Reinforcement Learning in Low-rank MDPs with Density Features

Audrey Huang [*] [1] **Jinglin Chen** [*] [1] **Nan Jiang** [1]

Abstract

MDPs with low-rank transitions—that is, the transition matrix can be factored into the product of two matrices, left and right—is a highly representative structure that enables tractable learning. The left matrix enables expressive function approximation for value-based learning and has been studied extensively. In this work, we instead investigate sample-efficient learning with density features, i.e., the right matrix, which induce powerful models for state-occupancy distributions. This setting not only sheds light on leveraging unsupervised learning in RL, but also enables plug-in solutions for settings like convex RL. In the offline setting, we propose an algorithm for off-policy estimation of occupancies that can handle non-exploratory data. Using this as a subroutine, we further devise an online algorithm that constructs exploratory data distributions in a level-by-level manner. As a central technical challenge, the additive error of occupancy estimation is incompatible with the multiplicative definition of data coverage. In the absence of strong assumptions like reachability, this incompatibility easily leads to exponential error blow-up, which we overcome via novel technical tools. Our results also readily extend to the representation learning setting, when the density features are unknown and must be learned from an exponentially large candidate set.

1. Introduction

The theory of reinforcement learning (RL) in large state spaces has seen fast development. In the model-free regime, how to use powerful function approximation to learn *value functions* has been extensively studied in both the online and the offline settings (Jiang et al., 2017; Jin et al., 2020b,c; Xie et al., 2021), which also builds the theoretical foundations that connect RL with (discriminative) supervised learning. On the other hand, generative models for unsupervised/self-supervised learning—which define a sampling distribution explicitly or implicitly—are becoming increasingly powerful (Devlin et al., 2018; Goodfellow et al., 2020), yet how to leverage them to address the key challenges in RL remains under-investigated. While prior works on RL with unsupervised-learning oracles exist (Du et al., 2019; Feng et al., 2020), they often consider models such as block MDPs, which are more restrictive than typical model structures considered in the value-based setting such as low-rank MDPs.

In this paper, we study model-free RL in low-rank MDPs with density features for state occupancy estimation. In a low-rank MDP, the transition matrix can be factored into the product of two matrices, and the left matrix is known to serve as powerful features for value-based learning (Jin et al., 2020b), as it can be used to approximate the Bellman backup of any function. On the other hand, the *right* matrix can be used to represent the policies' state-occupancy distributions, yet how to leverage such *density features* (without the knowledge of the left matrix) in offline or online RL is unknown. To this end, our main research question is:

Is sample-efficient offline/online RL with density features possible in low-rank MDPs?

We answer this question in the positive, and below is a summary of our contributions:

1. **Offline:** Section 3 provides an algorithm for off-policy occupancy estimation. It bears similarity to existing algorithms for estimating *importance weights* (Hallak and Mannor, 2017; Gelada and Bellemare, 2019), but our setting gives rise to a number of novel challenges. Most importantly, our algorithm enjoys guarantees under *arbitrary* offline data distributions, when the standard notion of importance weights are not even well-defined. We introduce a novel notion of *recursively clipped occupancy* and show that it can be learned in a sample-efficient manner. The recursively clipped occupancy always lower bounds the true occupancy, and the two notions coincide when the data has sufficient coverage. Such a guarantee immediately enables an offline policy

[*] Equal contribution [1] Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA. Correspondence to: Nan Jiang <nanjiang@illinois.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

learning result that only requires "single-policy concentrability", which is comparable to the recent advances in value-based offline RL (Jin et al., 2020c; Xie et al., 2021).

2. **Online:** Using the offline algorithm as a subroutine, in Section 4, we design an *online* algorithm that builds an exploratory data distribution (or "policy cover" (Du et al., 2019)) from scratch in a level-by-level manner. At each level, we estimate each policy's state-occupancy distribution and construct an approximate cover by choosing the *barycentric spanner* of such distributions. A critical challenge here is that the additive ℓ_1 error in occupancy estimation destroys the multiplicative coverage guarantee of the barycentric spanner, so the constructed distribution is never perfectly exploratory. Worse still, standard algorithm designs and analyses for handling such a mismatch easily lead to *an exponential error blow-up*. We overcome this by a novel technique, where two inductive error terms are maintained and analyzed in parallel, with delicate interdependence that still allows for a polynomial error accumulation (Figure 1).

3. **Representation learning:** We also extend our offline and online results to the representation learning setting (Agarwal et al., 2020), where the true density features are not given but must also be learned from an exponentially large candidate feature set.

4. **Implications:** Our online algorithm is automatically reward-free (Jin et al., 2020a; Chen et al., 2022b) and deployment-efficient (Huang et al., 2022). Further, since we can accurately estimate the occupancy distribution for all candidate policies, our results enable plug-in solutions for settings such as convex RL (Mutti et al., 2022; Zahavy et al., 2021), where the objectives and/or constraints are functions over the entire state distributions (see Appendix C).

2. Preliminaries

Markov Decision Processes (MDPs) We consider a finite-horizon episodic MDP (without reward) defined as $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, H)$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $P = (P_0, \ldots, P_{H-1})$ with $P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ is the transition dynamics, H is the horizon, and $d_0 \in \Delta(\mathcal{X})$ is the known initial state distribution.[1] We assume that $\mathcal{X}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is finite with cardinality K. Each episode is a trajectory $\tau = (x_0, a_0, x_1, \ldots, x_{H-1}, a_{H-1}, x_H)$, where $x_0 \sim d_0$, the agent takes a sequence of actions $a_0, \ldots, a_{H-1}$, and $x_{h+1} \sim P_h(\cdot \mid x_h, a_h)$. We use $\pi = (\pi_0, \ldots, \pi_{H-1}) \in (\mathcal{X} \to \Delta(\mathcal{A}))^H$ to denote a (non-stationary) H-step Markov policy, which chooses $a_h \sim \pi_h(\cdot|x_h)$. (We will also omit the subscript h and write $\pi(\cdot|x_h)$ when it is clear from context.) We use ρ to refer to non-Markov policies that can choose a_h based on the history $x_{0:h}, a_{0:h-1}$, which often arises from the probability mixture of Markov policies at the beginning of an trajectory. Once a policy π is fixed, the MDP becomes an Markov chain, with $d_h^\pi(x_h)$ being its h-th step distribution. As a shorthand, we use the notation $[H]$ to denote $\{0, 1, \ldots, H-1\}$.

Low-rank MDPs We consider learning in a low-rank MDP, defined as:

Assumption 1 (Low-rank MDP)**.** *$\mathcal{M}$ is a low-rank MDP with dimension* d*, that is, $\forall h \in [H]$, there exist $\phi_h^* : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{\mathsf{d}}$ and $\mu_h^* : \mathcal{X} \to \mathbb{R}^{\mathsf{d}}$ such that $\forall x_h, x_{h+1} \in \mathcal{X}, a_h \in \mathcal{A} : P_h(x_{h+1}|x_h, a_h) = \langle \phi_h^*(x_h, a_h), \mu_h^*(x_{h+1}) \rangle$. Further, $\int \|\mu_h^*(x)\|_1 (\mathrm{d}x) \le B^\mu$ and $\|\phi_h^*(\cdot)\|_\infty \le 1$.*[2]

Notation We use the convention $\frac{0}{0} = 0$ when we define the ratio between two functions. Define $a \wedge b = \min(a, b)$, and we treat $\wedge$ as an operator with precedence between "$\times/$" and "$+-$". When clear from the context, $\{\Box_h\} = \{\Box_h\}_{h=0}^{H-1}$, and we refer to state "occupancies," "distributions," and "densities" interchangeably. Finally, letter "d" has a few different versions (with different fonts): d is the low-rank dimension, $d(x)$ is a density, and $(\mathrm{d}x)$ is the differential used in integration. Further, while d_h^π and d_h^D refer to true densities, d_h (without superscripts) is often used for optimization variables.

Learning setups We provide algorithms and guarantees under a number of different setups (e.g., offline vs. online). The result that connects all pieces together is the setting of online *reward-free* exploration with known density features $\mu^* = (\mu_0^*, \ldots, \mu_{H-1}^*)$ and a policy class $\Pi \subseteq (\mathcal{X} \to \Delta(\mathcal{A}))^H$ (Section 4). Here, the learner must explore the MDP and form accurate estimations of d_h^π for all $\pi \in \Pi$ and $h \in [H]$, that is, output $\{\widehat{d_h^\pi}\}_{h \in [H], \pi \in \Pi}$ such that with probability at least $1 - \delta$, $\forall \pi \in \Pi, h \in [H]$, $\|\widehat{d_h^\pi} - d_h^\pi\|_1 \le \varepsilon$, by only collecting $\mathrm{poly}(H, K, \mathsf{d}, \log(|\Pi|), 1/\varepsilon, \log(1/\delta))$ trajectories. Two remarks are in order:

1. Such a guarantee immediately leads to standard guarantees for return maximization when a reward function is specified. More concretely (with proof in Appendix F.2),

[1] We assume the known initial state distribution for simplicity. Our results easily extend to the unknown version.

[2] This is w.l.o.g. as the norm of ϕ_h^* can be absorbed into B^μ. In a natural special case of low-rank MDPs with "simplex features" (Jin et al., 2020b, Example 2.2), Assumption 1 holds with $B^\mu = \mathsf{d}$. Our sample complexities only have polylogarithmic dependence on B^μ which will be suppressed by $\widetilde{O}$.

Proposition 1. *Given any policy π and reward function*[3] $R = \{R_h\}$ *with* $R_h : \mathcal{X} \times \mathcal{A} \to [0,1]$, *define expected return as* $v_R^\pi := \mathbb{E}_\pi[\sum_{h=0}^{H-1} R_h(x_h, a_h)] = \sum_{h=0}^{H-1} \iint d_h^\pi(x_h) R_h(x_h, a_h) \pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$. *Then for* $\{\widehat{d}_h^\pi\}$ *such that* $\|\widehat{d}_h^\pi - d_h^\pi\|_1 \le \varepsilon/(2H)$ *for all* $\pi \in \Pi$ *and* $h \in [H]$, *we have* $v_R^{\widehat{\pi}_R} \ge \max_{\pi \in \Pi} v_R^\pi - \varepsilon$, *where* $\widehat{\pi}_R = \operatorname{argmax}_{\pi \in \Pi} \widehat{v}_R^\pi$, *and* $\widehat{v}_R^\pi$ *is the expected return calculated using* $\{\widehat{d}_h^\pi\}$.

Moreover, the result can be extended to more general settings, where the optimization objective is some function of the state (and action) distribution that cannot be written as cumulative expected rewards; e.g., entropy as in max-entropy exploration (Hazan et al., 2019), or $\|d_h^\pi - d_h^{\pi_E}\|_2^2$, where π_E is an expert policy, used in imitation learning (Abbeel and Ng, 2004). A detailed discussion is deferred to Appendix C.

2. The introduction of Π and the dependence on $K = |\mathcal{A}|$ are both necessary, since low-rank MDPs can emulate general contextual bandits where the density features μ^* become useless; see Appendix B for more details.

To enable such a result, a key component is to estimate d_h^π using offline data (Section 3). Later in Section 5, we also generalize our results to the *representation-learning* setting (Agarwal et al., 2020; Modi et al., 2021; Uehara et al., 2021b), where μ^* is not known but must be learned from an exponentially large candidate set.

3. Off-policy occupancy estimation

In this section, we describe our algorithm, FORC, which estimates the occupancy distribution d_h^π of any given policy π using an offline dataset. Note that this section serves both as an important building block for the online algorithm in Section 4 and a standalone offline-learning result in its own right, so we will make remarks from both perspectives.

We start by introducing our assumption on the offline data.

Assumption 2 (Offline data)**.** *Consider a dataset* $\mathcal{D}_{0:H-1} = \mathcal{D}_0 \bigcup \ldots \bigcup \mathcal{D}_{H-1}$, *where* $\mathcal{D}_h = \{(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)})\}_{i=1}^n$. *For any fixed h, we assume that tuples in* $\mathcal{D}_h$ *are sampled i.i.d. from* $\rho^{h-1} \circ \pi_h^D$, *where* $a_0, \ldots, a_{h-1} \sim \rho^{h-1}$ *is an arbitrary* $(h-1)$*-step (possibly non-Markov) policy*[4] *and* $a_h \sim \pi_h^D$ *is a single-step Markov policy. Further,* ρ_{h-1}, π_h^D *can be a function of* $\mathcal{D}_{0:h-1}$, *and* π_h^D *is known to the learner.*

The dataset consists of H parts, where the h-th part consists of (x_h, a_h, x_{h+1}) tuples, allowing us to reason about the transition dynamics at level h. In practice (as well as in Section 4), such tuples will be extracted from trajectory data. We use $d_h^D(x_h, a_h, x_{h+1}), d_h^D(x_h), d_h^{D,\dagger}(x_{h+1})$ to denote the joint and the marginal distributions, respectively. Importantly, we do *not* assume that $d_h^{D,\dagger}(x_{h+1}) = d_{h+1}^D(x_{h+1})$, i.e., the next-state distribution of $\mathcal{D}_h$ and the current-state distribution of $\mathcal{D}_{h+1}$ (which are both over $\mathcal{X}$) may not be the same, as we will need this flexibility in Section 4. The H parts can also sequentially depend on each other, though samples within each part are i.i.d. While this setup is sufficient for Section 4 and already weaker than the fully i.i.d. setting commonly adopted in the offline RL literature (Chen and Jiang, 2019; Yin and Wang, 2021), in Appendix D.1 we discuss how to relax it to handle more general situations in offline learning.

3.1. Occupancy estimation via importance weights

Recall that value functions satisfy the familiar Bellman equations, allowing us to learn them by approximating Bellman operators via squared-loss regression. The occupancy distributions $\{d_h^\pi\}$ also satisfy the Bellman flow equation: let $\mathbf{P}_h^\pi$ denote the Bellman flow operator, where for any given $d_h : \mathcal{X} \to \mathbb{R}$ and policy π, $(\mathbf{P}_h^\pi d_h)(x_{h+1}) := \iint P_h(x_{h+1}|x_h, a_h)\pi(a_h|x_h) d_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$.[5] d_h^π can be then recursively defined via the Bellman flow equation $d_h^\pi = \mathbf{P}_{h-1}^\pi d_{h-1}^\pi$, with the base case $d_0^\pi = d_0$. (One difference is that value functions are defined bottom-up, whereas occupancies are defined top-down.) Furthermore, in a low-rank MDP, $\mathbf{P}_h^\pi d_h$ is always linear in μ_h^* (Lemma 16), just like the image of Bellman operators for value is always in the linear span of ϕ_h^*.

Given the similarity, one might think that we can also approximate $\mathbf{P}_{h-1}^\pi$ by regressing directly onto the occupancies, hoping to obtain d_h^π via

$$\operatorname*{argmin}_{d} \mathbb{E}_{d_{h-1}^D}\left[\left(d(x_h) - d_{h-1}^\pi(x_{h-1})\tfrac{\pi_{h-1}(a_{h-1}|x_{h-1})}{\pi_{h-1}^D(a_{h-1}|x_{h-1})}\right)^2\right] \quad (1)$$

where $\frac{\pi_{h-1}(a_{h-1}|x_{h-1})}{\pi_{h-1}^D(a_{h-1}|x_{h-1})}$ is the standard importance weighting to correct the mismatch on actions between π_{h-1} and data policy π_{h-1}^D. Unfortunately, this does not work due to the "time-reversed" nature of flow operators (Liu et al., 2018). In fact, the Bayes-optimal solution of Eq. (1) is

$$d_h(x_h) = \frac{(\mathbf{P}_{h-1}^\pi(d_{h-1}^D d_{h-1}^\pi))(x_h)}{d_{h-1}^{D,\dagger}(x_h)} \ne (\mathbf{P}_{h-1}^\pi d_{h-1}^\pi)(x_h).$$

However, the fractional form of the solution indicates that we may instead aim to learn a related function—the impor-

[3] We assume known and deterministic rewards, and can easily handle unknown/stochastic versions (Appendix D.2).

[4] h on the superscript of a policy distinguishes identities and does not refer to the h-th step component (which is indicated by the subscript), that is, ρ^h and $\rho^{h'}$ for $h' \ne h$ can be completely unrelated policies.

[5] In this definition, we do not require d_h to be a valid distribution. Even π is allowed to be unnormalized; see the definition of pseudo-policy in Definition 1.

tance weight, or density ratio (Hallak and Mannor, 2017). If we use $w^\pi_{h-1} = d^\pi_{h-1}/d^D_{h-1}$ to replace d^π_{h-1} as the regression target in Eq. (1), the population solution would be

$$\frac{(\mathbf{P}^\pi_{h-1} d^\pi_{h-1})(x_h)}{d^{D,\dagger}_{h-1}(x_h)} = \frac{d^\pi_h(x_h)}{d^{D,\dagger}_{h-1}(x_h)} =: w^\pi_h(x_h).$$

The occupancy can then be straightforwardly extracted from the weight via elementwise multiplication, i.e., $d^\pi_h = w^\pi_h \cdot d^{D,\dagger}_{h-1}$, where $d^{D,\dagger}_{h-1}$ can be estimated via MLE from the dataset itself.

While this is promising, the approach uses importance weight $w^\pi_h(x_h)$ as an intermediate variable, whose very existence and boundedness rely on the assumption that the data distribution $d^{D,\dagger}_{h-1}$ is exploratory and provides sufficient coverage over d^π_h. We next consider the scenario where such an assumption does *not* hold. Perhaps surprisingly, although we would like to construct exploratory datasets in Section 4 and feed them into the offline algorithm, being able to handle non-exploratory data turns out to be crucial to the online setting, and also yields novel offline guarantees of independent interest.

3.2. Handling insufficient data coverage

Because we make no assumptions about data coverage, the true occupancy d^π_h may be completely unsupported by data, in which case there is no hope to estimate it well. What kind of learning guarantees can we still obtain?

To answer this question, we introduce one of our main conceptual contributions, a novel learning target for occupancy estimation under arbitrary data distributions.

Definition 1 (Pseudo-policy and recursively clipped occupancy)**.** *Given a Markov policy π, data distributions $\{d^D_h\}$, and state and action clipping thresholds $\{C^{\mathbf{x}}_h\}$, $\{C^{\mathbf{a}}_h\}$, the recursively clipped occupancy, $\{\overline{d}^\pi_h\}$, is defined as follows. Let $\overline{d}^\pi_0 := d^\pi_0 = d_0$. Define $\overline{\pi}_h(a_h|x_h) := \pi_h(a_h|x_h) \wedge C^{\mathbf{a}}_h \pi^D_h(a_h|x_h)$ (or $\overline{\pi}_h = \pi_h \wedge C^{\mathbf{a}}_h \pi^D_h$ for short), and for $1 \le h \le H-1$, inductively set* [6]

$$\overline{d}^\pi_h(x_h) := \left(\mathbf{P}^{\overline{\pi}}_{h-1}\left(\overline{d}^\pi_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\right)\right)(x_h). \qquad (4)$$

We also call objects like $\overline{\pi}$ a pseudo-policy*, which can yield unnormalized distributions over actions.*

The above definition first clips the previous-level $\overline{d}^\pi_{h-1}$ to have at most $C^{\mathbf{x}}_{h-1}$ ratio over the data distribution d^D_{h-1} and the policy π to have at most $C^{\mathbf{a}}_{h-1}$ ratio over π^D_{h-1}, then applies the Bellman flow operator. This guarantees that $\overline{d}^\pi_h$ is always supported on the data distribution (unlike d^π_h), and $\overline{d}^\pi_h \le d^\pi_h$ because poorly-supported mass is removed from every level (and hence $\overline{d}^\pi_h$ is generally an unnormalized distribution). Further, when we do have data coverage and the original importance weights on states and actions are always bounded by $\{C^{\mathbf{x}}_h\}$ and $\{C^{\mathbf{a}}_h\}$, it is easy to see that $\overline{d}^\pi_h = d^\pi_h$, since the clipping operations will have no effects and Definition 1 simply coincides with the Bellman flow equation for $\{d^\pi_h\}$.

As we will see below in Section 3.3, $\{\overline{d}^\pi_h\}$ becomes a learnable target and the ℓ_1 estimation error of our algorithm goes to 0 when the sample size $n \to \infty$. The thresholds $\{C^{\mathbf{x}}_h\}$ and $\{C^{\mathbf{a}}_h\}$ reflect a bias-variance trade-off: higher thresholds ensure that less "mass" is clipped away (i.e., $\overline{d}^\pi_h$ will be closer to d^π_h), but result in a worse sample complexity as the algorithm will need to deal with larger importance weights. Below we provide more fine-grained characterization on the bias part, i.e., how $\overline{d}^\pi_h$ is related to d^π_h, and the proof is deferred to Appendix E.2.

Proposition 2 (Properties of $\overline{d}^\pi_h$)**.**

1. $\overline{d}^\pi_h \le d^\pi_h$.

2. $\overline{d}^\pi_h = d^\pi_h$ *when data covers π (i.e., $\forall h' < h$ we have $d^\pi_{h'} \le C^{\mathbf{x}}_{h'} d^D_{h'}$ and $\pi_{h'} \le C^{\mathbf{a}}_{h'} \pi^D_{h'}$).*

3. $\|\overline{d}^\pi_h - d^\pi_h\|_1 \le \|\overline{d}^\pi_{h-1} - d^\pi_{h-1}\|_1 + \|\overline{d}^\pi_{h-1} - \overline{d}^\pi_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\|_1 + \|\mathbf{P}^\pi_{h-1} d^\pi_{h-1} - \mathbf{P}^{\overline{\pi}}_{h-1} d^\pi_{h-1}\|_1$.

The 3rd claim shows how the bias term $\|\overline{d}^\pi_h - d^\pi_h\|_1$ (i.e., how much mass $\overline{d}^\pi_h$ is missing from d^π_h) accumulates over the horizon: the RHS of the bound consists of 3 terms, where the first is missing mass from the previous level, and the other terms correspond to the mass being clipped away from states and actions, respectively, at the current level.

3.3. Algorithm and analyses

We are now ready to introduce our algorithm, FORC, with its analyses and guarantees. See pseudocode in Algorithm 1. The overall structure of the algorithm largely follows the sketch in Section 3.1: we use squared-loss regression to iteratively learn the importance weights (line 5), and convert them to densities by multiplying with the data distributions (line 6) estimated via MLE (line 4).

The major difference is that we introduce clipping in line 5 (in the same way as Definition 1) to guarantee that the regression target is always well-behaved and bounded, and below we show that this makes $\widehat{d}^\pi_h$ a good estimation of $\overline{d}^\pi_h$. In particular, we will bound the *regression error* $\|\widehat{d}^\pi_h - \overline{d}^\pi_h\|_1$ as a function of sample size n_{reg}. A key lemma that enables such a guarantee is the following error propagation result:

[6] Note that $\overline{d}^\pi_h$ depends on hyperparameters $C^{\mathbf{x}}_h$ and $C^{\mathbf{a}}_h$, which are omitted in the notation. Appendix E.1 discusses the relationship between $C^{\mathbf{x}}_h, C^{\mathbf{x}}_a$ and the missingness error, namely, that $\|d^\pi_h - \overline{d}^\pi_h\|_1$ is Lipschitz in, and thus insensitive to misspecifications of, the clipping thresholds.

Algorithm 1 Fitted **O**ccupancy Ite**r**ation with **C**lipping (FORC)

```
Input: policy $\pi$, density feature $\mu^*$, dataset $\mathcal{D}_{0:H-1}$, sample sizes $n_{\text{mle}}$ and $n_{\text{reg}}$, clipping thresholds $\{C^{\mathbf{x}}_h\}$ and $\{C^{\mathbf{a}}_h\}$.
Initialize $\widehat{d}^\pi_0 = d_0$.
for $h = 1, \dots, H$ do
    Randomly split $\mathcal{D}_{h-1}$ to two folds $\mathcal{D}^{\text{mle}}_{h-1}$ and $\mathcal{D}^{\text{reg}}_{h-1}$ with sizes $n_{\text{mle}}$ and $n_{\text{reg}}$, respectively.
    Estimate marginal data distributions $\widehat{d}^D_{h-1}(x_{h-1})$ and $\widehat{d}^{D,\dagger}_{h-1}(x_h)$ by MLE on dataset $\mathcal{D}^{\text{mle}}_{h-1}$:
        $\widehat{d}^D_{h-1} = \operatorname*{argmax}_{d_{h-1} \in \mathcal{F}_{h-1}} \frac{1}{n_{\text{mle}}} \sum_{i=1}^{n_{\text{mle}}} \log\left(d_{h-1}(x^{(i)}_{h-1})\right)$ and $\widehat{d}^{D,\dagger}_{h-1} = \operatorname*{argmax}_{d_h \in \mathcal{F}_h} \frac{1}{n_{\text{mle}}} \sum_{i=1}^{n_{\text{mle}}} \log\left(d_h(x^{(i)}_h)\right)$,   (2)
    where $\mathcal{F}_h = \left\{d_h = \langle \mu^*_{h-1}, \theta_h \rangle : d_h \in \Delta(\mathcal{X}), \theta_h \in \mathbb{R}^{\mathsf{d}}, \|\theta_h\|_\infty \le 1\right\}$.   # $\|\theta_h\|_\infty \le 1$ guarantees $d^D_h \in \mathcal{F}_h$
    Define $\mathcal{L}_{\mathcal{D}^{\text{reg}}_{h-1}}(w_h, w_{h-1}, \overline{\pi}_{h-1}) := \frac{1}{n_{\text{reg}}} \sum_{i=1}^{n_{\text{reg}}} \left(w_h(x^{(i)}_h) - w_{h-1}(x^{(i)}_{h-1}) \frac{\overline{\pi}_{h-1}(a^{(i)}_{h-1}|x^{(i)}_{h-1})}{\pi^D_{h-1}(a^{(i)}_{h-1}|x^{(i)}_{h-1})}\right)^2$, and estimate
        $\widehat{w}^\pi_h = \operatorname*{argmin}_{w_h \in \mathcal{W}_h} \mathcal{L}_{\mathcal{D}^{\text{reg}}_{h-1}}\left(w_h, \frac{\widehat{d}^\pi_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}}{\widehat{d}^D_{h-1}}, \pi_{h-1} \wedge C^{\mathbf{a}}_{h-1} \pi^D_{h-1}\right)$,   (3)
    where $\mathcal{W}_h = \left\{w_h = \frac{\langle \mu^*_{h-1}, \theta^{\text{up}}_h \rangle}{\langle \mu^*_{h-1}, \theta^{\text{down}}_h \rangle} : \|w_h\|_\infty \le C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1}, \theta^{\text{up}}_h, \theta^{\text{down}}_h \in \mathbb{R}^{\mathsf{d}}\right\}$.
    Set the estimate $\widehat{d}^\pi_h = \widehat{w}^\pi_h \, \widehat{d}^{D,\dagger}_{h-1}$.
end for
Output: estimated state occupancies $\{\widehat{d}^\pi_h\}_{h \in [H]}$.
```

Lemma 1. *For every $h \in [H]$, the error between estimates $\widehat{d}^\pi_h$ from Algorithm 1 and the clipped target $\overline{d}^\pi_h$ is decomposed recursively as*

$$\left\|\widehat{d}^\pi_h - \overline{d}^\pi_h\right\|_1 \le \left\|\widehat{d}^\pi_{h-1} - \overline{d}^\pi_{h-1}\right\|_1$$
$$+ 2C^{\mathbf{x}}_{h-1} \left\|\widehat{d}^D_{h-1} - d^D_{h-1}\right\|_1 + C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} \left\|\widehat{d}^{D,\dagger}_{h-1} - d^{D,\dagger}_{h-1}\right\|_1$$
$$+ \sqrt{2} \left\|\widehat{w}^\pi_h - \mathbf{E}^{\overline{\pi}}_{h-1}\left(d^D_{h-1} \frac{\widehat{d}^\pi_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}}{\widehat{d}^D_{h-1}}\right)\right\|_{2, d^{D,\dagger}_{h-1}},$$

where $(\mathbf{E}^\pi_h d_h) := (\mathbf{P}^\pi_h d_h)/d^{D,\dagger}_h$.

The proof can be found in Appendix E.2. The bound consists of 3 parts: the first line is the error at the previous level $h-1$, showing that the regression error accumulatives *linearly* over the horizon. The second line captures errors due to imperfect estimation of the data distributions, since we use the estimated $\widehat{d}^D_{h-1}$ and $\widehat{d}^{D,\dagger}_{h-1}$, instead of the groundtruth distributions, to set up the weight regression problem and extract the density; these errors can be reduced by simply using larger n_{mle}. The last line represents the finite-sample error in regression, which is the difference between the estimated weight $\widehat{w}^\pi_h$ and the Bayes-optimal predictor. We set the constraints in the hypothesis class in a way to guarantee the Bayes-optimal predictor is in the class (see the definition of $\mathcal{W}_h$ below Eq. (3)), so the regression is realizable.

Bounding the complexities of $\mathcal{F}_h$ and $\mathcal{W}_h$ The last challenge is in controlling the statistical complexities of the function classes used in learning, $\mathcal{F}_h$ and $\mathcal{W}_h$, both of which are infinite classes. For $\mathcal{F}_h$, we construct an optimistic covering to bound its covering number (Chen et al., 2022a). For $\mathcal{W}_h$, however, its hypothesis takes the form of ratio between linear functions, $\frac{\langle \mu^*_{h-1}, \theta^{\text{up}}_h \rangle}{\langle \mu^*_{h-1}, \theta^{\text{down}}_h \rangle}$, where standard covering arguments, which discretize θ^{up}_h and θ^{down}_h, run into sensitivity issues, as θ^{down}_h is on the denominator where small perturbations can lead to large changes in the ratio. We overcome this by recalling a technique from Bartlett and Tewari (2006): we bound the pseudo-dimension of $\mathcal{W}_h$, which is equal to the VC-dimension of the corresponding thresholding class. Then, using Goldberg and Jerrum (1993), the VC-dimension is bounded by the syntactic complexity of the classification rule, written as a Boolean formula of polynomial inequality predicates. The pseudo-dimension of $\mathcal{W}_h$ further implies ℓ_1 covering number bounds, for which Dong et al. (2020); Modi et al. (2021) provide fast-rate regression guarantees.

