\le|\Phi_{h+1}|^2\left(\frac{2}{\gamma}\right)^{d_\eta}\left(\frac{1}{\gamma}\right)^{2d_\gamma}\quad\textit{and}\quad N_\infty(\mathcal{F}^2_{\eta,h},\gamma)\le|\Phi_{h+1}|\left(\frac{4}{\gamma}\right)^{d_\eta}\left(\frac{2c}{\gamma}\right)^{d_\eta}.$$

**Proof of Lemma F.14.** For $\mathcal{F}^1_{\eta,h}$, note that the size of a $\frac{\gamma}{2}$-cover of $\{w\in\mathbb{R}^{d_\eta}\mid\|w\|_\infty\le 1\}$ is bounded by $\left(\frac{2}{\gamma}\right)^{d_\eta}$, and the size of a $\frac{\gamma}{2}$-cover of Lip is bounded by $\left(\frac{1}{\gamma}\right)^{2d_\gamma}$ (Wainwright, 2019), and taking union with $\Phi\times\Phi$ we complete the calculation of the covering number of $\mathcal{F}^1_{\eta,h}$. The calculation of the covering number of $\mathcal{F}^2_{\eta,h}$ is similar by calculating the size of $\frac{\gamma}{2}$-cover of $\{w\in\mathbb{R}^{d_\eta}\mid\|w\|_\infty\le 2\}$ and $\{w\in\mathbb{R}^{d_\eta}\mid\|w\|_\infty\le c\}$. □

### F.4. Proof of Proposition 4.1

**Background on multi-step inverse kinematics.** The multistep inverse kinematics approach (Lamb et al., 2023; Mhammedi et al., 2023b) involves predicting the action $a_h \sim \mathsf{unif}(\mathcal{A})$ at time $h$ from the observation $x_h$ and a future observation $x_t$ for fixed $t > h$ under a roll-out policy $\pi$. When the action space is finite, one can show that the optimal population-level objective takes the form

$$(x_h, a_h, x_t) \mapsto \frac{\mathbb{P}^\pi(\phi^\star(x_t) \mid \phi^\star(x_h), a_h)}{\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(\phi^\star(x_t) \mid \phi^\star(x_h), a)}. \tag{27}$$

Similar to contrastive learning, the a key property of this objective is that it depends on the observation only through the corresponding latent state, a central property used by (Mhammedi et al., 2023b). However, analogously to contrastive learning, this property alone is not sufficient for sample-efficient learning in the RichCLD, because we need to construct a low-complexity function class to express the optimal predictor (27) in this latent space. Unfortunately, the optimal predictor for multistep inverse kinematics objective may not be a Lipschitz function of the latent state, even when though transition dynamics themselves are Lipschitz.

**Proof of Proposition 4.1.** To begin, we focus on multi-step inverse kinematics (Lamb et al., 2023; Mhammedi et al., 2023b). Consider the case where $|\mathcal{A}| = \{\mathfrak{a}, \mathfrak{b}\}$ (two actions will suffice, as the difficulty arises from continuity of the latent state space). Let $t > h$ be fixed. Consider a setting where we sample $x_h \sim \rho_h$ (an arbitrary roll-in distribution), $a_h \sim \pi^{\mathsf{unif}}$, and sample $x_t$ by executing a given policy $\pi$ from steps $h+1, \ldots, t$. The multi-step inverse kinematics objective performs conditional density estimation under this process:

$$(\widehat{P}, \phi) = \operatorname*{arg\,max}_{P\in\mathcal{P}, \phi\in\Phi} \widehat{\mathbb{E}}[\log(P(a_t \mid \phi(x_t), \phi(x_h))]$$

for a function class $\mathcal{P}$. If $\mathcal{P}$ is unconstrained and $\phi^\star \in \Phi$, the population-level optimizer for this objective is

$$P_{\mathsf{musik}}(a_h \mid x_t, x_h) := \frac{\mathbb{P}^\pi(\phi^\star(x_t) \mid \phi^\star(x_h), a_h)}{\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(\phi^\star(x_t) \mid \phi^\star(x_h), a)}.$$

Now let us focus on the case where $t = h + 1$, i.e., the one-step inverse kinematics, which is captured by the multi-step inverse kinematics objective. And the one-step inverse kinematics optimizer is

$$P_{\mathsf{musik}}(a_h \mid x_{h+1}, x_h) := \frac{\mathbb{P}^\pi(\phi^\star(x_{h+1}) \mid \phi^\star(x_h), a_h)}{\sum_{a\in\mathcal{A}} P_h(\phi^\star(x_{h+1}) \mid \phi^\star(x_h), a)},$$

where $P_h$ is the latent dynamics at timestep $h$. Now recall the latent state space $\mathcal{S}$ which is a metric space with metric $D_{\mathcal{S}}$. Then for each $x \in \mathcal{X}$, let $s = \phi^\star(x)$ be the corresponding latent state according to the ground truth decoder. Then we can rewrite the one-step inverse kinematics optimizer as $P_{\mathsf{musik}}(a_h \mid x_{h+1}, x_h) = P_{\mathsf{musik}}(a_h \mid \phi^\star(x_{h+1}), \phi^\star(x_h))$ for

$$P_{\mathsf{musik}}(a_h \mid s_{h+1}, s_h) := \frac{\mathbb{P}^\pi(s_{h+1} \mid s_h, a_h)}{\sum_{a\in\mathcal{A}} P_h(s_{h+1} \mid s_h, a)}.$$