Sample complexity of FORC We now provide the guarantee for FORC, with its proof deferred to Appendix E.2.

Theorem 2 (Offline d^π estimation)**.** *Fix $\delta \in (0, 1)$. Suppose Assumption 1 and Assumption 2 hold, and μ^* is known. Then, given an evaluation policy π, by setting*[7]

[7] While it may appear that we need to set the value of n_{mle} and n_{reg} in a delicate manner, this is not the case and we can simply set $n_{\text{mle}} = n_{\text{reg}} = n/2$ and suffer at most a constant blow-up in the error guarantee. The values given in the theorem statements

$n_{\text{mle}} = \tilde{O}(\mathsf{d}(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h)^2 \log(1/\delta)/\varepsilon^2)$ and $n_{\text{reg}} = \tilde{O}(\mathsf{d}(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h)^2 \log(1/\delta)/\varepsilon^2)$, with probability at least $1-\delta$, FORC *(Algorithm 1) returns state occupancy estimates $\{\widehat{d}^{\pi}_h\}_{h=0}^{H-1}$ satisfying*

$$\|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\|_1 \le \varepsilon, \forall h \in [H].$$

The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}H\left(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h\right)^2 \log(1/\delta)/\varepsilon^2\right).$$

This result can also be used to establish a guarantee for $\|\widehat{d}^{\pi}_h - d^{\pi}_h\|_1$, simply by decomposing $\|\widehat{d}^{\pi}_h - d^{\pi}_h\|_1 \le \|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\|_1 + \|\overline{d}^{\pi}_h - d^{\pi}_h\|_1$. The regression error in the first term is controlled by Theorem 2. The second term is a *one-sided missingness* error due to insufficient coverage of data, which we have characterized in Proposition 2. Note that we split $\|\widehat{d}^{\pi}_h - d^{\pi}_h\|_1$ into two terms using $\overline{d}^{\pi}_h$ as an intermediate quantity and analyze how their errors accumulate over the horizon separately; alternatively, one can directly try to analyze how $\|\widehat{d}^{\pi}_h - d^{\pi}_h\|_1$ depends on $\|\widehat{d}^{\pi}_{h-1} - d^{\pi}_{h-1}\|_1$. In general, we find the latter can yield significantly worse bounds—in fact, *exponentially worse*, as will be seen in Section 4.

Offline policy optimization Theorem 2 provides learning guarantees for $\overline{d}^{\pi}_h$, which is a point-wise lower bound of d^{π}_h. When we consider standard return maximization with a given reward function, having access to $\widehat{d}^{\pi}_h \approx \overline{d}^{\pi}_h$ immediately enables *pessimistic* policy evaluation (Jin et al., 2020c; Xie et al., 2021), and we are only ε-suboptimal compared to the maximal value computed over covered parts of the data, i.e., with respect to $\overline{d}^{\pi}_h$. The immediate implication is that we can compete with the best policy fully covered by data (satisfying property 2 of Proposition 2); see Appendix E.3 for the full statement and proof.

Theorem 3 (Offline policy optimization)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1 and Assumption 2 hold, and μ^* is known. Given a policy class Π, let $\{\widehat{d}^{\pi}_h\}_{h\in[H],\pi\in\Pi}$ be the output of running Algorithm 1. Then with probability at least $1-\delta$, for any reward function R and policy selected as $\widehat{\pi}_R = \operatorname{argmax}_{\pi\in\Pi} \widehat{v}^{\pi}_R$, we have*

$$v^{\widehat{\pi}_R}_R \ge \max_{\pi\in\Pi} \overline{v}^{\pi}_R - \varepsilon,$$

where v^{π}_R and $\widehat{v}^{\pi}_R$ are defined in Proposition 1, and $\overline{v}_R$ is defined similarly for $\{\overline{d}^{\pi}_h\}$. The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}H^3\left(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h\right)^2 \log(|\Pi|/\delta)/\varepsilon^2\right).$$

Computation We remark that our policy optimization result only enjoys statistical efficiency and does not guarantee computational efficiency, as Theorem 3 assumes that we can enumerate over candidate policies and run FORC for each of them; similar comments apply to our later online algorithm as well. Since the optimization variable is a policy, the most promising approach is to come up with off-policy policy-gradient (OPPG) algorithms to approximate the objective. However, existing model-free OPPG methods all rely on value-function approximation (Nachum et al., 2019b; Liu et al., 2019), which is not available in our setting. Studying OPPG with only density(-ratio) approximation will be a pre-requisite for investigating the computational feasibility of our problem, which we leave for future work.

4. Online policy cover construction

We now consider the online setting where the learner explores the MDP to collect its own data. The hope is that we will collect exploratory datasets that provide sufficient coverage for *all* policies in Π (so that we can estimate their occupancies accurately), which is measured by the standard definition of concentrability.

Definition 2 (Concentrability Coefficient (CC))**.** *Given a policy class Π and any distribution $d \in \Delta(\mathcal{X})$, the concentrability coefficient at level h relative to d is*

$$\mathrm{CC}_h(d) = \inf\left\{c \in \mathbb{R} : \max_{\pi\in\Pi}\left\|\frac{d^{\pi}_h}{d}\right\|_\infty \le c\right\}.$$

To achieve this goal, we first recall the following result, which shows the existence of an exploratory data distribution that satisfies the above criterion and hints at how to construct it.

Proposition 3 (Adapted from Chen and Jiang (2019), Prop. 10)**.** *Given a policy class Π and h, let $\{d^{\pi^{h,i}_*}_h\}_{i=1}^{\mathsf{d}}$ be the barycentric spanner (Definition 4 in Appendix I.2) of $\{d^{\pi}_h\}_{\pi\in\Pi}$. Then, $\mathrm{CC}_h\left(\frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} d^{\pi^{h,i}_*}_h\right) \le \mathsf{d}$.*

Proposition 3 shows that for each level h, an exploratory distribution that has d concentrability always exists. It is simply the mixture of $\{d^{\pi^{h,i}_*}_h\}$ for $i \in [\mathsf{d}]$, which can be identified if we have access to d^{π}_h for all $\pi \in \Pi$. Of course, we can only estimate d^{π}_h if we have exploratory data, so the estimation of d^{π}_h and the identification of $\{\pi^{h,i}_*\}$ need to be interleaved to overcome this “chicken-and-egg” problem (Agarwal et al., 2020; Modi et al., 2021): suppose we have already constructed policy cover at $h-1$. We can construct it for the next level as follows:

1. Collect a dataset $\mathcal{D}_{h-1}$ by rolling in to level $h-1$ with the policy cover, with $\mathrm{CC}_{h-1}(d^{D}_{h-1}) \le \mathsf{d}$, then taking a uniformly random action, thereby $\mathrm{CC}_h(d^{D,\dagger}_{h-1}) \le \mathsf{d}K$.

are the most “natural” values based on the the analysis.

2. Use FORC to estimate d_h^{π} for all $\pi \in \Pi$ based on $\mathcal{D}_{h-1}$.
3. Choose their barycentric spanner as the policy cover for level h, with $\mathrm{CC}_h(d_h^D) \leq \mathsf{d}$.

The idea is that, since we have an exploratory distribution at level $h-1$, taking a uniform action afterwards will give us an exploratory distribution at level h, though the degree of exploration will be diluted by a factor of K. We collect data from this distribution to estimate d_h^{π} and compute the barycentric spanner for level h, which will bring the concentrability coefficient back to d, so that the process can repeat inductively.

The above reasoning makes an idealized assumption that d_h^{π} can be estimated perfectly. In such a case, the constructed distribution will provide perfect coverage, so that the clipping introduced in Section 3 becomes completely unnecessary: all clipping operations would be inactive (by setting $C_h^{\mathbf{x}} = \mathsf{d}$ and $C_h^{\mathbf{a}} = K$), and $\overline{d}_h^{\pi} \equiv d_h^{\pi}$. Unfortunately, when the estimation error of d_h^{π} is taken into consideration, the reasoning breaks down seriously.

The first problem is that our estimate $\widehat{d}_h^{\pi}$ from FORC is not necessarily linear due to its product form. However, that is not a concern as we can linearize it (corresponding to line 7 in Algorithm 2); we also have an alternative procedure for FORC that directly produces linear $\widehat{d}_h^{\pi}$ (see Appendix D.3), so in this section we will ignore this issue and pretend that $\widehat{d}_h^{\pi}$ is linear (thus is the same as $\widetilde{d}_h^{\pi}$ in Algorithm 2) for ease of presentation.

4.1. Taming error exponentiation

Now that the issue of (non-)linear $\widehat{d}_h^{\pi}$ is out of the way, we are ready to see where the real trouble is: note that the barycentric spanner computed from $\{\widehat{d}_h^{\pi}\}_{\pi \in \Pi}$ satisfies

$$\left\| \frac{\widehat{d}_h^{\pi}}{\frac{1}{\mathsf{d}} \sum_{i=1}^{\mathsf{d}} \widehat{d}_h^{\pi^{h,i}}} \right\|_{\infty} \leq \mathsf{d}, \quad \forall \pi \in \Pi. \tag{5}$$

However, the actual distribution induced by the policy cover $\{\pi^{h,i}\}_{i=1}^{\mathsf{d}}$ is $d_h^D = \frac{1}{\mathsf{d}} \sum_{i=1}^{\mathsf{d}} d_h^{\pi^{h,i}}$. Suppose for now we have $n_{\text{mle}} = \infty$ for perfect estimation of d_h^D; even then, the regression target in Eq. (3) will no longer be bounded without clipping, as the boundedness of $\widehat{d}/\widehat{d}$ does not imply that of $\widehat{d}/d$, and the latter can be very large or even infinite.

While the unbounded regression target can be easily controlled by clipping, analyzing the algorithm and bounding its error still prove to be very challenging. A natural strategy is to inductively bound $\|\widehat{d}_h^{\pi} - d_h^{\pi}\|_1$ using $\|\widehat{d}_{h-1}^{\pi} - d_{h-1}^{\pi}\|_1$. Unfortunately, this approach fails miserably, as directly analyzing $\|\widehat{d}_h^{\pi} - d_h^{\pi}\|_1$ yields

$$\|\widehat{d}_h^{\pi} - d_h^{\pi}\|_1 \leq (1+\mathsf{d})\|\widehat{d}_{h-1}^{\pi} - d_{h-1}^{\pi}\|_1 + \cdots, \tag{6}$$

implying an $O(\mathsf{d})^H$ exponential error blow-up. (The concrete reason for this failure will be made clear shortly.) In Appendix D.4, we also discuss an alternative approach that "pretends" data to be perfectly exploratory, which only addresses the problem superficially and still suffers $O(\mathsf{d})^H$ error exponentiation, just in a different way. Issues that bear high-level similarities are commonly encountered in level-by-level exploration algorithms, which often demand the so-called reachability assumption (Du et al., 2019, Definition 2.1), which we do not need.

As all the earlier hints allude to, the key to breaking error exponentiation is to split the error using $\overline{d}_h^{\pi}$ into its two sources with very different natures: a "two-sided" regression error $\|\widehat{d}_h^{\pi} - \overline{d}_h^{\pi}\|_1$, and a "one-sided" missingness error $\|\overline{d}_h^{\pi} - d_h^{\pi}\|_1$ (in the sense that $\overline{d}_h^{\pi} \leq d_h^{\pi}$). Because the offline occupancy estimation module of Algorithm 2 is the same as that of Algorithm 1, Lemma 1 still holds (left ×1 chain of Figure 1), implying that $\|\widehat{d}_h^{\pi} - \overline{d}_h^{\pi}\|_1$ can be bounded *irrespective of the data distribution*.

This observation disentangles the regression error from the rest of the analysis, allowing us to focus on bounding the missingness error. For the latter, Proposition 2 also exhibits linear error propagation, as it takes the form of $A_h \leq A_{h-1} + B_{h-1}$ where $A_h = \|\overline{d}_h^{\pi} - d_h^{\pi}\|_1$. However, it still remains to show that the additional error ("B_{h-1}") has no dependence on the inductive error ("A_{h-1}"), otherwise we would still have error exponentiation.[9] This is shown in the following key lemma:

Lemma 4. *For any $h \in [H]$ and $\pi \in \Pi$ in Algorithm 2,*

$$\|\overline{d}_h^{\pi} - d_h^{\pi}\|_1 \leq \|\overline{d}_{h-1}^{\pi} - d_{h-1}^{\pi}\|_1 + 4\mathsf{d} \max_{\pi' \in \Pi} \|\widehat{d}_{h-1}^{\pi'} - \overline{d}_{h-1}^{\pi'}\|_1.$$

To understand this lemma, recall that the additional error in Proposition 2 characterizes the mass clipped away at the current level. This mass can be bounded by the regression error of the previous level ($\max_{\pi' \in \Pi} \|\widehat{d}_{h-1}^{\pi'} - \overline{d}_{h-1}^{\pi'}\|_1$): intuitively, had we had perfect estimation of $\widehat{d}_{h-1}^{\pi'} = \overline{d}_{h-1}^{\pi'}$, our barycentric spanner would also be perfect and we would not need any clipping at all in level h, implying 0 additional error in the bound. More generally, the closer $\widehat{d}_{h-1}^{\pi'}$ is to $\overline{d}_{h-1}^{\pi'}$, the less mass we need to clip away.

That said, this term is not instantaneous and depends inductively on quantities in the previous time step, still raising concerns of error exponentiation. To see why this is not a problem, we visualize error propagation in Figure 1: it can be clearly seen that such a dependence corresponds to a "cross-edge", and appears at most once along any long chain. This also explains the destined failure of directly

[9] For example, if B_{h-1} can only be bounded as $B_{h-1} \leq A_{h-1}$, we would still have $A_h \leq 2A_{h-1}$.

Algorithm 2 **FORC**-guided **E**xploration (FORCE)

```
Input: policy class $\Pi$, density feature $\mu^*$, $n = n_{\text{mle}} + n_{\text{reg}}$.
Initialize $\widehat{d}_0^\pi = d_0$ and $\widetilde{d}_0^\pi = d_0, \forall \pi \in \Pi$.
for $h = 1, \ldots, H$ do
    Construct $\{\widetilde{d}_{h-1}^{\pi^{h-1,i}}\}_{i=1}^{\mathsf{d}}$ as the barycentric spanner of $\{\widetilde{d}_{h-1}^\pi\}_{\pi\in\Pi}$, and set $\Pi_{h-1}^{\text{expl}} = \{\pi^{h-1,i}\}_{i=1}^{\mathsf{d}}$.
    Draw a tuple dataset $\mathcal{D}_{h-1} = \{(x_{h-1}^{(i)}, a_{h-1}^{(i)}, x_h^{(i)})\}_{i=1}^n$ using unif$(\Pi_{h-1}^{\text{expl}}) \circ$ unif$(\mathcal{A})$.
    for $\pi \in \Pi$ do
        Estimate $\widehat{d}_h^\pi$ using the $h$-level loop[8] of Algorithm 1 (lines 4-6) with $\mathcal{D}_{h-1}$, $\widehat{d}_{h-1}^\pi$, $C_{h-1}^{\mathbf{x}} = \mathsf{d}$, $C_{h-1}^{\mathbf{a}} = K$.
        Find the closest linear approximation $\widetilde{d}_h^\pi = \langle \mu_{h-1}^*, \widetilde{\theta}_h \rangle$ where $\widetilde{\theta}_h = \operatorname{argmin}_{\theta_h \in \mathbb{R}^{\mathsf{d}}} \|\langle \mu_{h-1}^*, \theta_h \rangle - \widehat{d}_h^\pi\|_1$.
    end for
end for
Output: estimated state occupancy measure $\{\widehat{d}_h^\pi\}_{h\in[H],\pi\in\Pi}$.
```

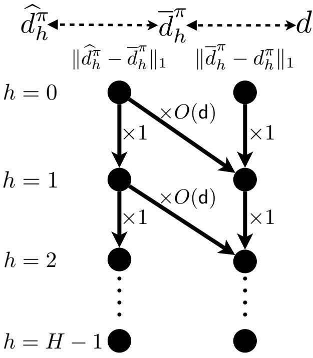

Figure 1. Error propagation diagram for FORCE. "$\bullet \rightarrow \bullet$" with $\times c$ means $(\bullet) \le c \times (\bullet)$ + (other instantaneous errors that do not accumulate over horizon), and multiple incoming arrows imply sum of errors. The left $\times 1$ chain is from Lemma 1, the right $\times 1$ chain from Proposition 2, and the $\times O(\mathsf{d})$ edges from Lemma 4.

analyzing $\|\widehat{d}_h^\pi - d_h^\pi\|_1$ in Eq. (6), as that corresponds to merging the two chains into one, where every edge along the only chain acquires an $O(\mathsf{d})$ multiplicative factor.

With this, we can now state the formal guarantee for our algorithm, FORCE. See Algorithm 2 for its pseudo-code, and the proof of the guarantee is deferred to Appendix F.1.

Theorem 5 (Online d^π estimation)**.** *Fix $\delta \in (0,1)$ and consider an MDP $\mathcal{M}$ that satisfies Assumption 1, and μ^* is known. Then by setting $n_{\text{mle}} = \widetilde{O}\left(\mathsf{d}^3 K^2 H^4 \log(1/\delta)/\varepsilon^2\right)$, $n_{\text{reg}} = \widetilde{O}\left(\mathsf{d}^5 K^2 H^4 \log(|\Pi|/\delta)/\varepsilon^2\right)$, $n = n_{\text{mle}} + n_{\text{reg}}$, with probability at least $1-\delta$,* FORCE *returns state occupancy estimates $\{\widehat{d}_h^\pi\}_{h\in[H],\pi\in\Pi}$ satisfying that*

$$\left\|\widehat{d}_h^\pi - d_h^\pi\right\|_1 \le \varepsilon, \forall h \in [H], \pi \in \Pi.$$

The total number of episodes required by the algorithm is

$$\widetilde{O}(nH) = \widetilde{O}\left(\mathsf{d}^5 K^2 H^5 \log(|\Pi|/\delta)/\varepsilon^2\right).$$

Theorem 5 also immediately translates to a policy optimization guarantee when combined with Proposition 1:

Theorem 6 (Online policy optimization)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1 and Assumption 2 hold, and μ^* is known. Given a policy class Π, let $\{\widehat{d}_h^\pi\}_{h\in[H],\pi\in\Pi}$ be the output of running* FORCE*. Then with probability at least $1-\delta$, for any reward function R and policy selected as $\widehat{\pi}_R = \operatorname{argmax}_{\pi\in\Pi} \widehat{v}_R^\pi$, we have*

$$v_R^{\widehat{\pi}_R} \ge \max_{\pi\in\Pi} v_R^\pi - \varepsilon,$$

where v_R^π and $\widehat{v}_R^\pi$ are defined in Proposition 1. The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}^5 K^2 H^7 \log(|\Pi|/\delta)/\varepsilon^2\right).$$

The proof is deferred to Appendix F.2. We remark that Theorem 6 is a *reward-free* learning guarantee (Jin et al., 2020a; Chen et al., 2022b,a), and it is easy to see that Algorithm 2 is deployment efficient (Huang et al., 2022).

5. Representation learning

In this section, we extend the offline (Section 3) and online (Section 4) results to the representation learning setting. Here, the true density feature μ^* is unknown, but the learner has access to a realizable density feature class Υ, defined formally below. For simplicity, we consider finite and normalized Υ, as is standard in the literature (Agarwal et al., 2020; Modi et al., 2021; Uehara et al., 2021b).

Assumption 3. *We have a finite density feature class $\Upsilon = \bigcup_{h\in[H]} \Upsilon_h$ such that $\mu_h^* \in \Upsilon_h$ for each $h \in [H]$, thus $\mu^* \in \Upsilon$. Further, for any $\mu_h \in \Upsilon_h$, we have $\int \|\mu_h(x)\|_1 (\mathrm{d}x) \le B^\mu$.*

The algorithms and analyses for the representation learning case mostly follow the same template as the known feature case, so we restrict our discussion to their differences. Recall that, in order to have realizable function

[9] MLE only needs to be done once and not for every $\pi \in \Pi$.

classes for regression and MLE in Section 3, we constructed $\mathcal{F}_h, \mathcal{W}_h$ using functions linear in the known μ^*_{h-1}. In order to maintain this realizability when μ^*_{h-1} is unknown, we instead construct $\mathcal{F}_h, \mathcal{W}_h$ using the union of all functions linear in some candidate $\mu_{h-1} \in \Upsilon_{h-1}$, i.e., $\bigcup_{\mu_{h-1}\in\Upsilon_{h-1}}\{\langle \mu_{h-1}, \theta_h\rangle, \theta_h \in \mathbb{R}^{\mathsf{d}}\}$ (see Eq. (28) and Eq. (29) for their formal definitions).

While such union classes allow most of Section 3 and Section 4 to straightforwardly extend to the representation learning setting, a nontrivial modification must be made to the online algorithm. Recall in line 7 of Algorithm 2, we constructed our policy cover using the barycentric spanner of $\{\widetilde{d}^\pi_h\}_{\pi\in\Pi}$, the set of linearized approximations to the density estimates. Importantly, this guaranteed a concentrability coefficient of d because all $\widetilde{d}^\pi_h$ are linear in the same feature μ^*_{h-1}. This is no longer the case with unknown features because, if linearized in the same way (but over all feasible $\mu_{h-1} \in \Upsilon_{h-1}$), each $\widetilde{d}^\pi_h$ can be composed of a different μ_{h-1} feature, resulting in a CC linear in $|\Pi|$. To overcome this issue, we replace line 7 with the following "joint linearization" step (see line 8 in Algorithm 4):

$$\widehat{\mu}_{h-1} = \min_{\mu_{h-1}\in\Upsilon_{h-1}} \max_{\pi\in\Pi} \min_{\theta_h\in\mathbb{R}^{\mathsf{d}}} \|\langle \mu_{h-1}, \theta_h\rangle - \widehat{d}^\pi_h\|_1,$$

where all density estimates are linearized using a single feature $\widehat{\mu}_{h-1}$, whose linear span approximates all $\widehat{d}^\pi_h$ well. We provide theorems for offline/online d^π estimation with representation learning below.

Theorem 7 (Offline d^π estimation with representation learning)**.** *Fix $\delta \in (0,1)$. Suppose Assumption 1, Assumption 2, and Assumption 3 hold. Given evaluation policy π, by setting $n_{\text{mle}} = \tilde{O}(\mathsf{d}(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h)^2 \log(|\Upsilon|/\delta)/\varepsilon^2)$ and $n_{\text{reg}} = \tilde{O}(\mathsf{d}(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h)^2 \log(|\Upsilon|/\delta)/\varepsilon^2)$, with probability at least $1-\delta$,* ForcRl *(Algorithm 3) returns state occupancy estimates $\{\widehat{d}^\pi_h\}_{h=0}^{H-1}$ satisfying that*

$$\left\|\widehat{d}^\pi_h - \overline{d}^\pi_h\right\|_1 \le \varepsilon, \forall h \in [H].$$

The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}H\left(\textstyle\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h\right)^2 \log(|\Upsilon|/\delta)/\varepsilon^2\right).$$

Theorem 8 (Online d^π estimation with representation learning)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1 and Assumption 3 hold. By setting $n = n_{\text{mle}} + n_{\text{reg}}$, $n_{\text{mle}} = \widetilde{O}(\mathsf{d}^3 K^2 H^4 \log(|\Upsilon|/\delta)/\varepsilon^2)$, $n_{\text{reg}} = \widetilde{O}(\mathsf{d}^5 K^2 H^4 \log(|\Pi||\Upsilon|/\delta)/\varepsilon^2)$, with probability at least $1-\delta$,* ForcRle *(Algorithm 4) returns state occupancy estimates $\{\widehat{d}^\pi_h\}_{h=0}^{H-1}$ satisfying that*

$$\|\widehat{d}^\pi_h - d^\pi_h\|_1 \le \varepsilon, \forall h \in [H], \pi \in \Pi.$$

The total number of episodes required by the algorithm is

$$\widetilde{O}\left(\mathsf{d}^5 K^2 H^5 \log(|\Pi||\Upsilon|/\delta)/\varepsilon^2\right).$$

The detailed proofs of these two theorems are given in Appendix G. We also present the theorems and proofs for offline/online policy optimization with representation learning as well as the formal representation learning algorithms in Appendix G.

6. Conclusion

We have shown how to leverage density features for statistically efficient state occupancy estimation and reward-free exploration in low-rank MDPs, culminating in policy optimization guarantees. An important open problem lies in investigating the computational efficiency of our algorithms (e.g., through off-policy policy gradient).

Acknowledgements

The authors thank Akshay Krishnamurthy and Dylan Foster for discussions related to MLE generalization error bounds. NJ acknowledges funding support from NSF IIS-2112471 and NSF CAREER IIS-2141781.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship Learning via Inverse Reinforcement Learning. In *Proceedings of the 21st International Conference on Machine learning*, page 1. ACM, 2004.

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

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in Neural Information Processing Systems*, 2020.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Peter Bartlett and Ambuj Tewari. Sample complexity of policy search with known dynamics. *Advances in Neural Information Processing Systems*, 19, 2006.

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022a.

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

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

Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. In *Advances in Neural Information Processing Systems*, 2022b.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. $\sqrt{n}$-regret for learning in Markov decision processes with function approximation and low Bellman rank. In *Conference on Learning Theory*, 2020.

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

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

Fei Feng, Ruosong Wang, Wotao Yin, Simon S Du, and Lin Yang. Provably efficient exploration for reinforcement learning using unsupervised learning. *Advances in Neural Information Processing Systems*, 33:22492–22504, 2020.

Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

Paul Goldberg and Mark Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 361–369, 1993.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pages 1372–1383. PMLR, 2017.

Elad Hazan, Sham M Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Audrey Huang and Nan Jiang. Beyond the return: Off-policy function estimation under user-specified error-measuring distributions. In *Advances in Neural Information Processing Systems*, 2022.

Jiawei Huang, Jinglin Chen, Li Zhao, Tao Qin, Nan Jiang, and Tie-Yan Liu. Towards deployment-efficient reinforcement learning: Lower bound and optimality. In *International Conference on Learning Representations*, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020a.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020c.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv:2102.07035*, 2021.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. *arXiv preprint arXiv:2202.01511*, 2022.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline rl with general function approximation. *arXiv preprint arXiv:2212.13861*, 2022.

Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021a.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2021b.

Sara A Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Vladimir Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34: 4065–4078, 2021.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

Tong Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

A. Related works

In this section, we discuss a few lines of related work in detail.

First, the closest related works involve RL with unsupervised-learning oracles (Du et al., 2019; Feng et al., 2020). Instead of investigating low-rank MDPs, they consider more restricted block MDPs and need stronger assumptions such as reachability, identifiability, and separatability (we refer the reader to their works for the definitions). Their notion of "decoder" looks like density features in low-rank MDPs, but they are incomparable. The crucial property of "decoder" is that it is a map from the $\mathcal{X}$ space to the low d dimensional space. This map itself no longer exists in low-rank MDPs. In addition, the density feature serves a different purpose in our paper, as its primary purpose is for constructing the weight function class.

A second line of related work is model-based representation learning in low-rank MDPs (Agarwal et al., 2020; Uehara et al., 2021b; Ren et al., 2022), which assumes that both a realizable left feature class $\Phi \ni \phi^*$ and realizable density (right) feature class $\Upsilon \ni \mu^*$ are given to the learner, essentially inducing a realizable dynamics model class. The learned model (features) are subsequently used for downstream planning. In comparison, we utilize a much weaker inductive bias as we only require a realizable density feature class Υ, and we do not try to learn a dynamics model. Though we additionally need a policy class Π, this is a very basic and natural function class to include. It can be immediately obtained from the (Q-)value function class in the value-based approach, and from the dynamics model class (given a reward function) in the model-based approach above. In terms of the algorithm design, we also use MLE, but for a different objective (the data distribution, instead of the dynamics model).

The importance weight (density-ratio) learning used within our algorithms is related to the marginalized importance sampling of the offline RL algorithms in Nachum et al. (2019a); Lee et al. (2021); Uehara et al. (2021a); Zhan et al. (2022); Chen and Jiang (2022); Huang and Jiang (2022); Ozdaglar et al. (2022). These works do not make the low-rank MDP assumption and study the problem in general MDPs, and require both a weight function class and value function class for learning. We leverage the true density μ^* or density feature class Υ to construct the realizable weight function class, allowing us to achieve statistically faster rates in the low-rank MDP setting. We do not need a value function class and instead only need a weaker (as discussed in the previous paragraph) policy class Π. Lastly, we note that the aforementioned works all learn weights, while our goal is to learn the densities. Extracting the densities from the weights allows us to efficiently explore the MDP using its low-dimensional structure, and additionally enables our return maximization guarantees of Proposition 1 by separating them from the underlying data distribution.

B. Hardness result without the policy class

In this section, we show that without policy class Π, learning in low-rank MDPs (or an easier simplex feature setting) is provably hard even when the true density feature μ^* is known to the learner. The crux is that low-rank MDPs can readily emulate a fully general contextual bandit problem, where μ^* is useless. For the hardness result, we adapt Theorem 2 of Dann and Brunskill (2015) to our case by only keeping their second to third level to get a contextual bandit problem.

To provide specifics for the reward and transition functions, we first note that the subscript of the reward/transition function denotes which level it applies to (e.g., P_0 are the transitions to x_1 from x_0). Level $h = 0$ is composed of $|\mathcal{X}| - 3$ states with zero reward, i.e., $x_0 \in \{1, \ldots, |\mathcal{X}| - 3\}$ and $R_0(i) = 0, \forall i \in \{1, \ldots, |\mathcal{X}| - 3\}$. Level $h = 1$ is composed of 2 states, i.e., $x_1 \in \{+, -\}$, where $R_1(+) = 1$ and $R_1(-) = 0$. Lastly, at level $h = 2$ we have a single null absorbing state x_2.

For the transition functions, in level $h = 0$ the transitions P_0 are Bernoulli distributions where for any state $i \in \{1, \ldots, |\mathcal{X}| - 3\}$ and action $a_0 \in \mathcal{A}$, we have $P_0(+|i, a_0) = \frac{1}{2} + \varepsilon'_i(a_0)$ and $P_0(-|i, a_0) = \frac{1}{2} - \varepsilon'_i(a_0)$. Here, ε'_i is defined in a per-state manner given a parameter ε. We have $\varepsilon'_i(a_0) = \varepsilon/2$ if $a_0 = a_0^*$, where a_0^* is a fixed action; $\varepsilon'_i(a_0) = \varepsilon$ if $a_0 = a_0^{i,*}$ where $a_0^{i,*}$ is an unknown action defined per state i; and $\varepsilon'_i(a_0) = 0$ otherwise. In level $h = 1$, the transitions P_1 simply transmit deterministically to the absorbing state x_2, i.e., $P_1(x_2|x_1, a_1) = 1$ for all $x_1 \in \{+, -\}$ and $a_1 \in \mathcal{A}$.

It is easy to see that the dynamics of this contextual bandit can be modeled using simplex features, thus it is an instantiation of low-rank MDPs. Since we only have two levels ($H = 2$), we only need to verify that P_0 and P_1 can be written in the desired form (Assumption 1). In level $h = 0$, we add two latent states corresponding to the rewarding and non-rewarding state, thus $\mathsf{d} = 2$. Then in level $h = 0$, we have right features $\mu_0^*(+) = [1, 0]$ and $\mu_0^*(-) = [0, 1]$, and left features $\phi_0^*(x_0, a_0) = [P_1(+|x_0, a_0), P_1(-|x_0, a_0)]$ for any (x_0, a_0), corresponding to the original Bernoulli distribution. It is easy to see that this satisfies Assumption 1, i.e., for any (x_0, a_0, x_1) we have $P_0(x_1|x_0, a_0) = \langle \phi_0^*(x_0, a_0), \mu_0^*(x_1) \rangle$. In level $h = 1$ we can simply set a single latent state representing the singleton x_2, and observe that Assumption 1 is trivially satisfied with $\mu_1^*(x_2) = 1$, and $\phi_1^*(x_1, a_1) = 1$ for any (x_1, a_1).

Finally, from Theorem 2 of Dann and Brunskill (2015), we know that the sample complexity of learning in this contextual bandit problem is $\Omega(|\mathcal{X}|)$, demonstrating that efficient learning is impossible in low-rank MDPs (or the simplex feature setting) given only μ^*.

The necessity of $K = |\mathcal{A}|$ dependence It is well known that learning contextual bandits with just a policy class requires a dependence on $|\mathcal{A}|$ in regret and sample complexity; see Agarwal et al. (2014) and the references therein. This can also be reproduced in the above hardness result: first, we can scale up the construction by adding more actions, and show an $\Omega(|\mathcal{X}|K)$ lower bound. Second, we now provide the learner with a policy class that contains all Markov deterministic policies. The size of the class is $O(K^{|\mathcal{X}|})$, and the log-size is $O(|\mathcal{X}|\log(K))$. Given the logarithmic dependence on K, no polynomial dependence on $\log(|\Pi|)$ can explain away the linear-in-K dependence in the lower bound, and we must introduce K as a separate factor in the sample complexity.

C. RL with objectives on state distributions

Proposition 1 also extends to general optimization objectives $f(\{d_h\})$ that are Lipschitz in the input $\{d_h\}$ (note the Lipschitz property does not require the input to be a valid distribution). This Lipschitzness property is key for many recent results in convex RL (Zahavy et al., 2021; Mutti et al., 2022), and also holds for return maximization where $f(\{d_h^\pi\}) = v_R^\pi$, in which case the Lipschitz constant is related to the maximum reward $\max_{h,x,a} R_h(x, a)$. While we write the objective $f(\{d_h\})$ using state densities $d_h(x_h)$ as input for simplicity, it is straightforward to instead use state-action densities $d_h(x_h)\pi(a_h|x_h)$ formed by directly composing the state density d_h with the policy π. If f is Lipschitz in state-action densities, it will still be Lipschitz in the state-action densities in the ℓ_1 norm, which is the exactly the case in return maximization, since any input density will be composed with same π. Lastly, we note that constraints can also be added to the objective and to result in a similar statement.

Proposition 4. *Suppose the optimization objective is $f(\{d_h\})$, where f is Lipschitz in $\{d_h\}$ under the ℓ_1 norm, i.e., there exists a constant $L > 0$ such that for any $\{d'_h\}$ and $\{d''_h\}$*

$$|f(\{d'_h\}) - f(\{d''_h\})| \le L \sum_{h\in[H]} \|d'_h - d''_h\|_1.$$

Then for $\{\widehat{d_h^\pi}\}$ such that $\|\widehat{d_h^\pi} - d_h^\pi\|_1 \le \frac{\varepsilon}{2H}$ for all $\pi \in \Pi$ and $h \in [H]$, and $\widehat{\pi}$ maximizing the plug-in estimate of the objective:

$$\widehat{\pi} = \operatorname*{argmax}_{\pi\in\Pi} f(\{\widehat{d_h^\pi}\}),$$

we have

$$f(\{d_h^{\widehat{\pi}}\}) \ge \max_{\pi\in\Pi} f(\{d_h^\pi\}) - L\varepsilon.$$

Proof. For any $\pi \in \Pi$, from the Lipschitz assumption,

$$\left| f(\{d_h^\pi\}) - f(\{\widehat{d_h^\pi}\}) \right| \le L \sum_{h\in[H]} \|d_h^\pi - \widehat{d_h^\pi}\|_1 \le L\varepsilon/2.$$

Then, letting $\pi^* = \operatorname{argmax}_{\pi\in\Pi} f(\{d_h^\pi\})$ denote the maximizer of the true objective and using the above inequality,

$$f(\{d_h^{\widehat{\pi}}\}) - f(\{d_h^{\pi^*}\}) = f(\{d_h^{\widehat{\pi}}\}) - f(\{\widehat{d_h^{\widehat{\pi}}}\}) + f(\{\widehat{d_h^{\widehat{\pi}}}\}) - f(\{\widehat{d_h^{\pi^*}}\}) + f(\{\widehat{d_h^{\pi^*}}\}) - f(\{d_h^{\pi^*}\}) \ge -L\varepsilon. \qquad \square$$

On $\widehat{d_h^\pi}$ being invalid distributions One potential issue is that some of the objective functions f considered in the literature are only well defined for valid probability distributions (e.g., entropy). This is easy to deal with in the online setting, as we can simply project $\widehat{d_h^\pi}$ onto the probability simplex, which picks up a multiplicative factor of 2 in $\|\widehat{d_h^\pi} - d_h^\pi\|_1$ (c.f. the analysis of the linearization step in Algorithm 2).

For the offline setting, however, the situation can be trickier. For example, the above projection idea is clearly bad for return maximization, since after projection all $\widehat{d_h^\pi}$ satisfy $\|\widehat{d_h^\pi}\|_1 = 1$ and we lose pessimism. From an analytical point of view, pessimistic approaches (e.g., Theorem 3) only pays one factor of the missingness error $\|\overline{d}_h^\pi - d_h^\pi\|_1$ by leveraging its one-sidedness, and a factor of 2 introduced by projection is simply unacceptable. Therefore, the question is whether we can generalize the pessimism in Theorem 3 to general objective functions.

We answer this question with a rough sketch without detailed proofs: since we know $\|\widehat{d}_h^\pi - \overline{d}_h^\pi\|_1 \le \varepsilon'$ (for some appropriate value of ε' from our analysis), we can form a version space for d_h^π as (see also Appendix D.3.6 for a tighter approach to forming version spaces):

$$d_h^\pi \in \{d_h = \langle \mu_{h-1}^*, \theta_h \rangle : d_h \in \Delta(\mathcal{X}),\ \|[\widehat{d}_h^\pi - d_h]_+\|_1 \le \varepsilon'\} \tag{7}$$

where $[\cdot]_+ := \max(\,\cdot\,, 0)$ is used to capture the part that $\widehat{d}_h^\pi$ "exceeds" d_h, choosing normalized d_h that is approximately a pointwise upper bound of d_h. d_h^π is in the set because $\|[\widehat{d}_h^\pi - d_h^\pi]_+\|_1 \le \|[\widehat{d}_h^\pi - \overline{d}_h^\pi]_+\|_1 \le \|\widehat{d}_h^\pi - \overline{d}_h^\pi\|_1 \le \varepsilon'$, and the inequalities here show that $\|[(\cdot)]_+\|_1$ behaves like a one-sided (and hence assymetric) version of ℓ_1 error between unnormalized distributions. Then we can simply come up with pessimistic evaluation of $f(\{d_h^\pi\})$ by minimizing $f(\{d_h\})$ over the above set. It is not hard to see that such an approach will provide similar guarantees to Theorem 3 when applied to return maximization.

D. Alternative setups, algorithm designs, and analyses

D.1. Offline data assumptions

As mentioned in Section 3, our offline data assumption allows sequentially dependent batches, where in-batch tuples are i.i.d. samples. This is already weaker than the standard fully i.i.d. settings considered in the offline RL literature, and here we further comment on how to handle various extensions.

Trajectory data One simple setting is when data are i.i.d. trajectories sampled from a fixed policy. (This setting does not fit our need for the online algorithm, but is a representative setup for the purpose of offline learning.) While our protocol directly handles it (we can simply split the data in H chunks and call them $\mathcal{D}_0, \mathcal{D}_1, \ldots$), it seems somewhat wasteful as we only extract 1 transition tuple per trajectory, potentially worsening the sample complexity by a factor of H. This is because in our analysis of the regression step (Algorithm 1, line 5), we treat the regression target (which depends on $\widehat{d}_h^\pi$) as fixed and independent of the current dataset. If we want to use all the data, we would need to union bound over the target as well; see similar considerations in the work of Fan et al. (2020). A slow-rate analysis follows straightforwardly, and we leave the investigation of fast-rate analysis to future work. We also remark that our current offline setup (Assumption 2) is the most natural protocol for the data collected from the online algorithm (Section 4), and using full trajectory data does not seem to improve the theoretical guarantees of the online setting.

Fully adaptive data A more general setting than Assumption 2 is that the data is fully adaptive, i.e., each trajectory is allowed to depend on all trajectories that before it. To handle such a case, we will need to replace the i.i.d. concentration inequalities with their martingale versions. Some special treatment in the concentration bounds will also be needed to handle the random data-splitting step in Algorithm 1, line 3 (c.f. Mohri and Rostamizadeh, 2008); alternatively, if we union bound over regression targets (see previous paragraph), the data splitting step will no longer be needed.

Unknown and/or non-Markov π^D In Assumption 2 we assume that the last-step policy in the data-collecting policy is Markov and known, as we need it to form the importance weights on actions. When π^D is still Markov and unknown, we can use behavior cloning to back it out from data, which would require some additional assumptions (e.g., having access to a policy class that realizes π^D), and we do not further expand on such an analysis. When π^D is non-Markov, it is well known that the action in the data tuple (x_h, a_h, x_{h+1}) can be still treated as if it were generated from a Markov policy—one can compute the state-action occupancy for (x_h, a_h) (which is well-defined even if π^D is non-Markov) and then obtain the equivalent Markov policy by conditioning on x_h. Incidentally, the algorithmic solution is the same as the case of unknown Markov π^D, i.e., behavior cloning.

D.2. Stochastic and/or unknown reward functions

When the reward function is stochastic but still known, Proposition 1 and all policy optimization guarantees extend straightforwardly, since we can still directly compute the return. The more nontrivial case is when the reward function R is unknown and comes as part of the data, i.e., we have the usual format of data tuples that include (possibly) stochastic reward signals, $\{(x_h^{(i)}, a_h^{(i)}, r_h^{(i)})\}_{i=1}^{n_{\text{ret}}} \sim d_h^D$. Then given estimates $\{\widehat{d}_h^D\}$ (from MLE) and $\{\widehat{d}_h^\pi\}$ (from Algorithm 1 or Algorithm 2), the expected return can be estimated by reweighting the rewards according to the importance weight $\widehat{d}_h^\pi / \widehat{d}_h^D$, and assuming this ratio is well-defined:

$$\widehat{v}_R^\pi = \frac{1}{n_{\text{ret}}} \sum_{i=1}^{n_{\text{ret}}} \sum_{h \in [H]} \frac{\widehat{d}_h^\pi(x_h^{(i)})}{\widehat{d}_h^D(x_h^{(i)})} \frac{\pi_h(a_h^{(i)}|x_h^{(i)})}{\pi_h^D(a_h^{(i)}|x_h^{(i)})} r_h^{(i)}.$$

It can be shown that we then have $|\widehat{v}_R^\pi - v_R^\pi| \leq \varepsilon +$ (additive terms), where the additive terms correspond to the statistical error of return and MLE estimation, which is $O((n_{\text{ret}})^{-1/2})$. If $\widehat{d}_h^D$ does not cover $\widehat{d}_h^\pi$, which may generally be the case, clipping (e.g., according to thresholds $C_h^{\mathbf{x}}, C_h^{\mathbf{a}}$) can again be used, which will lead to additional error corresponding to clipped mass.

D.3. Algorithm design and analyses

In this section, we discuss alternative designs of the offline density learning algorithm (Algorithm 1), as well as their downstream impacts on the online and representation learning algorithms, which use the offline module in their inner loops. For simplicity, most discussions are in the case of offline density learning with known representation μ^*.

D.3.1. Point estimate in denominator

First, we discuss alternative parameterizations of the weight function class. To enable more "elementary" ℓ_∞ covering arguments, one may consider instead parameterizing the weight function class as a ratio of linear functions over a fixed function $v_h : \mathcal{X} \to \mathbb{R}$, specifically

$$\mathcal{W}_h(v_h) = \left\{ w_h = \frac{\langle \mu_{h-1}^*, \theta_h \rangle}{v_h} : \|w_h\|_\infty \leq C_{h-1}^{\mathbf{x}} C_{h-1}^{\mathbf{a}}, \theta_h \in \mathbb{R}^{\mathsf{d}} \right\}.$$

When μ^* consists of simplex features, it can be shown that an ℓ_∞ covering with scale γ of size $(1/\gamma)^{\mathsf{d}}$ can be constructed for $\mathcal{W}_h(v_h)$, because it can be induced by an ℓ_∞ covering of the low-dimensional parameter space that has scale adaptively chosen according to how much the weight can be perturbed with respect to the denominator, thus fixed size. It is unclear how to construct such ℓ_∞ coverings for "linear-over-linear" function classes such as $\mathcal{W}_h$ of Algorithm 1. One may consider compositions of standard ℓ_∞ coverings generated separately for the linear numerator and denominator, but bounding the covering error is challenging due to sensitivity of the denominator to perturbations.

As we will see, however, the key issue with such fixed-denominator parameterizations is that the Bayes-optimal solution is no longer realizable. To handle this in the analysis, we can introduce an additional *approximation error* (similar to Chen and Jiang (2019, Assumption 3) in the value learning setting) that will appear in the final bound, corresponding to how well the Bayes-optimal solution is approximated by the function class. Depending on the choice of denominator, the approximation error may not be controlled, or may lead to a slower rate of estimation; loosely, it is defined as

$$\varepsilon_h^{\text{approx}} = \max_{w_{h-1}: \|w_{h-1}\|_\infty \leq C_{h-1}^{\mathbf{x}}} \min_{w_h \in \mathcal{W}_h(v_h)} \left\| w_h - \mathbf{E}_{h-1}^\pi(d_{h-1}^D w_{h-1}) \right\|_{2, d_{h-1}^{D,\dagger}}.$$

One obvious choice for the fixed denominator is $v_h = \widehat{d}_{h-1}^{D,\dagger}$, since it is immediately available from the MLE data estimation step, plus the linear numerator can then be extracted exactly through the elementwise multiplication $\widehat{d}_h^\pi = \widehat{w}_h^\pi \widehat{d}_{h-1}^{D,\dagger}$. However, the Bayes-optimal predictor $\mathbf{E}_{h-1}^\pi(d_{h-1})$ is no longer realizable, since $\mathbf{E}_{h-1}^\pi(d_{h-1}) = \mathbf{P}_{h-1}^\pi(d_{h-1})/d_{h-1}^{D,\dagger}$ is a linear function over the true data distribution $d_{h-1}^{D,\dagger}$. In this case, using Lemma 19 gives a more interpretable upper bound on the approximation error involves the difference between the ratio of any linear d_h covered on $d_{h-1}^{D,\dagger}$ and the corresponding ratio over $\widehat{d}_{h-1}^{D,\dagger}$:

$$\varepsilon_h^{\text{approx}} \leq \max_{\substack{d_h = \langle \mu_{h-1}^*, \theta_h \rangle : \\ d_h \leq C_{h-1}^{\mathbf{x}} C_{h-1}^{\mathbf{a}} d_{h-1}^{D,\dagger}}} \left\| \frac{d_h}{\widehat{d}_{h-1}^{D,\dagger}} - \frac{d_h}{d_{h-1}^{D,\dagger}} \right\|_{2, d_{h-1}^{D,\dagger}}.$$

However such approximation error may be difficult to control even with small data estimation error due to sensitivity of the denominator (for example if $\|\widehat{d}_{h-1}^{D,\dagger} - d_{h-1}^{D,\dagger}\|_1 \leq \varepsilon_{\text{mle}}$ but they have disjoint support).

D.3.2. Barycentric spanner in denominator

To avoid the above support issue and control the approximation error, we can instead consider a denominator function upon which $d^{D,\dagger}_{h-1}$ is supported. This is satisfied by the barycentric spanner of the version space of the estimate $\widehat{d}^{D,\dagger}_{h-1}$,

$$\mathcal{V}_h = \left\{ v_h = \langle \mu^*_{h-1}, \theta_h \rangle : \|v_h - \widehat{d}^{D,\dagger}_{h-1}\|_1 \le \varepsilon_{\text{mle}}, \theta_h \in \mathbb{R}^{\mathsf{d}} \right\},$$

noting that $d^{D,\dagger}_{h-1} \in \mathcal{V}_h$ with high probability due to the MLE guarantee. Then letting $\widetilde{v}_h$ denote the spanner, Lemma 15 guarantees that $\frac{d^{D,\dagger}_{h-1}}{\widetilde{v}_h} \le \mathsf{d}$, and the approximation error of $\mathcal{W}_h(\widetilde{v}_h)$ can be controlled by the error of MLE estimation, since for any $d_h \le C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} d^{D,\dagger}_{h-1}$ we have

$$\begin{aligned}
\left\| \frac{d_h}{\widetilde{v}_h} - \frac{d_h}{d^{D,\dagger}_{h-1}} \right\|^2_{2,d^{D,\dagger}_{h-1}} &\le (C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1})^2 \int \frac{d^{D,\dagger}_{h-1}(x)}{\widetilde{v}_h(x)} \left(1 + \frac{d^{D,\dagger}_{h-1}(x)}{\widetilde{v}_h(x)}\right) \left| \widetilde{v}_h(x) - d^{D,\dagger}_{h-1}(x) \right| (\mathrm{d}x) \\
&\le 2(C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} \mathsf{d})^2 \|\widetilde{v}_h - d^{D,\dagger}_{h-1}\|_1
\end{aligned}$$

which implies that $\varepsilon^{\text{approx}}_h \le 2 C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} \mathsf{d} \sqrt{\varepsilon_{\text{mle}}}$ by the definition of $\mathcal{V}_h$. However, since ε_{mle} is $O(n^{-1/2}_{\text{mle}})$, this results in a slow rate of $1/\varepsilon^4$ total sample complexity for offline density estimation, and from a computational standpoint, introduces another barycentric spanner construction step in the algorithm which can be expensive. The representation learning setting has the additional challenge that there will be approximation error if the wrong representation $\widehat{\mu}_{h-1} \in \Upsilon_{h-1}$ is chosen for $\widehat{d}^{D,\dagger}_{h-1}$, since $d^{D,\dagger}_{h-1} \notin \mathcal{V}_h(\widehat{\mu}_h)$ (we extend the definition to $\mathcal{V}_h(\mu_{h-1}) = \left\{ v_h = \langle \mu_{h-1}, \theta_h \rangle : \|v_h - \widehat{d}^{D,\dagger}_{h-1}\|_1 \le \varepsilon_{\text{mle}}, \theta_h \in \mathbb{R}^{\mathsf{d}} \right\}$), which, as in the first case above, may be difficult to bound.

D.3.3. Clipped function class with point estimate in denominator

Generalizing and improving upon the previous analyses, using a clipped version of the function class $\mathcal{W}_h(v_h)$

$$\mathcal{W}^{\text{clip}}_h(v_h) = \left\{ w_h = \frac{\langle \mu^*_{h-1}, \theta_h \rangle \wedge C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} v_h}{v_h} : \theta_{h+1} \in \mathbb{R}^{\mathsf{d}} \right\}$$

will allow us to bound the approximation error for general denominator functions v_h. For any d_h such that $d_h \le C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} d^{D,\dagger}_{h-1}$, we can approximate the ratio $\frac{d_h}{d^{D,\dagger}_{h-1}}$ with $\frac{d_h \wedge C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} v_h}{v_h} \in \mathcal{W}^{\text{clip}}_h(v_h)$, and separate the approximation error into two terms, based on whether $d^{D,\dagger}_{h-1}$ is covered by v_h according to a threshold $C \ge 1$:

$$\begin{aligned}
&\left\| \frac{d_h \wedge C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} v_h}{v_h} - \frac{d_h}{d^{D,\dagger}_{h-1}} \right\|^2_{2,d^{D,\dagger}_{h-1}} \\
&\le \left\| \left(\frac{d_h \wedge C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} v_h}{v_h} - \frac{d_h}{d^{D,\dagger}_{h-1}} \right) \cdot \mathbf{1}\left[\frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)} \le C \right] \right\|^2_{2,d^{D,\dagger}_{h-1}} && \text{(``covered'')} \\
&\quad + \left\| \left(\frac{d_h \wedge C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} v_h}{v_h} - \frac{d_h}{d^{D,\dagger}_{h-1}} \right) \cdot \mathbf{1}\left[\frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)} > C \right] \right\|^2_{2,d^{D,\dagger}_{h-1}} && \text{(``not covered'')}
\end{aligned}$$

Bounding the two terms individually, for the “covered” term, we have

$$\begin{aligned}
\text{(``covered'')} &\le \int_{x: \frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)} \le C} d^{D,\dagger}_{h-1}(x) \left(\frac{d_h(x)}{v_h(x)} - \frac{d_h(x)}{d^{D,\dagger}_{h-1}(x)} \right)^2 (\mathrm{d}x) \\
&\le (C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1})^2 \int_{x: \frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)} \le C} \frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)} \frac{(d^{D,\dagger}_{h-1}(x) - v_h(x))^2}{v_h(x)} (\mathrm{d}x)
\end{aligned}$$

$$\le (C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2 C(1+C)\int_{x:\frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)}\le C}\left|d^{D,\dagger}_{h-1}(x)-v_h(x)\right|$$
$$\le (C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2 C(1+C)\left\|d^{D,\dagger}_{h-1}-v_h\right\|_1 .$$

For the “not covered” term, noticing that both parenthesized ratios are bounded on $[0, C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}]$, we have

$$(\text{“not covered”}) \le (C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2 \int d^{D,\dagger}_{h-1}(x)\cdot \mathbf{1}\left[\frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)} > C\right](\mathrm{d}x)$$
$$\le (C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2 \left(1-\frac{1}{C}\right)^{-1}\left\|d^{D,\dagger}_{h-1}-v_h\right\|_1 ,$$

where the second inequality is because

$$\left(1-\frac{1}{C}\right)\int_{x:\frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)}>C} d^{D,\dagger}_{h-1}(x)(\mathrm{d}x) < \int_{x:\frac{d^{D,\dagger}_{h-1}(x)}{v_h(x)}>C}(d^{D,\dagger}_{h-1}(x)-v_h(x))(\mathrm{d}x) \le \left\|d^{D,\dagger}_{h-1}-v_h\right\|_1$$

since $\frac{d^{D,\dagger}_{h-1}}{C} > v_h$. Thus in total, we have

$$\varepsilon^{\text{approx}}_h \le C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}\left(C + C^2 + \frac{C}{C-1}\right)\sqrt{\left\|d^{D,\dagger}_{h-1}-v_h\right\|_1} .$$

The bound depends on how close the point estimate v_h is to the true $d^{D,\dagger}_{h-1}$, as well as the threshold C. In the case where $v_h = \widehat{d}^{D,\dagger}_{h-1}$ is the point estimate, we are now able to bound $\varepsilon^{\text{approx}}_h \le C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}(C + C^2 + \frac{C}{C-1})\sqrt{\varepsilon_{\text{mle}}}$, which results in a slower rate than our results in the main text. If $v_h = \widetilde{v}_h$ is the barycentric spanner of the version space, then it suffices to set $C = \mathsf{d}$, in which case only the “covered” part of the error is nonzero, and we recover the analysis in the previous paragraph.

In general, the best choice of threshold C is not obvious because $d^{D,\dagger}_{h-1}$ is not known, and will trade off between the two errors. When C is large, the “covered” error will be large since it is proportional to C^2, while if C is too small (too close to 1), the “not-covered” error will be large since it is proportional to $\frac{C}{C-1}$.

D.3.4. Direct extraction of the estimate

Putting aside the discussion of point estimates in the denominator, we now present an alternative to pointwise multiplication + linearization used to extract $\widehat{d}^\pi_h$ from Algorithm 1. Instead, we can directly extract the numerator, which will already be a linear function (in μ^*), from weight ratio and use it as the estimate for $\widehat{d}^\pi_h$. The regression objective might then be (replacing line 5 in Algorithm 1)

$$\widehat{d}^\pi_h = \operatorname*{argmin}_{d_h\in\mathcal{F}_h(v_h)}\ \min_{v_h\in\mathcal{V}_h}\ \mathcal{L}_{\mathcal{D}^{\text{reg}}_{h-1}}\left(\frac{d_h}{v_h}, \frac{\widehat{d}^\pi_{h-1}\wedge C^{\mathbf{x}}_{h-1}\widehat{d}^D_{h-1}}{\widehat{d}^D_{h-1}}, \pi_{h-1}\wedge C^{\mathbf{a}}_{h-1}\pi^D_{h-1}\right),$$

where the version space of denominator functions $\mathcal{V}_h$ is defined above, and $\mathcal{F}_h(v_h) = \{d_h = \langle \mu^*_{h-1}, \theta_h\rangle : \|d_h/v_h\|_\infty \le C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}, \theta_h \in \mathbb{R}^{\mathsf{d}}\}$ represents linear numerator functions covered by v_h. It is necessary to constrain the denominator functions to the version space in order to ensure that the numerator is close to the true density, since regression only guarantees quality of estimated weight. For example, even if $\widehat{w}^\pi_h = w^\pi_h$, if the denominator function is $c \cdot d^{D,\dagger}_{h-1}$ then the numerator will be $c \cdot d^\pi_h$, leading to large $\widehat{d}^\pi_h$ estimation error. In terms of the analysis, this is quantified as the error between the denominator and true $d^{D,\dagger}_{h-1}$ in Eq. (11), which is controlled by ε_{mle} when the denominator is constrained to the version space $\mathcal{V}_h$, and will result in the same guarantee as we have for Algorithm 1 and Algorithm 2 in the known feature setting.

In the online setting with known features, direct extraction has the advantage of no longer requiring the linearization step (line 7 in Algorithm 2), though it is computationally more expensive because the function classes are jointly optimized, and the version space must be maintained. This advantage is lost in the representation learning setting because the estimates $\{\widehat{d}^\pi_h\}_{\pi\in\Pi}$ must be jointly re-linearized with the same representation in order to construct the policy cover (line 9 of

Algorithm 4). As another related advantage, this approach will relax the expressivity assumptions in the offline setting to a form of "completeness" (Uehara et al., 2021a), that we have function classes that are closed under the operators explicitly defined in Eq. (4) (i.e., that maps $\overline{d}^{\pi}_{h-1}$ to $\overline{d}^{\pi}_{h}$).

D.3.5. MLE instead of regression

An alternative to using regression to estimate the occupancy is instead using MLE-type estimation. Along similar veins as the regression algorithm, (a clipped version of) the previous-level estimate $\widehat{d}^{\pi}_{h-1}$ must be reused to reweight the data distribution in order to estimate $\widehat{d}^{\pi}_{h}$:

$$\widehat{d}^{\pi}_{h} = \operatorname*{argmin}_{f_h \in \mathcal{F}_h} \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^{D}_{h-1}}{\widehat{d}^{D}_{h-1}} \frac{\pi_{h-1} \wedge C^{\mathbf{a}}_{h-1} \pi^{D}_{h-1}}{\pi^{D}_{h-1}} \log(f_h).$$

where $\mathcal{F}_h$ is some linear function class. One possible advantage of such an approach is that a linear density estimate can be directly learned, but establishing formal guarantees for an MLE-type algorithm remains future work. After separating the missingness error $\|d^{\pi}_{h} - \overline{d}^{\pi}_{h}\|_1$ in the same way as in Section 3, similar methods as classical MLE analysis (Appendix H) might be used to control $\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\|_1$. The challenge is that such MLE analyses require $\mathcal{F}_h$ to include only valid densities $\in \Delta(\mathcal{X})$, but this is at odds with reweighted MLE objectives such as the one above, since the weights $\frac{\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^{D}_{h-1}}{\widehat{d}^{D}_{h-1}}$ generally will not induce a valid density when multiplied with the data distribution.