To prove the result, it suffices to show that $P_{\mathsf{musik}}$ is not Lipschitz with respect to the parameter $s_h$, i.e., for some fixed $s_{h+1}$ and $a_h$, there exists $s_h^1$ and $s_h^2$ such that

$$\left|P_{\mathsf{musik}}(a_h \mid s_{h+1}, s_h^1) - P_{\mathsf{musik}}(a_h \mid s_{h+1}, s_h^2)\right| > D_{\mathcal{S}}(s_h^1, s_h^2). \tag{28}$$

Now let us fix $s_{h+1}$ and $a_h = \mathfrak{a}$. Consider a pair of states $s_h^1$ and $s_h^2$ with $D_{\mathcal{S}}(s_h^1, s_h^2) = \delta$ for a given parameter $\delta > 0$. Let us define the dynamics such that (i) $P_h(s_{h+1} \mid s_h^1, \mathfrak{a}) = 2\delta$, (ii) $P_h(s_{h+1} \mid s_h^2, \mathfrak{a}) = \delta$, (iii) $P_h(s_{h+1} \mid s_h^1, \mathfrak{b}) = P_h(s_h \mid s_h^2, \mathfrak{b}) = \delta$, (iv) there is one state $\widetilde{s}_{h+1}$ such that $P_h(\widetilde{s}_{h+1} \mid s_h^1, \mathfrak{a}) = \delta$, and $P_h(\widetilde{s}_{h+1} \mid s_h^2, \mathfrak{a}) = 2\delta$, (v) $P_h(s'_{h+1} \mid s_h^1, a) = P_h(s'_{h+1} \mid s_h^2, a)$, for all $a \in \mathcal{A}, s'_{h+1} \notin \{s_{h+1}, \widetilde{s}_{h+1}\}$. We can check that $\left\|P_h(\cdot \mid s_h^1, a) - P_h(\cdot \mid s_h^2, a)\right\|_{\mathsf{TV}} = \frac{1}{2}\left\|P_h(\cdot \mid s_h^1, a) - P_h(\cdot \mid s_h^2, a)\right\|_1 = \delta = D_{\mathcal{S}}(s_h^1, s_h^2)$ and thus the construction satisfies the Lipschitz latent dynamics condition. However,

$$\frac{\mathbb{P}^\pi(s_{h+1} \mid s_h^1, \mathfrak{a})}{\sum_a \mathbb{P}^\pi(s_{h+1} \mid s_h^1, a)} - \frac{\mathbb{P}^\pi(s_{h+1} \mid s_h^2, \mathfrak{a})}{\sum_a \mathbb{P}^\pi(s_{h+1} \mid s_h^2, a)} = \frac{2\delta}{2\delta + \delta} - \frac{\delta}{\delta + \delta} = \frac{1}{6},$$

but we can take $\delta$ arbitrarily small. This proves Eq. (28).

For contrastive learning (Misra et al., 2020), recall from the main text that for $h \in [H]$, the optimal classifier takes the form $P_{\mathsf{cl}}(x_{h+1}, x_h, a_h) = P_{\mathsf{cl}}(\phi^\star(x_{h+1}), \phi^\star(x_h), a_h)$, where

$$P_{\mathsf{cl}}(s_{h+1}, s_h, a_h) := \frac{P_h(s_{h+1} \mid s_h, a_h)}{P_h(s_{h+1} \mid s_h, a_h) + \widetilde{\rho}_{h+1}(s_{h+1})},$$

for the process in which we sample $s_h \sim \rho_h$ for a data collection distribution $\rho_h$; $\widetilde{\rho}_{h+1}$ denotes the law of $s_{h+1}$ when we sample $s_h \sim \rho_h$ and $a_h \sim \pi^{\mathsf{unif}}$. For this objective, we can follow exactly the same construction as above, where we define the dynamics as (i) $P_h(s_{h+1} \mid s_h^1, \mathfrak{a}) = 2\delta$, (ii) $P_h(s_{h+1} \mid s_h^2, \mathfrak{a}) = \delta$, (iii) $P_h(s_{h+1} \mid s_h^1, \mathfrak{b}) = P_h(s_h \mid s_h^2, \mathfrak{b}) = \delta$, (iv) there is one state $\widetilde{s}_{h+1}$ such that $P_h(\widetilde{s}_{h+1} \mid s_h^1, \mathfrak{a}) = \delta$, and $P_h(\widetilde{s}_{h+1} \mid s_h^2, \mathfrak{a}) = 2\delta$, (v) $P_h(s'_{h+1} \mid s_h^1, a) = P_h(s'_{h+1} \mid s_h^2, a)$, for all $a \in \mathcal{A}, s'_{h+1} \notin \{s_{h+1}, \widetilde{s}_{h+1}\}$. Now consider the data collection distribution that puts all the mass on $s_h^2$, then by construction we have $\tilde{\rho}_{h+1}(s_{h+1}) = \delta$. Plugging everything in we get

$$P_{\mathsf{cl}}(\mathfrak{a} \mid s_{h+1}, s_h^1) - P_{\mathsf{cl}}(\mathfrak{a} \mid s_{h+1}, s_h^2) = \frac{2\delta}{2\delta + \delta} - \frac{\delta}{\delta + \delta} = \frac{1}{6},$$

and we prove that the optimal classifier for contrastive learning is not Lipschitz with respect to $s_h$ as well. □