D.3.6. Version space for d^{π}_{h}

Most algorithmic ideas presented in the paper for estimating d^{π}_{h} have a top-down manner, which resembles the standard bottom-up structure of dynamic-programming algorithms for value-function estimation. On the other hand, Bellman-residual minimization (Antos et al., 2008) learns value functions by checking whether each candidate function is temporally self-consistent based on the data, and is very useful for producing version spaces of the functions of interest (Xie et al., 2021) and is statistically more superior to dynamic-programming algorithms in various situations (Xie and Jiang, 2020; Uehara et al., 2021a). Here we describe a method to produce a version space for d^{π}_{h}. In Appendix C we also described how to form version space based on $\overline{d}^{\pi}_{h}$ (Eq. (7)); in contrast, the method below will not estimate $\overline{d}^{\pi}_{h}$ but instead directly produce a version space that will be generally tighter than Eq. (7).

Similar to the case of value functions, the key to forming tight version spaces is to check whether a candidate function $\{d_h\}$ is temporally self-consistent. We do so by the following criterion: (we assume all candidate $\{d_h\}$ agree on d_0) $\forall h \geq 1$,

$$\left\| \left[\mathbf{P}^{\overline{\pi}}_{h-1} \left(d_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^{D}_{h-1} \right) - d_h \right]_+ \right\|_1 \leq \varepsilon'. \tag{8}$$

Inside $[\cdot]_+$, the term $\mathbf{P}^{\overline{\pi}}_{h-1} \left(d_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^{D}_{h-1} \right)$ corresponds to pushing d_{h-1} to the next level with clipped dynamics, which is exactly the kind of object FORC learns in each step. $\{d_h\} = \{d^{\pi}_{h}\}$ satisfies the criterion with $\varepsilon' = 0$: when data covers $\{d^{\pi}_{h}\}$, the LHS becomes $\|[\mathbf{P}^{\pi}_{h-1} d^{\pi}_{h-1} - d^{\pi}_{h}]_+\|_1 = 0$ as $\mathbf{P}^{\pi}_{h-1} d^{\pi}_{h-1} = d^{\pi}_{h}$; when data does not provide sufficient coverage, $\mathbf{P}^{\overline{\pi}}_{h-1} \left(d^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^{D}_{h-1} \right) \leq d^{\pi}_{h}$, and the LHS is still 0 since $[\cdot]_+$ only considers the positive part of the difference. The above reasoning assumes $\mathbf{P}^{\overline{\pi}}_{h-1} \left(d^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^{D}_{h-1} \right)$ is known, but in practice we need to estimate it from data; therefore, ε' cannot be set as 0 and must be instead set to the estimation error of $\mathbf{P}^{\overline{\pi}}_{h-1} \left(d_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^{D}_{h-1} \right)$ to guarantee $\{d^{\pi}_{h}\}$ is not eliminated.

Given the criterion, we can form a version space of $\{d^{\pi}_{h}\}$ that includes all normalized $\{d_h\}$ from the function class that satisfies Eq. (8). Offline policy learning follows straightforwardly (Appendix C), and this version-space-based approach produces generally less conservative estimate than using $\{\overline{d}^{\pi}_{h}\}$ (Theorem 3). The online case is trickier as we still need to produce a point estimate $\{d'_h\}$ for policy cover construction (the role of $\widetilde{d}^{\pi}_{h-1}$ in FORCE), and our analysis requires the chosen $\{d'_h\}$ to be approximately a point-wise lower bound of $\{d^{\pi}_{h}\}$ (which we informally denote as $d'_h \lesssim d^{\pi}_{h}$, meaning that $\|[d'_h - d^{\pi}_{h}]_+\|_1$ is small). To handle this problem, we can construct two version spaces, one that only includes normalized distributions for reasoning about d^{π}_{h} (which we call VS_π) and one that includes unnormalized distributions for selecting $\{d'_h\}$ (which we call VS'_π). It is easy to see that $\{d^{\pi}_{h}\} \in \mathrm{VS}_\pi \subseteq \mathrm{VS}'_\pi$, and any member $\{d'_h\}$ of VS'_π satisfies $d'_h \gtrsim \overline{d}^{\pi}_{h}$. Given the two version spaces, we can choose any $\{d'_h\} \in \mathrm{VS}'_\pi$ that satisfies $d'_h \lesssim d_h$, $\forall \{d_h\} \in \mathrm{VS}_\pi$. $\{\overline{d}^{\pi}_{h}\} \in \mathrm{VS}'_\pi$ is always a viable choice, but in general there may be better choices of $\{d'_h\}$ that have significantly larger norm $\|d'_h\|_1$ than $\|\overline{d}^{\pi}_{h}\|_1$, thus preserving more mass in the online algorithm.

D.4. Discussion of other approaches for controlling error exponentiation in the online setting

Barycentric spanner in regression target (without clipping) In Section 4 we controlled the error exponentiation arising from having only approximately exploratory data by first clipping the regression target $\widehat{d}_h^\pi / \widehat{d}_h^D$ (since the MLE estimate $\widehat{d}_h^D$ does not necessarily cover $\widehat{d}_h^\pi$), then separating the error $\|\widehat{d}_h^\pi - d_h^\pi\|_1$ into the "two-sided regression error" and "one-sided missingness error". It will be instructive to also look at an alternative approach that avoids clipping and "pretends" that data is perfectly exploratory, which provides interesting insights on the underlying issue and the delicacy of error propagation in our problem from a different perspective.

The seemingly feasible solution is based on the observation that $\frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} \widehat{d}_h^{\pi^{h,i}}$, the barycentric spanner of $\{\widehat{d}_h^\pi\}_{\pi\in\Pi}$ in the denominator of Eq. (5), *is* a good approximation of d_h^D. So instead of using MLE to estimate $d_h^D = \frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} d_h^{\pi^{h,i}}$, we could simply use $\frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} \widehat{d}_h^{\pi^{h,i}}$, which will keep the regression target bounded in Algorithm 1 without any clipping.

However, a closer look reveals that this only sweeps the issue under the rug. The problem does not go away, and only appears in a different form: recall from Lemma 1 that the bound includes a term of $2\mathsf{d}\left\|\widehat{d}_h^D - d_h^D\right\|_1$, and when we use $\frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} \widehat{d}_h^{\pi^{h,i}}$ to replace $\widehat{d}_h^D$, we obtain

$$\left\|\widehat{d}_h^D - d_h^D\right\|_1 = \left\|\frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} \widehat{d}_h^{\pi^{h,i}} - \frac{1}{\mathsf{d}}\sum_{i=1}^{\mathsf{d}} d_h^{\pi^{h,i}}\right\|_1 \le \max_{\pi\in\Pi} \|\widehat{d}_h^\pi - d_h^\pi\|_1$$

which, in addition to merging the two inductive chains, gives us $\|\widehat{d}_h^\pi - d_h^\pi\|_1 \le (1+\mathsf{d})\max_{\pi\in\Pi}\|\widehat{d}_h^\pi - d_h^\pi\|_1 + \ldots$, resulting in $O(\mathsf{d})^H$ error. In other words, because the error of the denominator distribution depends on the quality of regression, even with full coverage we will suffer the same error exponentiation issues.

Reachability-based approach Error exponentiation can be avoided if a *reachability* assumption (Du et al., 2019; Modi et al., 2021) is satisfied in the underlying MDP. Formally, this assumption requires that there exists a constant $\eta_{\min}$ such that $\forall h \in [H], z \in \mathcal{Z}_{h+1}$ we have $\max_{\pi\in\Pi} \mathbb{P}_\pi[z_{h+1} = z] \ge \eta_{\min}$, where $\mathcal{Z}_{h+1}$ correspond to the latent states of the MDP. For example, in the case where μ_h^* is full-rank and composed of simplex features, $\mathcal{Z}_{h+1} = \{1, \ldots, \mathsf{d}\}$ and $\theta_h[i]$ directly corresponds to $\mathbb{P}_\pi[z_{h+1} = i]$ for $i \in \{1, \ldots, \mathsf{d}\}$. The direct implication is that we can construct a fully exploratory policy cover that reaches all latent states (and thus covers all $\pi \in \Pi$) as long as we find, for each latent state, the policy that reaches it with probability at least $\eta_{\min}$. This policy can be found as long as $\widehat{d}_h^\pi$ is estimated sufficiently well, which when backed up implies the latent state visitation is estimated sufficiently well.

Specifically, in the offline module used in Algorithm 2, we can instead set n_{reg} such that $\|\widehat{d}_h^\pi - d_h^\pi\|_1 \le \sigma_{\min}(\mu_{h-1}^*)\eta_{\min}/4$ for all $\pi \in \Pi$, which implies that when backed up to latent states the error of estimation is $\|\widehat{\theta}_h^\pi - \theta_h^\pi\|_\infty \le \eta_{\min}/4$. Then the exploratory policy cover can be chosen as $\Pi_h^{\text{expl}} = \{\pi^{h,i}\}_{i=1}^{\mathsf{d}}$ where for each $i \in \{1, \ldots, \mathsf{d}\}$, $\pi^{h,i}$ is such that $\widehat{\theta}_h^{\pi^{h,i}}[i] \ge \eta_{\min}/4$, which implies $\theta_h^{\pi^{h,i}}[i] \ge \eta_{\min}/2$ with high probability, and such a policy is guaranteed to exist from the reachability assumption. Since the policy cover is fully exploratory, a single induction chain in the error analysis (instead of the two in Figure 1) will suffice.

E. Off-policy occupancy estimation proofs (Section 3)

E.1. Discussion of clipping thresholds for $\overline{d}_h^\pi$

As we have previously mentioned, the clipped occupancy $\overline{d}_h^\pi$ depends on clipping thresholds $\{C_h^{\mathbf{x}}\}$ and $\{C_h^{\mathbf{a}}\}$ that are hyperparameter inputs to the offline estimation algorithm (Algorithm 1). To better understand the effects of $C_h^{\mathbf{x}}, C_h^{\mathbf{a}}$ on $\overline{d}_h^\pi$ and downstream analysis, we highlight three properties below, which we have written only for $C_h^{\mathbf{x}}$ (but that take analogous forms for $C_h^{\mathbf{a}}$).

Importantly, property 3 shows that the missingness error $\|\overline{d}_h^\pi - d_h^\pi\|_1$ is Lipschitz in the clipping thresholds $\{C_h^{\mathbf{x}}\}$, indicating that small changes in $C_h^{\mathbf{x}}$ will only lead to small changes in the missingness error, and thus the result of Theorem 2. For practical purposes, this serves as a reassurance that, within some limit, misspecifications of $C_h^{\mathbf{x}}, C_h^{\mathbf{a}}$ in the algorithm do not have catastrophic consequences.

Proposition 5. *For two sets of clipping thresholds $\{C_h^{\mathbf{x}}\}, \{(C_h^{\mathbf{x}})'\}$, following Definition 1, for each $h = 1, \ldots, H$ let their* *corresponding clipped occupancies be defined recursively as*

$$\overline{d}_h^\pi = \mathbf{P}_{h-1}^{\overline{\pi}} \left(\overline{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} d_{h-1}^D\right)$$
$$(\overline{d}_h^\pi)' = \mathbf{P}_{h-1}^{\overline{\pi}} \left((\overline{d}_{h-1}^\pi)' \wedge (C_{h-1}^{\mathbf{x}})' d_{h-1}^D\right)$$

with $\overline{d}_0^\pi = (\overline{d}_0^\pi)' = d_0$. Then the following two properties hold for each $h \in [H]$:

1. *(Monotonicity) $\overline{d}_h^\pi \leq (\overline{d}_h^\pi)'$ if $C_{h'}^{\mathbf{x}} \leq (C_{h'}^{\mathbf{x}})'$ for all $h' < h$. The relationship also holds in the other direction, i.e., replacing "$\leq$" with "$>$".*
2. *(Clipped occupancy Lipschitz in thresholds) $\|(\overline{d}_h^\pi)' - \overline{d}_h^\pi\|_1 \leq \sum_{h'<h} |(C_{h'}^{\mathbf{x}})' - C_{h'}^{\mathbf{x}}|$.*
3. *(Missingness error Lipschitz in thresholds) $\left|\|d_h^\pi - (\overline{d}_h^\pi)'\|_1 - \|d_h^\pi - \overline{d}_h^\pi\|_1\right| \leq \sum_{h'<h} |(C_{h'}^{\mathbf{x}})' - C_{h'}^{\mathbf{x}}|$.*

Proof. We prove these three claims one by one.

Proof of Claim 1 We will prove Claim 1 via induction. Suppose $\overline{d}_{h'-1}^\pi \leq (\overline{d}_{h'-1}^\pi)'$ for some $h' \leq h$. This holds for the base case $h' = 1$ since $\overline{d}_0^\pi = (\overline{d}_0^\pi)'$. Then since $C_{h'-1}^{\mathbf{x}} \leq (C_{h'-1}^{\mathbf{x}})'$,

$$\overline{d}_{h'}^\pi = \mathbf{P}_{h'-1}^{\overline{\pi}} \left(\overline{d}_{h'-1}^\pi \wedge C_{h'-1}^{\mathbf{x}} d_{h'-1}^D\right) \leq \mathbf{P}_{h'-1}^{\overline{\pi}} \left((\overline{d}_{h'-1}^\pi)' \wedge (C_{h'-1}^{\mathbf{x}})' d_{h'-1}^D\right) = (\overline{d}_{h'}^\pi)'.$$

Then by induction we have that $\overline{d}_h^\pi \leq (\overline{d}_h^\pi)'$.

Proof of Claim 2 For Claim 2, using Lemma 20, we have

$$\begin{aligned}
&\|(\overline{d}_h^\pi)' - \overline{d}_h^\pi\|_1 \\
\leq& \left\|\left(\overline{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} d_{h-1}^D\right) - \left((\overline{d}_{h-1}^\pi)' \wedge (C_{h-1}^{\mathbf{x}})' d_{h-1}^D\right)\right\|_1 \\
\leq& \left\|\left(\overline{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} d_{h-1}^D\right) - \left((\overline{d}_{h-1}^\pi)' \wedge C_{h-1}^{\mathbf{x}} d_{h-1}^D\right)\right\|_1 + \left\|\left((\overline{d}_{h-1}^\pi)' \wedge C_{h-1}^{\mathbf{x}} d_{h-1}^D\right) - \left((\overline{d}_{h-1}^\pi)' \wedge (C_{h-1}^{\mathbf{x}})' d_{h-1}^D\right)\right\|_1 \\
\leq& \left\|\overline{d}_{h-1}^\pi - (\overline{d}_{h-1}^\pi)'\right\|_1 + \left\|C_{h-1}^{\mathbf{x}} d_{h-1}^D - (C_{h-1}^{\mathbf{x}})' d_{h-1}^D\right\|_1 \\
=& \left\|\overline{d}_{h-1}^\pi - (\overline{d}_{h-1}^\pi)'\right\|_1 + \left|C_{h-1}^{\mathbf{x}} - (C_{h-1}^{\mathbf{x}})'\right|.
\end{aligned}$$

Unfolding this recursion from level $h-1$ through level 0 gives the result.

Proof of Claim 3 For Claim 3, we have

$$\begin{aligned}
\left|\|d_h^\pi - (\overline{d}_h^\pi)'\|_1 - \|d_h^\pi - \overline{d}_h^\pi\|_1\right| &= \left|\int |d_h^\pi(x) - (\overline{d}_h^\pi)'(x)| - |d_h^\pi(x) - \overline{d}_h^\pi(x)|(\mathrm{d}x)\right| \\
&\leq \int \left||d_h^\pi(x) - (\overline{d}_h^\pi)'(x)| - |d_h^\pi(x) - \overline{d}_h^\pi(x)|\right|(\mathrm{d}x) \\
&\leq \int \left|(\overline{d}_h^\pi)'(x) - \overline{d}_h^\pi(x)\right|(\mathrm{d}x) && \text{(since } ||x| - |y|| \leq |x-y|\text{)} \\
&= \left\|(\overline{d}_h^\pi)' - \overline{d}_h^\pi\right\|_1.
\end{aligned}$$

Then applying Claim 2 gives the stated claim. □

E.2. Proof of occupancy estimation

Proposition (Restatement of Proposition 2). *We have the following properties for $\overline{d}_h^\pi$:*

1. $\overline{d}^{\pi}_h \leq d^{\pi}_h$.

2. $\overline{d}^{\pi}_h = d^{\pi}_h$ *when data covers* π*, i.e.,* $\forall h' < h$ *we have* $d^{\pi}_{h'} \leq C^{\mathbf{x}}_{h'} d^D_{h'}$ *and* $\pi_{h'} \leq C^{\mathbf{a}}_{h'} \pi^D_{h'}$.

3. $\|\overline{d}^{\pi}_h - d^{\pi}_h\|_1 \leq \|\overline{d}^{\pi}_{h-1} - d^{\pi}_{h-1}\|_1 + \|\overline{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\|_1 + \|\mathbf{P}^{\pi}_{h-1} d^{\pi}_{h-1} - \mathbf{P}^{\overline{\pi}}_{h-1} d^{\pi}_{h-1}\|_1$.

Proof. We prove these three claims one by one.

Proof of Claim 1 Firstly, we have $\overline{d}^{\pi}_h = d^{\pi}_h = d_0$. Assuming the claim holds for $h' - 1$, then we have $\overline{d}^{\pi}_{h'} = \mathbf{P}^{\overline{\pi}}_{h'-1}(\overline{d}^{\pi}_{h'-1} \wedge C^{\mathbf{x}}_{h'-1} d^D_{h'-1}) \leq \mathbf{P}^{\pi}_{h'-1}(\overline{d}^{\pi}_{h'-1} \wedge C^{\mathbf{x}}_{h'-1} d^D_{h'-1}) \leq \mathbf{P}^{\pi}_{h'-1}(d^{\pi}_{h'-1} \wedge C^{\mathbf{x}}_{h'-1} d^D_{h'-1}) \leq \mathbf{P}^{\pi}_{h'-1} d^{\pi}_{h'-1} = d^{\pi}_{h'}$. By induction, we complete the proof.

Proof of Claim 2 It is easy to see that $d^{\pi}_{h'} \leq C^{\mathbf{x}}_{h'} d^D_{h'}$ together with Claim 1 implies $\overline{d}^{\pi}_{h'} \leq C^{\mathbf{x}}_{h'} d^D_{h'}$, thus $\|\overline{d}^{\pi}_{h'} - \overline{d}^{\pi}_{h'} \wedge C^{\mathbf{x}}_{h'} d^D_{h'}\|_1 = 0$. In addition, $\pi_{h'} \leq C^{\mathbf{a}}_{h'} \pi^D_{h'}$ gives us $\pi_{h'} = \overline{\pi}_{h'}$, therefore $\left\|\mathbf{P}^{\pi}_{h'-1} d^{\pi}_{h'-1} - \mathbf{P}^{\overline{\pi}}_{h'-1} d^{\pi}_{h'-1}\right\|_1 = 0$. Now we can prove Claim 2 inductively. For $h' = 0$, we know the claim holds since $\overline{d}^{\pi}_0 = d^{\pi}_0 = d_0$. Assuming the claim holds for $h' - 1$, by Claim 3 we have that

$$0 \leq \|\overline{d}^{\pi}_{h'} - d^{\pi}_{h'}\|_1 \leq \|\overline{d}^{\pi}_{h'-1} - d^{\pi}_{h'-1}\|_1 + \|\overline{d}^{\pi}_{h'-1} - \overline{d}^{\pi}_{h'-1} \wedge C^{\mathbf{x}}_{h'-1} d^D_{h'-1}\|_1 + \|\mathbf{P}^{\pi}_{h'-1} d^{\pi}_{h'-1} - \mathbf{P}^{\overline{\pi}}_{h'-1} d^{\pi}_{h'-1}\|_1 = 0.$$

This means the claim holds for h'. By induction, we complete the proof.

Proof of Claim 3 For the third part, we have the following decomposition

$$\begin{aligned}
\|\overline{d}^{\pi}_h - d^{\pi}_h\|_1 &= \left\|\mathbf{P}^{\overline{\pi}}_{h-1} \left(\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\right) - \mathbf{P}^{\pi}_h d^{\pi}_{h-1}\right\|_1 \\
&\leq \left\|\mathbf{P}^{\overline{\pi}}_{h-1} \left(\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\right) - \mathbf{P}^{\overline{\pi}}_{h-1} d^{\pi}_{h-1}\right\|_1 + \left\|\mathbf{P}^{\overline{\pi}}_{h-1} d^{\pi}_{h-1} - \mathbf{P}^{\pi}_h d^{\pi}_{h-1}\right\|_1 \\
&\leq \left\|\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1} - d^{\pi}_{h-1}\right\|_1 + \left\|\mathbf{P}^{\overline{\pi}}_{h-1} d^{\pi}_{h-1} - \mathbf{P}^{\pi}_h d^{\pi}_{h-1}\right\|_1 && \text{(Lemma 20)} \\
&\leq \left\|\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1} - \overline{d}^{\pi}_{h-1}\right\|_1 + \left\|\overline{d}^{\pi}_{h-1} - d^{\pi}_{h-1}\right\|_1 + \left\|\mathbf{P}^{\overline{\pi}}_{h-1} d^{\pi}_{h-1} - \mathbf{P}^{\pi}_h d^{\pi}_{h-1}\right\|_1 . && \square
\end{aligned}$$

Lemma (Restatement of Lemma 1)**.** *For every* $h \in [H]$*, the error between estimates* $\widehat{d}^{\pi}_h$ *from Algorithm 1 and the clipped target* $\overline{d}^{\pi}_h$ *is decomposed recursively as*

$$\begin{aligned}
\left\|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\right\|_1 \leq{}& \left\|\widehat{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1}\right\|_1 + 2C^{\mathbf{x}}_{h-1} \left\|\widehat{d}^D_{h-1} - d^D_{h-1}\right\|_1 + C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1} \left\|\widehat{d}^{D,\dagger}_{h-1} - d^{D,\dagger}_{h-1}\right\|_1 \\
&+ \sqrt{2} \left\|\widehat{w}^{\pi}_h - \mathbf{E}^{\overline{\pi}}_{h-1} \left(d^D_{h-1} \frac{\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}}{\widehat{d}^D_{h-1}}\right)\right\|_{2, d^{D,\dagger}_{h-1}},
\end{aligned}$$

where $(\mathbf{E}^{\pi}_h d_h) := (\mathbf{P}^{\pi}_h d_h)/d^{D,\dagger}_h$.

Proof. We start by separating out the recursive term

$$\begin{aligned}
&\left\|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\right\|_1 = \left\|\widehat{d}^{\pi}_h - \mathbf{P}^{\overline{\pi}}_{h-1} \left(\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\right)\right\|_1 \\
\leq{}& \left\|\widehat{d}^{\pi}_h - \mathbf{P}^{\overline{\pi}}_{h-1} \left(\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right)\right\|_1 + \left\|\mathbf{P}^{\overline{\pi}}_{h-1} \left(\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right) - \mathbf{P}^{\overline{\pi}}_{h-1} \left(\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right)\right\|_1 \\
&+ \left\|\mathbf{P}^{\overline{\pi}}_{h-1} \left(\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right) - \mathbf{P}^{\overline{\pi}}_{h-1} \left(\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\right)\right\|_1 \\
\leq{}& \left\|\widehat{d}^{\pi}_h - \mathbf{P}^{\overline{\pi}}_{h-1} \left(\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right)\right\|_1 + \left\|\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1} - \overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right\|_1 \\
&+ \left\|\overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1} - \overline{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} d^D_{h-1}\right\|_1 \\
\leq{}& \left\|\widehat{d}^{\pi}_h - \mathbf{P}^{\overline{\pi}}_{h-1} \left(\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1} \widehat{d}^D_{h-1}\right)\right\|_1 + \left\|\widehat{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1}\right\|_1 + C^{\mathbf{x}}_{h-1} \left\|\widehat{d}^D_{h-1} - d^D_{h-1}\right\|_1 . \qquad (9)
\end{aligned}$$

Here, we apply Lemma 20 in the second inequality. The last inequality is due to $|\min(x,y)-\min(x,z)| \le |y-z|$ for $x,y,z \in \mathbb{R}$.

Now, we consider the first term in Eq. (9) and get

$$
\begin{aligned}
&\left\|\widehat{d}_h^\pi - \mathbf{P}_{h-1}^{\overline{\pi}}\ \left(\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D\right)\right\|_1 \\
\le &\left\|\widehat{d}_h^\pi - \mathbf{P}_{h-1}^{\overline{\pi}} \left(\frac{\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D}{\widehat{d}_{h-1}^D}\, d_{h-1}^D\right)\right\|_1 \\
&\quad + \left\|\mathbf{P}_{h-1}^{\overline{\pi}} \left(\frac{\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D}{\widehat{d}_{h-1}^D}\, d_{h-1}^D\right) - \mathbf{P}_{h-1}^{\overline{\pi}} \left(\frac{\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D}{\widehat{d}_{h-1}^D}\, \widehat{d}_{h-1}^D\right)\right\|_1 \\
\le &\left\|\widehat{d}_h^\pi - \mathbf{P}_{h-1}^{\overline{\pi}} \left(\frac{\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D}{\widehat{d}_{h-1}^D}\, d_{h-1}^D\right)\right\|_1 + C_{h-1}^{\mathbf{x}} \left\|d_{h-1}^D - \widehat{d}_{h-1}^D\right\|_1 .
\end{aligned} \tag{10}
$$

In the last inequality, we notice $\left\|\frac{\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D}{\widehat{d}_{h-1}^D}\right\|_\infty \le C_{h-1}^{\mathbf{x}}$ by our convention $\frac{0}{0} = 0$ and apply Lemma 20 again.

Let $\widetilde{w}_{h-1} := \frac{\widehat{d}_{h-1}^\pi \wedge C_{h-1}^{\mathbf{x}} \widehat{d}_{h-1}^D}{\widehat{d}_{h-1}^D}$ for short. Since $\|\widetilde{w}_{h-1}\|_\infty \le C_{h-1}^{\mathbf{x}}$, Lemma 19 guarantees $\frac{\left(\mathbf{P}_{h-1}^{\overline{\pi}}(d_{h-1}^D \widetilde{w}_{h-1})\right)}{d_{h-1}^{D,\dagger}} \le C_{h-1}^{\mathbf{x}} C_{h-1}^{\mathbf{a}}$, thus the ratio is well-defined. Then we can further upper-bound the first term in Eq. (10) as

$$
\begin{aligned}
\left\|\widehat{d}_h^\pi - \mathbf{P}_{h-1}^{\overline{\pi}} \left(d_{h-1}^D \widetilde{w}_{h-1}\right)\right\|_1 &= \left\|\widehat{w}_h^\pi\, \widehat{d}_{h-1}^{D,\dagger} - \frac{\mathbf{P}_{h-1}^{\overline{\pi}} \left(d_{h-1}^D \widetilde{w}_{h-1}\right)}{d_{h-1}^{D,\dagger}}\, d_{h-1}^{D,\dagger}\right\|_1 \\
&\le \left\|\widehat{w}_h^\pi\, \widehat{d}_{h-1}^{D,\dagger} - \widehat{w}_h^\pi\, d_{h-1}^{D,\dagger}\right\|_1 + \left\|\widehat{w}_h^\pi\, d_{h-1}^{D,\dagger} - \frac{\mathbf{P}_{h-1}^{\overline{\pi}} \left(d_{h-1}^D \widetilde{w}_{h-1}\right)}{d_{h-1}^{D,\dagger}}\, d_{h-1}^{D,\dagger}\right\|_1 \\
&= \left\|\widehat{w}_h^\pi\, \widehat{d}_{h-1}^{D,\dagger} - \widehat{w}_h^\pi\, d_{h-1}^{D,\dagger}\right\|_1 + \left\|\widehat{w}_h^\pi - \frac{\mathbf{P}_{h-1}^{\overline{\pi}} \left(d_{h-1}^D \widetilde{w}_{h-1}\right)}{d_{h-1}^{D,\dagger}}\right\|_{1,d_{h-1}^{D,\dagger}} \\
&\le \|\widehat{w}_h^\pi\|_\infty \left\|\widehat{d}_{h-1}^{D,\dagger} - d_{h-1}^{D,\dagger}\right\|_1 + \left\|\widehat{w}_h^\pi - \frac{\mathbf{P}_{h-1}^{\overline{\pi}} \left(d_{h-1}^D \widetilde{w}_{h-1}\right)}{d_{h-1}^{D,\dagger}}\right\|_{1,d_{h-1}^{D,\dagger}} \\
&\le C_h^{\mathbf{x}} C_h^{\mathbf{a}} \left\|\widehat{d}_{h-1}^{D,\dagger} - d_{h-1}^{D,\dagger}\right\|_1 + \left\|\widehat{w}_h^\pi - \frac{\mathbf{P}_{h-1}^{\overline{\pi}} \left(d_{h-1}^D \widetilde{w}_{h-1}\right)}{d_{h-1}^{D,\dagger}}\right\|_{2,d_{h-1}^{D,\dagger}} .
\end{aligned} \tag{11}
$$

Combining Eq. (9), Eq. (10), and Eq. (11) and noticing the definition of $\mathbf{E}_h^\pi$ and $\widetilde{w}_{h-1}$ completes the proof. □

Theorem (Restatement of Theorem 2)**.** *Fix $\delta \in (0,1)$. Suppose Assumption 1 and Assumption 2 hold, and μ^* is known. Then, given an evaluation policy π, by setting*

$$
n_{\mathrm{mle}} = \tilde{O}\left(\mathsf{d}\left(\sum_{h\in[H]} C_h^{\mathbf{x}} C_h^{\mathbf{a}}\right)^2 \log(1/\delta)/\varepsilon^2\right) \text{ and } n_{\mathrm{reg}} = \tilde{O}\left(\mathsf{d}\left(\sum_{h\in[H]} C_h^{\mathbf{x}} C_h^{\mathbf{a}}\right)^2 \log(1/\delta)/\varepsilon^2\right),
$$

with probability at least $1-\delta$, FORC *(Algorithm 1) returns state occupancy estimates $\{\widehat{d}_h^\pi\}_{h=0}^{H-1}$ satisfying*

$$
\left\|\widehat{d}_h^\pi - \overline{d}_h^\pi\right\|_1 \le \varepsilon, \forall h \in [H].
$$

The total number of episodes required by the algorithm is

$$
\tilde{O}\left(\mathsf{d}H\left(\sum_{h\in[H]} C_h^{\mathbf{x}} C_h^{\mathbf{a}}\right)^2 \log(1/\delta)/\varepsilon^2\right).
$$

Proof. We first make two claims on MLE estimation and error propagation.

Claim 1 Our estimated data distributions satisfy that with probability $1-\delta/2$, for any $h \in [H]$

$$\left\|\widehat{d}^{D}_{h} - d^{D}_{h}\right\|_1 \le \varepsilon_{\mathrm{mle}} \text{ and } \left\|\widehat{d}^{D,\dagger}_{h} - d^{D,\dagger}_{h}\right\|_1 \le \varepsilon_{\mathrm{mle}}, \tag{12}$$

where

$$\varepsilon_{\mathrm{mle}} := 6\sqrt{\frac{\mathsf{d}\log(16HB^{\mu}n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}}.$$

Claim 2 Under the high-probability event that Eq. (12) holds, we further have that with probability at least $1-\delta/2$, for any $1 \le h \le H$,

$$\left\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\right\|_1 \le \left\|\widehat{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1}\right\|_1 + 3C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}\varepsilon_{\mathrm{mle}} + \sqrt{2}\varepsilon_{\mathrm{reg,h-1}}, \tag{13}$$

where

$$\varepsilon_{\mathrm{reg,h-1}} := \sqrt{\frac{221184\mathsf{d}(C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2\log\left(2Hn_{\mathrm{reg}}/\delta\right)}{n_{\mathrm{reg}}}}.$$

Now we establish the final error bound with these two claims. Notice that the total failure probability is less than δ. Unfolding Eq. (13) from $h' = h$ to $h' = 1$ and noticing that $\widehat{d}^{\pi}_0 = \overline{d}^{\pi}_0 = d_0$ yields that for any $h \in [H]$

$$\left\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\right\|_1 \le \sum_{h'=0}^{h-1}\left(3C^{\mathbf{x}}_{h'}C^{\mathbf{a}}_{h'}\varepsilon_{\mathrm{mle}} + \sqrt{2}\varepsilon_{\mathrm{reg,h'}}\right). \tag{14}$$

Substituting in the expressions for $\varepsilon_{\mathrm{mle}}$ and $\varepsilon_{\mathrm{reg},}$, we have

$$\left\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\right\|_1 \le \sum_{h'=0}^{h-1}\left(18C^{\mathbf{x}}_{h'}C^{\mathbf{a}}_{h'}\sqrt{\frac{\mathsf{d}\log(16HB^{\mu}n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}} + 666C^{\mathbf{x}}_{h'}C^{\mathbf{a}}_{h'}\sqrt{\frac{\mathsf{d}\log\left(2Hn_{\mathrm{reg}}/\delta\right)}{n_{\mathrm{reg}}}}\right). \tag{15}$$

It is easy to see that if we set

$$n_{\mathrm{mle}} = \tilde{O}\left(\mathsf{d}\left(\sum_{h\in[H]}C^{\mathbf{x}}_{h}C^{\mathbf{a}}_{h}\right)^2\log(1/\delta)/\varepsilon^2\right) \text{ and } n_{\mathrm{reg}} = \tilde{O}\left(\mathsf{d}\left(\sum_{h\in[H]}C^{\mathbf{x}}_{h}C^{\mathbf{a}}_{h}\right)^2\log(1/\delta)/\varepsilon^2\right),$$

then we have

$$\left\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\right\|_1 \le \varepsilon, \forall h \in [H].$$

In the following, we provide the proof of these two claims respectively.

Proof of Claim 1 We start with a fixed $h \in [H]$ and bounding $\|\widehat{d}^{D}_{h} - d^{D}_{h}\|_1$, where we recall that $\widehat{d}^{D}_{h}$ is the MLE solution in Eq. (2). By Lemma 22, we know that function class $\mathcal{F}_h$ has an ℓ_1 optimistic cover with scale $1/n_{\mathrm{mle}}$ of size $(2\lceil B^{\mu}n_{\mathrm{mle}}\rceil)^{\mathsf{d}}$. It is easy to see that the true marginal distribution $d^{D}_{h} \in \mathcal{F}_h$ from Lemma 17 and any $d_h \in \mathcal{F}_h$ is a valid probability distribution over $\mathcal{X}$. From Assumption 2, we know that once conditioned on prior dataset $\mathcal{D}_{0:h-1}$, the current dataset $\mathcal{D}^{\mathrm{mle}}_{h}$ is drawn i.i.d. from the fixed distribution denoted as d^{D}_{h}. Thus, Lemma 12 tells us that when conditioned on $\mathcal{D}_{0:h-1}$, with probability at least $1-\delta/(4H)$

$$\begin{aligned}\|\widehat{d}^{D}_{h} - d^{D}_{h}\|_1 &\le \frac{1}{n_{\mathrm{mle}}} + \sqrt{\frac{12\log(4H\left(2\lceil B^{\mu}n_{\mathrm{mle}}\rceil\right)^{\mathsf{d}}/\delta)}{n_{\mathrm{mle}}} + \frac{6}{n_{\mathrm{mle}}}} \\ &\le \frac{1}{n_{\mathrm{mle}}} + \sqrt{\frac{12\mathsf{d}\log(16HB^{\mu}n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}} + \frac{6}{n_{\mathrm{mle}}}}\end{aligned} \tag{16}$$

$$\leq 6\sqrt{\frac{\mathsf{d}\log(16HB^{\mu}n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}} = \varepsilon_{\mathrm{mle}}. \tag{17}$$

Since Eq. (16) holds for any such fixed $\mathcal{D}_{0:h-1}$, applying the law of total expectation gives us this that Eq. (16) holds with probability $1-\delta/(4H)$ without conditioning on $\mathcal{D}_{0:h-1}$.

Similarly, with probability at least $1-\delta/(4H)$, for the MLE solution $\widehat{d}^{D,\dagger}_{h}$ we have $\|\widehat{d}^{D,\dagger}_{h} - d^{D,\dagger}_{h}\|_1 \leq \varepsilon_{\mathrm{mle}}$. Union bounding these two high-probability events and further union bounding over $h \in [H]$ gives us that Eq. (12) holds with probability $1-\delta/2$.

Proof of Claim 2 Notice that the proof in this part is under the high-probability event that Eq. (12) holds. We consider a fixed $h \in [H]$. From Lemma 1, we have the error propagation result that

$$\begin{aligned}\left\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\right\|_1 \leq{}& \left\|\widehat{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1}\right\|_1 + 2C^{\mathbf{x}}_{h-1}\left\|\widehat{d}^{D}_{h-1} - d^{D}_{h-1}\right\|_1 + C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}\left\|\widehat{d}^{D,\dagger}_{h-1} - d^{D,\dagger}_{h-1}\right\|_1 \\ &+ \sqrt{2}\left\|\widehat{w}^{\pi}_{h} - \frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}}\right\|_{2,d^{D,\dagger}_{h-1}},\end{aligned} \tag{18}$$

where $\widetilde{w}_{h-1} := \frac{\widehat{d}^{\pi}_{h-1} \wedge C^{\mathbf{x}}_{h-1}\widehat{d}^{D}_{h-1}}{\widehat{d}^{D}_{h-1}}$.

Since $\widehat{w}^{\pi}_{h} \in \mathcal{W}_h$, we have $\|\widehat{w}^{\pi}_{h}\|_\infty \leq C^{\mathbf{x}}_{h}C^{\mathbf{a}}_{h}$. The last term on RHS isolates the finite-sample error of regression, involving the difference between the empirical minimizer $\widehat{w}^{\pi}_{h}$ and the population minimizer $\frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}}$ of the regression objective. To bound this error, we apply Lemma 13 and Lemma 14, which give us that, with probability at least $1-\delta/(2H)$,

$$\begin{aligned}&\left\|\widehat{w}^{\pi}_{h} - \frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}}\right\|^2_{2,d^{D,\dagger}_{h-1}} \\ &= \mathbb{E}\left[\mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(\widehat{w}^{\pi}_{h}, \widetilde{w}_{h-1}, \overline{\pi}\right)\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(\frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}}, \widetilde{w}_{h-1}, \overline{\pi}\right)\right] \\ &\leq 2\left(\mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(\widehat{w}^{\pi}_{h}, \widetilde{w}_{h-1}, \overline{\pi}\right) - \mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(\frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}}, \widetilde{w}_{h-1}, \overline{\pi}\right)\right) + 2\varepsilon^2_{\mathrm{reg,h-1}}\end{aligned} \tag{19}$$

where

$$\varepsilon_{\mathrm{reg,h-1}} := \sqrt{\frac{221184 \cdot \mathsf{d}(C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2 \log\left(2Hn_{\mathrm{reg}}/\delta\right)}{n_{\mathrm{reg}}}}$$

The first term in Eq. (19) compares the empirical regression loss of the empirical minimizer $\widehat{w}^{\pi}_{h}$ against the population solution. In order to show that this is ≤ 0, we first need to check that $\frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}} \in \mathcal{W}_h$. As we have previously seen, we have $\frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}} \leq C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}$ from Lemma 19, thus satisfying the norm constraints of $\mathcal{W}_h$. Further, Lemma 16 guarantees that both the numerator and denominator are linear functions of μ^*_{h-1}, i.e., $\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right) = \langle \mu^*_{h-1}, \theta^{\mathrm{up}}_{h}\rangle$ and $d^{D,\dagger}_{h-1} = \langle \mu^*_{h-1}, \theta^{\mathrm{down}}_{h}\rangle$ for some $\theta^{\mathrm{up}}_{h}, \theta^{\mathrm{down}}_{h} \in \mathbb{R}^d$. Then since $\widehat{w}^{\pi}_{h}$ minimzes the empirical regression loss Eq. (3), we have

$$\mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(\widehat{w}^{\pi}_{h-1}, \widetilde{w}_{h-1}, \overline{\pi}\right) - \mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(\frac{\mathbf{P}^{\overline{\pi}}_{h-1}\left(d^{D}_{h-1}\widetilde{w}_{h-1}\right)}{d^{D,\dagger}_{h-1}}, \widetilde{w}_{h-1}, \overline{\pi}\right) \leq 0. \tag{20}$$

Combining Eq. (18), Eq. (19), Eq. (20) with the MLE bound of Eq. (12), with probability at least $1-\delta/(2H)$ we have

$$\|\widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h}\|_1 \leq \|\widehat{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1}\|_1 + 2C^{\mathbf{x}}_{h-1}\varepsilon_{\mathrm{mle}} + C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}\varepsilon_{\mathrm{mle}} + \sqrt{2}\varepsilon_{\mathrm{reg,h-1}}$$

$$\leq \|\widehat{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1}\|_1 + 3C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}\varepsilon_{\mathrm{mle}} + \sqrt{2}\varepsilon_{\mathrm{reg,h-1}}.$$

Finally, union bounding over $h \in [H]$, plugging in the definition of $\varepsilon_{\mathrm{mle}}$, and rearranging gives that Eq. (13) holds with probability at least $1 - \delta/2$. □

E.3. Proof of offline policy optimization

Theorem (Restatement of Theorem 3)**.** *Fix $\delta \in (0, 1)$ and suppose Assumption 1 and Assumption 2 hold. Given a policy class Π, let $\{\widehat{d}^{\pi}_h\}_{h\in[H],\pi\in\Pi}$ be the output of running Algorithm 1. Then with probability at least $1-\delta$, for any deterministic reward function R and policy selected as $\widehat{\pi}_R = \mathrm{argmax}_{\pi\in\Pi}\,\widehat{v}^{\pi}_R$, we have*

$$v^{\widehat{\pi}_R}_R \geq \max_{\pi\in\Pi} \overline{v}^{\pi}_R - \varepsilon,$$

where $\widehat{v}^{\pi}_R := \sum_{h=0}^{H-1} \iint \widehat{d}^{\pi}_h(x_h)R(x_h, a_h)\pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$ and $\overline{v}_R$ is defined similarly for $\{\overline{d}^{\pi}_h\}$. The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}H^3\left(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h\right)^2 \log(|\Pi|/\delta)/\varepsilon^2\right).$$

Additionally, define the set of policies fully covered by the data to be

$$\Pi^{\mathrm{covered}} = \left\{\pi \in \Pi : d^{\pi}_h = \overline{d}^{\pi}_h, \forall h \in [H]\right\}.$$

Then under the above guarantee, we also have

$$v^{\widehat{\pi}_R}_R \geq \max_{\pi\in\Pi^{\mathrm{covered}}} v^{\pi}_R - \varepsilon.$$

Proof. Firstly, Theorem 2 states that, with probability at least $1 - \delta/|\Pi|$, $\tilde{O}\left(\mathsf{d}H^3\left(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h\right)^2 \log(|\Pi|/\delta)/\varepsilon^2\right)$ samples are sufficient for learning $\{\widehat{d}^{\pi}_h\}$ such that $\|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\|_1 \leq \frac{\varepsilon}{2H}$ for all $h \in [H]$ and each $\pi \in \Pi$. Taking a union bound over $\pi \in \Pi$, with probability at least $1 - \delta$, we have that for all $h \in [H], \pi \in \Pi$,

$$\|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\|_1 \leq \frac{\varepsilon}{2H}.$$

Then since the R is bounded on $[0, 1]$, for any $\pi \in \Pi$ we have

$$\begin{aligned}
|\widehat{v}^{\pi}_R - \overline{v}^{\pi}_R| &= \sum_{h=0}^{H-1} \iint (\widehat{d}^{\pi}_h(x_h) - \overline{d}^{\pi}_h(x_h))R(x_h, a_h)\pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h) \\
&\leq \sum_{h=0}^{H-1} \int |\widehat{d}^{\pi}_h(x_h) - \overline{d}^{\pi}_h(x_h)| \left(\int \pi(a_h|x_h)(\mathrm{d}a_h)\right)(\mathrm{d}x_h) \\
&= \sum_{h=0}^{H-1} \|\widehat{d}^{\pi}_h - \overline{d}^{\pi}_h\|_1 \leq \varepsilon/2.
\end{aligned}$$

Denote $\overline{\pi}^*_R = \max_{\pi\in\Pi}\overline{v}^{\pi}_R$, and recall that we pick $\widehat{\pi}_R = \mathrm{argmax}_{\pi\in\Pi}\,\widehat{v}^{\pi}_R$. Then

$$v^{\widehat{\pi}_R}_R - \max_{\pi\in\Pi}\overline{v}^{\pi}_R = v^{\widehat{\pi}_R}_R - \overline{v}^{\overline{\pi}^*_R}_R \geq \overline{v}^{\widehat{\pi}_R}_R - \overline{v}^{\overline{\pi}^*_R}_R = \overline{v}^{\widehat{\pi}_R}_R - \widehat{v}^{\widehat{\pi}_R}_R + \widehat{v}^{\widehat{\pi}_R}_R - \widehat{v}^{\overline{\pi}^*_R}_R + \widehat{v}^{\overline{\pi}^*_R}_R - \overline{v}^{\overline{\pi}^*_R}_R \geq -\varepsilon,$$

where the first inequality follows from the fact that $d^{\pi}_h \geq \overline{d}^{\pi}_h$, thus $v^{\pi}_R \geq \overline{v}^{\pi}_R$. The second inequality results from the fact that $\widehat{v}^{\widehat{\pi}_R}_R \geq \widehat{v}^{\overline{\pi}^*_R}_R$ and $|\widehat{v}^{\pi}_R - \overline{v}^{\pi}_R| \leq \varepsilon/2$ for all $\pi \in \Pi$.

The result for Π^{covered} is straightforward from the observation that $\max_{\pi\in\Pi}\overline{v}^{\pi}_R \geq \max_{\pi\in\Pi^{\mathrm{covered}}} v^{\pi}_R$, since $\overline{v}^{\pi}_R = v^{\pi}_R$ for each covered policy. □

F. Online policy cover construction proofs (Section 4)

F.1. Proof of occupancy estimation

Lemma (Restatement of Lemma 4)**.** *For any $h \in [H]$ and $\pi \in \Pi$ in Algorithm 2,*

$$\left\|\overline{d}_h^{\pi} - d_h^{\pi}\right\|_1 \le \left\|\overline{d}_{h-1}^{\pi} - d_{h-1}^{\pi}\right\|_1 + 4\mathsf{d} \max_{\pi' \in \Pi} \left\|\widehat{d}_{h-1}^{\pi'} - \overline{d}_{h-1}^{\pi'}\right\|_1 .$$

Proof. Firstly, from the third claim of Proposition 2, we have that for any $h \in [H], \pi \in \Pi$

$$\|\overline{d}_h^{\pi} - d_h^{\pi}\|_1 \le \|\overline{d}_{h-1}^{\pi} - d_{h-1}^{\pi}\|_1 + \|\overline{d}_{h-1}^{\pi} - \overline{d}_{h-1}^{\pi} \wedge C_{h-1}^{\mathbf{x}} d_{h-1}^{D}\|_1 + \|\mathbf{P}_{h-1}^{\pi} d_{h-1}^{\pi} - \mathbf{P}_{h-1}^{\overline{\pi}} d_{h-1}^{\pi}\|_1. \tag{21}$$

Now we further simplify the latter two error terms on the RHS of Eq. (21) by noticing that $C_h^{\mathbf{x}} = \mathsf{d}$ and $C_h^{\mathbf{a}} = K$ for all $h \in [H]$. For the last term, $\pi^D = \mathrm{unif}(\mathcal{A})$ gives us

$$\overline{\pi}(a_{h-1}|x_{h-1}) = \min\{\pi(a_{h-1}|x_{h-1}), C_{h-1}^{\mathbf{a}} \pi^D(a_{h-1}|x_{h-1})\} = \min\{\pi(a_{h-1}|x_{h-1}), 1)\} = \pi(a_{h-1}|x_{h-1})$$

and thus $\left\|\mathbf{P}_{h-1}^{\pi} d_{h-1}^{\pi} - \mathbf{P}_{h-1}^{\overline{\pi}} d_{h-1}^{\pi}\right\|_1 = 0$. For the middle term, we expand the expression as

$$\left\|\overline{d}_{h-1}^{\pi} - \overline{d}_{h-1}^{\pi} \wedge \mathsf{d} d_{h-1}^{D}\right\|_1 = \int \overline{d}_{h-1}^{\pi}(x_{h-1}) - \left(\overline{d}_{h-1}^{\pi} \wedge \mathsf{d} d_{h-1}^{D}\right)(x_{h-1})(\mathrm{d}x_{h-1}).$$

Consider a fixed $x_{h-1} \in \mathcal{X}$. Note that $\overline{d}_{h-1}^{\pi}(x_{h-1}) - \left(\overline{d}_{h-1}^{\pi} \wedge \mathsf{d} d_{h-1}^{D}\right)(x_{h-1})$ is nonzero only if $\mathsf{d} d_{h-1}^{D}(x_{h-1}) < \overline{d}_{h-1}^{\pi}(x_{h-1})$, for which we have

$$\begin{aligned}
&\overline{d}_{h-1}^{\pi}(x_{h-1}) - \left(\overline{d}_{h-1}^{\pi} \wedge \mathsf{d} d_{h-1}^{D}\right)(x_{h-1}) = \overline{d}_{h-1}^{\pi}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1}) \\
&\le \widehat{d}_{h-1}^{\pi}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + \left|\overline{d}_{h-1}^{\pi}(x_{h-1}) - \widehat{d}_{h-1}^{\pi}(x_{h-1})\right| .
\end{aligned}$$

To bound $\widehat{d}_{h-1}^{\pi}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1})$, we have

$$\begin{aligned}
&\widehat{d}_{h-1}^{\pi}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1}) \\
&\le \widetilde{d}_{h-1}^{\pi}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + \left|\widehat{d}_{h-1}^{\pi}(x_{h-1}) - \widetilde{d}_{h-1}^{\pi}(x_{h-1})\right| \\
&\le \sum_{i=1}^{\mathsf{d}} \left|\widetilde{d}_{h-1}^{\pi^{h-1,i}}(x_{h-1})\right| - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + \left|\widehat{d}_{h-1}^{\pi}(x_{h-1}) - \widetilde{d}_{h-1}^{\pi}(x_{h-1})\right| \\
&\le \sum_{i=1}^{\mathsf{d}} \left|\widehat{d}_{h-1}^{\pi^{h-1,i}}(x_{h-1})\right| - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + (\mathsf{d}+1) \max_{\pi' \in \Pi} \left|\widehat{d}_{h-1}^{\pi'}(x_{h-1}) - \widetilde{d}_{h-1}^{\pi'}(x_{h-1})\right| \\
&\le \sum_{i=1}^{\mathsf{d}} \left|\overline{d}_{h-1}^{\pi^{h-1,i}}(x_{h-1})\right| - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + (\mathsf{d}+1) \max_{\pi' \in \Pi} \left|\widehat{d}_{h-1}^{\pi'}(x_{h-1}) - \widetilde{d}_{h-1}^{\pi'}(x_{h-1})\right| \\
&\quad + \mathsf{d} \max_{\pi' \in \Pi} \left|\widehat{d}_{h-1}^{\pi'}(x_{h-1}) - \overline{d}_{h-1}^{\pi'}(x_{h-1})\right| \\
&= \sum_{i=1}^{\mathsf{d}} \overline{d}_{h-1}^{\pi^{h-1,i}}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + (\mathsf{d}+1) \max_{\pi' \in \Pi} \left|\widehat{d}_{h-1}^{\pi'}(x_{h-1}) - \widetilde{d}_{h-1}^{\pi'}(x_{h-1})\right| \\
&\quad + \mathsf{d} \max_{\pi' \in \Pi} \left|\widehat{d}_{h-1}^{\pi'}(x_{h-1}) - \overline{d}_{h-1}^{\pi'}(x_{h-1})\right| \\
&\le \sum_{i=1}^{\mathsf{d}} d_{h-1}^{\pi^{h-1,i}}(x_{h-1}) - \mathsf{d} d_{h-1}^{D}(x_{h-1}) + (\mathsf{d}+1) \max_{\pi' \in \Pi} \left|\widehat{d}_{h-1}^{\pi'}(x_{h-1}) - \widetilde{d}_{h-1}^{\pi'}(x_{h-1})\right|
\end{aligned}$$

$$
\begin{aligned}
&+\ \mathsf{d} \max_{\pi' \in \Pi} \left| \widehat{d}^{\pi'}_{h-1}(x_{h-1}) - \overline{d}^{\pi'}_{h-1}(x_{h-1}) \right| \\
&= (\mathsf{d}+1) \max_{\pi' \in \Pi} \left| \widehat{d}^{\pi'}_{h-1}(x_{h-1}) - \widetilde{d}^{\pi'}_{h-1}(x_{h-1}) \right| + \mathsf{d} \max_{\pi' \in \Pi} \left| \widehat{d}^{\pi'}_{h-1}(x_{h-1}) - \overline{d}^{\pi'}_{h-1}(x_{h-1}) \right| .
\end{aligned}
$$

In the second inequality, we use that $\Pi^{\mathrm{expl}}_{h-1} = \{\pi^{h-1,1}, \ldots, \pi^{h-1,\mathsf{d}}\}$ are the policies corresponding to the barycentric spanner, which Lemma 15 guarantees to be of cardinality no larger than d. The first equality is because $\overline{d}^{\pi}_{h-1}(x_{h-1}) \geq 0, \forall \pi$, which can be seen by the induction definition in Eq. (4) and the non-negativity of d_0. The fifth inequality is due to $\overline{d}^{\pi}_{h-1}(x_{h-1}) \leq d^{\pi}_{h-1}(x_{h-1}), \forall \pi$, which can be shown inductively by noticing $\overline{d}^{\pi}_{0} \leq d^{\pi}_{0}$ and the definition of $\overline{d}^{\pi}_{h}$ in Eq. (4). The last equality can be seen from that $d^{D}_{h-1}(x_{h-1})$ is the marginal distribution of $\mathcal{D}_{h-1}$ and $\mathcal{D}_{h-1}$ is rolled in with $\mathrm{unif}(\Pi^{\mathrm{expl}}_{h-1})$.

Integrating over x_{h-1} yields

$$
\left\| \overline{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1} \wedge \mathsf{d} d^{D}_{h-1} \right\|_1 \leq (\mathsf{d}+1) \max_{\pi' \in \Pi} \left\| \widehat{d}^{\pi'}_{h-1} - \widetilde{d}^{\pi'}_{h-1} \right\|_1 + (\mathsf{d}+1) \max_{\pi' \in \Pi} \left\| \widehat{d}^{\pi'}_{h-1} - \overline{d}^{\pi'}_{h-1} \right\|_1 .
$$

Since $\overline{d}^{\pi'}_{h-1} = \mathbf{P}^{\pi'}_{h-2}(\overline{d}^{\pi}_{h-2} \wedge C^{\mathbf{x}}_{h-2} d^{D}_{h-2}) = \mathbf{P}^{\pi'}_{h-2}(\overline{d}^{\pi}_{h-2} \wedge \mathsf{d} d^{D}_{h-2})$ is linear in the features μ^{*}_{h-2} (Lemma 16), and $\widetilde{d}^{\pi'}_{h-1}$ is the closest linear approximation in the ℓ_1 norm to $\widehat{d}^{\pi'}_{h-1}$ (line 7), for any $\pi' \in \Pi$ we have

$$
\left\| \widehat{d}^{\pi'}_{h-1} - \widetilde{d}^{\pi'}_{h-1} \right\|_1 \leq \left\| \widehat{d}^{\pi'}_{h-1} - \overline{d}^{\pi'}_{h-1} \right\|_1 \tag{22}
$$

and thus

$$
\left\| \overline{d}^{\pi}_{h-1} - \overline{d}^{\pi}_{h-1} \wedge \mathsf{d} d^{D}_{h-1} \right\|_1 \leq 2(\mathsf{d}+1) \max_{\pi' \in \Pi} \left\| \widehat{d}^{\pi'}_{h-1} - \overline{d}^{\pi'}_{h-1} \right\|_1 . \tag{23}
$$

Then combining Eq. (21) with Eq. (23) gives

$$
\left\| \overline{d}^{\pi}_{h} - d^{\pi}_{h} \right\|_1 \leq \left\| \overline{d}^{\pi}_{h-1} - d^{\pi}_{h-1} \right\|_1 + 4\mathsf{d} \max_{\pi' \in \Pi} \left\| \widehat{d}^{\pi'}_{h-1} - \overline{d}^{\pi'}_{h-1} \right\|_1 . \qquad \square
$$

Theorem (Restatement of Theorem 5). *Fix $\delta \in (0,1)$ and consider an MDP $\mathcal{M}$ that satisfies Assumption 1, where the right feature μ^* is known. Then by setting*

$$
n_{\mathrm{mle}} = \widetilde{O}\left(\frac{\mathsf{d}^3 K^2 H^4 \log(1/\delta)}{\varepsilon^2} \right), n_{\mathrm{reg}} = \widetilde{O}\left(\frac{\mathsf{d}^5 K^2 H^4 \log(|\Pi|/\delta)}{\varepsilon^2} \right), n = n_{\mathrm{mle}} + n_{\mathrm{reg}},
$$

with probability at least $1-\delta$, FORCE *returns state occupancy estimates $\{\widehat{d}^{\pi}_{h}\}_{h=0}^{H-1}$ satisfying that*

$$
\|\widehat{d}^{\pi}_{h} - d^{\pi}_{h}\|_1 \leq \varepsilon, \forall h \in [H], \pi \in \Pi.
$$

The total number of episodes required by the algorithm is

$$
\widetilde{O}(nH) = \widetilde{O}\left(\frac{\mathsf{d}^5 K^2 H^5 \log(|\Pi|/\delta)}{\varepsilon^2} \right).
$$

Proof. From Algorithm 2, we know that dataset $\mathcal{D}_{0:H-1}$ satisfies Assumption 2 and for each $\pi \in \Pi$, $\widehat{d}^{\pi}_{h}$ is estimated in the same way as that in Algorithm 1. Therefore, we can follow the same steps as the proof of Theorem 2. By setting $C^{\mathbf{x}}_{h} = \mathsf{d}$ and $C^{\mathbf{a}}_{h} = K$ for all $h \in [H]$ in Eq. (15), with probability at least $1-\delta$, for any policy $\pi \in \Pi$, we get that

$$
\left\| \widehat{d}^{\pi}_{h} - \overline{d}^{\pi}_{h} \right\|_1 \leq 18h\mathsf{d}^{3/2}K\sqrt{\frac{\log(16HB^{\mu}n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}} + 666h\mathsf{d}^{3/2}K\sqrt{\frac{\log\left(2|\Pi|Hn_{\mathrm{reg}}/\delta\right)}{n_{\mathrm{reg}}}}. \tag{24}
$$

The primary difference between the above results and the corresponding statements in Theorem 2 is that the regression error in Eq. (24) includes an additional union bound over all $\pi \in \Pi$. This is because Algorithm 2 performs estimation for all policies, while Algorithm 1 only concerns a single fixed policy. We note that this change in the proof occurs only through application of Lemma 14, which is stated generally and already includes a union bound over all policies of interest. Because MLE estimation occurs only for the data distribution and is policy-agnostic, the MLE error (second term) does not require such a union bound.

Next, to bound the missingness error, from Lemma 4, we have

$$\left\|\overline{d}_h^\pi - d_h^\pi\right\|_1 \le \left\|\overline{d}_{h-1}^\pi - d_{h-1}^\pi\right\|_1 + 4\mathsf{d} \max_{\pi'\in\Pi} \left\|\widehat{d}_{h-1}^{\pi'} - \overline{d}_{h-1}^{\pi'}\right\|_1. \tag{25}$$

Unfolding Eq. (25) yields

$$\left\|\overline{d}_h^\pi - d_h^\pi\right\|_1 \le 4\mathsf{d} \sum_{h'=0}^{h-1} \max_{\pi'\in\Pi} \left\|\widehat{d}_{h'}^{\pi'} - \overline{d}_{h'}^{\pi'}\right\|_1. \tag{26}$$

Plugging the bound for $\left\|\widehat{d}_{h'}^{\pi'} - \overline{d}_{h'}^{\pi'}\right\|_1$ from Eq. (24) into Eq. (26) gives

$$\left\|\overline{d}_h^\pi - d_h^\pi\right\|_1 \le 72h^2\mathsf{d}^{3/2}K\sqrt{\frac{\log(16HB^\mu n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}} + 2664h^2\mathsf{d}^{5/2}K\sqrt{\frac{\log\left(2|\Pi|Hn_{\mathrm{reg}}/\delta\right)}{n_{\mathrm{reg}}}}. \tag{27}$$

Combining Eq. (24) and Eq. (27) via triangle inequality and simplifying, we have

$$\left\|\widehat{d}_h^\pi - d_h^\pi\right\|_1 \le 90h^2\mathsf{d}^{3/2}K\sqrt{\frac{\log(16HB^\mu n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}} + 3330h^2\mathsf{d}^{5/2}K\sqrt{\frac{\log\left(2|\Pi|Hn_{\mathrm{reg}}/\delta\right)}{n_{\mathrm{reg}}}}.$$

Finally, noticing that $n_{\mathrm{mle}} = \widetilde{O}\left(\frac{\mathsf{d}^3K^2H^4\log(1/\delta)}{\varepsilon^2}\right)$, $n_{\mathrm{reg}} = \widetilde{O}\left(\frac{\mathsf{d}^5K^2H^4\log(|\Pi|/\delta)}{\varepsilon^2}\right)$, $n = n_{\mathrm{mle}} + n_{\mathrm{reg}}$ completes the proof. □

F.2. Proof of online policy optimization

First, we prove Proposition 1, from which our online policy optimization guarantee (Theorem 6) follows when combined with Theorem 5.

Proposition 6 (Restatement of Proposition 1)**.** *Given any policy π and reward function*[10] *$R = \{R_h\}$ with $R_h : \mathcal{X}\times\mathcal{A}\to[0,1]$, define expected return as $v_R^\pi := \mathbb{E}_\pi[\sum_{h=0}^{H-1} R_h(x_h,a_h)] = \sum_{h=0}^{H-1}\iint d_h^\pi(x_h)R_h(x_h,a_h)\,\pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$. Then for $\{\widehat{d}_h^\pi\}$ such that $\|\widehat{d}_h^\pi - d_h^\pi\|_1 \le \varepsilon/(2H)$ for all $\pi\in\Pi$ and $h\in[H]$, and policy chosen as*

$$\widehat{\pi}_R = \operatorname*{argmax}_{\pi\in\Pi} \widehat{v}_R^\pi,$$

we have

$$v_R^{\widehat{\pi}_R} \ge \max_{\pi\in\Pi} v_R^\pi - \varepsilon,$$

where $\widehat{v}_R^\pi = \sum_{h=0}^{H-1}\iint \widehat{d}_h^\pi(x_h)R_h(x_h,a_h)\pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$ is the expected return calculated using $\{\widehat{d}_h^\pi\}$.

Proof. Since the R is bounded on $[0,1]$, for any $\pi\in\Pi$ we have

$$\begin{aligned}
|\widehat{v}_R^\pi - v_R^\pi| &= \sum_{h=0}^{H-1}\iint(\widehat{d}_h^\pi(x_h) - d_h^\pi(x_h))R(x_h,a_h)\pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h) \\
&\le \sum_{h=0}^{H-1}\int|\widehat{d}_h^\pi(x_h) - d_h^\pi(x_h)|\left(\int\pi(a_h|x_h)(\mathrm{d}a_h)\right)(\mathrm{d}x_h)
\end{aligned}$$

[10] We assume known & deterministic rewards, and can easily handle unknown/stochastic versions (Appendix D.2).

$$= \sum_{h=0}^{H-1} \|d_h^\pi - \widehat{d}_h^\pi\|_1 \le \varepsilon/2.$$

Next, recall we pick $\widehat{\pi}_R = \operatorname{argmax}_{\pi\in\Pi} \widehat{v}_R^\pi$, and denote $\pi_R^* = \operatorname{argmax}_{\pi\in\Pi} v_R^\pi$. Then using the above inequality, we have

$$v_R^{\widehat{\pi}_R} - \max_{\pi\in\Pi} v_R^\pi = v_R^{\widehat{\pi}_R} - v_R^{\pi_R^*} = v_R^{\widehat{\pi}_R} - \widehat{v}_R^{\widehat{\pi}_R} + \widehat{v}_R^{\widehat{\pi}_R} - \widehat{v}_R^{\pi_R^*} + \widehat{v}_R^{\pi_R^*} - v_R^{\pi_R^*} \ge -\varepsilon$$

since $\widehat{v}_R^{\widehat{\pi}_R} \ge \widehat{v}_R^{\pi_R^*}$, completing the proof. □

Theorem 9 (Restatement of Theorem 6)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1 and Assumption 2 hold, and μ^* is known. Given a policy class Π, let $\{\widehat{d}_h^\pi\}_{h\in[H],\pi\in\Pi}$ be the output of running* FORCE*. Then with probability at least $1-\delta$, for any reward function R and policy selected as $\widehat{\pi}_R = \operatorname{argmax}_{\pi\in\Pi} \widehat{v}_R^\pi$, we have*

$$v_R^{\widehat{\pi}_R} \ge \max_{\pi\in\Pi} v_R^\pi - \varepsilon,$$

where v_R^π and $\widehat{v}_R^\pi$ are defined in Proposition 1. The total number of episodes required by the algorithm is

$$\tilde{O}\left(\frac{\mathsf{d}^5 K^2 H^7 \log(|\Pi|/\delta)}{\varepsilon^2}\right).$$

Proof. The proof takes similar steps as the proof of Theorem 3. From Theorem 5, w.p. $\ge 1-\delta$, we obtain estimates $\{\widehat{d}_h^\pi\}$ such that $\|d_h^\pi - \widehat{d}_h^\pi\|_1 \le \frac{\varepsilon}{2H}$ for all $\pi \in \Pi$ with $\tilde{O}\left(\frac{\mathsf{d}^5 K^2 H^7 \log(|\Pi|/\delta)}{\varepsilon^2}\right)$ total number of samples, where we use the union bound over $\pi \in \Pi$. Combining this with Proposition 1 gives the result. □

G. Representation learning

In this section, we present the detailed algorithms and results for the representation learning setting (Section 5), where the true density features are not given but must also be learned from an exponentially large candidate feature set. The algorithms and analyses mostly follow that of the known density feature case (Section 3 and Section 4), therefore, we mainly discuss the difference here.

G.1. Off-policy occupancy estimation

We start with describing our algorithm FORCRL (Algorithm 3), which estimates the occupancy distribution d_h^π of any given policy π using an offline dataset $\mathcal{D}_{0:H-1}$ when the true density feature μ^* is unknown and the learner is given a realizable density feature class $\Upsilon \ni \mu^*$ (see Assumption 3).

As discussed in Section 5, instead of using μ^* to construct the function classes, a natural choice here is to use the union of all linear function classes. Since now the feature comes from candidate feature classes $\Upsilon_{h-2}, \Upsilon_{h-1}$, in line 4 of Algorithm 3, we use different function classes $\mathcal{F}_{h-1}(\Upsilon_{h-2}), \mathcal{F}_h(\Upsilon_{h-1})$ as defined in Eq. (28) for the MLE objective. In addition, in line 5 of Algorithm 3, now we run regression with a different function class $\mathcal{W}_h(\Upsilon_{h-1})$ as defined in Eq. (29).

Similar as in the known feature case counterpart (Theorem 2), we have the following guarantee for estimating d^π.

Theorem (Restatement of Theorem 7)**.** *Fix $\delta \in (0,1)$. Suppose Assumption 1, Assumption 2, and Assumption 3 hold. Then, given an evaluation policy π, by setting*

$$n_{\text{mle}} = \tilde{O}\left(\mathsf{d}\left(\sum_{h\in[H]} C_h^{\mathbf{x}} C_h^{\mathbf{a}}\right)^2 \log(|\Upsilon|/\delta)/\varepsilon^2\right) \text{ and } n_{\text{reg}} = \tilde{O}\left(\mathsf{d}\left(\sum_{h\in[H]} C_h^{\mathbf{x}} C_h^{\mathbf{a}}\right)^2 \log(|\Upsilon|/\delta)/\varepsilon^2\right),$$

with probability at least $1-\delta$, FORCRL *(Algorithm 3) returns state occupancy estimates $\{\widehat{d}_h^\pi\}_{h=0}^{H-1}$ satisfying that*

$$\left\|\widehat{d}_h^\pi - \overline{d}_h^\pi\right\|_1 \le \varepsilon, \forall h \in [H].$$

Algorithm 3 Fitted **O**ccupancy Ite**r**ation with **C**lipping and Representation Learning (FORCRL)

```
Input: policy $\pi$, density feature class $\Upsilon$, dataset $\mathcal{D}_{0:H-1}$, sample sizes $n_{\mathrm{mle}}$ and $n_{\mathrm{reg}}$, clipping thresholds $\{C^{\mathbf{x}}_h\}$ and $\{C^{\mathbf{a}}_h\}$.
Initialize $\widehat{d}^\pi_0 = d_0,\ \forall \pi \in \Pi$.
for $h = 1, \ldots, H$ do
    Randomly split $\mathcal{D}_{h-1}$ to two folds $\mathcal{D}^{\mathrm{mle}}_{h-1}$ and $\mathcal{D}^{\mathrm{reg}}_{h-1}$ with sizes $n_{\mathrm{mle}}$ and $n_{\mathrm{reg}}$ respectively.
    Estimate marginal data distributions $\widehat{d}^D_{h-1}(x_{h-1})$ and $\widehat{d}^{D,\dagger}_{h-1}(x_h)$ by MLE with dataset $\mathcal{D}^{\mathrm{mle}}_{h-1}$.
        $$\widehat{d}^D_{h-1} = \mathop{\mathrm{argmax}}_{d_{h-1}\in\mathcal{F}_{h-1}(\Upsilon_{h-2})} \frac{1}{n_{\mathrm{mle}}}\sum_{i=1}^{n_{\mathrm{mle}}} \log\left(d_{h-1}(x^{(i)}_{h-1})\right) \text{ and } \widehat{d}^{D,\dagger}_{h-1} = \mathop{\mathrm{argmax}}_{d_h\in\mathcal{F}_h(\Upsilon_{h-1})} \frac{1}{n_{\mathrm{mle}}}\sum_{i=1}^{n_{\mathrm{mle}}} \log\left(d_h(x^{(i)}_h)\right)$$
    where
        $$\mathcal{F}_h(\Upsilon_{h-1}) = \left\{d_h = \langle \mu_{h-1}, \theta_h\rangle : d_h \in \Delta(\mathcal{X}), \mu_{h-1} \in \Upsilon_{h-1}, \theta_h \in \mathbb{R}^{\mathsf{d}}, \|\theta_h\|_\infty \le 1\right\}. \quad (28)$$
    Define $\mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}(w_h, w_{h-1}, \overline{\pi}_{h-1}) := \frac{1}{n_{\mathrm{reg}}}\sum_{i=1}^{n_{\mathrm{reg}}}\left(w_h(x^{(i)}_h) - w_{h-1}(x^{(i)}_{h-1})\frac{\overline{\pi}_{h-1}(a^{(i)}_{h-1}|x^{(i)}_{h-1})}{\pi^D_{h-1}(a^{(i)}_{h-1}|x^{(i)}_{h-1})}\right)^2$ and estimate
        $$\widehat{w}^\pi_h = \mathop{\mathrm{argmin}}_{w_h\in\mathcal{W}_h(\Upsilon_{h-1})} \mathcal{L}_{\mathcal{D}^{\mathrm{reg}}_{h-1}}\left(w_h, \frac{\widehat{d}^\pi_{h-1}\wedge C^{\mathbf{x}}_{h-1}\widehat{d}^D_{h-1}}{\widehat{d}^D_{h-1}}, \pi_{h-1}\wedge C^{\mathbf{a}}_{h-1}\pi^D_{h-1}\right)$$
    where
        $$\mathcal{W}_h(\Upsilon_{h-1}) = \left\{w_h = \frac{\langle \mu_{h-1}, \theta^{\mathrm{up}}_h\rangle}{\langle \mu_{h-1}, \theta^{\mathrm{down}}_h\rangle} : \|w_h\|_\infty \le C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}, \mu_{h-1}\in\Upsilon_{h-1}, \theta^{\mathrm{up}}_h, \theta^{\mathrm{down}}_h \in \mathbb{R}^{\mathsf{d}}\right\}. \quad (29)$$
    Set the estimate $\widehat{d}^\pi_h = \widehat{w}^\pi_h\, \widehat{d}^{D,\dagger}_{h-1}$.
end for
Output: estimated state occupancies $\{\widehat{d}^\pi_h\}_{h\in[H]}$.
```

The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}H\left(\sum_{h\in[H]} C^{\mathbf{x}}_h C^{\mathbf{a}}_h\right)^2 \log(|\Upsilon|/\delta)/\varepsilon^2\right).$$

Proof. The proof for this theorem largely follows its counterpart for the known feature case (Theorem 2), and we mainly discuss the different steps here. We now make the following two slightly different claims on MLE estimation and error propagation. Based on them, the final error bound is obtained in the same way as Theorem 2.

Claim 1 Our estimated data distributions satisfy that with probability $1-\delta/2$, for any $h\in[H]$

$$\left\|\widehat{d}^D_h - d^D_h\right\|_1 \le \varepsilon_{\mathrm{mle}} \text{ and } \left\|\widehat{d}^{D,\dagger}_h - d^{D,\dagger}_h\right\|_1 \le \varepsilon_{\mathrm{mle}}, \tag{30}$$

where

$$\varepsilon_{\mathrm{mle}} := 6\sqrt{\frac{\mathsf{d}\log(16H|\Upsilon|B^\mu n_{\mathrm{mle}}/\delta)}{n_{\mathrm{mle}}}}.$$

Claim 2 Under the high-probability event that Eq. (30) holds, we further have with probability at least $1-\delta/2$, for any $1\le h\le H$, we have

$$\left\|\widehat{d}^\pi_h - \overline{d}^\pi_h\right\|_1 \le \left\|\widehat{d}^\pi_{h-1} - \overline{d}^\pi_{h-1}\right\|_1 + 3C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1}\varepsilon_{\mathrm{mle}} + \sqrt{2}\varepsilon_{\mathrm{reg,h-1}},$$

where

$$\varepsilon_{\text{reg,h}-1} := \sqrt{\frac{221184\mathsf{d}(C^{\mathbf{x}}_{h-1}C^{\mathbf{a}}_{h-1})^2\log\left(2H|\Upsilon|n_{\text{reg}}/\delta\right)}{n_{\text{reg}}}}. \tag{31}$$

Proof of Claim 1 Notice that for the term ε_{mle} in Eq. (30), we now have an additional $|\Upsilon|$ factor inside the $\log$. The reason is that here we use $\mathcal{F}_{h-1}(\Upsilon_{h-2}), \mathcal{F}_h(\Upsilon_{h-1})$ instead of $\mathcal{F}_{h-1}, \mathcal{F}_h$. By Lemma 22, the two function classes considered here have ℓ_1 optimistic covers with scale $1/n_{\text{mle}}$ of size $|\Upsilon|\left(2\lceil B^\mu n_{\text{mle}}\rceil\right)^{\mathsf{d}}$. In addition, we still have that $d^D_{h-1} \in \mathcal{F}_{h-1}(\Upsilon_{h-2}), d^{D,\dagger}_{h-1} \in \mathcal{F}_h(\Upsilon_{h-1})$ from Lemma 18, and any $d_{h-1} \in \mathcal{F}_{h-1}(\Upsilon_{h-2}), \mathcal{F}_h(\Upsilon_{h-1})$ is a valid probability distribution over $\mathcal{X}$.

Proof of Claim 2 This proof mostly follows the proof of Claim 2 in Theorem 2. The difference is that the function class $\mathcal{W}_h(\Upsilon_{h-1})$ now consists of all features in Υ_{h-1} instead of only the true feature μ^*_{h-1}. Therefore, in Eq. (31), the term $\varepsilon_{\text{reg,h}-1}$ has an additional $|\Upsilon|$ inside the $\log$, which is from the counterpart of Eq. (19). It is also easy to see that $\frac{\mathbf{P}^{\overline{\pi}}_{h-1}(d^D_{h-1}\widetilde{w}_{h-1})}{d^{D,\dagger}_{h-1}} \in \mathcal{W}_h(\Upsilon_{h-1})$ by following the same logic before. Further noticing that $\mu^*_{h-1} \in \Upsilon_{h-1}$, we again have Eq. (20) holds here. □

Theorem 10 (Offline policy optimization with representation learning)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1, Assumption 2, and Assumption 3 hold. Given a policy class Π, let $\{\widehat{d}^\pi_h\}_{h\in[H],\pi\in\Pi}$ be the output of running Algorithm 3. Then with probability at least $1-\delta$, for any deterministic reward function R and policy selected as $\widehat{\pi}_R = \operatorname{argmax}_{\pi\in\Pi}\widehat{v}^\pi_R$, we have*

$$v^{\widehat{\pi}_R}_R \geq \max_{\pi\in\Pi}\overline{v}^\pi_R - \varepsilon,$$

where v^π_R and $\widehat{v}^\pi_R$ are defined in Proposition 1, and $\overline{v}_R$ is defined similarly for $\{\overline{d}^\pi_h\}$. The total number of episodes required by the algorithm is

$$\tilde{O}\left(\mathsf{d}H^3\left(\sum_{h\in[H]}C^{\mathbf{x}}_hC^{\mathbf{a}}_h\right)^2\log(|\Pi||\Upsilon|/\delta)/\varepsilon^2\right).$$

Additionally, define the set of policies fully covered by the data to be

$$\Pi^{\text{covered}} = \left\{\pi\in\Pi : d^\pi_h = \overline{d}^\pi_h, \forall h\in[H]\right\}.$$

Then with the same total number of episodes required by the algorithm, for any reward function R and policy selected as $\widehat{\pi}_R = \operatorname{argmax}_{\pi\in\Pi^{\text{covered}}}\widehat{v}^\pi_R$, with probability at least $1-\delta$, we have

$$v^{\widehat{\pi}_R}_R \geq \max_{\pi\in\Pi^{\text{covered}}} v^\pi_R - \varepsilon.$$

Proof. The proof follows the same steps as that of Theorem 3. Notice that now we will apply Theorem 7 rather than Theorem 2 to get the bound $\|\widehat{d}^\pi_h - \overline{d}^\pi_h\|_1$, which leads to the additional $\log(|\Upsilon|)$ factor. □

G.2. Online policy cover construction

Now we present the algorithm ForcRle (Algorithm 4), which estimates the occupancy distribution d^π_h of any given policy π with the access of online interaction. Again the true density feature μ^* is unknown and the learner is given a realizable density feature class Υ ($\mu^* \in \Upsilon$).

Similar as the know feature case online algorithm (Algorithm 2), we use the offline algorithm (Algorithm 3) as a submodule. However, as discussed in the main text, the crucial different step is to select a representation $\widehat{\mu}_{h-1}$ in Eq. (32) in line 8 before setting $\widetilde{d}^\pi_h$. This guarantee the cardinality of the barycentric spanner is at most d. Then the state occupancy $\widetilde{d}^\pi_h$ is set as the linear estimate using $\widehat{\mu}_{h-1}$ (rather than using μ^*_{h-1} in the known feature case) in line 9.

Similar as in the known feature case counterpart (Theorem 5), we have the following guarantee for estimating d^π.

Algorithm 4 FORCRL-guided Exploration (FORCRLE)

```
Input: policy class $\Pi$, density feature class $\Upsilon$, $n = n_{\text{mle}} + n_{\text{reg}}$
 1: Initialize $\widehat{d}_0^\pi = d_0$ and $\widetilde{d}_0^\pi = d_0$, $\forall \pi \in \Pi$.
 2: for $h = 1, \ldots, H$ do
 3:   Construct $\{\widetilde{d}_{h-1}^{\pi^{h-1,i}}\}_{i=1}^{\mathsf{d}}$ as the barycentric spanner of $\{\widetilde{d}_{h-1}^\pi\}_{\pi \in \Pi}$, and set $\Pi_{h-1}^{\text{expl}} = \{\pi^{h-1,i}\}_{i=1}^{\mathsf{d}}$.
 4:   Draw a tuple dataset $\mathcal{D}_{h-1} = \{(x_{h-1}^{(i)}, a_{h-1}^{(i)}, x_h^{(i)})\}_{i=1}^n$ using $\text{unif}(\Pi_{h-1}^{\text{expl}}) \circ \text{unif}(\mathcal{A})$.
 5:   for $\pi \in \Pi$ do
 6:     Estimate $\widehat{d}_h^\pi$ using the $h$-level loop[11] of Algorithm 3 (lines 4-6) with $\mathcal{D}_h$, $\widehat{d}_{h-1}^\pi$, $C_h^{\mathbf{x}} = \mathsf{d}$, $C_h^{\mathbf{a}} = K$.
 7:   end for
 8:   Select feature $\widehat{\mu}_{h-1}$ according to
$$\widehat{\mu}_{h-1} = \min_{\mu_{h-1} \in \Upsilon_{h-1}} \max_{\pi \in \Pi} \min_{\theta_h \in \mathbb{R}^{\mathsf{d}}} \|\langle \mu_{h-1}, \theta_h \rangle - \widehat{d}_h^\pi\|_1. \tag{32}$$
 9:   For all $\pi \in \Pi$, set the closest linear approximation to $\widehat{d}_h^\pi$ with feature $\widehat{\mu}_{h-1}$ as $\widetilde{d}_h^\pi = \langle \widehat{\mu}_{h-1}, \widetilde{\theta}_h \rangle$, where $\widetilde{\theta}_h = \text{argmin}_{\theta_h \in \mathbb{R}^{\mathsf{d}}} \|\langle \widehat{\mu}_{h-1}, \theta_h \rangle - \widehat{d}_h^\pi\|_1$.
10: end for
Output: estimated state occupancy measure $\{\widehat{d}_h^\pi\}_{h \in [H], \pi \in \Pi}$.
```

Theorem (Restatement of Theorem 8)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1 and Assumption 3 hold. Then by setting*

$$n_{\text{mle}} = \widetilde{O}\left(\frac{\mathsf{d}^3 K^2 H^4 \log(|\Upsilon|/\delta)}{\varepsilon^2}\right), n_{\text{reg}} = \widetilde{O}\left(\frac{\mathsf{d}^5 K^2 H^4 \log(|\Pi||\Upsilon|/\delta)}{\varepsilon^2}\right), n = n_{\text{mle}} + n_{\text{reg}},$$

with probability at least $1 - \delta$, FORCRLE *(Algorithm 4) returns state occupancy estimates $\{\widehat{d}_h^\pi\}_{h=0}^{H-1}$ satisfying that*

$$\|\widehat{d}_h^\pi - d_h^\pi\|_1 \le \varepsilon, \forall h \in [H], \pi \in \Pi.$$

The total number of episodes required by the algorithm is

$$\widetilde{O}(nH) = \widetilde{O}\left(\frac{\mathsf{d}^5 K^2 H^5 \log(|\Pi||\Upsilon|/\delta)}{\varepsilon^2}\right).$$

Proof. The proof for this theorem largely follows its counterpart for the known feature case (Theorem 5), and we only discuss the different steps here.

Firstly, Lemma 4 still holds. However, since we use "joint linearization" in line 8 and line 9, we need to modify the proof of Eq. (22) as the following. Again, we have $\overline{d}_{h-1}^{\pi'} = \mathbf{P}_{h-2}^{\pi'}(\overline{d}_{h-2}^\pi \wedge C_{h-2}^{\mathbf{x}} d_{h-2}^D) = \mathbf{P}_{h-2}^{\pi'}(\overline{d}_{h-2}^\pi \wedge \mathsf{d} d_{h-2}^D)$ is linear in the true feature μ_{h-2}^* (Lemma 16). Together with the feature selection criteria Eq. (32), we have that

$$\max_{\pi' \in \Pi} \|\widetilde{d}_{h-1}^{\pi'} - \widehat{d}_{h-1}^{\pi'}\|_1 = \max_{\pi' \in \Pi} \min_{\theta_{h-1} \in \mathbb{R}^d} \|\langle \widehat{\mu}_{h-2}, \theta_{h-1} \rangle - \widehat{d}_{h-1}^{\pi'}\|_1$$
$$\le \max_{\pi' \in \Pi} \min_{\theta_{h-1} \in \mathbb{R}^d} \|\langle \mu_{h-2}^*, \theta_{h-1} \rangle - \widehat{d}_{h-1}^{\pi'}\|_1 \le \max_{\pi' \in \Pi} \|\overline{d}_{h-1}^{\pi'} - \widehat{d}_{h-1}^{\pi'}\|_1.$$

For Eq. (24), we will have an additional $|\Upsilon|$ factor inside the log as

$$\varepsilon_{\text{mle}} := 6\sqrt{\frac{\mathsf{d} \log(16H|\Upsilon|B^\mu n_{\text{mle}}/\delta)}{n_{\text{mle}}}}.$$

The reason is that here we use $\mathcal{F}_{h-1}(\Upsilon_{h-2}), \mathcal{F}_h(\Upsilon_{h-1})$ instead of $\mathcal{F}_{h-1}, \mathcal{F}_h$. By Lemma 22, the two function classes considered here have ℓ_1 optimistic covers with scale $1/n_{\text{mle}}$ of size $|\Upsilon|\left(2\lceil B^\mu n_{\text{mle}} \rceil\right)^{\mathsf{d}}$. In addition, we still have that $d_{h-1}^D \in \mathcal{F}_{h-1}(\Upsilon_{h-2}), d_{h-1}^{D,\dagger} \in \mathcal{F}_h(\Upsilon_{h-1})$ Lemma 18, and any $d_{h-1} \in \mathcal{F}_{h-1}(\Upsilon_{h-2}), \mathcal{F}_h(\Upsilon_{h-1})$ is a valid probability distribution over $\mathcal{X}$.

The remaining part of the proof is the same as that of Theorem 5. □

Theorem 11 (Online policy optimization with representation learning)**.** *Fix $\delta \in (0,1)$ and suppose Assumption 1 and Assumption 3 hold. Given a policy class Π, let $\{\widehat{d}^\pi_h\}_{h\in[H],\pi\in\Pi}$ be the output of running Algorithm 4. Then with probability at least $1-\delta$, for any deterministic reward function R (as per Proposition 1) and policy selected as $\widehat{\pi}_R = \mathrm{argmax}_{\pi\in\Pi}\, \widehat{v}^\pi_R$, we have*

$$v_R^{\widehat{\pi}_R} \geq \max_{\pi\in\Pi} v_R^\pi - \varepsilon,$$

where $\widehat{v}^\pi_R := \sum_{h=0}^{H-1} \iint \widehat{d}^\pi_h(x_h)R(x_h,a_h)\pi(a_h|x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$. The total number of episodes required by the algorithm is

$$\tilde{O}\left(\frac{\mathsf{d}^5 K^2 H^7 \log(|\Pi||\Upsilon|/\delta)}{\varepsilon^2}\right).$$

Proof. The proof follows the same steps as that of Theorem 6. Notice that now we will apply Theorem 8 rather than Theorem 5 to get the bound $\|d^\pi_h - \widehat{d}^\pi_h\|$, which leads to the additional $\log(|\Pi|)$ factor. □

H. Maximum likelihood estimation

In this section, we adapt the standard i.i.d. results of maximum likelihood estimation (Van de Geer, 2000) to our setting, and in particular, to our (infinite) linear function class. We consider the problem of estimating a probability distribution over the instance space $\mathcal{X}$, and note that we abuse some notations (e.g., $n, \mathcal{L}, \mathcal{D}, \mathcal{F}$) in this section, as they have different meanings in other parts of the paper. Given an i.i.d. sampled dataset $\mathcal{D} = \{x^{(i)}\}_{i=1}^n$ and a function class $\mathcal{F}$, we optimize the MLE objective

$$\widehat{f} = \underset{f\in\mathcal{F}}{\mathrm{argmin}}\, \frac{1}{n}\sum_{i=1}^{n} \log\left(f(x^{(i)})\right). \tag{33}$$

We consider the function class $\mathcal{F}$ to be infinite, and as is common in statistical learning, our result will depends on its structural complexity. In particular, this will be quantified using the ℓ_1 optimistic cover, defined below:

Definition 3 (ℓ_1 optimistic cover)**.** *For a function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, we call function class $\overline{\mathcal{F}}$ an ℓ_∞ optimistic cover of $\mathcal{F}$ with scale γ, if for any $f \in \mathcal{F}$ there exists $\overline{f} \in \overline{\mathcal{F}}$, such that $\|f - \overline{f}\|_1 \leq \gamma$ and $f(x) \leq \overline{f}(x)$, $\forall x \in \mathcal{X}$. Notice that here we do not require the cover to be proper, i.e., we allow $\overline{\mathcal{F}} \not\subseteq \mathcal{F}$.*

Now we are ready to state the MLE guarantee formally.

Lemma 12 (MLE guarantee)**.** *Let $\mathcal{D} = \{x^{(i)}\}_{i=1}^n$ be a dataset, where $x^{(i)}$ are drawn i.i.d. from some fixed probability distribution f^* over $\mathcal{X}$. Consider a function class $\mathcal{F}$ that satisfies: (i) $f^* \in \mathcal{F}$, (ii) each function $f \in \mathcal{F}$ is a valid probability distribution over $\mathcal{X}$ (i.e., $f \in \Delta(\mathcal{X})$), and (iii) $\mathcal{F}$ has a finite ℓ_1 optimistic cover (Definition 3) $\overline{\mathcal{F}}$ with scale γ and $\overline{\mathcal{F}} \subseteq (\mathcal{X} \to \mathbb{R}_{\geq 0})$. Then with probability at least $1-\delta$, the MLE solution $\widehat{f}$ in Eq. (33) has an ℓ_1 error guarantee*

$$\|\widehat{f} - f^*\|_1 \leq \gamma + \sqrt{\frac{12\log(|\overline{\mathcal{F}}|/\delta)}{n} + 6\gamma}.$$

Proof. Our proof is based on Zhang (2006); Agarwal et al. (2020); Liu et al. (2022) and is simpler since we assume the $\mathcal{D}$ here is drawn i.i.d. instead of adaptively. We first define $\mathcal{L}(f,\mathcal{D}) = \frac{1}{2}\sum_{i=1}^n \log\left(\frac{f(x^{(i)})}{f^*(x^{(i)})}\right)$. By Chernoff's method, for a fixed $f \in \overline{\mathcal{F}}$ we have that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{L}(f,\mathcal{D}) - \log(\mathbb{E}_{\mathcal{D}}[\exp(\mathcal{L}(f,\mathcal{D}))]) \geq \log(|\overline{\mathcal{F}}|/\delta)\right) \\
\leq\ & \exp(-\log(|\overline{\mathcal{F}}|/\delta))\mathbb{E}_{\mathcal{D}}\left[\exp\left(\mathcal{L}(f,\mathcal{D}) - \log(\mathbb{E}_{\mathcal{D}}[\exp(\mathcal{L}(f,\mathcal{D}))])\right)\right] \\
=\ & \delta/|\overline{\mathcal{F}}|.
\end{aligned}$$

Union bounding over $f \in \overline{\mathcal{F}}$, with probability at least $1-\delta$, for any $f \in \overline{\mathcal{F}}$ we have

$$-\log(\mathbb{E}_{\mathcal{D}}[\exp(\mathcal{L}(f,\mathcal{D}))]) \leq -\mathcal{L}(f,\mathcal{D}) + \log(|\overline{\mathcal{F}}|/\delta). \tag{34}$$

Let $\overline{f} \in \overline{\mathcal{F}}$ be the γ-close ℓ_1 optimistic approximator of the MLE solution $\widehat{f} \in \mathcal{F}$. Since $\overline{f}(x) \geq \widehat{f}(x)$, $\forall x \in \mathcal{X}$ due to the optimistic covering construction and $\widehat{f}$ is the MLE estimator, for the RHS of Eq. (34). we have

$$-\mathcal{L}(\overline{f}, \mathcal{D}) = \frac{1}{2}\sum_{i=1}^{n} \log\left(\frac{f^*(x^{(i)})}{\overline{f}(x^{(i)})}\right) \leq \frac{1}{2}\sum_{i=1}^{n} \log\left(\frac{f^*(x^{(i)})}{\widehat{f}(x^{(i)})}\right) = \frac{1}{2}\left(\sum_{i=1}^{n} \log(f^*(x^{(i)})) - \sum_{i=1}^{n} \log(\widehat{f}(x^{(i)}))\right) \leq 0.$$

Next, consider the LHS of Eq. (34). From the definition of dataset $\mathcal{D}$ and $\mathcal{L}(\overline{f}, \mathcal{D})$, we get

$$\begin{aligned} -\log(\mathbb{E}_{\mathcal{D}}[\exp(\mathcal{L}(\overline{f}, \mathcal{D}))]) &= -\log\left(\mathbb{E}_{\mathcal{D}}\left[\exp\left(\frac{1}{2}\sum_{i=1}^{n} \log\left(\frac{\overline{f}(x^{(i)})}{f^*(x^{(i)})}\right)\right)\right]\right) \\ = -n\log\left(\mathbb{E}_{\mathcal{D}}\left[\exp\left(\frac{1}{2}\log\left(\frac{\overline{f}(x)}{f^*(x)}\right)\right)\right]\right) &= -n\log\left(\mathbb{E}_{\mathcal{D}}\left[\sqrt{\frac{\overline{f}(x)}{f^*(x)}}\right]\right). \end{aligned}$$

Furthermore, by $-\log(y) \geq 1 - y$, ℓ_1 optimistic cover definition, and f^*, $\widehat{f}$ are valid distributions over $x \in \mathcal{X}$, we have

$$\begin{aligned} &-n\log\left(\mathbb{E}_{\mathcal{D}}\left[\sqrt{\frac{\overline{f}(x)}{f^*(x)}}\right]\right) \geq n\left(1 - \mathbb{E}_{\mathcal{D}}\left[\sqrt{\frac{\overline{f}(x)}{f^*(x)}}\right]\right) = n\left(1 - \int \sqrt{\overline{f}(x)f^*(x)}(\mathrm{d}x)\right) \\ &= \frac{n}{2}\int\left(\sqrt{f^*(x)} - \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x) + \frac{n}{2}\left(1 - \int \overline{f}(x)(\mathrm{d}x)\right) \\ &= \frac{n}{2}\int\left(\sqrt{f^*(x)} - \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x) + \frac{n}{2}\int\left(\widehat{f}(x) - \overline{f}(x)\right)(\mathrm{d}x) \\ &\geq \frac{n}{2}\int\left(\sqrt{f^*(x)} - \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x) - \frac{n\gamma}{2}. \end{aligned}$$

Then notice that $\int\left(\sqrt{f^*(x)} + \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x) \leq 2\int\left(f^*(x) + \overline{f}(x)\right)(\mathrm{d}x) \leq 2\int(f^*(x) + \widehat{f}(x) + |\overline{f}(x) - \widehat{f}(x)|)(\mathrm{d}x) \leq 6$ and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} &\frac{n}{2}\int\left(\sqrt{f^*(x)} - \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x) - \frac{n\gamma}{2} \\ &\geq \frac{n}{12}\left(\int\left(\sqrt{f^*(x)} - \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x)\right)\left(\int\left(\sqrt{f^*(x)} + \sqrt{\overline{f}(x)}\right)^2(\mathrm{d}x)\right) - \frac{n\gamma}{2} \\ &\geq \frac{n}{12}\left(\int|\overline{f}(x) - f^*(x)|(\mathrm{d}x)\right)^2 - \frac{n\gamma}{2} = \frac{n}{12}\|\overline{f} - f^*\|_1^2 - \frac{n\gamma}{2}. \end{aligned}$$

Combining the above inequalities and rearranging yields

$$\|\overline{f} - f^*\|_1^2 \leq \frac{12\log(|\overline{\mathcal{F}}|/\delta)}{n} + 6\gamma.$$

Finally, by the triangle inequality and the definition of the ℓ_1 optimistic cover, we get

$$\|\widehat{f} - f^*\|_1 \leq \|\widehat{f} - \overline{f}\|_1 + \|\overline{f} - f^*\|_1 \leq \gamma + \sqrt{\frac{12\log(|\overline{\mathcal{F}}|/\delta)}{n} + 6\gamma}\,,$$

which completes the proof. □

I. Auxiliary lemmas

In this section, we provide detailed proofs for auxiliary lemmas.

I.1. Squared loss regression results

Lemma 13 (Squared loss decomposition)**.** *For any $w_h, w_{h+1} : \mathcal{X} \to \mathbb{R}$, dataset $\mathcal{D}_h^{\mathrm{reg}} = \{(x_h, a_h, x_{h+1})\} \sim d_h^D$, and a pseudo-policy π, we have*

$$\left\| w_{h+1} - \frac{\mathbf{P}_h^\pi \left(d_h^D w_h\right)}{d_h^{D,\dagger}} \right\|_{2,d_h^{D,\dagger}}^2 = \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\mathrm{reg}}}(w_{h+1}, w_h, \pi)\right] - \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\mathrm{reg}}}\left(\frac{\mathbf{P}_h^\pi \left(d_h^D w_h\right)}{d_h^{D,\dagger}}, w_h, \pi\right)\right]. \tag{35}$$

Proof. We introduce a new notation

$$\begin{aligned}(\mathbf{E}_h^\pi w_h)(x_{h+1}) &:= \frac{\left(\mathbf{P}_h^\pi \left(d_h^D w_h\right)\right)(x_{h+1})}{d_h^{D,\dagger}(x_{h+1})} \\ &= \frac{\iint P_h(x_{h+1}|x_h, a_h)\pi(a_h|x_h) d_h^D(x_h) w_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)}{d_h^{D,\dagger}(x_{h+1})},\end{aligned} \tag{36}$$

which represents the conditional expectation. Then we have the decomposition

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\mathrm{reg}}}(w_{h+1}, w_h, \pi)\right] \\
=& \iiint d_h^D(x_h, a_h, x_{h+1}) \left(w_{h+1}(x_{h+1}) - \frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)} w_h(x_h)\right)^2 (\mathrm{d}x_h)(\mathrm{d}a_h)(\mathrm{d}x_{h+1}) \\
=& \iiint d_h^D(x_h, a_h, x_{h+1}) \left(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}) + (\mathbf{E}_h^\pi w_h)(x_{h+1}) - \frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)} w_h(x_h)\right)^2 \\
&\qquad (\mathrm{d}x_h)(\mathrm{d}a_h)(\mathrm{d}x_{h+1}) \\
=& \int d_h^{D,\dagger}(x_{h+1})(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}))^2 (\mathrm{d}x_{h+1}) \\
&+ \iiint d_h^D(x_h, a_h, x_{h+1}) \left((\mathbf{E}_h^\pi w_h)(x_{h+1}) - \frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)} w_h(x_h)\right)^2 (\mathrm{d}x_h)(\mathrm{d}a_h)(\mathrm{d}x_{h+1}) \\
&+ 2 \iiint d_h^D(x_h, a_h, x_{h+1})(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1})) \left((\mathbf{E}_h^\pi w_h)(x_{h+1}) - \frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)} w_h(x_h)\right) \\
&\qquad (\mathrm{d}x_h)(\mathrm{d}a_h)(\mathrm{d}x_{h+1}) \\
=& \|w_{h+1} - (\mathbf{E}_h^\pi w_h)\|_{2,d_h^{D,\dagger}}^2 + \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\mathrm{reg}}}(\mathbf{E}_h^\pi w_h, w_h, \pi)\right] \\
&+ 2 \int d_h^{D,\dagger}(x_{h+1})(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}))(\mathbf{E}_h^\pi w_h)(x_{h+1})(\mathrm{d}x_{h+1}) \\
&- 2 \int d_h^{D,\dagger}(x_{h+1})(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1})) \\
&\qquad \cdot \left(\iint d_h^D(x_h, a_h | x_{h+1}) \frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)} w_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)\right)(\mathrm{d}x_{h+1}) \\
=& \|w_{h+1} - (\mathbf{E}_h^\pi w_h)\|_{2,d_h^{D,\dagger}}^2 + \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\mathrm{reg}}}(\mathbf{E}_h^\pi w_h, w_h, \pi)\right] \\
&+ 2 \int d_h^{D,\dagger}(x_{h+1})(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}))((\mathbf{E}_h^\pi w_h)(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}))(\mathrm{d}x_{h+1}) \\
=& \|w_{h+1} - (\mathbf{E}_h^\pi w_h)\|_{2,d_h^{D,\dagger}}^2 + \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\mathrm{reg}}}(\mathbf{E}_h^\pi w_h, w_h, \pi)\right].
\end{aligned}$$

□

Lemma 14 (Deviation bound for regression with squared loss)**.** *For $h \in [H]$, consider a dataset $\mathcal{D}_{0:h}$ that satisfies Assumption 2 and a function $w_h : \mathcal{X} \to [0, C_h^{\mathbf{x}}]$ that only depends on $\mathcal{D}_{0:h-1} \bigcup \mathcal{D}_h^{\mathrm{mle}}$. Consider a finite feature class Υ_h and a* *finite policy class* Π' *such that any* $\pi \in \Pi'$ *is a pseudo-policy (Definition 1) satisfying* $\pi_h(a_h|x_h) \leq C_h^{\mathbf{a}} \pi_h^D(a_h|x_h), \forall x_h \in \mathcal{X}, a_h \in \mathcal{A}$. *Then with probability* $1-\delta$*, for any* $w_{h+1} \in \mathcal{W}_{h+1}(\Upsilon_h)$ *and* $\pi \in \Pi'$*, we have*

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\text{reg}}}\left(w_{h+1}, w_h, \pi\right) - \mathcal{L}_{\mathcal{D}_h^{\text{reg}}}(\mathbf{E}_h^\pi w_h, w_h, \pi)\right] - \left(\mathcal{L}_{\mathcal{D}_h^{\text{reg}}}\left(w_{h+1}, w_h, \pi\right) - \mathcal{L}_{\mathcal{D}_h^{\text{reg}}}(\mathbf{E}_h^\pi w_h, w_h, \pi)\right)\right| \\
\leq{}& \frac{1}{2}\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_h^{\text{reg}}}\left(w_{h+1}, w_h, \pi\right) - \mathcal{L}_{\mathcal{D}_h^{\text{reg}}}(\mathbf{E}_h^\pi w_h, w_h, \pi)\right] + \frac{221184\mathsf{d}(C_h^{\mathbf{x}} C_h^{\mathbf{a}})^2 \log\left(n_{\text{reg}}|\Pi'||\Upsilon_h|/\delta\right)}{n_{\text{reg}}}
\end{aligned}$$

where the function class $\mathcal{W}_{h+1}(\Upsilon_h)$ *is defined in Algorithm 1 as in Eq. (28) and the operator* $\mathbf{E}_h^\pi$ *is defined in Eq. (36).*

Proof. We first fix the datasets $\mathcal{D}_{0:h-1} \bigcup \mathcal{D}_h^{\text{mle}}$ and prove the desired bound when conditioned on these datasets, in which case $w_h, d_h^{D,\dagger}, \pi^D$ are fixed. In the following, the expectation $\mathbb{E}$ and variance $\mathbb{V}$ are w.r.t. $(x_h, a_h, x_{h+1}) \sim d_h^D$, i.e., the data distribution from which the samples in $\mathcal{D}_h^{\text{reg}}$ are drawn i.i.d. from (Assumption 2), when conditioned on $\mathcal{D}_{0:h-1} \bigcup \mathcal{D}_h^{\text{mle}}$.

Consider a single $\pi \in \Pi'$ and feature $\mu_h \in \Upsilon_h$, and consider the hypothesis class

$$\mathcal{Y}(\mathcal{W}_{h+1}(\mu_h), w_h, \pi) = \left\{Y(w_{h+1}, w_h, \pi) : w_{h+1} \in \mathcal{W}_{h+1}(\mu_h)\right\}.$$

where the random variable $Y(w_{h+1}, w_h, \pi)$ (suppressing the dependence on the (x_h, a_h, x_{h+1}) tuple) is defined for convenience as

$$Y(w_{h+1}, w_h, \pi) := \left(w_{h+1}(x_{h+1}) - w_h(x_h)\frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)}\right)^2 - \left((\mathbf{E}_h^\pi w_h)(x_{h+1}) - w_h(x_h)\frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)}\right)^2,$$

and we use $Y_i(w_{h+1}, w_h, \pi)$ to denote its realization on the i-th tuple data $(x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)}) \in \mathcal{D}_h^{\text{reg}}$. The function class $\mathcal{W}_{h+1}(\mu_h)$ is defined as in Eq. (29), i.e.,

$$\mathcal{W}_{h+1}(\mu_h) = \left\{w_{h+1} = \frac{\langle \mu_h, \theta_{h+1}^{\text{up}}\rangle}{\langle \mu_h, \theta_{h+1}^{\text{down}}\rangle} : \|w_{h+1}\|_\infty \leq C_h^{\mathbf{x}} C_h^{\mathbf{a}}, \theta_{h+1}^{\text{up}}, \theta_{h+1}^{\text{down}} \in \mathbb{R}^{\mathsf{d}}\right\}.$$

It can be seen that $|Y(w_{h+1}, w_h, \pi)| \leq 4(C_h^{\mathbf{x}} C_h^{\mathbf{a}})^2$ from the following. From their respective definitions, we know $\|w_h\|_\infty \leq C_h^{\mathbf{x}}$, $\|\frac{\pi}{\pi^D}\|_\infty \leq C_h^{\mathbf{a}}$, and $\|w_{h+1}\|_\infty \leq C_h^{\mathbf{x}} C_h^{\mathbf{a}}$. We also have $(\mathbf{E}_h^\pi w_h)(x_{h+1}) = \frac{\left(\mathbf{P}_h^\pi\left(d_h^D w_h\right)\right)(x_{h+1})}{d_h^{D,\dagger}(x_{h+1})} \in [0, C_h^{\mathbf{x}} C_h^{\mathbf{a}}]$ from Lemma 19.

Further, for any $Y(w_{h+1}, w_h, \pi) \in \mathcal{Y}(\mathcal{W}_{h+1}(\mu_h), w_h, \pi)$, we can bound the variance $\mathbb{V}[Y(w_{h+1}, w_h, \pi)]$ as

$$\begin{aligned}
&\mathbb{V}[Y(w_{h+1}, w_h, \pi)] \leq \mathbb{E}\left[Y(w_{h+1}, w_h, \pi)^2\right] \\
={}& \mathbb{E}\left[\left(\left(w_{h+1}(x_{h+1}) - w_h(x_h)\frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)}\right)^2 - \left((\mathbf{E}_h^\pi w_h)(x_{h+1}) - w_h(x_h)\frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)}\right)^2\right)^2\right] \\
={}& \mathbb{E}\left[(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}))^2\left(w_{h+1}(x_{h+1}) - 2w_h(x_h)\frac{\pi(a_h|x_h)}{\pi^D(a_h|x_h)} + (\mathbf{E}_h^\pi w_h)(x_{h+1})\right)^2\right] \\
\leq{}& 4(C_h^{\mathbf{x}} C_h^{\mathbf{a}})^2 \mathbb{E}\left[(w_{h+1}(x_{h+1}) - (\mathbf{E}_h^\pi w_h)(x_{h+1}))^2\right] \\
={}& 4(C_h^{\mathbf{x}} C_h^{\mathbf{a}})^2 \mathbb{E}\left[Y(w_{h+1}, w_h, \pi)\right]. && \text{(Lemma 13)}
\end{aligned}$$

Next, we show that the uniform covering number $\mathcal{N}_1(\gamma, \mathcal{Y}(\mathcal{W}_{h+1}(\mu_h), w_h, \pi), m)$ (see Definition 7) for any $\gamma \in \mathbb{R}, m \in \mathbb{N}$ can be bounded by the covering number of $\mathcal{W}_{h+1}(\mu_h)$. Let $Z^m = (x_h^{(i)}, a_h^{(i)}, x_{h+1}^{(i)})_{i=1}^m$ denote m i.i.d. samples from d_h^D, and denote $X^m = (x_{h+1}^{(i)})_{i=1}^m$ the corresponding x_{h+1} samples. For any Z^m and $Y(w_{h+1}, w_h, \pi), Y(w'_{h+1}, w_h, \pi) \in \mathcal{Y}(\mathcal{W}_{h+1}(\mu_h), w_h, \pi)$,

$$\frac{1}{m}\sum_{i=1}^m \left|Y_i(w_{h+1}, w_h, \pi) - Y_i(w'_{h+1}, w_h, \pi)\right|$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i=1}^{m}\left|\left(w_{h+1}(x_{h+1}^{(i)}) - w_h(x_h^{(i)})\frac{\pi(a_h^{(i)}|x_h^{(i)})}{\pi^D(a_h^{(i)}|x_h^{(i)})}\right)^2 - \left(w'_{h+1}(x_{h+1}^{(i)}) - w_h(x_h^{(i)})\frac{\pi(a_h^{(i)}|x_h^{(i)})}{\pi^D(a_h^{(i)}|x_h^{(i)})}\right)^2\right| \\
&= \frac{1}{m}\sum_{i=1}^{m}\left|w_{h+1}(x_{h+1}^{(i)}) - 2w_h(x_h^{(i)})\frac{\pi(a_h^{(i)}|x_h^{(i)})}{\pi^D(a_h^{(i)}|x_h^{(i)})} + w'_{h+1}(x_{h+1}^{(i)})\right| \cdot \left|w_{h+1}(x_{h+1}^{(i)}) - w'_{h+1}(x_{h+1}^{(i)})\right| \\
&\le \frac{4C_h^{\mathbf{x}}C_h^{\mathbf{a}}}{m}\sum_{i=1}^{m}\left|w_{h+1}(x_{h+1}^{(i)}) - w'_{h+1}(x_{h+1}^{(i)})\right|.
\end{aligned}
$$

Thus any $\gamma/(4C_h^{\mathbf{x}}C_h^{\mathbf{a}})$-covering of $\mathcal{W}_{h+1}|_{X^m}$ in ℓ_1 is a γ-covering of $Y(\mathcal{W}_{h+1}, w_h, \pi)|_{Z^m}$ in ℓ_1, and

$$
\mathcal{N}_1(\gamma, Y(\mathcal{W}_{h+1}(\mu_h), w_h, \pi), Z^m) \le \mathcal{N}_1(\gamma/(4C_h^{\mathbf{x}}C_h^{\mathbf{a}}), \mathcal{W}_{h+1}(\mu_h), X^m)
$$

which implies the same relationship for the uniform covering numbers:

$$
\begin{aligned}
\mathcal{N}_1(\gamma, Y(\mathcal{W}_{h+1}(\mu_h), w_h, \pi), m) &= \max_{Z^m}\mathcal{N}_1(\gamma, Y(\mathcal{W}_{h+1}(\mu_h), w_h, \pi), Z^m) \\
&\le \max_{X^m}\mathcal{N}_1(\gamma/(4C_h^{\mathbf{x}}C_h^{\mathbf{a}}), \mathcal{W}_{h+1}(\mu_h), X^m) = \mathcal{N}_1(\gamma/(4C_h^{\mathbf{x}}C_h^{\mathbf{a}}), \mathcal{W}_{h+1}(\mu_h), m).
\end{aligned}
$$

Then using this inequality and $b = 4(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2$ in Lemma 26 and conditioning on $\mathcal{D}_{0:h-1}\bigcup\mathcal{D}_h^{\text{mle}}$, for any $w_{h+1} \in \mathcal{W}_{h+1}(\mu_h)$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|\mathbb{E}[Y(w_{h+1}, w_h, \pi)] - \frac{1}{n_{\text{reg}}}\sum_{i=1}^{n} Y_i(w_{h+1}, w_h, \pi)\right| \ge \varepsilon\right) \\
&\le 36\mathcal{N}_1\left(\frac{\varepsilon^3}{10240(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}, \mathcal{Y}(\mathcal{W}_{h+1}(\mu_h), w_h, \pi), \frac{640 n_{\text{reg}}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}{\varepsilon^2}\right) \\
&\qquad \cdot \exp\left(-\frac{n_{\text{reg}}\varepsilon^2}{128\mathbb{V}[Y(w_{h+1}, w_h, \pi)] + 2048\varepsilon(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2}\right) \\
&\le 36\mathcal{N}_1\left(\frac{\varepsilon^3}{40960(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^5}, \mathcal{W}_{h+1}(\mu_h), \frac{640 n_{\text{reg}}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}{\varepsilon^2}\right) \\
&\qquad \cdot \exp\left(-\frac{n_{\text{reg}}\varepsilon^2}{512(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\mathbb{E}[Y(w_{h+1}, w_h, \pi)] + 2048\varepsilon(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2}\right).
\end{aligned}
$$

Then setting the RHS equal to δ', we have

$$
n_{\text{reg}} = \frac{512(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\left(\mathbb{E}[Y(w_{h+1}, w_h, \pi)] + 4\varepsilon\right)\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{40960(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^5}, \mathcal{W}_{h+1}(\mu_h), \frac{640 n_{\text{reg}}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}{\varepsilon^2}\right)/\delta'\right)}{\varepsilon^2}
$$

implying

$$
\begin{aligned}
\varepsilon \le &\sqrt{\frac{512(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\mathbb{E}[Y(w_{h+1}, w_h, \pi)]\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{40960(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^5}, \mathcal{W}_{h+1}(\mu_h), \frac{640 n_{\text{reg}}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}{\varepsilon^2}\right)/\delta'\right)}{n_{\text{reg}}}} \\
&+ \frac{2048(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{40960(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^5}, \mathcal{W}_{h+1}(\mu_h), \frac{640 n_{\text{reg}}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}{\varepsilon^2}\right)/\delta'\right)}{n_{\text{reg}}}.
\end{aligned}
$$

From Lemma 23 and Lemma 25, and noting that $n_{\text{reg}} \ge \frac{2048(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2}{\varepsilon}$, we have that

$$
\log\left(36\mathcal{N}_1\left(\frac{\varepsilon^3}{40960(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^5}, \mathcal{W}_{h+1}(\mu_h), \frac{640 n_{\text{reg}}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^4}{\varepsilon^2}\right)/\delta'\right)
$$

$$
\begin{aligned}
&\le 4(\mathsf{d}+1)\log(8e)\log\left(\frac{655360e^2(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^6}{\varepsilon^3\delta'}\right)\\
&\le 96\mathsf{d}\log\left(\frac{n_{\mathrm{reg}}}{\delta'}\right).
\end{aligned}
$$

Thus with probability at least $1-\delta'$,

$$
\begin{aligned}
&\left|\mathbb{E}[Y(w_{h+1},w_h,\pi)] - \frac{1}{n_{\mathrm{reg}}}\sum_{i=1}^{n_{\mathrm{reg}}} Y_i(w_{h+1},w_h,\pi)\right|\\
&\le \sqrt{\frac{49152\mathsf{d}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\mathbb{E}[Y(w_{h+1},w_h,\pi)]\log\left(\frac{n_{\mathrm{reg}}}{\delta'}\right)}{n_{\mathrm{reg}}}} + \frac{196608\mathsf{d}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\log\left(\frac{n_{\mathrm{reg}}}{\delta'}\right)}{n_{\mathrm{reg}}}.
\end{aligned}
$$

Then invoking the AM-GM inequality,

$$
\begin{aligned}
&\left|\mathbb{E}[Y(w_{h+1},w_h,\pi)] - \frac{1}{n_{\mathrm{reg}}}\sum_{i=1}^{n_{\mathrm{reg}}} Y_i(w_{h+1},w_h,\pi)\right|\\
&\le \frac{1}{2}\mathbb{E}[Y(w_{h+1},w_h,\pi)] + \frac{221184\cdot\mathsf{d}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\log\left(\frac{n_{\mathrm{reg}}}{\delta'}\right)}{n_{\mathrm{reg}}}.
\end{aligned}
$$

Recall that this result holds for a fixed π and $\mathcal{W}_{h+1}(\mu_h)$ defined using a fixed μ_h. Then setting $\delta' = \frac{\delta}{|\Pi'||\Upsilon_h|}$ and taking a union bound over Π and Υ_h, we have that with probability at least $1-\delta$ that for any $\pi\in\Pi'$ and $w_{h+1}\in\mathcal{W}_{h+1}(\Upsilon_h)$ that

$$
\begin{aligned}
&\left|\mathbb{E}[Y(w_{h+1},w_h,\pi)] - \frac{1}{n_{\mathrm{reg}}}\sum_{i=1}^{n_{\mathrm{reg}}} Y_i(w_{h+1},w_h,\pi)\right|\\
&\le \frac{1}{2}\mathbb{E}[Y(w_{h+1},w_h,\pi)] + \frac{221184\mathsf{d}(C_h^{\mathbf{x}}C_h^{\mathbf{a}})^2\log\left(\frac{n_{\mathrm{reg}}|\Pi'||\Upsilon_h|}{\delta}\right)}{n_{\mathrm{reg}}}.
\end{aligned}
$$

Finally, since this result holds for any fixed $\mathcal{D}_{0:h-1}\bigcup\mathcal{D}_h^{\mathrm{mle}}$, by the law of total expectation, it also holds with probability at least $1-\delta'$ without conditioning on $\mathcal{D}_{0:h-1}\bigcup\mathcal{D}_h^{\mathrm{mle}}$. Using Lemma 13 with the definitions of $Y(w_{h+1},w_h,\pi)$ and $Y_i(w_{h+1},w_h,\pi)$ completes the proof. □

I.2. Barycentric spanner

In this section we first define the barycentric spanner (Awerbuch and Kleinberg, 2008, Definition 2.1), then prove that a spanner of size d always exists for a set of functions linear in a feature μ_{h-1}, from which Proposition 3 follows straightforwardly. The proof is adapted from Awerbuch and Kleinberg (2008, Proposition 2.2), which only applies to square matrices, and we extend it to rectangular matrices for completeness. We close with a discussion of the computational complexity of finding the barycentric spanner.

Definition 4 (Barycentric spanner)**.** *Let V be a vector space over the real numbers, and $S\subseteq V$ a subset whose linear span is a m-dimensional subspace of V. A set $X=\{x_1,\ldots,x_m\}\subseteq S$ is a barycentric spanner of S if every $x\in S$ may be expressed as a linear combination of elements of X using coefficients in $[-1,+1]$.*

Lemma 15 (Barycentric spanner for linear functions)**.** *For a feature $\mu_{h-1}\in\Upsilon_{h-1}$ with rank d, any compact set of linear functions $\mathcal{U}\subseteq\{\langle\mu_{h-1},\theta_h\rangle:\theta_h\in\mathbb{R}^{\mathsf{d}}\}$ has a barycentric spanner of cardinality at most d.*

Proof. We prove the proposition when $\mathrm{rank}(\mu_{h-1})=\mathsf{d}$ is full rank (the argument should be the same when $\mathrm{rank}(\mu_{h-1})<d$). Because $\mathcal{U}$ is linear in μ_{h-1}, its linear span is a d-dimensional subspace of $\mathbb{R}^{|\mathcal{X}|}$, and any $u\in\mathcal{U}$ can be written as the linear combination of a subspace basis.

We claim the barycentric spanner is any subset $B=\{b_1,\ldots,b_{\mathsf{d}}\}\subseteq\mathcal{U}$ with $B\in\mathbb{R}^{\mathsf{d}\times|\mathcal{X}|}$ that maximizes the volume $|\det(BB^\top)|$. By compactness, the maximum is obtained by at least one subset of $\mathcal{U}$. Since $\det(BB^\top)=(\prod_{i=1}^{\mathsf{d}}\sigma_i(B))^2$, the maximizing B will have d singular values and full row rank (otherwise the determinant will be 0). As a result, any $u \in \mathcal{U}$ will be a linear combination of the rows of B, i.e., there exists $\{c_i\}_{i=1}^{\mathsf{d}}$ such that $u = \sum_{i=1}^{\mathsf{d}} c_i b_i$. We will prove that $|c_i| \le 1$ by contradiction.

W.l.o.g, suppose there exists u with coefficient $|c_1| > 1$. Then consider a new matrix $\widetilde{B} = \{u, b_2, \dots, b_{\mathsf{d}}\}$, which can be expressed as $\widetilde{B} = CB$, where $C \in \mathbb{R}^{\mathsf{d}\times\mathsf{d}}$ is the coefficient matrix. Then $\widetilde{B}$ has determinant

$$|\det(\widetilde{B}\widetilde{B}^\top)| = |\det(C)|^2 |\det(BB^\top)| = |c_1|^2 |\det(BB^\top)| \ge \det(BB^\top).$$

Then we have a contradiction because B was volume-maximizing, and $|c_i| \le 1$. □

Computation of barycentric spanner Lastly, we discuss computation of the barycentric spanner. In the main results of the paper we assume that we can perfectly compute the barycentric spanner in an efficient manner. When this is not the case, the algorithm in Figure 2 in Awerbuch and Kleinberg (2008) (with similar adaptations to handle rectangular matrices as in the proof of Lemma 15) can be used to compute a C-approximate barycentric spanner, where $C > 1$, with $O(\mathsf{d}^2 \log_C \mathsf{d})$ calls to a linear optimization oracle (Awerbuch and Kleinberg, 2008, Proposition 2.5). A C-approximate barycentric spanner is defined similarly as Definition 4, except that the coefficients are in the range $[-C, +C]$. This will only change our main results by increasing them by a factor of C, and we may simply set $C = 2$ with minimal effects on our sample complexity guarantees.

I.3. Properties of low-rank MDPs

Lemma 16. *In the low-rank MDP (Assumption 1), for any $h \in [H]$, function $d_{h-1} : \mathcal{X} \to \mathbb{R}$, and pseudo-policy $\overline{\pi}$ (Definition 1), we have*

$$(\mathbf{P}_h^{\overline{\pi}} d_h)(x_{h+1}) = \iint P_h(x_{h+1}|x_h, a_h)\overline{\pi}_h(a_h|x_h) d_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h) = \langle \mu_h^*(x_{h+1}), \theta_{h+1}\rangle$$

for some $\theta_{h+1} \in \mathbb{R}^{\mathsf{d}}$ with $\|\theta_{h+1}\|_\infty \le \|d_h\|_1$.

Proof. By the definition of low-rank MDPs (Assumption 1), we have

$$\begin{aligned}
\mathbf{P}_h^{\overline{\pi}} d_h &= \iint P_h(x_{h+1}|x_h, a_h)\overline{\pi}_h(a_h|x_h) d_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h) \\
&= \iint \langle \mu_h^*(x_{h+1}), \phi_h^*(x_h, a_h)\rangle \overline{\pi}_h(a_h|x_h) d_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h) \\
&= \langle \mu_h^*(x_{h+1}), \theta_{h+1}\rangle,
\end{aligned}$$

where $\theta_{h+1} = \iint \phi_h^*(x_h, a_h)\overline{\pi}_h(a_h|x_h) d_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h) \in \mathbb{R}^{\mathsf{d}}$. In addition,

$$\begin{aligned}
\|\theta_{h+1}\|_\infty &\le \iint \|\phi_h^*(x_h, a_h)\|_\infty \overline{\pi}_h(a_h|x_h) |d_h(x_h)|(\mathrm{d}x_h)(\mathrm{d}a_h) \\
&\le \int \left(\int \overline{\pi}_h(a_h|x_h)(\mathrm{d}a_h) \right) |d_h(x_h)|(\mathrm{d}x_h) \\
&\le \int |d_h(x_h)|(\mathrm{d}x_h) = \|d_h\|_1
\end{aligned}$$

where we use Lemma 21 in the last inequality. □

Lemma 17. *In low-rank MDPs (Assumption 1), given a dataset $\mathcal{D}_h$ satisfying Assumption 2 for $h \in [H]$, let d_h^D and $d_h^{D,\dagger}$ be the corresponding current-state and next-state data distributions. Then for the function class*

$$\mathcal{F}_h = \left\{ d_h = \langle \mu_{h-1}^*, \theta_h \rangle : d_h \in \Delta(\mathcal{X}), \theta_h \in \mathbb{R}^{\mathsf{d}}, \|\theta_h\|_\infty \le 1 \right\},$$

we have that $d_h^D \in \mathcal{F}_h$ and $d_h^{D,\dagger} \in \mathcal{F}_{h+1}$.

Proof. Recall that under Assumption 2, $\mathcal{D}_h$ is collected by $\rho^{h-1} \circ \pi_h^D$ where $a_{0:h-1} \sim \rho^{h-1}$, an $(h-1)$-step non-Markov policy, and $a_h \sim \pi_h^D$, a Markov policy.

First we prove the lemma statement for $d_h^{D,\dagger}$. Since d_h^D is a valid distribution and π_h^D is a valid Markov policy, from Lemma 16 we know that $d_h^{D,\dagger} = \mathbf{P}_h^{\pi_h^D}(d_h^D)$ can be written as $\langle \mu_h^*, \theta_{h+1} \rangle$ with $\|\theta_{h+1}\|_\infty \le 1$. Finally, since $d_h^{D,\dagger}$ is a valid marginal distribution, $d_h^{D,\dagger} \in \Delta(\mathcal{X})$, thus satisfying all constraints of $\mathcal{F}_{h+1}$.

To prove the lemma statement for d_h^D, we first prove a variant of Lemma 16 for non-Markov policies. With some overload of notation, let $d_{h-1}^D(x_{h-1})$ denote the marginal distribution of x_{h-1} induced by rolling the non-Markov policy ρ^{h-1} to level $h-1$. Then

$$d_h^D(x_h) = \iint P_h(x_h|x_{h-1}, a_{h-1}) \rho^{h-1}(a_{h-1}|x_{0:h-1}) d_{h-1}^D(x_{h-1}) (\mathrm{d}x_{h-1})(\mathrm{d}a_{h-1}).$$

Using similar steps as the proof of Lemma 16, we have that

$$\begin{aligned} d_h^D(x_h) &= \iint P_h(x_h|x_{h-1}, a_{h-1}) \rho^{h-1}(a_{h-1}|x_{0:h-1}) d_{h-1}^D(x_{h-1}) (\mathrm{d}x_{h-1})(\mathrm{d}a_{h-1}) \\ &= \iint \langle \phi_{h-1}^*(x_{h-1}, a_{h-1}), \mu_{h-1}^*(x_h) \rangle \rho^{h-1}(a_{h-1}|x_{0:h-1}) d_{h-1}^D(x_{h-1}) (\mathrm{d}x_{h-1})(\mathrm{d}a_{h-1}) \\ &= \langle \mu_{h-1}^*(x_h), \theta_h \rangle, \end{aligned}$$

where $\theta_h = \iint \phi_{h-1}^*(x_{h-1}, a_{h-1}) \rho^{h-1}(a_{h-1}|x_{0:h-1}) d_{h-1}^D(x_{h-1}) (\mathrm{d}x_{h-1})(\mathrm{d}a_{h-1}) \in \mathbb{R}^{\mathsf{d}}$. Since d_{h-1}^D and $\rho^{h-1}(\cdot|x_{0:h-1})$ are valid probability distributions over states x_h and actions a_h, respectively, it is easy to see that

$$\|\theta_h\|_\infty \le \iint \|\phi_{h-1}^*(x_{h-1}, a_{h-1})\|_\infty \rho^{h-1}(a_{h-1}|x_{0:h-1}) d_{h-1}^D(x_{h-1}) (\mathrm{d}x_{h-1})(\mathrm{d}a_{h-1}) \le 1$$

since $\|\phi_{h-1}^*(\cdot)\|_\infty \le 1$ from Assumption 1. Finally, since d_h^D is a valid distribution, we have $d_h^D \in \mathcal{F}_h$. □

Lemma 18. *In low-rank MDPs (Assumption 1), given a dataset $\mathcal{D}_h$ satisfying Assumption 2 for $h \in [H]$, let d_h^D and $d_h^{D,\dagger}$ be the corresponding current-state and next-state data distributions. Then for the function class*

$$\mathcal{F}_h(\Upsilon_{h-1}) = \left\{ d_h = \langle \mu_{h-1}, \theta_h \rangle : d_h \in \Delta(\mathcal{X}), \mu_{h-1} \in \Upsilon_{h-1}, \theta_h \in \mathbb{R}^{\mathsf{d}}, \|\theta_h\|_\infty \le 1 \right\},$$

we have that $d_h^D \in \mathcal{F}_h(\Upsilon_{h-1})$ and $d_h^{D,\dagger} \in \mathcal{F}_{h+1}(\Upsilon_h)$.

Proof. From Lemma 17 we know that $d_h^D \in \mathcal{F}_h$ (where $\mathcal{F}_h$ is linear in the true features μ_{h-1}^*, as defined in the Lemma 17), and $d_h^{D,\dagger} \in \mathcal{F}_{h+1}$. Noting that $\mathcal{F}_h \subseteq \mathcal{F}_h(\Upsilon_{h-1})$ and $\mathcal{F}_{h+1} \subseteq \mathcal{F}_{h+1}(\Upsilon_h)$ completes the proof. □

Lemma 19. *For $h \in [H]$, suppose we have a dataset $\mathcal{D}_h$ satisfying Assumption 2, with corresponding data distributions d_h^D and $d_h^{D,\dagger}$. Given a function $w_h : \mathcal{X} \to [-C_h^{\mathbf{x}}, C_h^{\mathbf{x}}]$ and pseudo-policy $\overline{\pi}$ (Definition 1) with $\frac{\overline{\pi}_h(a|x)}{\pi_h^D(a|x)} \le C_h^{\mathbf{a}}, \forall x \in \mathcal{X}, a \in \mathcal{A}$, we have*

$$\left\| \frac{\mathbf{P}_h^{\overline{\pi}}(d_h^D w_h)}{d_h^{D,\dagger}} \right\|_\infty \le C_h^{\mathbf{x}} C_h^{\mathbf{a}}.$$

Proof. For any $x_{h+1} \in \mathcal{X}$, we have

$$\begin{aligned} \left(\mathbf{P}_h^{\overline{\pi}}\left(d_h^D w_h\right)\right)(x_{h+1}) &\le C_h^{\mathbf{x}} \left(\mathbf{P}_h^{\overline{\pi}} d_h^D\right)(x_{h+1}) \\ &= C_h^{\mathbf{x}} \iint P_h(x_{h+1}|x_h, a_h) \overline{\pi}_h(a_h|x_h) d_h^D(x_h) (\mathrm{d}x_h)(\mathrm{d}a_h) \\ &\le C_h^{\mathbf{x}} C_h^{\mathbf{a}} \iint P_h(x_{h+1}|x_h, a_h) \pi_h^D(a_h|x_h) d_h^D(x_h) (\mathrm{d}x_h)(\mathrm{d}a_h) \\ &= C_h^{\mathbf{x}} C_h^{\mathbf{a}} d_h^{D,\dagger}(x_{h+1}). \end{aligned}$$

The last equality follows from the Bellman flow equation and Assumption 2. The convention that $\frac{0}{0} = 0$ gives the lemma statement. □

Lemma 20. *For any two state distributions d_h, d'_h and a pseudo-policy π (Definition 1), we have the following inequality*

$$\|\mathbf{P}^\pi_h d_h - \mathbf{P}^\pi_h d'_h\|_1 \le \|d_h - d'_h\|_1,$$

where we recall that $(\mathbf{P}^\pi_h d_h)(x_{h+1}) = \iint P_h(x_{h+1}|x_h, a_h)\pi(a_h|x_h)d_h(x_h)(\mathrm{d}x_h)(\mathrm{d}a_h)$.

Proof. From definition of $\mathbf{P}^\pi_h$ and Lemma 21, we have

$$\begin{aligned}\|\mathbf{P}^\pi_h d_h - \mathbf{P}^\pi_h d'_h\|_1 &= \iint \left|P_h(x_{h+1}|x_h, a_h)\pi(a_h|x_h)\left(d_h(x_h) - d'_h(x_h)\right)(\mathrm{d}x_h)(\mathrm{d}a_h)\right|(\mathrm{d}x_{h+1}). \\ &\le \int \left(|d_h(x_h) - d'_h(x_h)| \left(\iint \pi(a_h|x_h)P_h(x_{h+1}|x_h, a_h)(\mathrm{d}x_{h+1})(\mathrm{d}a_h)\right)\right)(\mathrm{d}x_h) \\ &\le \int |d_h(x_h) - d'_h(x_h)|(\mathrm{d}x_h) = \|d_h - d'_h\|_1.\end{aligned}$$

□

Lemma 21. *For any pseudo-policy $\overline{\pi}$ (Definition 1), we have*

$$\int \overline{\pi}_h(a_h|x_h)(\mathrm{d}a_h) \le 1 \quad \forall x_h \in \mathcal{X}, h \in [H].$$

Proof. Recall $\overline{\pi}_h(a_h|x_h) = \min\left\{\pi_h(a_h|x_h), C^{\mathbf{a}}_h \pi^D_h(a_h|x_h)\right\}$ where π_h is a valid Markov policy. Then

$$\int \overline{\pi}_h(a_h|x_h)(\mathrm{d}a_h) = \int \min\left\{\pi_h(a_h|x_h), C^{\mathbf{a}}_h \pi^D_h(a_h|x_h)\right\}(\mathrm{d}a_h) \le \int \pi_h(a_h|x_h)(\mathrm{d}a_h) = 1.$$

□

I.4. Covering lemmas

In this subsection, we provide the ℓ_1 optimistic cover lemma used in MLE (Lemma 22) and pseudo-dimension bound for the weight function class (Lemma 23) respectively.

Lemma 22. *Suppose Assumption 3 holds. Then for the function class*

$$\mathcal{F}_h(\Upsilon_{h-1}) = \{d_h = \langle \mu_{h-1}, \theta_h \rangle : \mu_{h-1} \in \Upsilon_{h-1}, \theta_h \in \mathbb{R}^{\mathsf{d}}, \|\theta_h\|_\infty \le 1, d_h \in \Delta(\mathcal{X})\},$$

there exists an ℓ_1 optimistic cover $\overline{\mathcal{F}}_h(\Upsilon_{h-1})$ (according to Definition 3) with scale γ of size $|\Upsilon_{h-1}|\left(2\lceil B^\mu/\gamma\rceil\right)^{\mathsf{d}}$ and $\overline{\mathcal{F}}_h(\Upsilon_{h-1}) \subseteq (\mathcal{X} \to \mathbb{R}_{\ge 0})$.

Proof. The ideas of this proof are adapted from the proof of Proposition H.15 in (Chen et al., 2022a). Let $\Theta_h = \{\theta_h : \exists \mu_{h-1} \in \Upsilon_{h-1},\ \text{s.t.,}\ \langle \mu_{h-1}, \theta_h \rangle \in \mathcal{F}_h(\Upsilon_{h-1})\} \subseteq \{\theta_h : \theta_h \in \mathbb{R}^{\mathsf{d}}, \|\theta_h\|_\infty \le 1\}$ be the set of θ_h parameters associated with $\mathcal{F}_h(\Upsilon_{h-1})$. Then any $d_h \in \mathcal{F}_h(\Upsilon_{h-1})$ can be written as $\langle \mu_{h-1}, \theta_h \rangle$ for some $\mu_{h-1} \in \Upsilon_h$ and $\theta_h \in \Theta_h$. Define the γ'-neighborhood of θ_h to be $\mathcal{B}(\theta_h, \gamma') := \gamma' \lfloor \theta_h/\gamma' \rfloor + [0, \gamma']^{\mathsf{d}}$, and construct the optimistic covering function for each $d_h = \langle \mu_{h-1}, \theta_h \rangle$ as

$$f_{\mu_{h-1},\theta_h}(x) = \max_{\overline{\theta} \in \mathcal{B}(\theta_h, \gamma')} \langle \mu_{h-1}(x), \overline{\theta} \rangle \quad \forall x \in \mathcal{X}.$$

Note that $f_{\mu_{h-1},\theta_h} \ge d_h$ pointwise, thus $f_{\mu_{h-1},\theta_h} \ge 0$, though it is not necessarily a valid distribution. Further,

$$\begin{aligned}\|f_{\mu_{h-1},\theta_h} - d_h\|_1 &\le \int \max_{\overline{\theta} \in \mathcal{B}(\theta_h, \gamma')} |\langle \overline{\theta} - \theta_h, \mu_{h-1}(x) \rangle|(\mathrm{d}x) \\ &\le \int \max_{\overline{\theta} \in \mathcal{B}(\theta_h, \gamma')} \|\overline{\theta} - \theta_h\|_\infty \|\mu_{h-1}(x)\|_1 (\mathrm{d}x) \\ &\le \gamma' \int \|\mu_{h-1}(x)\|_1 (\mathrm{d}x) \\ &\le \gamma' B^\mu\end{aligned}$$

using Assumption 3 in the last line. Observe that there are at most $\left(2\lceil 1/\gamma' \rceil\right)^{\mathsf{d}}$ unique γ'-neighborhoods in the set $\{\mathcal{B}(\theta_h, \gamma')\}_{\theta_h \in \Theta_h}$. This implies that there are at most $|\Upsilon_{h-1}|\left(2\lceil 1/\gamma' \rceil\right)^{\mathsf{d}}$ unique functions in the set $\{f_{\mu_{h-1},\theta_h}\}_{\langle \mu_{h-1}, \theta_h \rangle \in \mathcal{F}_h(\Upsilon_{h-1})}$, which forms an ℓ_1-optimistic cover of $\mathcal{F}_h(\Upsilon_{h-1})$ of scale γ'. Finally, setting $\gamma' = \gamma/B^\mu$ gives us an ℓ_1-optimistic covering of $\mathcal{F}_h(\Upsilon_{h-1})$ of scale γ with size $|\Upsilon_{h-1}|\left(2\lceil B^\mu/\gamma \rceil\right)^{\mathsf{d}}$. □

Lemma 23. *For any $h \in [H]$ and density feature $\mu_{h-1} \in \Upsilon_{h-1}$, the function class*

$$\mathcal{W}_h(\mu_{h-1}) = \left\{ w_h = \frac{\langle \mu_{h-1}, \theta_h^{\mathrm{up}} \rangle}{\langle \mu_{h-1}, \theta_h^{\mathrm{down}} \rangle} : \|w_h\|_\infty \leq C^{\mathbf{x}}_{h-1} C^{\mathbf{a}}_{h-1}, \theta_h^{\mathrm{up}}, \theta_h^{\mathrm{down}} \in \mathbb{R}^{\mathsf{d}} \right\}.$$

has pseudo-dimension (Definition 6) bounded as $\mathrm{Pdim}(\mathcal{W}_h(\mu_{h-1})) \leq 4(\mathsf{d}+1)\log(8e)$.

Proof. For any h and μ_h, consider the unconstrained version $\mathcal{W}'_h(\mu_{h-1})$ of $\mathcal{W}_h(\mu_{h-1})$:

$$\mathcal{W}'_h(\mu_{h-1}) = \left\{ w = \frac{\langle \mu_{h-1}, \theta_h^{\mathrm{up}} \rangle}{\langle \mu_{h-1}, \theta_h^{\mathrm{down}} \rangle} : \theta_h^{\mathrm{up}}, \theta_h^{\mathrm{down}} \in \mathbb{R}^{\mathsf{d}} \right\}.$$

Clearly, $\mathcal{W}_h(\mu_{h-1}) \subseteq \mathcal{W}'_h(\mu_{h-1})$, thus $\mathrm{Pdim}(\mathcal{W}_h(\mu_{h-1})) \leq \mathrm{Pdim}(\mathcal{W}'_h(\mu_{h-1}))$, and $\mathrm{Pdim}(\mathcal{W}'_h(\mu_{h-1})) = \mathrm{VCdim}(\mathcal{H}_{\mathcal{W}'_h(\mu_{h-1})})$, where $\mathcal{H}_{\mathcal{W}'_h(\mu_{h-1})} = \{h = \mathrm{sign}(w-c) : w \in \mathcal{W}'_h(\mu_{h-1}), c \in \mathbb{R}\}$. We will use Lemma 24 to bound $\mathrm{VCdim}(\mathcal{H}_{\mathcal{W}'_h(\mu_{h-1})})$. Any $h(x) \in \mathcal{H}_{\mathcal{W}'_h(\mu_{h-1})}$ may be written as the following Boolean formula

$$\begin{aligned}\Phi &= \mathbf{1}\left[\frac{\langle \mu_{h-1}(x), \theta_h^{\mathrm{up}} \rangle}{\langle \mu_{h-1}(x), \theta_h^{\mathrm{down}} \rangle} - c \geq 0 \right] \\ &= \left(\mathbf{1}\left[\sum_{i=1}^{\mathsf{d}} \mu_{h-1}(x)[i]\theta_h^{\mathrm{up}}[i] - c\sum_{i=1}^{\mathsf{d}} \mu_{h-1}(x)[i]\theta_h^{\mathrm{down}}[i] \geq 0 \right] \mathbf{1} \wedge \left[\sum_{i=1}^{\mathsf{d}} \mu_{h-1}(x)[i]\theta_h^{\mathrm{down}}[i] \geq 0 \right] \right) \\ &\quad \vee \left(\mathbf{1}\left[\sum_{i=1}^{\mathsf{d}} \mu_{h-1}(x)[i]\theta_h^{\mathrm{up}}[i] - c\sum_{i=1}^{\mathsf{d}} \mu_{h-1}(x)[i]\theta_h^{\mathrm{down}}[i] \leq 0 \right] \wedge \mathbf{1}\left[\sum_{i=1}^{\mathsf{d}} \mu_{h-1}(x)[i]\theta_h^{\mathrm{down}}[i] < 0 \right] \right)\end{aligned}$$

which involves $k = 2\mathsf{d}+1$ real variables, a polynomial degree of at most $l = 1$ in these variables, and $s = 4$ atomic predicates. Then from Lemma 24, $\mathrm{Pdim}(\mathcal{W}_h(\mu_{h-1}))) \leq \mathrm{VCdim}(\mathcal{H}_{\mathcal{W}'_h(\mu_{h-1})}) \leq 4(\mathsf{d}+1)\log(8e)$. □

Lemma 24 (Theorem 2.2 of Goldberg and Jerrum (1993))**.** *Let $\mathcal{C}_{k,m}$ be a concept class where concepts and instances are represented by k and m real values, respectively. Suppose that the membership test for any instance c in any concept C of $\mathcal{C}_{k,m}$ can be expressed as a Boolean formula $\Phi_{k,m}$ containing s distinct atomic predicates, each predicate being a polynomial inequality over $k+m$ variables of degree at most l. Then the VC dimension of $\mathcal{C}_{k,m}$ is bounded as* $\mathrm{VCdim}(\mathcal{C}_{k,m}) \leq 2k\log(8els)$.

I.5. Probabilistic tools

In this section, we define standard tools from statistical learning theory (Anthony and Bartlett, 2009; Vapnik, 1998) that we use in our proofs. We note that, for convenience, we may override some notations from the main paper, e.g., ε does not refer to the same thing as in other sections.

Definition 5 (VC-dimension)**.** *Let $\mathcal{F} \subseteq \{-1,+1\}^{\mathcal{X}}$ and $x_1^m = (x_1, \dots, x_m) \in \mathcal{X}^m$. We say x_1^m is shattered by $\mathcal{F}$ if $\forall \mathbf{b} \in \{-1,+1\}^m$, $\exists f_{\mathbf{b}} \in \mathcal{F}$ such that $(f_{\mathbf{b}}(x_1), \dots, f_{\mathbf{b}}(x_m)) = (b_1, \dots, b_m) \in \mathbb{R}^m$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$ is the cardinality of the largest set of points in $\mathcal{X}$ that can be shattered by $\mathcal{F}$, that is, $\dim(\mathcal{F}) = \max\{m \in \mathbb{N} \mid \exists x_1^m \in \mathcal{X}^m$, s.t. x_1^m is shattered by $\mathcal{F}\}$.*

Definition 6 (Pseudo-dimension)**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and $x_1^m = (x_1, \dots, x_m) \in \mathcal{X}^m$. We say x_1^m is pseudo-shattered by $\mathcal{F}$ if $\exists \mathbf{c} = (c_1, \dots, c_m) \in \mathbb{R}^m$ such that $\forall \mathbf{y} = (y_1, \dots, y_m) \in \{-1,+1\}^m$, $\exists f_{\mathbf{y}} \in \mathcal{F}$ such that $\mathrm{sign}(f_{\mathbf{y}}(x_i - c_i) = y_i \ \forall i \in [m]$. The pseudo-dimension of $\mathcal{F}$ is the cardinality of the largest set of points in $\mathcal{X}$ that can be pseudo-shattered by $\mathcal{F}$, that is, $\mathrm{Pdim}(\mathcal{F}) = \max\{m \in \mathbb{N} \mid \exists x_1^m \in \mathcal{X}^m$, s.t. x_1^m is pseudo-shattered by $\mathcal{F}\}$.*

Definition 7 (Uniform covering number)**.** *For $p = 1, 2, \infty$, the uniform covering number of $\mathcal{H}$ w.r.t. the norm $\|\cdot\|_p$ is define as*

$$\mathcal{N}_p(\varepsilon, \mathcal{H}, m) = \max_{x_1^m \in \mathcal{X}^m} \mathcal{N}_p(\varepsilon, \mathcal{H}, x_1^m)$$

where $\mathcal{N}_p(\varepsilon, \mathcal{H}, x_1^m)$ is the ε-covering number of $\mathcal{H}|_{x_1^m}$ w.r.t. $\|\cdot\|_p$, that is, the cardinality of the smallest set S such that for every $h \in \mathcal{H}|_{x_1^m}$, $\exists s \in S$ such that $\|h - s\|_p < \varepsilon$.

Lemma 25 (Bounding uniform covering number by pseudo-dimension, Corollary 42 of (Modi et al., 2021))**.** *Given a hypothesis class $\mathcal{H} \subseteq (\mathcal{Z} \to [a, b])$, for any $m \in \mathbb{N}$ we have*

$$\mathcal{N}_1(\varepsilon, \mathcal{H}, m) \leq \left(\frac{4e^2(b-a)}{\varepsilon}\right)^{\mathrm{Pdim}(\mathcal{H})}.$$

Lemma 26 (Uniform deviation bound using covering number, adapted from Corollary 39 of Modi et al. (2021))**.** *For $b \geq 1$, let $\mathcal{H} \subseteq (\mathcal{Z} \to [-b, b])$ be a hypothesis class and $Z^n = (z_1, \ldots, z_n)$ be i.i.d. samples drawn from some distribution $\mathbb{P}(z)$ supported on $\mathcal{Z}$. Then*

$$\mathbb{P}\left(\left|\mathbb{E}[h(z)] - \frac{1}{n}\sum_{i=1}^{n} h(z_i)\right| \geq \varepsilon\right) \leq 36\mathcal{N}_1\left(\frac{\varepsilon^3}{640b^2}, \mathcal{H}, \frac{40nb^2}{\varepsilon^2}\right)\exp\left(-\frac{n\varepsilon^2}{128\mathbb{V}[h(z)] + 512\varepsilon b}\right).$$